Sampling, Testing and Estimation with Stein Reproducing Kernels

Xing Liu

IMPERIAL

Submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy
in the Department of Mathematics, Faculty of Natural Sciences,
Imperial College London.
October 2024

Abstract

Statistical models have grown increasingly complex to better capture the diverse data types present in modern applications. However, this complexity has created significant challenges for statistical inference. A key obstacle is that many of these models are *unnormalized*, meaning that the model likelihood can only be evaluated up to an intractable normalizing constant. In such cases, traditional methods that rely on the exact model likelihood—such as likelihood-ratio tests for model evaluation or maximum likelihood estimation for model fitting—are no longer feasible.

To address the challenges posed by unnormalized models, one promising tool is the *kernel Stein discrepancy* (KSD). KSD is a statistical divergence that leverages the *Stein reproducing kernel*—a reproducing kernel based on Stein's method—to allow tractable computation without requiring the normalizing constant of the model likelihood. Consequently, KSD has facilitated the development of novel algorithms that can be efficiently applied to unnormalized models across various statistical tasks, including sample approximation, goodness-of-fit testing, and parameter estimation.

This thesis focuses on identifying the limitations of existing KSD-based methods and proposing novel solutions. First, we focus on KSD-based methods for sample approximation. We address the curse of dimensionality inherent in a widely used KSD-based sample approximation method, and propose an extension that mitigates this issue by using linear projections. Second, we turn our attention to goodness-of-fit tests based on KSD. We identify two key limitations of such tests: their lack of test power against multi-modal alternatives, and their lack of robustness under data contamination. For each limitation, we introduce new solutions. Finally, we explore the use of KSD in parameter estimation. We compare the relative statistical efficiency of KSD-based estimators to alternative methods, providing insights into how the kernel choice can affect their efficiency.

Statement of Originality

I, Xing Liu, certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged below in accordance with the standard referencing practices of the discipline.

1. Liu, X., Zhu, H., Ton, J.-F., Wynne, G., and Duncan, A. (2022). Grassmann Stein variational gradient descent. In *International Conference on Artificial Intelligence and Statistics*, volume 151, pages 2002–2021. PMLR.

    - Most of the methodology sections and numerical experiments were done by myself. Harrison Zhu and Jean-François Ton helped with formalizing the main idea of the proposed methodology, as well as with revising the experiment section. George Wynne contributed to Lemma 3.1 and Theorem 3.2. Andrew Duncan provided helpful suggestions to improve the manuscript, especially the section on how to find optimal projections.

2. Liu, X., Duncan, A. B., and Gandy, A. (2023). Using perturbation to improve goodness-of-fit tests based on kernelized Stein discrepancy. In *International Conference on Machine Learning*, pages 21527–21547. PMLR.

    - Axel Gandy and I contributed equally to the methodology and theory sections. I led the designing of the project, and the numerical experiments were done by myself. Andrew Duncan proofread and revised the manuscript.

3. Liu, X. and Briol, F.-X. (2024). On the robustness of kernel goodness-of-fit tests. *arXiv preprint* arXiv:2408.05854.

    - This manuscript is under review by JMLR. I contributed to most of the manuscript. François-Xavier Briol provided valuable suggestions to improve the theory, the proofs and the numerical experiments.

4. Chapter 6 is based on an ongoing work. Both the theory and the numerical experiments were done by myself, with Andrew Duncan and Axel Gandy providing feedback to early versions of the manuscript.

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-NonCommercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Acknowledgements

I want to thank my supervisors, Axel Gandy and Andrew Duncan, for their support throughout my PhD journey and for making me a more mature researcher. Axel's detailed feedback on my work was invaluable, and his insights on how to approach research problems were always enlightening. Andrew has always been resourceful and inspiring, and I am particularly grateful to him for introducing me to the worlds of Stein's method and kernel methods, which ultimately shaped my research focus. I was lucky to have both of you as my supervisors.

I want to thank my collaborators, Harrison Zhu, François-Xavier Briol, Kevin Huang, Jean-François Ton, and George Wynne, and my internship mentors, Niek Tax, Dima Karamshuk, and Ido Guy, for making those collaborations a fun and memorable experience. I especially appreciate Dr. Briol for welcoming me into their reading group as a visiting student, which allowed me to engage with many amazing researchers and fellow students. From all of you, I have learned how to become a better researcher.

I want to thank my office mates, CDT mates and friends, Alex Larionov, Andrea Brizzi, Paula Cordero Encinar, Linying Yang, Yu Chen, Yijin Zeng, and Xiaoxue Shen, for all the conversations and laughter we shared. Your company made my PhD journey much more enjoyable, and I will especially miss the Friday afternoon chats we had in HXLY 617.

Lastly, I want to thank my partner Yiran Chen, my loving parents and families, my best friend Edison Mok, as well as my cat and "research assistant" Shroedy, for their unconditional support during the ups and downs in the past few years, for being there for me throughout my time as a student, and for everything we have shared in life. You made this experience more meaningful, and I am so glad to have you all in my life.

Contents

List of Tables 9

List of Figures 10

Principal Notation 12

1 Introduction 13

2 Background 17

2.1 Reproducing Kernel Hilbert Spaces . . . 18

2.1.1 Characterization of RKHSs . . . 18

2.1.2 Properties of RKHSs . . . 20

2.1.3 Examples of RKHSs . . . 21

2.2 Kernel Stein Discrepancy . . . 22

2.2.1 Integral Probability Metrics . . . 23

2.2.2 Stein Discrepancies . . . 25

2.2.3 Langevin Stein Operator . . . 26

2.2.4 Langevin Kernel Stein Discrepancy . . . 27

2.2.5 Estimation of Langevin Kernel Stein Discrepancy . . . 29

2.2.6 Stein Reproducing Kernels . . . 30

2.2.7 Other Theoretical Properties . . . 30

2.3 Applications to Computational Statistics . . . 31

2.3.1 Sample Approximation . . . 31

2.3.2 Goodness-of-Fit Testing . . . 34

2.3.3 Parameter Estimation . . . 37

3 Sample Approximation: Grassmann Stein Variational Gradient Descent 39

3.1 Introduction . . . 39

3.2 Background . . . 41

3.3 Related Work . . . 44

3.4 Grassmann Kernel Stein Discrepancy . . . 45

3.5 Grassmann Stein Variational Gradient Descent . . . 47
3.6 Experiments . . . 50
3.7 Conclusion and Discussion . . . 54
3.A Further Background Materials on the Grassmann Manifold . . . 54
3.B Proofs of Theoretical Results . . . 56
3.B.1 Proof of Theorem 3.2 . . . 56
3.B.2 Proof of Proposition 3.3 . . . 58
3.C Experimental Details . . . 61
3.D Supplementary Figures . . . 62
3.D.1 Synthetic Experiments . . . 62
3.D.2 Conditioned Diffusion Process . . . 63
3.D.3 Ablation Studies . . . 63

4 Goodness-of-Fit Testing Part I: Using Perturbation to Improve KSD Tests 66
4.1 Introduction . . . 66
4.2 Background . . . 68
4.3 Related Work . . . 70
4.4 Limitations of KSD Test . . . 70
4.5 KSD Test with Perturbation . . . 72
4.5.1 KSD with a Single Perturbation Kernel . . . 72
4.5.2 KSD with Multiple Perturbation Kernels . . . 73
4.5.3 GOF Testing with spKSD . . . 74
4.6 A Transition Kernel for Multi-Modal Alternatives . . . 75
4.6.1 Choosing the Proposal Density . . . 75
4.6.2 Understanding the Source of Test Power . . . 76
4.6.3 Choosing the Set of Perturbations $\mathcal{S}$ . . . 78
4.6.4 Estimating Mode Vectors and Hessians . . . 78
4.6.5 Choosing the Number of Transitions T . . . 79
4.7 Experiments . . . 80
4.8 Conclusion and Discussion . . . 84
4.A Proofs of Theoretical Results . . . 85
4.A.1 Proof of Proposition 4.1 . . . 85
4.A.2 Proof of Theorem 4.2 . . . 87
4.A.3 Proof of Proposition 4.4 . . . 88
4.A.4 Validity of the Accept-Reject Rule . . . 90
4.A.5 Proof of Proposition 4.5 . . . 92
4.B Details on Sampling from Gaussian-Bernoulli RBMs . . . 95

5 Goodness-of-Fit Testing Part II: On the Robustness of KSD Tests **96**

5.1 Introduction . . . 96
5.2 Background . . . 98
5.2.1 The Kernel Stein Discrepancy . . . 99
5.2.2 Kernel Goodness-of-Fit Testing . . . 100
5.2.3 Robustness for Goodness-of-Fit Testing and Related Work . . . 101
5.3 The (Lack of) Robustness of Existing Kernel Goodness-of-fit Tests . . . 102
5.3.1 Existing KSD Tests with Stationary Kernels are not Qualitatively Robust . . . 102
5.3.2 Existing KSD Tests with Tilted Kernels are Qualitatively Robust . . . 103
5.4 Robust Kernel Goodness-of-fit Tests for KSD-Ball Uncertainty Sets . . . 105
5.4.1 A Robust KSD Test . . . 105
5.4.2 Choosing the Uncertainty Radius . . . 106
5.5 Experiments . . . 107
5.5.1 Toy Gaussian Model . . . 108
5.5.2 Gaussian-Bernoulli Restricted Boltzmann Machines . . . 111
5.5.3 A Kernel Exponential Family Model for Density Estimation . . . 112
5.5.4 Limitations for Multimodal Models . . . 114
5.6 Conclusion and Discussion . . . 114
5.A Proof of Theoretical Results . . . 115
5.A.1 Proof of Theorem 5.1 and Related Preliminary Results . . . 116
5.A.2 Proof of Lemma 5.2 . . . 126
5.A.3 Proof of Theorem 5.3 and Related Preliminary Results . . . 127
5.A.4 P-Targeted Kernel Stein Discrepancy . . . 138
5.A.5 Validity of the Bootstrap Approach . . . 141
5.A.6 Connections Between KSD Balls and Contamination Models . . . 145
5.B Complementary Experiments . . . 148
5.B.1 Decline Rate of Contamination Ratio . . . 148
5.B.2 Implementation Details for the Robust MMD Tests . . . 148
5.B.3 Approximating the Supremum of the Stein Kernel . . . 150
5.B.4 Ablation Study for Kernel Bandwidth in the Standard KSD Test . . . 151
5.B.5 Gaussian Mean-Shift Experiment . . . 151
5.B.6 Scalability with Dimension . . . 152
5.B.7 Weighted Bootstrap and Wild Bootstrap . . . 153
5.C Extension to U-statistics . . . 153
5.D A Robust KSD Test based on Deviation Bounds . . . 154

6 Parameter Estimation: On the Efficiency of Minimum KSD Estimation **156**

6.1 Introduction . . . 156
6.2 Background . . . 158
6.2.1 Minimum Divergence Estimators . . . 158
6.2.2 Consistency and Asymptotic Normality of MDE . . . 158
6.2.3 MDE for Unnormalized Models . . . 159
6.2.4 Minimum KSD Estimators . . . 159
6.2.5 Minimum Score-Matching Estimators . . . 161
6.3 Related Work . . . 162
6.4 Efficiency Bounds via a Kernel Decomposition . . . 163
6.4.1 KSD with Overly Smoothing Kernels is Inefficient . . . 166
6.4.2 KSD with Less Smoothing kernels is Comparably Efficient as SM . . . 168
6.5 Conclusion and Discussion . . . 170
6.A Asymptotic Theory for Minimum Score-based Estimators . . . 171
6.A.1 Asymptotic Normality for Minimum KSD Estimators . . . 171
6.A.2 Asymptotic Normality for SM Estimators . . . 172
6.B Proofs of Theoretical Results . . . 173
6.B.1 Preliminary Results . . . 173
6.B.2 Proof for Theorem 6.1 . . . 177
6.B.3 Proof of Corollary 6.1.1 . . . 178
6.B.4 Proof for Theorem 6.2 . . . 179
6.B.5 Proof of Corollary 6.2.1 . . . 181

7 Summary and Future Directions **182**

Bibliography **183**

List of Tables

1.1 Summary of contributions in the thesis 15

4.1 Results for the sensor network localization experiment. 82

5.1 Summary of theoretical contributions to the robustness of kernel GOF tests 98

List of Figures

3.1 Summary of SVGD algorithms and our contribution. 41

3.2 Illustration of gradient descent on the Grassmann manifold 44

3.3 Results of GSVGD on a Gaussian target. 51

3.4 Results of GSVGD on multimodal mixture and X-shaped targets. 51

3.5 Performance metrics of GSVGD on multimodal mixture and X-shaped targets. 52

3.6 Results of GSVGD on the conditioned diffusion process experiment. 52

3.7 Results of GSVGD in the Covertype experiment. 54

3.8 Covariance matrices of GSVGD particles in the Covertype experiment. 55

3.9 Assessment of the convergence of particles in the synthetic experiments 63

3.10 Variance estimates in the Gaussian experiment. 64

3.11 Performance metrics of GSVGD in the conditioned diffusion process experiment 64

3.12 Ablation study of GSVGD on the number of projectors. 64

3.13 Run time of GSVGD against the dimension of the target. 64

4.1 Power of pKSD tests in the mixture of Gaussians experiment. 67

4.2 Connections between KSD and the proposed family of pKSD. 67

4.3 Illustration of the original and perturbed densities. 77

4.4 Rejection probability and target densities in the Gaussian mixture experiment. 80

4.5 Rejection probability in the mixture of t and banana distributions experiment. 80

4.6 Rejection probability in the multivariate Gaussian experiment. 81

4.7 Illustration of the sensor localization experiment. 83

4.8 Rejection probability in the GB-RBM experiment with the standard parameter setting. 84

4.9 Rejection probability in the GB-RBM experiment with the multi-modal parameter setting 84

5.1 Stein reproducing kernel and rejection probability with different tilted kernels. 108

5.2 Rejection probability of robust-KSD test with different bandwidths. 108

5.3 Rejection probability in the outlier-contaminated Gaussian experiment. 109

5.4 Density functions and rejection probability in the heavy-tailed experiment. 110

5.5 Data, density functions and rejection probability in the Gaussian-Bernoulli RBM experiment. 111

5.6 Density functions and rejection probability in the kernel exponential family experiment. 112

5.7 Data and rejection probability in the mixture-of-Gaussian experiment. 114

5.8 Rejection probability of standard KSD test with different contamination ratios. 148

5.9 Rejection probability of standard KSD test against kernel bandwidth. 148

5.10 Estimation of upper bound on Stein reproducing kernel. 151

5.11 Rejection probability of robust-KSD test with different uncertainty radii. 152

5.12 Rejection probability of robust-KSD test against dimensions. 152

5.13 Comparing the robust-KSD tests with weighted or wild booststrap. 153

5.14 Comparing the robust-KSD test with the robust-KSD-Dev test. 155

6.1 Eigendecomposition of the squared-exponential kernel. 164

6.2 Illustration of the sufficient statistic. 164

6.3 Inefficiency of KSD estimation in the augmented Gaussian-location model. 167

6.4 Efficiency of KSD estimation in the Gaussian-location model. 170

Principal Notation

$\mathbb{R}^d$	Euclidean space of dimension d, where d is a fixed, positive integer.
$\mathbb{S}^{d-1}$	Unit sphere in $\mathbb{R}^d$ defined as $\mathbb{S}^{d-1} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 = 1\}$
$\mathcal{P}(\mathcal{X})$	Set of Borel probability measures on $\mathcal{X} \subseteq \mathbb{R}^d$.
$\mathbb{1}\{A\}$	Indicator function for an event A
$\delta_{\mathbf{x}}$	Dirac measure with point mass at $\mathbf{x} \in \mathbb{R}^d$
x	Deterministic scalar (italic lowercase letters)
X	Scalar random variable (italic uppercase letters)
$\mathbf{x}$	Deterministic vector (bold italic lowercase letters).
$\mathbf{X}$	Vector random variable (bold italic uppercase letters)
$\mathbf{s}_p$	The (Hyvärinen) score function of a probability density function p, defined as $\mathbf{s}_p : \mathbb{R}^d \to \mathbb{R}^d$ with $\mathbf{s}_p(\mathbf{x}) = \nabla_{\mathbf{x}} \log p(\mathbf{x})$.
$\mathbb{E}[\cdot]$	Expectation of a random variable
$\mathrm{Var}(\cdot)$	Variance of a random variable.
k	A real-valued kernel on $\mathcal{X} \times \mathcal{X}$
$\mathcal{H}_k$	Reproducing kernel Hilbert space associated with kernel k
$\mathcal{C}^r$	Set of r-times continuously differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$.
$\mathcal{C}_b^r$	Subset of $\mathcal{C}^r$ containing functions with bounded r-th order partial derivatives
$\mathcal{C}^{(1,1)}$	Set of functions $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that $(\mathbf{x}, \mathbf{y}) \mapsto \partial_{1j}\partial_{2j}k(\mathbf{x}, \mathbf{y})$ exists and is continuous for all $j = 1, \ldots, d$, where $\partial_{ij}k(\mathbf{x}, \mathbf{y})$ denotes the partial derivative of k with respect to the j-th coordinate of the i-th argument.
$\mathcal{C}_b^{(1,1)}$	Subset of $\mathcal{C}^{(1,1)}$ containing functions k such that $(\mathbf{x}, \mathbf{y}) \mapsto \partial_{1j}\partial_{2j}k(\mathbf{x}, \mathbf{y})$ is bounded, for all $j = 1, \ldots, d$.
$L_p(Q)$	Space of functions $f : \mathbb{R}^d \to \mathbb{R}$ that are L_p-integrable with respect to a probability measure $Q \in \mathcal{P}(\mathbb{R}^d)$, i.e., $\int_{\mathbb{R}^d} \|f(\mathbf{x})\|^p Q(\mathrm{d}\mathbf{x}) < \infty$.
$\|\cdot\|_{\mathrm{op}}$	Operator norm defined as $\|A\|_{\mathrm{op}} = \sup_{\mathbf{x} \in \mathbb{S}^{d-1}} \|A\mathbf{x}\|_2$ for a matrix $A \in \mathbb{R}^{s \times d}$.

Chapter 1

Introduction

Statistical models permeate all scientific disciplines, serving as an essential tool for describing observed data and making inferences about the underlying data-generating process. Thanks to the rapid development of computational resources, statistical models have grown increasingly complex, allowing for more accurate and flexible data representations. However, this added flexibility also introduces new challenges for downstream tasks in statistical inference.

One of such challenges stems from the intractability of the model likelihood. Many statistical inference procedures are likelihood-based procedures that rely on the ability to evaluate the model likelihood for a given datum and parameter. Examples of such procedures include likelihood-ratio tests (Hogg et al., 1977, Chapter 8.7) for goodness-of-fit testing and maximum likelihood estimation (van der Vaart, 2000, Chapter 5) for model fitting. However, as statistical models become more complex, the normalizing constant of the model likelihood often becomes intractable and also challenging to estimate accurately. As a result, most likelihood-based methods are no longer feasible. Developing alternative tools for such models has been a major research focus in computational statistics.

This thesis aims to tackle this challenge of intractability in three areas of computational statistics: sample approximation, goodness-of-fit testing, and parameter estimation. These fundamental tasks arise frequently in practice and are essential for extracting meaningful information from data, both in the frequentist and in the Bayesian frameworks (Casella and Berger, 2002; Green et al., 2015; Gelman et al., 2020). Specifically, we focus on the following tasks:

1. **Sample approximation** concerns the approximation of a given probabilistic distribution by a finite set of samples. For example, in Bayesian analysis, one often wishes to make inference with the posterior distribution, which typically requires computing its moments, such as the posterior mean or variance. In frequentist analysis, quantiles of a distribution are often required for constructing confidence intervals. However, these expectations or quantiles are rarely available in closed form except in special cases. This necessitates the use of Monte Carlo methods to obtain sample-based approximations of the target distribution.

2. **Goodness-of-fit testing** evaluates how well a given distribution aligns with observed data. This process typically involves hypothesizing that the data were generated from a particular

probabilistic distribution, and rejecting this hypothesis if significant deviations are observed. This step allows for the detection of substantial violations of model assumptions, thereby ensuring the reliability of subsequent inferences.

3. **Parameter estimation** focuses on fitting a statistical model to observed data. In frequentist analysis, this often involves specifying a family of candidate probability distributions, possibly parameterized by some model parameters, and identifying the distribution within the family that best describes the data according to a pre-specified criterion, often through optimizing the model parameters. Parameter estimation is an integral part of statistical modelling and is important for many downstream tasks, such as making predictions or identifying hidden patterns of the observed data.

We focus on statistical models that admit explicit density functions with a potentially intractable normalizing constant. In this setting, the aforementioned tasks become particularly challenging. Such models are common in frequentist analysis, including most exponential family models (Barndorff-Nielsen 1978, Part III; Bishop 2006, Section 2.4; Efron 2022), large graphical models (Jordan and Weiss, 2002), or deep neural networks (Zhang et al., 2018; Song and Kingma, 2021). Unnormalized models are also ubiquitous in Bayesian analysis (Robert et al., 2007), where, by the Bayes' rule, the normalizing constant of the posterior distribution involves integration over the entire parameter space, which is typically unavailable in closed form unless the model is designed to be conjugate to its prior. However, unnormalized models exclude those specified only by a generative process, such as generative adversarial networks (Goodfellow et al., 2014) or simulation-based models (Lintusaari et al., 2017).

One viable approach to addressing the challenge of intractable normalizing constants is through the use of *kernel Stein discrepancy* (KSD). KSD is a statistical divergence based on *Stein reproducing kernels* (Oates et al., 2017), a family of reproducing kernels constructed using Stein's method (Stein, 1972) to allow tractable computations without needing to know the normalizing constant of the model density. Consequently, a key advantage of KSD over other statistical divergences such as the Kullback-Leibler divergence is that it can be computed efficiently even when the normalizing constant of the model is intractable. This makes KSD a popular tool in computational statistics and machine learning, inspiring many novel methods for a variety of applications, including sample approximation (Liu and Wang, 2016), goodness-of-fit testing (Chwialkowski et al., 2016; Liu et al., 2016), variance reduction (Oates et al., 2017), parameter estimation (Barp et al., 2019), generalized Bayesian inference (Matsubara et al., 2022), and many others; see Anastasiou et al. (2023) for a review.

Outline Despite the success of KSD, many problems with the methods it has inspired remain unsolved. This thesis aims to identify and address some of the limitations in applying kernel Stein discrepancy to various problems. A brief summary of the key contributions is provided in Table 1.1. The rest of this thesis is structured as follows.

1. Chapter 2 provides motivation and background. We start by introducing unnormalized models and explaining the challenges they give rise to. We then provide the background on kernel Stein discrepancy, including prerequisites on kernel methods, integral probability metrics, and Stein's method. Next, we review how KSD has been applied to address the intractability issue in three

Table 1.1: Summary of contributions in this thesis. Each chapter focuses on a different task in computational statistics and addresses a specific problem of existing methods that are based on kernel Stein discrepancy.

	Task	Problem to solve	Contributions	Publication
Chapter 3	Sample approximation	Curse of dimensionality	Grassmann SVGD	Liu et al. (2022)
Chapter 4	Goodness-of-fit testing	Blindness against multi-modal alternatives	pKSD tests	Liu et al. (2023)
Chapter 5	Goodness-of-fit testing	Lack of robustness	Robust-KSD tests	Liu and Briol (2024)
Chapter 6	Parameter estimation	Comparison of statistical efficiency with competitors	Theoretical results and empirical examples	Work in progress

computational statistical tasks, namely sample approximation (Section 2.3.1), goodness-of-fit testing (Section 2.3.2), and parameter estimation (Section 2.3.3). These three tasks will be explored in detail in the following chapters. Our contributions to each task are briefly summarized at the end of Section 2.3.1, Section 2.3.2 and Section 2.3.3, respectively.

2. Chapter 3 focuses on **sample approximation**. We begin by reviewing the Stein Variational Gradient Descent (SGVD) algorithm of Liu and Wang (2016), which is a method based on kernel Stein discrepancy and is designed for generating finite samples that can approximate a target distribution. A well-known limitation of this method is its poor scalability with dimension (Ba et al., 2019). Motivated by this, we propose a novel variant capable of mitigating this issue by using linear projections. Through a range of empirical assessments, we demonstrate the superior performance of our proposed method, especially when the target distribution exhibits a sparse structure.

3. Chapter 4 and 5 focus on **goodness-of-fit testing**. Assessing the goodness-of-fit (GOF) of a statistical model can be challenging when the model likelihood has an unknown normalizing constant, since traditional methods like likelihood-ratio tests or Kolmogorov-Smirnov tests are no longer applicable. We review how KSD has been used to construct goodness-of-fit tests applicable to such models, leading to a family of kernel GOF tests called KSD tests (Chwialkowski et al., 2016; Liu et al., 2016). We then identify two limitations of these tests. The first limitation, discussed in Chapter 4, concerns the lack of test power of these tests against multi-modal alternatives. We address this issue by proposing to boost the test power by using stochastic perturbations via Markov processes. The second limitation, articulated in Chapter 5, concerns the lack of robustness of these tests under data contamination or model misspecification. We subsequently propose the first robust kernel goodness-of-fit test, which is a direct generalization of the KSD test and guarantees robustness under no extra computational cost than the standard KSD test.

4. Chapter 6 focuses on **parameter estimation**. Parameter estimators are often constructed by minimizing some divergence between observed data and the statistical model (Pardo, 2005; Basu et al., 2011), whereby the maximum likelihood estimator corresponds to choosing the Kullback-Leibler (KL) divergence. However, when the model likelihood has an unknown normalizing

constant, the KL divergence is no longer applicable, and other options such as score matching loss (Hyvärinen, 2005) and KSD (Barp et al., 2019) have been proposed as promising alternatives. We compare the statistical efficiency of these methods, providing insights into how the performance of KSD-based estimators can be affected by the choice of kernels. Our results highlight the inefficiency of the KSD-based estimators compared with the score matching counterpart.

5. Chapter 7 concludes with a summary of the main contributions and a discussion on open research questions.

Chapter 2

Background

Our primary interest is in performing inference with statistical models that involve an unknown normalizing constant. Let $P \in \mathcal{P}(\mathbb{R}^d)$ be a probability measure on $\mathbb{R}^d$. We assume P admits a probability density function p with respect to the Lebesgue measure, taking the form

$$p(\mathbf{x}) = \frac{p^*(\mathbf{x})}{Z},$$

where p^* is an *unnormalized* density function, and $Z = \int_{\mathbb{R}^d} p^*(\mathbf{x})\,\mathrm{d}\mathbf{x}$ is a positive constant called the *normalizing constant*. Typically, pointwise evaluation of p^* is feasible, but the normalizing constant Z lacks a closed-form expression, as the integral of p^* is often infeasible to compute. We call such models *unnormalized models*.

Unnormalized models arise frequently in statistics and machine learning. In frequentist analysis, common examples include many modern density estimation models (Canu and Smola, 2006; Sriperumbudur et al., 2017), exponential families (Barndorff-Nielsen, 1978), energy-based models (Cho et al., 2013; Zhang et al., 2018), and large graphical models (Jordan and Weiss, 2002). In Bayesian inference, Bayesian posteriors are typical unnormalized, except in special cases where the model is designed to be conjugate.

Unnormalized models pose challenges to many tasks in statistical inference. For example, to construct finite-sample approximation of unnormalized models, Markov chain Monte Carlo (MCMC) methods (Meyn and Tweedie, 1993; Gilks et al., 1995; Robert and Casella, 2004) are often used, but these methods tend to be slow (Blei et al., 2017) and can struggle in high dimensions (Levin and Peres, 2017, Chapter 7). When testing for goodness-of-fit of unnormalized models, classic methods such as likelihood-ratio tests (Hogg et al., 1977, Chapter 8.7) and Kolmogorov-Smirnov tests (Kolmogorov, 1933) are no longer applicable, since they either demand the normalization constant Z to have a closed-form expression in order to compute the likelihood ratio, or require access to the cumulative distribution function of P or a sample-approximation of it, neither of which is readily available for unnormalized models. In parameter estimation, fitting a statistical model using maximum likelihood estimation requires that the likelihood can be evaluated for any given values of data and parameters; however, with unnormalized models, the normalization constant Z of the likelihood is unknown and also depends on the parameters to optimize, making maximum likelihood estimation impractical.

In some cases, one can still approach the aforementioned tasks by resorting to approximations of $p(\mathbf{x})$. This can be done by approximating $p(\mathbf{x})$ with *pseudo likelihoods* (Besag, 1974) or *profile likelihoods* (Murphy and Van der Vaart, 2000). On the other hand, if the model is *generative* (Mohamed and Lakshminarayanan, 2016), meaning that generating samples from P is straightforward, then it might also be feasible to replace $p(\mathbf{x})$ by finite-sample approximations, leading to the field of *likelihood-free inference*. A rich literature exists in this field, encompassing MCMC methods (Murray et al., 2012), distribution testing (Gretton et al., 2012; Gerber and Polyanskiy, 2024), parameter estimation (Bernton et al., 2019; Briol et al., 2019), and approximate Bayesian computation (Marin et al., 2012; Lintusaari et al., 2017; De Visscher et al., 2024). However, all these methods inevitably introduce approximation errors, potentially resulting in loss of statistical efficiency. Another limitation is that constructing accurate approximations can be computationally demanding, especially because many unnormalized models are *not* generative and sampling from them can be non-trivial.

An alternative solution is to develop tools that can be implemented *without* the normalizing constant. More precisely, this can be done by leveraging *statistical divergences* that are computable even with unnormalized models. One such divergence is the *kernel Stein discrepancy* (KSD), a family of statistical divergences based on kernel methods and Stein's method. KSD has been successfully applied to various statistical inference tasks; a review is given by Anastasiou et al. (2023).

Outline of the chapter In this chapter, we will introduce KSD and explain how it can address the challenges posed by unnormalized models. Section 2.1 covers basic concepts and results in the theory of reproducing kernel Hilbert spaces. Section 2.2 provides the necessary background on integral probability metrics and Stein discrepancies, before introducing kernel Stein discrepancy. Section 2.3 reviews how KSD has been applied to tackle the challenges of unnormalized models in three areas of computational statistics: sample approximation, goodness-of-fit testing, and parameter estimation. For each area, we also provide a brief summary of our contributions.

2.1 Reproducing Kernel Hilbert Spaces

We briefly review the theory of reproducing kernel Hilbert spaces (RKHS). A full treatment is given by Berlinet and Thomas-Agnan (2004) and Steinwart and Christmann (2008). We will focus on RKHSs of real-valued functions $f : \mathcal{X} \to \mathbb{R}$ defined on some Euclidean domain $\mathcal{X} \subseteq \mathbb{R}^d$, although extensions to complex- or vector-valued functions and non-Euclidean domains also exist (Carmeli et al., 2006, 2010).

2.1.1 Characterization of RKHSs

A reproducing kernel Hilbert spaces is a Hilbert space that possess a *reproducing kernel*, defined next.

Definition 2.1 (Reproducing kernel)**.** *Let $\mathcal{H}$ be a Hilbert space of real-valued functions on $\mathcal{X}$ equipped with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a* reproducing kernel *of the Hilbert space $\mathcal{H}$ if*

(i) $\forall \mathbf{x} \in \mathcal{X}, \quad k(\cdot, \mathbf{x}) \in \mathcal{H}$, *and*

(ii) $\forall \mathbf{x} \in \mathcal{X}, \ \forall f \in \mathcal{H}, \quad \langle f, k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}} = f(\mathbf{x})$.

An RKHS is often denoted as $\mathcal{H}_k$ to emphasize its dependence on the reproducing kernel k. Condition *(ii)* in the above definition is called the *reproducing property*, since it states that the value of the function f at any point $\mathbf{x}$ can be reproduced by the inner product of f with $k(\cdot, \mathbf{x})$. The reproducing property is a key property of RKHS that makes them extremely useful in machine learning and statistics.

RKHSs can be characterized as Hilbert spaces where the evaluation of functions at a fixed point is a continuous mapping. This characterization is specified by the following *Riesz's representation theorem*.

Theorem 2.1 (Riesz's representation theorem; Berlinet and Thomas-Agnan, 2004, Theorem 1)**.** *A Hilbert space $\mathcal{H}$ of functions on $\mathcal{X}$ has a reproducing kernel if and only if all the evaluations functionals $e_{\mathbf{x}} : \mathcal{H} \to \mathbb{R}$ defined by $e_{\mathbf{x}}(f) = f(\mathbf{x})$, for $\mathbf{x} \in \mathcal{X}$, are continuous on $\mathcal{H}$.*

Having characterized RKHSs, we now focus on the characterization of reproducing kernels, which would allow us to identify when a given function is a reproducing kernel. To do so, we introduce the class of *positive definite* functions. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called positive definite (or *positive type*) if

$$\forall n \geq 1,\ \forall a_1, \ldots, a_n \in \mathbb{R},\ \forall \mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}, \qquad \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(\mathbf{x}_i, \mathbf{x}_j) \geq 0 \ .$$

The above condition is equivalent to the positive semi-definiteness of the matrix $(k(\mathbf{x}_i, \mathbf{x}_j))_{1 \leq i,j \leq n}$, known as the *Gram matrix*, for any $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and any $n \geq 1$. It can be shown that positive definite kernels characterize the class of reproducing kernels—the reproducing kernel of any RKHS is positive definite, and, conversely, any positive definite kernel k induces a unique RKHS $\mathcal{H}_k$. The converse implication is known as the *Moore–Aronszajn theorem*, due to Moore (1935) and Aronszajn (1950).

Theorem 2.2 (Moore–Aronszajn theorem; Berlinet and Thomas-Agnan, 2004, Theorem 3)**.** *Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel. Define $\mathcal{H}_0$ to be the space spanned by the functions $k(\cdot, \mathbf{x})$, for $\mathbf{x} \in \mathcal{X}$, and endow $\mathcal{H}_0$ with the inner product*

$$\langle f,\ g \rangle_{\mathcal{H}} \ = \ \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j k(\mathbf{x}_i, \mathbf{y}_j) \ ,$$

where $f = \sum_{i=1}^{n} a_i k(\cdot, \mathbf{x}_i)$ and $g = \sum_{j=1}^{m} b_j k(\cdot, \mathbf{y}_j)$. Let $\mathcal{H}$ be the set of functions on E that are pointwise limits of Cauchy sequences in $\mathcal{H}_0$. Then $\mathcal{H}$ is the unique Hilbert space of functions on $\mathcal{X}$ with k as reproducing kernel.

The Moore-Aronszajn theorem also reveals the close connection between the properties of a reproducing kernel and of the functions in its associated RKHS—from the form of f and g in the above theorem, it is not surprising that many properties of the kernel k, such as boundedness and smoothness, are inherited by the functions in $\mathcal{H}_k$. Moreover, this theorem suggests that an RKHS can be constructed by specifying a positive definite kernel k. The next result then characterize *all* reproducing kernels on a domain $\mathcal{X}$.

Theorem 2.3 (Kernel trick; Berlinet and Thomas-Agnan, 2004, Theorem 4)**.** *A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel or a positive definite kernel if and only if there exists a mapping $\phi : \mathcal{X} \to \mathcal{H}_0$, where $\mathcal{H}_0$ is a Hilbert space equipped with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$, such that $k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\mathcal{H}_0}$, for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.*

This theorem claims that, to verify if a function is a reproducing kernel, one can check whether it can be expressed as an inner product in some Hilbert space. Intuitively, the function ϕ maps an element $\mathbf{x}$ in $\mathcal{X}$ to some potentially higher-dimensional representation $\phi(\mathbf{x})$ in the Hilbert space $\mathcal{H}_0$. Consequently, in machine learning, ϕ is commonly known as the *feature map* and the space $\mathcal{H}_0$ is called the *feature space*. Unlike the unique correspondence between RKHSs and reproducing kernels as stated in Theorem 2.2, the relationship between reproducing kernels and feature maps is not one-to-one (Steinwart and Christmann, 2008, Chapter 4.1).

Theorem 2.3 is known as the *kernel trick* and is a direct consequence of Riesz representation theorem. The kernel trick is a powerful tool in machine learning and statistics, as it states that to evaluate inner products of the feature map ϕ, one only needs to evaluate the associated kernel function. This provides significant flexibility by allowing one to work with high-dimensional feature spaces without needing to explicitly compute the feature map and the inner product, while retaining linearity due to the ability of representing kernel evaluations by inner products in the feature space. Using the kernel trick, many algorithms that involve inner products have been "kernelized", meaning that the inner products are replaced by reproducing kernels. Examples include support vector machines (Schölkopf and Smola, 2002), algorithms for pattern recognition (Shawe-Taylor and Cristianini, 2004), and many other methods in machine learning (Hofmann et al., 2008). As we will elaborate in Section 2.2, the kernel trick is also essential in constructing statistical divergences that both have computational tractability with unnormalized distributions and exhibit desirable statistical properties.

2.1.2 Properties of RKHSs

In the previous section, we have shown that RKHSs and its associated kernels are uniquely determined, and properties of the functions in an RKHS are closely connected with the properties of the kernel. These motivate practitioners to design appropriate kernels so that the functions contained in the RKHSs possess desirable properties for the downstream task. We now review some kernel properties and their implications for the functions in an RKHS. Examples are provided in the next section.

Universality Universality refers to the capacity of an RKHS to approximate any function in some function class of interest to an arbitrary precision. Intuitively, it describes whether an RKHS is "rich" enough to contain sufficiently many functions. Several definitions of universality exist in the literature, each differing by the choice of the domain $\mathcal{X}$, the function to be approximated, and the notion used to measure the approximation error. One example is the notion of *cc-universality* (Carmeli et al., 2010; Sriperumbudur et al., 2010a), which states that the RKHS $\mathcal{H}$ induced by a kernel k is dense in $\mathcal{C}_0(\mathcal{X})$ endowed with the topology of compact convergence, where $\mathcal{C}_0(\mathcal{X})$ is the space of all continuous, real-valued functions on $\mathcal{X}$. Examples of *cc*-universal kernels include squared-exponential kernels, Laplacian kernels, inverse multi-quadric kernels and Matérn kernels; see Section 2.1.3 for the precise definition of some of these kernels. Other notions of universality include c_0-*universality* and L_p-*universality*; see Sriperumbudur et al. (2011) for a review.

Characteristicity Characteristicity concerns the ability of a kernel to embed probability measures into RKHSs (Sriperumbudur et al., 2010b). Formally, let $\mathcal{P}(\mathcal{X})$ be the set of probability measures

on $\mathcal{X}$ and let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a reproducing kernel. The *kernel mean embedding* of a probability measure $Q \in \mathcal{P}(\mathcal{X})$ is the function in $\mathcal{H}_k$ defined by $\mu_Q(\cdot) = \int_{\mathcal{X}} k(\cdot, \mathbf{x}) Q(\mathrm{d}\mathbf{x}) = \mathbb{E}_{\mathbf{X}\sim Q}[k(\cdot, \mathbf{X})]$. The kernel k is called *characteristic* if the kernel mean embedding exists for all probability measures in $\mathcal{P}(\mathcal{X})$ and is injective. In other words, characteristic kernels ensure that each probability measure has a *unique* representation in the RKHS. This allows us to compare probability measures by comparing their kernel mean embeddings. This principle forms the basis behind the construction of many notions of dissimilarity between probability measures, which are widely used in applications such as distribution testing (Gretton et al., 2006, 2012), independence testing (Gretton et al., 2007), and nonparameteric or generalized Bayesian inference (Fukumizu et al., 2013; Matsubara et al., 2022); see Muandet et al. (2017) for a review.

Differentiability Differentiability of a kernel can be inherited by functions in the associated RKHS. To state this formally, we first set up our notation. For $i \in \{1, 2\}$ and coordinate $j \in \{1, \ldots, d\}$, denote by $\partial_{ij} k(\mathbf{x}, \mathbf{y})$ the partial derivative of k with respect to the j-th coordinate of the i-th argument. Moreover, define a multi-index $\alpha = (\alpha_1, \ldots, \alpha_d)$ to be a vector of non-negative integers, and define its degree as $|\alpha| = \sum_{i=1}^{d} \alpha_i$. Given a multi-index α, we define the differential operator $\partial^{\alpha,\alpha} = \partial_{11}^{\alpha_1} \ldots \partial_{1d}^{\alpha_d} \partial_{21}^{\alpha_1} \ldots \partial_{2d}^{\alpha_d}$. Given an integer $m \geq 0$, a kernel k is called *m-times continuously differentiable* if $\partial^{\alpha,\alpha} k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ exists and is continuous for all multi-indices α with $|\alpha| \leq m$. The next result concerns the differentiability of functions in the RKHS associated with an m-differentiable kernel.

Theorem 2.4 (Steinwart and Christmann, 2008, Corollary 4.36)**.** *Let d be a positive integer and suppose $\mathcal{X}$ is an open subset of $\mathbb{R}^d$. Let $m \geq 0$ be an integer and let k be an m-times continuously differentiable kernel on $\mathcal{X}$ with RKHS $\mathcal{H}_k$. Then every $f \in \mathcal{H}_k$ is m-times continuously differentiable, and for any multi-index α with $|\alpha| \leq m$ and $\mathbf{x} \in \mathcal{X}$, we have*

$$|\partial^{\alpha} f(\mathbf{x})| \;\leq\; \|f\|_{\mathcal{H}} \cdot \left(\partial^{\alpha,\alpha} k(\mathbf{x}, \mathbf{x})\right)^{1/2},$$

where $\partial^{\alpha} f(\mathbf{x}) = \partial^{\alpha_1} \ldots \partial^{\alpha_d} f(\mathbf{x})$.

In this thesis, we will focus primarily on once continuously differentiable kernels, i.e., those $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for which $(\mathbf{x}, \mathbf{y}) \mapsto \partial_{1j} \partial_{2j} k(\mathbf{x}, \mathbf{y})$ exists and is continuous for all $j = 1, \ldots, d$. We will denote the set containing such functions by $\mathcal{C}^{(1,1)}$.

2.1.3 Examples of RKHSs

We finish the review for RKHSs by providing several examples of reproducing kernels. More examples can be found in Schölkopf and Smola (2002, Chapter 1).

1. The family of *polynomial* kernels is defined as

$$k(\mathbf{x}, \mathbf{y}) \;=\; \left(\mathbf{x}^{\top}\mathbf{y} + c\right)^{p},$$

   where $c > 0$ and $p \in \mathbb{N}$. The associated RKHS is a finite-dimensional vector space containing all real-valued p-th order polynomials on $\mathcal{X}$.

2. The family of *squard-exponential* (SE) kernels, also known as Gaussian kernels, is defined as

$$k(\mathbf{x}, \mathbf{y}) \;=\; \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{2\lambda^2}\right) ,$$

where $\lambda > 0$ is known as the bandwidth. Such kernels are infinitely differentiable, and so are the functions in the RKHS that they induce.

3. The family of *inverse multi-quadric* (IMQ) kernels is defined as

$$k(\mathbf{x}, \mathbf{y}) \;=\; \left(1+\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{2\lambda^2}\right)^{-b} ,$$

where $\lambda > 0$ and $b > 0$. The IMQ kernels have a polynomial decay rate, contrasting to the exponential decay of SE kernels. Consequently, the RKHS associated with IMQ kernels contains functions with slower decay rates than those in the RKHSs associated with SE kernels. This can be desirable since when using RKHSs to construct discrepancies for probability measures, in which case slowly decaying functions can improve the discrepancy's capability of discerning the difference in the tails of probability measures; see Section 2.2.7 for a detailed discussion.

Based on existing reproducing kernels, construction of new kernels is simple. Let $(k_n)_{n\geq 1}$ be a sequence of reproducing kernels on $\mathcal{X}$. It can be shown that the following functions are also reproducing kernels (see Steinwart and Christmann 2008, Lemma 4.5, Lemma 4.6, Corollary 4.17 and Saitoh et al. 2016, Corollary 2.5):

1. Sums of kernels: $k(\mathbf{x}, \mathbf{y}) = k_1(\mathbf{x}, \mathbf{y}) + k_2(\mathbf{x}, \mathbf{y})$ is a reproducing kernel.
2. Multiples of kernels: For any $\alpha \geq 0$, the function $k(\mathbf{x}, \mathbf{y}) = \alpha k_1(\mathbf{x}, \mathbf{y})$ is a reproducing kernel.
3. Products of kernels: $k(\mathbf{x}, \mathbf{y}) = k_1(\mathbf{x}, \mathbf{y})k_2(\mathbf{x}, \mathbf{y})$ is a reproducing kernel.
4. Tilted forms of kernels: For any $w : \mathcal{X} \to [0, \infty)$, the function $k(\mathbf{x}, \mathbf{y}) = w(\mathbf{x})k_1(\mathbf{x}, \mathbf{y})w(\mathbf{y})$ is a reproducing kernel.
5. Limits of kernels: If there exists a real-valued function k such that $k(\mathbf{x}, \mathbf{y}) = \lim_{n\to\infty} k_n(\mathbf{x}, \mathbf{y})$ pointwise for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, then k is a reproducing kernel.

In the main body of this thesis, we will primarily use either the SE or IMQ kernel (Chapter 3, 4, 6), or their tilted forms (Chapter 5).

2.2 Kernel Stein Discrepancy

This section introduces kernel Stein discrepancy, a class of statistical divergences computable with unnormalized models. A *statistical divergence* is a notion of dissimilarity between distributions. Formally, it is a function $D : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to [0, \infty)$ such that $D(Q, P) = 0$ if and only if $Q = P$, for all

probability measures $Q, P \in \mathcal{P}(\mathbb{R}^d)$. When D is also symmetric and satisfies a triangle inequality, it is also called a *distance* or *metric*. Examples of statistical divergences (that are not distances) include KL divergence and α-divergences (Cichocki and Amari, 2010), while distances include Total Variation distance, Wasserstein distance (Villani, 2008), and Kolmogorov-Smirnov distance (Kolmogorov, 1933).

The concept of statistical divergences dates back to Mahalanobis (1936), and since then, it has become fundamental in computational statistics and machine learning. Many well-known methods can be shown to be defined with an appropriate statistical divergence. Examples include the likelihood-ratio, chi-square and Kolmogorov-Smirnov tests for goodness-of-fit (Pardo, 2005, Chapter 3), non-parametric distribution tests (Sejdinovic et al., 2013; Bonnini et al., 2014), variational inference (Blei et al., 2017), maximum likelihood estimation (William, 1981), and generalized Bayes (Knoblauch et al., 2022).

However, when concerning distributions with an unknown normalizing factor, most statistical divergences become challenging to compute. This is where *kernel Stein discrepancy* (KSD) comes to rescue. KSD is a family of statistical divergences that can be computed efficiently with unnormalized distributions while enjoying desirable statistical properties. To motivate KSD, we first introduce a closely related class of dissimilarities called *Integral Probability Metrics* (IPM; Müller, 1997) in Section 2.2.1. We will then introduce Stein's discrepancies and Stein reproducing kernels, two main building blocks of KSD, in the subsequent subsections.

2.2.1 Integral Probability Metrics

IPMs measure the discrepancy between two distributions by the worst-case difference in the expectations of some test functions under these distributions. Given a class $\mathcal{F}$ of real-valued functions on $\mathbb{R}^d$, an IPM between two probability measures $Q, P \in \mathcal{P}(\mathbb{R}^d)$ is defined as

$$D_{\mathrm{IPM}}(Q, P; \mathcal{F}) \coloneqq \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{\mathbf{X} \sim Q}[f(\mathbf{X})] - \mathbb{E}_{\mathbf{Y} \sim P}[f(\mathbf{Y})] \right| = \sup_{f \in \mathcal{F}} \left| \int_{\mathbb{R}^d} f(\mathbf{x}) Q(\mathrm{d}\mathbf{x}) - \int_{\mathbb{R}^d} f(\mathbf{y}) P(\mathrm{d}\mathbf{y}) \right|. \tag{2.1}$$

The function class $\mathcal{F}$ fully characterizes the IPM and dictates its properties. A suitable $\mathcal{F}$ should ensure that the size of the IPM gives a meaningful notion of how far apart Q and P are. Intuitively, $\mathcal{F}$ should be rich enough to capture the differences between Q and P, but also simple enough so that the supremum can be computed efficiently.

By choosing an appropriate $\mathcal{F}$, IPMs can enjoy desirable properties. One such properties is *separation*, meaning that $D_{\mathrm{IPM}}(Q, P) = 0 \iff Q = P$ for all $Q, P \in \mathcal{P}(\mathbb{R}^d)$. In other words, IPMs can give a meaningful characterization of the equivalence of two probability measures. Technically, for (2.1) to be an IPM, the separation property is required, otherwise it would not be an metric[1]. In the general case where separation might not hold, (2.1) is sometimes called an *Integral Probability Pseudo-metric*.

Sometimes it is sufficient for the separation property to hold for a *fixed* probability measure P, i.e.,

[1] Given a set S, a metric on S is a function $\rho : S \times S \to [0, \infty)$ such that *(i)* $\forall x \in S, \rho(x, x) = 0$, *(ii)* $\forall x, y \in S, \rho(x, y) = \rho(y, x)$, *(iii)* $\forall x, y, z \in S, \rho(x, y) \leq \rho(x, z) + \rho(z, y)$, and *(iv)* $\rho(x, y) = 0 \implies x = y$. If *(i)-(iii)* hold, then ρ is called a *pseudo-metric*. That is, a pseudo-metric needs not satisfy *(iv)*.

$D_{\text{IPM}}(Q, P) = 0 \iff Q = P$ for all $Q \in \mathcal{P}(\mathbb{R}^d)$, in which case D_{IPM} is said to be *P-separating*.

Another property often required of IPMs is its ability to *metrize weak convergence*. An IPM D is said to metrize weak convergence if, for any sequence $\{Q_n\}_n \subset \mathcal{P}(\mathbb{R}^d)$, the convergence $D_{\text{IPM}}(Q_n, P; \mathcal{F}) \to 0$ holds if any only if $Q_n \to P$ weakly. Intuitively, this property states that weak convergence of probability measures can be characterized by convergence in the IPM, so that the relative size of the IPM can be used to quantify the closeness of two probability measures.

Many well-known metrics on probability measures can be expressed as IPMs. For example,

- The *1-Wasserstein distance* (Villani, 2008) corresponds to choosing $\mathcal{F} = \{f : \mathbb{R}^d \to \mathbb{R} : \mid M_L(f) \leq 1\}$, where $M_L(f) \coloneqq \sup_{\mathbf{x} \neq \mathbf{y}} |f(\mathbf{x}) - f(\mathbf{y})| / \|\mathbf{x} - \mathbf{y}\|_2$ denotes the Lipschitz constant of f. This is due to the well-known Monge-Kantorovich-Rubinstein duality (Villani, 2008, Theorem 5.10, Remark 6.5).
- The *Total Variation (TV) distance* corresponds to choosing $\mathcal{F} = \{f : \mathbb{R}^d \to \mathbb{R} \mid \|f\|_\infty \leq 1\}$ (Müller, 1997, Theorem 5.4).
- The *bounded Lipschitz* or *Dudley* metric corresponds to choosing $\mathcal{F} = \{f : \mathbb{R}^d \to \mathbb{R} \mid M_L(f) + \|f\|_\infty \leq 1\}$ (Müller, 1997, pp. 435).
- The *Maximum Mean Discrepancy* (MMD, Gretton et al., 2012) between two probability measures Q and P is defined as

$$\text{MMD}_k(Q, P) = \sup_{f \in \mathcal{H}_k; \|f\|_{\mathcal{H}_k} \leq 1} \left| \mathbb{E}_{\mathbf{X} \sim Q}[f(\mathbf{X})] - \mathbb{E}_{\mathbf{Y} \sim P}[f(\mathbf{Y})] \right|, \tag{2.2}$$

  where k is a reproducing kernel and $\mathcal{H}_k$ is its associated RKHS. It corresponds to choosing $\mathcal{F}$ in the IPM to be the unit ball of a Reproducing Kernel Hilbert Space, i.e., $\mathcal{F} = \mathcal{B}_k \coloneqq \{f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq 1\}$.

These IPMs have played a crucial role in the development of statistics and machine learning and there is an extensive literature on applying them to address challenges in computational statistics. However, when P is unnormalized and Q is an empirical measure formed by observed data, computing these IPMs can be challenging. Indeed, computing the IMP for a general $\mathcal{F}$ involves integrating over both Q and P, namely the terms $\mathbb{E}_{\mathbf{X} \sim Q}[f(\mathbf{X})]$ and $\mathbb{E}_{\mathbf{Y} \sim P}[f(\mathbf{Y})]$ in (2.1). While computing $\mathbb{E}_{\mathbf{X} \sim Q}[f(\mathbf{X})]$ is simple when Q is an empirical measure as the expectation reduces to a finite sum, the expectation over P typically cannot be computed in closed-form. One naive approach is to draw samples from P and approximate $\mathbb{E}_{\mathbf{Y} \sim P}[f(\mathbf{Y})]$ by Monte Carlo estimation; however, this is also not feasible since sampling from an unnormalized distribution P is often challenging. As a result, all aforementioned examples of IPMs, namely 1-Wasserstein distance, TV distance, Dudley metric and MMD, are not appropriate for this problem setup where P is unnormalized and Q is an empirical measure. Moreover, other well-known (non-IPM) statistical divergences such as Kullback-Leibler divergence (Csiszár et al., 2004; Van Erven and Harremos, 2014) or α-divergences (Cichocki and Amari, 2010) suffer from similar computational limitations, since they require the likelihood function of both Q and P to be known.

To circumvent this issue, a class of statistical divergences called *Stein discrepancies* has been proposed (Oates et al., 2017; Gorham and Mackey, 2015, 2017), whereby $\mathcal{F}$ is chosen to consist of functions whose integration over P is *known* in closed-form, thus allowing tractable computation of the integrals in the IMP objective (2.1). Such functions can be constructed by leveraging *Stein's method*, which provides a recipe for designing rich families of functions that integrate to zero under a given P that has appropriate regularity conditions. The next section reviews Stein's discrepancies.

2.2.2 Stein Discrepancies

A key ingredient of Stein's discrepancies is Stein's method (Stein, 1972; Chen et al., 2010). Stein's method is best known as a technique for distributional approximation, whereby the goal is to quantify the error made in approximating the distribution of some random variable by another distribution. It was first conceived by Charles Stein (Stein, 1972) for bounding the approximation error by normal distributions, and it has since been extended to other distributions including Poisson, exponential, and Geometric distributions. In the probability and statistics literature, Stein's method has been a fundamental tool for deriving the rate of convergence of distributional limit theorems. A detailed review of Stein's method is given by Ross (2011).

In recent years, Stein's method has received considerable interests from the communities of computational statistics and machine learning, widening its application to other problems besides distributional approximation. These novel applications include quality assessment of MCMC samples (Gorham and Mackey, 2015, 2017; Gorham et al., 2019), variance reduction (Oates et al., 2017; Sun et al., 2023b), sample compression (Riabiz et al., 2022; South et al., 2022), gradient flows for sample approximation (Liu and Wang, 2016; Korba et al., 2021; Duncan et al., 2023), goodness-of-fit testing (Liu et al., 2016; Chwialkowski et al., 2016), parameter estimation (Barp et al., 2019), and many others.

A core component of Stein's method is the construction of functions that integrate to a known quantity under a given probability measure P on $\mathbb{R}^d$. This is done by identifying *(i)* a class of functions $\mathcal{G}$ on $\mathbb{R}^d$, and *(ii)* a linear operator $\mathcal{T}$ acting on $\mathcal{G}$, such that

$$\mathbb{E}_{\mathbf{X}\sim P}[(\mathcal{T}g)(\mathbf{X})] \;=\; 0\;,\;\text{ for all } g \in \mathcal{G}\;. \tag{2.3}$$

When the above equality holds, the pair $(\mathcal{T}, \mathcal{G})$ is said to satisfy the *Stein identity* for P. The function class $\mathcal{G}$ is known a *Stein set*, and the operator $\mathcal{T}$ is called a *Stein operator*. Examples for $\mathcal{T}$ and $\mathcal{G}$ will be discussed later in Section 2.2.3 and Section 2.2.4. For the remainder of this subsection, we will assume that a pair $(\mathcal{T}, \mathcal{G})$ satisfying the Stein identity is given, and focus on the construction of a family of statistical divergences by leveraging this identity.

One specific use-case of the Stein identity is in constructing IPMs that are computable with unnormalized distributions. If the function class $\mathcal{F}$ in the IPM (2.1) is chosen to take the form $\mathcal{TG} \coloneqq \{\mathcal{T}g : g \in \mathcal{G}\}$, then $\mathbb{E}_{\mathbf{Y}\sim P}[f(\mathbf{Y})] = 0$ for all $f \in \mathcal{F}$ by the Stein identity; that is, the term involving the expectation over P vanishes, and the IPM becomes $D_{\text{IPM}}(Q, P; \mathcal{F}) = \sup_{f\in\mathcal{F}} |\mathbb{E}_{\mathbf{X}\sim Q}[f(\mathbf{X})]|$. If furthermore Q is an empirical measure, then the remaining expectation can be computed easily as it reduces to a finite

sum. Of course, to compute this IPM, one still needs to solve the supremum over $\mathcal{F}$, which can be challenging if $\mathcal{F}$ is "too large". We will defer a detailed discussion on the choice of $\mathcal{F}$ to Section 2.2.4.

This observation motivates a statistical divergence called *Stein discrepancy*. More formally, given a pair $(\mathcal{T}, \mathcal{G})$ that satisfies the Stein identity for P, the Stein discrepancy (Gorham et al., 2019) between a distribution $Q \in \mathcal{P}(\mathbb{R}^d)$ and P is defined as

$$\mathcal{S}(Q, P; \mathcal{G}) \;=\; \sup_{g \in \mathcal{G}} \left| \mathbb{E}_{\mathbf{X} \sim Q}[(\mathcal{T}g)(\mathbf{X})] \right| . \tag{2.4}$$

Stein discrepancies are not symmetric due to the asymmetric role of Q and P in the construction. Moreover, due to the Stein identity (2.3), it is clear that, regardless of the choice of the Stein set $\mathcal{G}$, the Stein discrepancy is zero when $Q = P$, namely $\mathcal{S}(P, P; \mathcal{G}) = 0$.

However, the Stein identity does not imply the converse, so that $\mathcal{S}(Q, P; \mathcal{G}) = 0$ does not necessarily imply $Q = P$. In many cases, we are interested in identifying a pair $(\mathcal{T}, \mathcal{G})$ that can fully characterize the distribution P in the sense that $\mathcal{S}(Q, P; \mathcal{G}) = 0 \iff Q = P$; in other words, we need the Stein discrepancy to be P-separating. Formally, this requires the pair $(\mathcal{T}, \mathcal{G})$ to satisfy the following condition known as *Stein characterization*:

$$Q \;=\; P \iff \mathbb{E}_{\mathbf{X} \sim Q}[(\mathcal{T}g)(\mathbf{X})] \;=\; 0 \text{ for all } g \in \mathcal{G} . \tag{2.5}$$

The Stein discrepancy defined using a pair $(\mathcal{T}, \mathcal{G})$ that satisfies the above Stein's characterization for P will have the desired P-separation property that $\mathcal{S}(Q, P; \mathcal{G}) = 0 \iff Q = P$. This property is required for most applications of Stein discrepancies as it ensures that the Stein discrepancy is a valid statistical divergence can distinguish when two distributions are different. In the following two subsections, we will provide an example of a Stein operator and a Stein set that satisfy the Stein characterization whenever the distribution P is sufficiently regular.

2.2.3 Langevin Stein Operator

In this section, we present a Stein operator for distributions P that admit a sufficiently smooth density function. Let P be a probability measure on $\mathbb{R}^d$ with Lebesgue density $p \in \mathcal{C}^1$, and assume p is supported on $\mathbb{R}^d$, i.e., $p(\mathbf{x}) > 0$ for $\mathbf{x} \in \mathbb{R}^d$. The *(Hyvärinen) score function* is defined as $\mathbf{s}_p(\mathbf{x}) \coloneqq \nabla \log p(\mathbf{x})$, where the gradient is with respect to the argument $\mathbf{x}$. The *Langevin Stein operator* $\mathcal{A}_p$ is a linear operator acting on vector-valued functions $g = (g_1, \ldots, g_d) \in \mathcal{C}^1(\mathbb{R}^d)$ and is defined as

$$(\mathcal{A}_p g)(\mathbf{x}) \;\coloneqq\; \langle \mathbf{s}_p(\mathbf{x}),\; g(\mathbf{x}) \rangle + \langle \nabla,\; g(\mathbf{x}) \rangle , \tag{2.6}$$

where $\langle \nabla,\; g(\mathbf{x}) \rangle = \sum_{j=1}^{d} \partial_j g_j(\mathbf{x})$. For the remainder of the thesis, we will also write $\nabla^\top g(\mathbf{x})$ as a shorthand for $\langle \nabla,\; g(\mathbf{x}) \rangle$.

There are two ways of showing that $\mathcal{A}_p$ is indeed a Stein operator. First, it can be related to the *infinitesimal generator* of the overdamped Langevin diffusion, which corresponds to replacing g in (2.6) by the gradient ∇u of a scalar-valued function u (Gorham and Mackey, 2015). Importantly, if $\mathcal{T}$

is the generator of a sufficiently regular Markov process with P as its equilibrium distribution, then $\mathbb{E}_{\mathbf{X}\sim Q}[(\mathcal{T}g)(\mathbf{X})] = 0$ for all sufficiently regular g (Barbour, 1988, 1990; Gotze, 1991). This approach of constructing Stein operators is also known as the *generator approach*. Second, that (2.6) integrates to zero under P can also be seen by applying the *Divergence Theorem* (see, e.g., Bourne, 1977, p.159), or *integration by parts*, which states that, for any bounded domains $\mathcal{X} \subset \mathbb{R}^d$ and any continuously differentiable function g,

$$\int_{\mathcal{X}} \langle \nabla,\ g(\mathbf{x})p(\mathbf{x}) \rangle \, \mathrm{d}\mathbf{x} \ = \ \oint_{\partial\mathcal{X}} \langle g(\mathbf{x})p(\mathbf{x}),\ \mathbf{n}(\mathbf{x}) \rangle S(\mathrm{d}\mathbf{x}) \ ,$$

where $\oint_{\partial\mathcal{X}}$ denotes the surface integral over the boundary $\partial\mathcal{X}$ of $\mathcal{X}$, the vector $\mathbf{n}(\mathbf{x})$ denotes the unit normal to the boundary $\mathcal{X}$, and $S(\mathrm{d}\mathbf{x})$ denotes the surface element at $\mathbf{x} \in \partial\mathcal{X}$. Using this observation, Oates et al. (2017, Lemma 1) and Chwialkowski et al. (2014, Lemma 5.1) showed that the Stein identity $\mathbb{E}_{\mathbf{X}\sim Q}[(\mathcal{T}g)(\mathbf{X})] = 0$ holds for the operator (2.6) whenever g is sufficiently regular, both when $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$ and when $\mathcal{X} = \mathbb{R}^d$.

2.2.4 Langevin Kernel Stein Discrepancy

Another design choice in Stein's discrepancies is the Stein set $\mathcal{G}$. As previously noted, $\mathcal{G}$ needs to be sufficiently rich to characterize P, but also simple enough so that the maximization objective in (2.4) can be efficiently computed. One approach to construct such $\mathcal{G}$ is to use Reproducing Kernel Hilbert Spaces (RKHS, Berlinet and Thomas-Agnan, 2004), leading to the *kernel Stein set*. Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a scalar-valued reproducing kernel with associated RKHS $\mathcal{H}_k$, and assume $k \in \mathcal{C}_b^{(1,1)} := \{\kappa \in \mathcal{C}^{(1,1)} : (\mathbf{x}, \mathbf{y}) \mapsto \partial_{1j}\partial_{2j}\kappa(\mathbf{x}, \mathbf{y}) \text{ is bounded}, \forall j = 1, \dots, d\}$. The kernel Stein set (Gorham and Mackey, 2017) is defined as

$$\mathcal{G}_k \ := \ \left\{ g = (g_1, \dots, g_d) \ : \ g \in \mathcal{H}_k,\ \sum_{j=1}^{d} \|g_j\|_{\mathcal{H}_k}^2 \leq 1 \right\} \ . \tag{2.7}$$

In words, $\mathcal{G}_k$ is the unit ball in the d-times Cartesian product $\mathcal{H}_k^d$ equipped with the norm $\|g\|_{\mathcal{H}_k^d} = \left(\sum_{j=1}^d \|g_j\|_{\mathcal{H}_k}^2\right)^{1/2}$ for $g = (g_1, \dots, g_d) \in \mathcal{H}_k^d$. Together with the Langevin Stein operator $\mathcal{A}_p$, the pair $(\mathcal{A}_p, \mathcal{G}_k)$ can be shown to satisfy the Stein identity for P whenever P finitely integrates $\|\mathbf{s}_p\|_2$ (Gorham and Mackey, 2017, Proposition 1). This is summarized in the following result.

Proposition 2.5 (Zero mean functions; Gorham and Mackey 2017, Proposition 1)**.** *Assume $k \in \mathcal{C}_b^{(1,1)}$. If $\mathbb{E}_{\mathbf{X}\sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2] < \infty$, then $\mathbb{E}_{\mathbf{X}\sim P}[(\mathcal{A}_p g)(\mathbf{X})] = 0$ for all $g \in \mathcal{G}_k$.*

Using the pair $(\mathcal{A}_p, \mathcal{G}_k)$, the Stein discrepancy is called *Langevin Kernel Stein Discrepancy* (Oates et al., 2017; Liu et al., 2016; Chwialkowski et al., 2016; Gorham and Mackey, 2017), where "Langevin" refers to the use of the Langevin Stein operator, and "kernel" emphasizes the use of RKHSs to construct the Stein set $\mathcal{G}_k$.

Definition 2.2 (Langevin kernel Stein discrepancy)**.** *Let $k \in \mathcal{C}_b^{(1,1)}$ be a scalar-valued reproducing kernel, and let P be a probability measure on $\mathbb{R}^d$ with Lebesgue density function $p \in \mathcal{C}_1$. The Langevin*

kernel Stein discrepancy between $Q \in \mathcal{P}(\mathbb{R}^d)$ *and* P *is defined as*

$$D(Q, P) \;=\; \sup_{g \in \mathcal{G}_k} \left| \mathbb{E}_{\mathbf{X} \sim Q}[(\mathcal{A}_p g)(\mathbf{X})] \right| . \tag{2.8}$$

Since $\mathcal{G}_k$ is symmetric in the sense that $g \in \mathcal{G}_k$ implies $-g \in \mathcal{G}_k$, the Langevin kernel Stein discrepancy can be defined equivalently without taking the absolute value in (2.8).

Moreover, the pair $(\mathcal{T}, \mathcal{G}_k)$ also characterize P under regularity conditions. The next result states this formally. It is a special case of Barp et al. (2022, Proposition 3.4); see also Gorham and Mackey (2017, Theorem 7). For a reproducing kernel of the form $k(\mathbf{x}, \mathbf{y}) = h(\mathbf{x} - \mathbf{y})$, where $h : \mathbb{R}^d \to \mathbb{R}$, Bochner's theorem (Berlinet and Thomas-Agnan, 2004, Theorem 20) shows that h is the Fourier transform of a unique bounded positive measure. We call such measure the *spectral measure* of h.

Theorem 2.6 (KSD P-separation)**.** *Suppose* k *is a reproducing kernel of the form* $k(\mathbf{x}, \mathbf{y}) = w(\mathbf{x})h(\mathbf{x} - \mathbf{y})w(\mathbf{y})$*, where* $h \in \mathcal{C}^2$ *has a fully supported spectral measure, and* $w : \mathbb{R}^d \to (0, \infty)$ *with* $w \in \mathcal{C}_b^1$*. Let* $P \in \mathcal{P}(\mathbb{R}^d)$ *have Lebesgue density function* $p \in \mathcal{C}^1$ *with* $p(\mathbf{x}) > 0$ *for* $\mathbf{x} \in \mathbb{R}^d$*, and assume* $\mathbb{E}_{\mathbf{X} \sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2] < \infty$*. Then for all* $Q \in \mathcal{P}(\mathbb{R}^d)$ *such that* $\mathbb{E}_{\mathbf{X} \sim Q}[\|\mathbf{s}_p(\mathbf{X})\|_2] < \infty$*, we have* $D(Q, P) = 0 \iff Q = P$ *.*

Proof of Theorem 2.6. The proof is a combination of Proposition 3.4 and Theorem 3.5 in Barp et al. (2022). Their Proposition 3.4 shows that the claim in the theorem holds when *(i)* $\mathcal{H}_k \subset \mathcal{C}_b^1(\mathbb{R}^d)$, where $\mathcal{H}_k$ is the RKHS associated with k, and *(ii)* k is $\mathscr{D}_{L^1}^1(\mathbb{R}^d)$-characteristic (Barp et al., 2022, Definition C.5). By Theorem 2.4, condition *(i)* is met whenever k has the assumed form and $h \in \mathcal{C}^2$. Moreover, combining Barp et al. (2022, Theorem 3.5.a) and the discussion before their Theorem 3.5, we conclude that the kernel $k(\mathbf{x}, \mathbf{y}) = w(\mathbf{x})h(\mathbf{x} - \mathbf{y})w(\mathbf{y})$ is $\mathscr{D}_{L^1}(\mathbb{R}^d)$-characteristic so long as the spectral measure of h is fully supported and $w \in \mathcal{C}_b^1$ is positively valued. □

Kernel assumptions The function $w(\mathbf{x})$ in the reproducing kernel is often called a *tilting function*. When $w(\mathbf{x}) \equiv 1$, the reproducing kernel takes the form $k(\mathbf{x}, \mathbf{y}) = h(\mathbf{x} - \mathbf{y})$, in which case it is called *stationary* or *translation-invariant*. Kernels of the form $k(\mathbf{x}, \mathbf{y}) = w(\mathbf{x})h(\mathbf{x} - \mathbf{y})w(\mathbf{y})$ are hence called *tilted stationary kernels*. In fact, Barp et al. (2022, Theorem 3.2) presents this result for more general kernels, and we made this restriction only for the sake of exposition. Nevertheless, tilted stationary kernels already cover most kernels used in practice, and all kernels used in the remainder of this thesis will take this form. Moreover, the assumptions on k in Theorem 2.6 are met by most commonly used kernels, such as SE kernels, IMQ kernels, and their tilted forms; see Section 2.1.3 for the definition.

KSDs with other Stein operators Langevin KSD is only one special instance of KSDs; other Stein operators can also be used to construct KSDs. One example is the *diffusion kernel Stein discrepancy* (Barp et al., 2019), which uses a Stein operator arising from the generator of a more general diffusion process. Nevertheless, the Langevin KSD has received more attention in the literature (Liu and Wang, 2016; Chwialkowski et al., 2016; Chen et al., 2018, 2019; Barp et al., 2022; Kanagawa et al., 2023).

In this thesis, we will focus exclusively on Langevin KSD, and therefore we will shorten Langevin KSD to KSD for brevity throughout the remainder of this thesis.

2.2.5 Estimation of Langevin Kernel Stein Discrepancy

A main advantage of the Langevin KSD over other statistical discrepancies is that it is computable even if p is unnormalized. This is because *(i)* the supremum in (2.8) admits a closed-form solution, and *(ii)* equation (2.8) does not depend on the normalizing constant of p.

The former has been shown in many works: when $\mathbf{s}_p \in L_1(P) \cap L_1(Q)$, the *squared* Langevin KSD $D^2(Q, P)$ admits the following double-integral form

$$D^2(Q,P) \;=\; \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}[u_P(\mathbf{X},\mathbf{X}')] \;=\; \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} u_P(\mathbf{x},\mathbf{x}')Q(\mathrm{d}\mathbf{x})Q(\mathrm{d}\mathbf{x}')\;, \tag{2.9}$$

$$u_P(\mathbf{x},\mathbf{x}') \;\coloneqq\; \mathbf{s}_p(\mathbf{x})^\top \mathbf{s}_p(\mathbf{x}')k(\mathbf{x},\mathbf{x}') + \mathbf{s}_p(\mathbf{x})^\top \nabla_2 k(\mathbf{x},\mathbf{x}') + \nabla_1 k(\mathbf{x},\mathbf{x}')^\top \mathbf{s}_p(\mathbf{x}') + \nabla_1^\top \nabla_2 k(\mathbf{x},\mathbf{x}')\;, \tag{2.10}$$

where $\nabla_1^\top \nabla_2 k(\mathbf{x},\mathbf{x}') = \sum_{j=1}^d \partial_{1j}\partial_{2j} k(\mathbf{x},\mathbf{x}')$, and $\partial_{1j}\partial_{2j}k(\mathbf{x},\mathbf{x}')$ is the partial derivative defined before Theorem 2.4, and $\nabla_i k(\mathbf{x},\mathbf{x}') = (\partial_{i1}k(\mathbf{x},\mathbf{x}'),\ldots,\partial_{id}k(\mathbf{x},\mathbf{x}'))^\top$ is the partial derivatives with respect to the i-th argument, respectively. For a proof, see, e.g., Barp et al. (2022, Corollary 2.10).

Importantly, u_P depends on p only through $\mathbf{s}_p$, which does *not* depend on the normalizing constant of p due to cancellation: for any constant $Z > 0$, we have $\nabla \log(Zp) = \nabla \log p + \nabla \log Z = \nabla \log p$. Consequently, u_P can be computed without needing to evaluate the normalizing constant Z of p. This is a key property that makes KSD a flexible tool that have led to many applications in computational statistics; see Anastasiou et al. (2023) for a review.

The double-integral form (2.9) suggests that the squared Langevin KSD can be efficiently estimated using Monte Carlo estimation. Given a random sample $\mathbb{X}_n \coloneqq \{\mathbf{X}_i\}_{i=1}^n$, where each $\mathbf{X}_i$ is drawn independently from Q, two natural estimators for $D^2(Q,P)$ are the *U-statistic*

$$D_u^2(\mathbb{X}_n) \;=\; \frac{1}{n(n-1)} \sum_{1\le i\ne j\le n} u_P(\mathbf{X}_i,\mathbf{X}_j)\;, \tag{2.11}$$

and the *V-statistic*

$$D^2(\mathbb{X}_n) \;=\; \frac{1}{n^2} \sum_{1\le i,j\le n} u_P(\mathbf{X}_i,\mathbf{X}_j)\;. \tag{2.12}$$

U-statistic versus V-statistic Both estimators are consistent. The U-statistic sums over *distinct* pairs of samples and is unbiased; consequently, it has been the more popular choice in the literature (Liu et al., 2016; Liu and Wang, 2016; Jitkrittum et al., 2017; Liu et al., 2023; Schrab et al., 2022). In comparison, the V-statistic sums over all pairs of samples and is a biased estimator, but it coincides with the squared KSD computed with the empirical measure $Q_n \coloneqq n^{-1}\sum_{i=1}^n \delta_{\mathbf{X}_i}$, i.e., $D^2(\mathbb{X}_n) = D^2(Q_n, P)$ (Gorham and Mackey, 2017), which can be a convenient property for some applications (Liu and Briol, 2024). In general, U- and V-statistics are known to have similar asymptotic properties (Grams and Serfling, 1973, Theorem 5.1) in the infinite sample limit. Leveraging the rich literature on U- and V-statistics, the asymptotic behavior of (2.11) and (2.12) can be characterized using well-established theory. A full treatment to U- and V-statistics can be found in Serfling (2009, Chapter 5 and 6).

2.2.6 Stein Reproducing Kernels

The function u_P in (2.10) is also a reproducing kernel, known as the *Stein reproducing kernel* (Oates et al., 2017). The KSD (2.8) is closely related to Maximum Mean Discrepancy (MMD) as it can be shown to be equivalent to an MMD with the Stein reproducing kernel (Hodgkinson et al., 2020; Barp et al., 2022); for the definition of MMD, see (2.2). This is summarized in the following result.

Theorem 2.7 (KSD as MMD; Barp et al. 2022, Theorem 2.8)**.** *Assume $P \in \mathcal{P}(\mathbb{R}^d)$ admits a Lebesgue density $p > 0$, and let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a positive definite kernel with associated RKHS $\mathcal{H}_k$ such that for any $g \in \mathcal{H}_k$, the function pg is differentiable. Then $\mathcal{A}_p(\mathcal{H}_k) \coloneqq \{\mathcal{A}_p g : \ g \in \mathcal{H}_k\} = \mathcal{H}_{u_P}$. If furthermore $\mathcal{A}_p(\mathcal{H}_k) \subset L_1(P)$ and $\mathbb{E}_{\mathbf{X}\sim P}[f(\mathbf{X})] = 0$ for all $f \in \mathcal{A}_p(\mathcal{H}_k)$, then*

$$D(Q, P) \ = \ \mathrm{MMD}_{u_P}(Q, P) \ , \qquad \textit{for all } Q \in \mathcal{P}(\mathbb{R}^d) \ .$$

This result highlights the close connections between KSD and MMD, and it underpines many theoretical developments in KSD. For example, this observation is used in showing the ability of KSD to control weak convergence of measures (Barp et al., 2022), in proving the convergence rate of KSD-based sample approximation methods (Shi and Mackey, 2024), and in deriving kernel goodness-of-fit tests that are robust to data contamination (Liu and Briol, 2024).

2.2.7 Other Theoretical Properties

P-separation The most fundamental property required of KSD is P-separation. One set of conditions for P-separation is presented in Theorem 2.6. These conditions are met by most kernels commonly used with KSD, including Gaussian kernel, IMQ, Laplacian, and Matérn. Furthermore, the conditions on the score function are satisfied for all targets P for which the KSD (2.8) is well-defined (Barp et al., 2022), including mixtures of Gaussians or models with strongly log-concave tails. The P-separation property is crucial for all applications of KSD, as it ensures that the KSD is a valid statistical divergence.

P-convergence control Also widely studied in the literature are the convergence properties of KSD. One such property is *P-convergence control*, meaning that, for any sequence of probability measures $\{Q_n\}_n$, the convergence $D(Q_n, P)$ implies $Q_n \to P$ weakly. Conditions under which KSD controls P-convergence are studied extensively in Barp et al. (2022); Gorham and Mackey (2017). Barp et al. (2022, Theorem 3.11) showed that, for sequences $\{Q_n\}_n$ that are *tight* (Billingsley, 1999, pp. 8), KSD based on standard translation-invariant kernels controls P-convergence if $\|\mathbf{s}_p\|_2$ grows at most root exponentially, i.e., $\mathbf{s}_p(\mathbf{x}) = \mathcal{O}(\exp(c\sum_{j=1}^d \sqrt{|x^j|}))$, for some constant $c > 0$. The P-convergence property is of interest because it is required for Stein variational gradient descent (SVGD), a technique for approximating a target distribution P by finite samples, to produce approximations that converges weakly to P (Liu et al., 2017).

Metrization of P-convergence If the converse of P-convergence control also holds, namely $D(Q_n, P) \to D(Q, P) \iff Q_n \to P$ weakly, then D is said to *metrize P-convergence*. Barp

et al. (2022, Theorem 4.8) shows that, under regularity conditions on P and with suitably constructed kernel k, the KSD *metrizes* P*-convergence*.

Convergence of polynomials Weak convergence is equivalent to convergence of the expectation of all bounded, continuous functions (Durrett, 2019, see, e.g., Theorem 3.10.1), which preclude unbounded functions such as polynomials. Polynomials are an important class of functions because their expectations are closely connected to moments of a probability measure, such as the mean and variance. Therefore, understanding how KSD controls the convergence of polynomials can be important for some applications where convergence of moments is desired, such as when using KSD to assess the approximation quality of MCMC samples (Gorham and Mackey, 2015, 2017). To this end, Kanagawa et al. (2022) studies conditions under which convergence in KSD implies convergence of polynomially growing functions.

2.3 Applications to Computational Statistics

Kernel Stein discrepancy has led to the development of numerous novel tools for tackling computational statistical problems (Anastasiou et al., 2023). This section provides a brief overview of its application in three key areas: goodness-of-fit testing, sample approximation, and parameter estimation. The remaining chapters will then be devoted to identifying and addressing challenges with these novel tools.

Throughout this section, we will restrict our attention to statistical models P that admit a Lebesgue density $p \in \mathcal{C}^1$. We will also assume regularity conditions are met so that the KSD is P-separating. The (unknown) data-generating distribution, however, needs not have a Lebesgue density.

2.3.1 Sample Approximation

Many tasks in statistical inference requires sampling from a target probability measure $P \in \mathcal{P}(\mathbb{R}^d)$. For example, in Bayesian analysis, samples drawn (approximately) from a Bayesian posterior are needed in order to estimate expectations over the posterior, such as posterior means or variances. In frequentist analysis, sampling is needed for approximating the quantiles of a distribution, or constructing confidence intervals. Since these expectations or quantiles are typically not available in closed-form, one often needs to resort to Monte Carlo approximations, where a finite sample $\{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^d$ is constructed so that the empirical measure $Q_n = n^{-1} \sum_{i=1}^n \delta_{\mathbf{x}_i}$ approximates the target measure P.

A common approach is Markov chain Monte Carlo (MCMC) algorithm (Hastings, 1970; Levin and Peres, 2017), which generate approximate samples by constructing a Markov chain that leaves the target distribution invariant. Despite being widely studied and applied, MCMC algorithms can sometimes incur considerable computational cost with complex models or large datasets (Blei et al., 2017).

An alternative is to use *optimization-based* algorithms. One well-known example is *variational inference* (VI, Blei et al., 2017), which approximates the target distribution P by distributions in a chosen class $\mathcal{Q}$ of distributions, called the *variational family*. Typically, $\mathcal{Q}$ is constructed to consist of simple distributions such that sampling from them is considerably easier. The best approximation $Q^* \in \mathcal{Q}$

can be found by minimizing a statistical divergence between Q and P. A common choice is the Kullback-Leibler (KL) divergence; that is, P is approximated by

$$Q^* \in \operatorname{argmin}_{Q \in \mathcal{Q}} D_{\mathrm{KL}}(Q, P) , \tag{2.13}$$

where $\mathbb{D}_{\mathrm{KL}}(Q, P)$ is the KL divergence defined as

$$D_{\mathrm{KL}}(Q, P) := \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} \, \mathrm{d}\mathbf{x} = \mathbb{E}_{\mathbf{X} \sim Q}[\log q(\mathbf{X})] - \mathbb{E}_{\mathbf{X} \sim Q}[\log p^*(\mathbf{X})] + \log Z ,$$

where we have assumed P and Q admit probability density functions p, q, respectively, with respect to the Lebesgue measure. In particular, the above decomposition implies that (2.13) does not depend on the normalization constant Z of $p = p^*/Z$.

Construction of the variational family $\mathcal{Q}$ can be non-trivial. It needs to be rich enough to provide a good approximation for the target P, while sufficiently simple so that sampling from any distribution in Q is efficient. One way to choose $\mathcal{Q}$ is to consider a set of distributions obtained by transporting some reference distribution via continuously differentiable mappings, i.e., $\mathcal{Q} = \{T_{\#}Q_0 : T \in \mathcal{F}\}$, where $\mathcal{F} \subset \{T : \mathbb{R}^d \to \mathbb{R}^d\}$ is some candidate class of mappings, $Q_0 \in \mathcal{P}(\mathbb{R}^d)$ is a reference distribution, and $T_{\#}Q_0$ denotes the *pushforward measure*, i.e., the distribution of $T(\mathbf{X})$ when $\mathbf{X} \sim Q_0$. The optimal map $T \in \mathcal{F}$ is then learned by solving (2.13).

If $\mathcal{F}$ is in some sense dense in the space of functions, this method can approximate any sufficiently regular target distribution, since it can be shown that for any two distributions Q_0, P satisfying regularity conditions, there always exists a map T such that $T_{\#}Q_0 = P$ (Villani, 2008). However, an overly large $\mathcal{F}$ would render the optimization problem in (2.13) computationally infeasible. An important question is hence how to choose $\mathcal{F}$ to balance the approximation accuracy while maintaining computational tractability? One approach is to restrict to a class of parametrized and invertible maps, leading to *Normalizing Flows* (NF, Rezende and Mohamed, 2015; Marzouk et al., 2016). However, even in this case, choosing a suitable parametric family to achieve both accuracy and computational efficiency is still non-trivial.

An alternative is proposed in Liu and Wang (2016), whereby $\mathcal{F}$ is chosen to be a *non-parametric* family. This algorithm, known as *Stein variational gradient descent* (SVGD), seeks a mapping of the form $T(\mathbf{x}) = T^{\phi}(\mathbf{x}) = \mathbf{x} + \epsilon\phi(\mathbf{x})$, where ϕ is selected to optimally transport towards the target p among all vector fields in a reproducing kernel Hilbert space (RKHS). SVGD is based on the key observation that, in this case, the optimal decay direction for the KL divergence coincides with the expectation $\mathbb{E}_{\mathbf{X} \sim Q}[(\mathcal{A}_p\phi)(\mathbf{X})]$ of the Langevin Stein operator. This is made precise in the following result; see also Anastasiou et al. (2023, Theorem 1).

Theorem 2.8 (Liu and Wang 2016, Theorem 3.1)**.** *Assume Q, P admit positive density functions q, p on $\mathbb{R}^d$, and $p \in \mathcal{C}^1$. For any $\epsilon \in \mathbb{R}$ and $\phi \in \mathcal{C}^1(\mathbb{R}^d)$ with $\sup_{\mathbf{x} \in \mathbb{R}^d} \|\nabla g(\mathbf{x})\|_2 < \infty$, where $\|\cdot\|_2$ denotes the spectral norm, define the function $T^{\phi}(\mathbf{x}) = \mathbf{x} + \epsilon\phi(\mathbf{x})$. Then*

$$-\frac{\mathrm{d}}{\mathrm{d}\epsilon} D_{\mathrm{KL}}(T^{\phi}_{\#}Q, P)\Big|_{\epsilon=0} = \mathbb{E}_{\mathbf{X} \sim Q}[(\mathcal{A}_p\phi)(\mathbf{X})] .$$

Algorithm 1 Stein Variational Gradient Descent (SVGD; Liu and Wang, 2016).

Inputs: n initialized particles $\{\mathbf{x}_i^0\}_{i=1}^n$; iteration number n_{epochs}; step size $\epsilon > 0$.
for $t = 0, 1, \ldots, n_{\text{epochs}} - 1$ **do**
    Compute $\mathbf{x}_i^{t+1} = \mathbf{x}_i^t + \epsilon\phi(\mathbf{x}_i^t)$ for $i = 1, \ldots, n$, where

$$\phi(\mathbf{x}_i^t) \;=\; \frac{1}{n}\sum_{j=1}^{n}\big[\mathbf{s}_p(\mathbf{x}_j^t)k(\mathbf{x}_j^t, \mathbf{x}_i^t) + \nabla_1 k(\mathbf{x}_j^t, \mathbf{x}_i^t)\big].$$

end for
Return: Particles $\{x_i^{n_{\text{epochs}}}\}_{i=1}^n$.

The LHS of the above equality is the decay rate of the KL divergence for an infinitesimally small movement along the vector field ϕ, while the LHS appears in the objective of Langevin kernel Stein discrepancy (2.8) (with g replaced by ϕ). Therefore, Theorem 2.8 outlines the connection between the KL divergence and the Langevin KSD—by maximizing this decay rate over all vector fields ϕ in the unit ball $\mathcal{G}_k$ of an RKHS defined in (2.7), we have

$$\sup_{\phi\in\mathcal{G}_k}\left\{-\frac{\mathrm{d}}{\mathrm{d}\epsilon}D_{\mathrm{KL}}(T_{\#}^{\phi}Q, P)\Big|_{\epsilon=0}\right\} \;=\; \sup_{\phi\in\mathcal{G}_k}\mathbb{E}_{\mathbf{X}\sim Q}[(\mathcal{A}_p\phi)(\mathbf{X})] \;=\; D(Q, P)\;,$$

where $D(Q, P)$ is the Langevin KSD defined in (2.8). This also suggests that the optimal vector field that attains the supremum coincides with the function that attains the supremum in the KSD objective, called the *witness function* of the KSD, which has the following closed-form expression

$$\phi^*(\cdot) \;\propto\; \mathbb{E}_{\mathbf{X}\sim Q}[\mathbf{s}_p(\mathbf{X})k(\mathbf{X}, \cdot) + \nabla_1 k(\mathbf{X}, \cdot)]\;.$$

In practice, $Q = n^{-1}\sum_{i=1}^n \delta_{\mathbf{x}_i}$ is an empirical measure of n particles $\{\mathbf{x}_i\}_{i=1}^n$, which can be initialized from any distributions, such as the standard Gaussian. SVGD then iteratively updates each particle $\mathbf{x}_i$ along the optimal transport direction ϕ^* by $\mathbf{x}_i + \epsilon\phi^*(\mathbf{x}_i)$, where the expectation in ϕ^* is replaced by a Monte Carlo estimation. This algorithm is summarized in Algorithm 1. Compared to NFs, SVGD does not require a parametric family of mappings to be explicitly specified, making it readily implementable even for non-experts in VI.

SVGD is a *deterministic* interacting particle system, as the update step is deterministic given the particles. The two terms in the transport map g^* play different roles. The first term $\mathbf{s}_p(\mathbf{X})k(\mathbf{X}, \cdot)$ drives the particles towards the high probability regions of P due to the gradient term $\mathbf{s}_p(\mathbf{X})$. The second term $\nabla_1 k(\mathbf{x}, \cdot)$ is a repulsive force keeping the particles dispersed. This is because the update $\mathbf{x} \mapsto \mathbf{x} + \epsilon\nabla_1 k(\mathbf{x}, \mathbf{y})$ would decrease the value of $k(\mathbf{x}, \mathbf{y})$, which can be viewed as a measure of similarity between $\mathbf{x}$ and $\mathbf{y}$. This repulsion mechanism is crucial for preventing mode collapse: without the second term and if only one particle is used, the SVGD update would reduce to finding the maximum a posteriori (MAP) point of p. Consequently, SVGD can also be viewed as an interpolation between the MAP and a full particle-based approximation (Liu and Wang, 2016).

Convergence of SVGD The convergence of SGVD as $n_{\text{epochs}} \to \infty$ has been well understood. Liu (2017, Theorem 3.3) first proved the convergence of SVGD to the target distribution, assuming

the algorithm is initialized with a *continuous* distribution rather than a finite sample. This result is later refined by Korba et al. (2020); Salim et al. (2022); Sun et al. (2023a), who established explicit rates of convergence, path-independent constants and weaker regularity conditions on the target P. Additionally, Duncan et al. (2023) investigated conditions for exponential convergence. However, these results are rather restricted because they assume an infinite sample size ($n \to \infty$), or an infinitesimal step size ($\epsilon \to 0$), or both. In the literature, the algorithm under these regimes are sometimes called the *infinite-particle, continuous-time SVGD* (Shi and Mackey, 2024; Balasubramanian et al., 2024), which is different from the *n-particle, discrete-time* algorithm used in practice. More recently, finite-sample, discrete-time rates under mild conditions on P are derived in Shi and Mackey (2024); Balasubramanian et al. (2024), with the latter showing a near-optimal rate of $1/\sqrt{n}$.

Extensions of SVGD Numerous extensions of SVGD have been proposed. Amortized SVGD (Feng et al., 2017) is a computationally efficient alternative when inference with multiple target distributions is needed. Gradient-free variants are developed for target densities whose score functions are unavailable or expensive to evaluate (Han and Liu, 2018; Yan and Zou, 2021). Other improvements include extensions for discrete spaces (Han et al., 2020), graphical models (Zhuo et al., 2018; Wang et al., 2018), and Riemannian manifolds (Liu and Zhu, 2018). SVGD has been applied to a wide range of problems, such as learning Variational Autoencoders (Pu et al., 2017), meta learning (Pu et al., 2024), reinforcement learning (Liu et al., 2017; Haarnoja et al., 2017; Liu et al., 2020), uncertainty quantification (Xie et al., 2024), anomaly detection (Chen et al., 2023), and science and engineering (Morelli et al., 2023; Liu et al., 2024b; Zhao et al., 2024; Liu et al., 2024a).

One limitation of SVGD is its poor performance when the target distribution P lies in a high-dimensional space. Specifically, it is found that in this case, the particles generated by SVGD can collapse onto the modes of the target, an issue known as *variance collapse* (Ba et al., 2022). In Chapter 3, we will dive into this problem and propose an extension that can mitigate this issue.

2.3.2 Goodness-of-Fit Testing

Goodness-of-fit (GOF) testing (Lehmann and Romano, 2022, Chapter 16; D'Agostino, Ralph B, 2017) aims to assess how well a model fits some given data. More formally, given a model P and observations drawn from some distribution Q, GOF testing compares the null hypothesis $H_0 : Q = P$ against the alternative hypothesis $H_1 : Q \neq P$. Addressing this question is crucial before using the model for predictions, decision-making, or providing probabilistic guarantees on key quantities of interest.

To test such hypotheses, one approach is to compute a statistical divergence $\mathbb{D}$ between the data-generating distribution Q and the model P. If the divergence satisfies $\mathbb{D}(Q, P) = 0 \iff Q = P$, then testing $H_0 : Q = P$ is equivalent to testing $H_0 : \mathbb{D}(Q, P) = 0$. If furthermore $\mathbb{D}(Q, P)$ can be estimated given observed data, and that the null distribution of its finite-sample estimator is available or can be approximated, then the GOF test can be performed whereby H_0 is rejected for large estimates of $\mathbb{D}(Q, P)$. One example of such GOF tests is the Kolmogorov-Smirnov test, where $\mathbb{D}$ is the Kolmogorov-Smirnov distance (Kolmogorov, 1933). Others include tests based on KL divergence (Vasicek, 1976; Song, 2002) or phi-divergences (Noughabi and Balakrishnan, 2016; Jager and Wellner, 2007).

Unfortunately, when P is unnormalized, many well-known statistical divergences cannot be estimated efficiently. This is the case for KL divergence, phi-divergences, Wasserstein distance (Del Barrio et al., 1999), and Maximum Mean Distance (Gretton et al., 2012). Tests based on these divergences require either a known normalizing constant, or a large number of samples to approximate the model P; see also the discussion in Section 2.2.1. However, when P is unnormalized, computing its normalizing constant can be challenging, while sampling from P can be computationally expensive and can introduce extra errors. Kolmogorov-Smirnov test, on the other hand, are restricted to univariate distributions.

One feasible choice for $\mathbb{D}$ is the KSD. As explained in Section 2.2.5, the KSD $D(Q, P)$ can be computed when P is unnormalized and samples from Q are available. Based on this observation, Liu et al. (2016); Chwialkowski et al. (2016) proposed a GOF test, called the *KSD test*, which we describe next.

GOF test based on KSD Given independent random variables $\mathbb{X}_n = \{\mathbf{X}_i\}_{i=1}^n$ drawn from some Q, the KSD test rejects $H_0 : Q = P$ for $D^2(\mathbb{X}_n) > \gamma$, where $D^2(\mathbb{X}_n)$ is the V-statistic estimate (2.12) and γ is a decision threshold chosen to control the Type-I error of the test. The value of γ can be determined by considering the null distribution of $D^2(\mathbb{X}_n)$, but this distribution is not usually tractable. Instead, bootstrap procedures are often used for approximation. One procedure uses the bootstrap samples of the form

$$D^2_{\mathbf{W}}(\mathbb{X}_n) \;=\; \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(W_i - 1)(W_j - 1)u_P(\mathbf{X}_i, \mathbf{X}_j)\;, \tag{2.14}$$

where $\mathbf{W} \coloneqq (W_1, \ldots, W_n) \sim \text{Multinomial}(n; 1/n, \ldots, 1/n)$. This bootstrap procedure proceeds by *(i)* drawing B i.i.d. copies $\{\mathbf{W}^b\}_{b=1}^B$ of $\mathbf{W}$, *(ii)* computing $D^2_{\mathbf{W}^b}(\mathbb{X}_n)$ for each $\mathbf{W}^b$, and *(iii)* approximating the p-value as

$$p(\mathbb{X}_n) \;=\; \frac{1}{B+1}\left(1 + \sum_{b=1}^{B}\mathbb{1}\{D^2_{\mathbf{W}^b}(\mathbb{X}_n) > D^2(\mathbb{X}_n)\}\right)\;, \tag{2.15}$$

where $\mathbb{1}\{A\}$ denotes the indicator function for the event A. This procedure is called *Efron's bootstrap* or *weighted bootstrapping* (Arcones and Gine, 1992; Janssen, 1994). Intuitively, it weighs each observation by a multinomial weight, which mimics recomputing the test statistic using data re-sampled from $\mathbb{X}_n$ without replacement. The KSD test is summarized in Algorithm 2.

Given a prescribed test level $\alpha \in (0, 1)$, the KSD test rejects H_0 if $p(\mathbb{X}_n) < \alpha$. This is equivalent to rejecting H_0 if $D(\mathbb{X}_n) > q^2_{B,1-\alpha}(\mathbb{X}_n)$, where the test threshold $q^2_{B,1-\alpha}(\mathbb{X}_n)$ is the $(1-\alpha)$-quantile of $\{D^2_{\mathbf{W}^b}(\mathbb{X}_n)\}_{b=1}^B$ conditional on $\mathbb{X}_n$, i.e.,

$$q^2_{B,1-\alpha}(\mathbb{X}_n) \;=\; \inf\left\{u \in \mathbb{R} :\; 1 - \alpha \le \frac{1}{B+1}\left(1 + \sum_{b=1}^{B}\mathbb{1}\{D^2_{\mathbf{W}_n^b}(\mathbb{X}_n) \le u\}\right)\right\}.$$

The KSD test is asymptotically well-calibrated for $H_0 : Q = P$ (Liu et al., 2016, Theorem 4.3), and when KSD is P-separating, this test is *consistent*, meaning that the probability of rejection approaches one as the sample size $n \to \infty$ whenever $Q \neq P$ (Chwialkowski et al., 2016; Liu et al., 2016).

Algorithm 2 KSD test for goodness-of-fit (Chwialkowski et al., 2016; Liu et al., 2016).

1: **Input:** Data $\mathbb{X}_n = \{\mathbf{x}_i\}_{i=1}^n$; bootstrap sample size B; test level $\alpha \in (0, 1)$.
2: Compute the test statistic $D^2(\mathbb{X}_n)$ by (2.12).
3: **for** $b = 1, 2, \ldots, B$ **do**
4: Draw $\mathbf{W}^b \sim \text{Multinomial}(n; 1/n, \ldots, 1/n)$ and compute the bootstrap statistic $D^2_{\mathbf{W}^b}(\mathbb{X}_n)$ by (2.14).
5: **end for**
6: Compute the p-value $p(\mathbb{X}_n)$ by (2.15), and reject $H_0 : Q = P$ if $p(\mathbb{X}_n) < \alpha$.

Other bootstrap procedures Other bootstrap procedures such as wild bootstraps (Leucht and Neumann, 2013; Shao, 2010) have also been used to approximate the p-value for KSD tests (Jitkrittum et al., 2017; Schrab et al., 2022; Liu et al., 2023). Compared with Efron's bootstrap, wild bootstraps have the advantage that they can be tuned to work for dependent data (Chwialkowski et al., 2016). However, for independent data, which will be the focus of this thesis, both Efron's bootstrap and wild bootstrap have similar empirical performance; see Section 5.B.7 for a comparison.

U- and V-statistics in KSD testing We have described the KSD test where the test statistic is the V-statistic (2.12). The same test procedure works with the U-statistic estimate (2.11) by replacing the indices in all double sums with *distinct* indices $i \neq j$. Both statistics are used in the literature of KSD testing; see Liu et al. (2016); Jitkrittum et al. (2017); Schrab et al. (2022) for U-statistics, and Chwialkowski et al. (2016); Schrab and Kim (2024); Liu and Briol (2024) for V-statistics. Empirically, the performance when using the U- or V-statistic is similar, as we will verify numerically in Section 5.C. From the theoretical view point, however, the non-negativity of the V-statistic might be convenient for some applications such as developing robust variants of the KSD test, which is discussed in detail in Chapter 5.

Applications KSD tests and their extensions have found success across a wide range of data types, including correlated MCMC samples (Chwialkowski et al., 2016; Liu et al., 2016), time-to-event data (Fernandez et al., 2020), directional data (Xu and Matsuda, 2020), Riemannian manifold data (Xu and Matsuda, 2021), discrete data (Yang et al., 2018), network data (Xu and Reinert, 2021), functional data (Wynne et al., 2022), point processes (Yang et al., 2019), and latent variable models (Kanagawa et al., 2023). Advances have also been made to address some of the limitations of KSD tests, including high computational cost for large samples (Jitkrittum et al., 2017; Huggins and Mackey, 2018), the need to evaluate the gradients of the model likelihood, which can be impractical for complex models (Afzali and Muthukumarana, 2023; Fisher and Oates, 2023), and the challenge of tuning kernel bandwidths (Schrab et al., 2022).

In Chapter 4 and 5, we identify two additional limitations of KSD tests, namely the lack of test power against multimodal alternatives and the lack of robustness to data contamination. Both scenarios are common in practice. Statistical models with multiple modes are ubiquitous in Bayesian modelling for localization problems (Ahn et al., 2013), astrophysics (Tak et al., 2018), and geophysics (de Figueiredo et al., 2019). On the other hand, real-world data often subject to contamination, such as outliers from erroneous measurements or noises from corrupted sensors. These two chapters will formalize these limitations and propose new variants of the KSD test to address these issues.

Algorithm 3 Minimum KSD estimation (Barp et al., 2019).

1: **Input:** Data $\mathbb{X}_n = \{\mathbf{x}_i\}_{i=1}^n$; parametrized model $\mathcal{M}(\Theta)$; reproducing kernel k.
2: Find an estimator $\theta_n^{\mathcal{L}}$ satisfying $\theta_n^{\mathcal{L}} \in \operatorname{argmin}_{\theta\in\Theta} \sum_{1\le i,j\le n} u_{P_\theta}(\mathbf{x}_i, \mathbf{x}_j)$, either by numerical optimization or by closed-form solution.

2.3.3 Parameter Estimation

A third application of statistical divergences is in parameter estimation. Let $\Theta \subset \mathbb{R}^m$ be a parameter space, where m is a positive integer, and let P_θ be a parametric model indexed by a parameter $\theta \in \Theta$. The goal of parameter estimation is to find $\theta^* \in \Theta$ such that P_{θ^*} best approximates the underlying distribution of some observed data.

One natural approach to find the optimal parameter θ^* is to minimize some statistical divergence between the data-generating distribution Q and the model. Often in practice, Q is observed only via finite samples, so a finite-sample estimate for the divergence, called a *loss function*, is used instead. This principle is known as *minimum divergence estimation* or *minimum distance estimation* (MDE, Pardo, 2005; Basu et al., 2011). Formally, given i.i.d. samples $\mathbb{X}_n = \{\mathbf{X}_i\}_{i=1}^n$ drawn from some $Q \in \mathcal{P}(\mathbb{R}^d)$ and given a loss function $\mathcal{L} : (\mathbb{R}^d)^n \times \Theta \to \mathbb{R}$, the MDE of a model $\{P_\theta : \theta \in \Theta\}$ is defined as an estimator $\theta_n^{\mathcal{L}}$ satisfying

$$\theta_n^{\mathcal{L}} \in \operatorname{argmin}_{\theta\in\Theta} \mathcal{L}(\mathbb{X}_n, \theta) .$$

For specific model families and loss functions, such an estimator can be derived using a closed-form solution. For more general cases, numerical optimization can be used to compute this estimator.

One famous example of MDE is the *maximum likelihood estimator* (MLE), which corresponds to choosing the Kullback-Leibler (KL) divergence as the statistical divergence. The loss function in this case is the negative log-likelihood $\mathcal{L}(\{\mathbf{X}_i\}_{i=1}^n, \theta) = -\sum_{i=1}^n \log p_\theta(\mathbf{X}_i)$. Other statistical divergences or distances used in MDE include the Maximum Mean Discrepancy (MMD, Dziugaite et al., 2015) and Wasserstein metric (Frogner et al., 2015).

One limitation of MDE using the aforementioned divergences is that they might no longer be feasible when P_θ belongs to an unnormalized model class. Indeed, to compute the KL divergence, the normalizing constant of P_θ needs to have a closed-form expression for any given parameter values, which is not the case for unnormalized models. On the other hand, MMD and Wasserstein metric require extra samples from the model, but generating samples from unnormalized models can be computationally challenging.

One solution is to use statistical divergences that depend on the model only through its score function $\mathbf{s}_{p_\theta}(\mathbf{x}) \coloneqq \nabla_{\mathbf{x}} \log p_\theta(\mathbf{x})$. As discussed in Section 2.2.5, the score $s_{p_\theta}(\mathbf{x})$ does not depend on the normalizing constant of p_θ. This allows minimum divergence estimators based on such statistical divergences to be computed even for unnormalized models, where conventional likelihood-based approaches like MLE are out of reach.

Two most representative score-based divergences are the *Fisher divergence*, also known as the *score matching* (SM) loss, and the kernel Stein discrepancy (KSD). Minimum divergence estimators based on

these divergences are called *score-matching estimators* (Hyvärinen, 2005, 2007) and *minimum kernel Stein estimators* (Barp et al., 2019), respectively. Various extensions of these estimators have also been developed. For example, Barp et al. (2019) extended the SM loss and the KSD by incorporating a preconditioning term called the *diffusion matrix*, which can enforce robustness of the resulting estimator. Song and Ermon (2019); Song et al. (2020) trained generative models using an annealed version of the SM loss, whereby the data and the model are convolved with Gaussian noises at different temperatures. Qin and Risteski (2023) proposed a generalized SM loss, which unifies the standard, preconditioned and annealed SM losses.

Algorithm 3 summarizes minimum KSD estimation, where the loss function is chosen to be the V-statistic defined in (2.12). Alternatively, the U-statistic estimate (2.11) can be used instead.

In Chapter 6, we will compare the relative *statistical efficiency* of the score-matching estimators and the minimum KSD estimators. The statistical efficiency of an estimator can be quantified by its asymptotic variance under a Central Limit Theorem (CLT), where a larger covariance indicates less efficiency. Several works have studied the statistical efficiency of SM estimators (Koehler et al., 2022; Qin and Risteski, 2023); however, the efficiency of the minimum KSD estimators has not been explored. In Chapter 6, we will provide both theoretical and numerical results to compare the efficiency of these two estimators. Our findings will demonstrate that KSD estimators can be significantly less efficient than SM estimators when the kernel is not properly chosen, highlighting the importance of kernel selection in using KSD for parameter estimation.

Chapter 3

Sample Approximation: Grassmann Stein Variational Gradient Descent

Stein variational gradient descent (SVGD) is a deterministic particle inference algorithm that provides an efficient alternative to Markov chain Monte Carlo. However, SVGD has been found to suffer from variance underestimation when the dimensionality of the target distribution is high. Recent developments proposed to sidestep this issue by using projections, where both the score function and the data are projected onto one-dimensional subspaces. In this chapter, we show empirically that this can severely overestimate the epistemic (model) uncertainty. We then propose *Grassmann Stein variational gradient descent* (GSVGD) as an alternative approach, which permits projections onto subspaces with an arbitrary dimension. GSVGD updates the projectors simultaneously for the score function and the data, and the optimal projectors are determined through a coupled Grassmann-valued diffusion process which explores favorable subspaces. Both our theoretical and experimental results suggest that GSVGD enjoys efficient state-space exploration in high-dimensional problems that have an intrinsic low-dimensional structure.

3.1 Introduction

Variational inference (VI) (Blei et al., 2017) is an optimization-centric framework for approximating complex distributions that are intractable (only known up to a scale factor): given any target distribution with probability density function p, VI searches over a user-defined class of distributions $\mathcal{Q}$ for an optimal $q \in \mathcal{Q}$ that is closest (in terms of a discrepancy or divergence) to p. Extensively applied in the field of Bayesian inference, VI has the attraction of being scalable to big datasets, although it leads to biased estimation unless $\mathcal{Q}$ is broad enough to include the target distribution. This is in contrast to Markov chain Monte Carlo (MCMC) algorithms (Gilks et al., 1995), which allows asymptotically exact sampling from the true target distribution, but does not scale well to big datasets and high dimensionality due to long mixing times (Levin and Peres, 2017).

The choice of $\mathcal{Q}$ in VI is crucial to guarantee a good approximation to p while retaining computational tractability. The classical mean-field approximation, which assumes that the distributions in $\mathcal{Q}$ have

independent marginals, can be overly simplistic in many cases. To address this, a growing line of work in VI with Normalizing Flows (NFs) (Rezende and Mohamed, 2015) seeks to construct an invertible map T so that the pushforward distribution of $T(\mathbf{X})$, with $\mathbf{X} \sim q$ and $q \in \mathcal{Q}$, will form a flexible approximation to p.

Stein variational gradient descent (SVGD) (Liu and Wang, 2016) was introduced as a particle-based variational inference method in which the map T seeks to move a set of particles along a vector field which is chosen within a reproducing kernel Hilbert space (RKHS) to optimally transport towards the target p. Building on the kernel Stein discrepancy (Liu and Wang, 2016), the SVGD transport both drives the particles to the high-probability regions of p and enforces repulsion between the particles to prevent mode-collapse. It hence has the advantage of being "particle-efficient" in that even a small number of particles can achieve good approximation of p. Although this has made SVGD a popular tool in a range of applications including meta-learning (Yoon et al., 2018) and learning diversified mixture models (Wang and Liu, 2019), it is found that the marginal variance of the resulting particles scales inversely with the dimension, resulting in under-estimation of the variance. This was studied analytically in Ba et al. (2022), in which the authors attribute this issue to *(i)* an uneven scale of variance for the two terms in the SVGD update, and *(ii)* a deterministic bias due to the dependence of the particle positions at each step.

Recently, Gong et al. (2021a) proposed a *sliced* version of kernel Stein discrepancy (KSD) analogous to other sliced discrepancies (Kolouri et al., 2019), and an associated *sliced SVGD* (S-SVGD). This variant of SVGD decomposes the kernel-based dynamics along a sequence of 1-dimensional projections, called *slices*. Empirical results (Gong et al., 2021a, Appendix J.5, Table 5) showed that the S-SVGD dynamic can mitigate the variance under-estimation issue of SVGD in high dimensions. However, S-SVGD is constrained to 1-dimensional projection, and only seeks the optimal slices for the data but keeps the slices for the score function deterministic. As we demonstrate in this chapter, these can lead to inflated variance of the S-SVGD estimation.

In this chapter, we introduce a unified approach where a non-uniform probability distribution over the projectors for *both the score function and the data* are adaptively updated to emphasize directions in which there is the largest discrepancy. This is achieved by introducing a stochastic diffusion process taking values in the Grassmann manifold (Bendokat et al., 2020) which evolves along with SVGD particles. By tuning the diffusion process parameters we can adjust the trade-off between exploration and exploitation of suitable projections. In addition, using projections onto higher dimensional spaces allows us to take into account the correlations/interactions between components and thus producing more accurate uncertainty estimates.

Contributions: Motivated by S-SVGD, we propose a novel algorithm, Grassmann SVGD (GSVGD), which employs m-dimensional projectors (compared with the 1-dimensional ones in S-SVGD) to evolve the particles towards the target distribution and to mitigate variance under-estimation. We show numerically that our method is competitive to SVGD and S-SVGD on high-dimensional synthetic and benchmark problems while more accurately estimating the epistemic uncertainty.

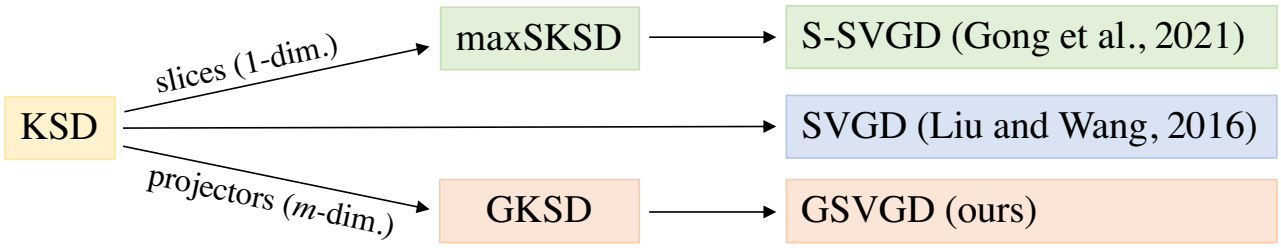


Figure 3.1: A summary of different SVGD algorithms. SVGD uses no projections, S-SVGD uses 1-dimensional projections, and GSVGD (proposed method) uses m-dimensional projections with an arbitrary $1 \leq m \leq d$, where d is the dimensionality of the problem.

3.2 Background

Stein Variational Gradient Descent Let $\mathcal{X} = \mathbb{R}^d$ and P be a probability measure over $\mathcal{X}$ with continuously differentiable density p which is positive on $\mathcal{X}$ and can be evaluated up to a normalization constant. We are interested in approximating P by transporting a known measure Q, defined over $\mathcal{X}$ and with continuously differentiable density q, to P via a sequence of maps that minimize a given loss.

Choosing the loss to be the Kullback-Leibler (KL) divergence defined as

$$\mathrm{KL}(Q, P) \;=\; \int_{\mathcal{X}} q(\mathbf{x}) \log(q(\mathbf{x})/p(\mathbf{x}))\, \mathrm{d}\mathbf{x}$$

allows us to minimize the discrepancy $\mathrm{KL}(T_{\#}Q, P)$ over maps $T\colon \mathcal{X} \to \mathcal{X}$ where $T_{\#}Q$ is the pushforward of Q with respect to T. In Liu and Wang (2016), the authors choose a specific parametrization $T(\mathbf{x}) = \mathbf{x} + \varepsilon\phi(\mathbf{x})$ where ϕ lies within the unit ball $\mathcal{B}_k^d := \{\phi \in \mathcal{H}_k^d : ||\phi||_{\mathcal{H}_k^d} \leq 1\}$ of the Cartesian product $\mathcal{H}_k^d := \{\phi = (\phi_1, \ldots, \phi_d) : \phi_i \in \mathcal{H}_k, \forall i\}$ of the RKHS $\mathcal{H}_k$ associated with kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The crucial observation of Liu and Wang (2016) is that the maximal rate of decay of the KL divergence with respect to ϵ is the kernel Stein discrepancy (KSD) (Chwialkowski et al., 2016; Liu et al., 2016), which is defined in (2.8) and recalled below

$$D(Q, P) \;=\; \sup_{\phi \in \mathcal{B}_k^d} \mathbb{E}_{\mathbf{X}\sim Q}[\mathcal{A}_p\phi(\mathbf{X})]\,, \tag{3.1}$$

where $\mathbf{X} \sim Q$, the function $\mathbf{s}_p(\mathbf{x}) = \nabla \log p(\mathbf{x}) = \nabla_{\mathbf{x}} \log p(\mathbf{x})$ is the *score function* and $\mathcal{A}_p\phi(\mathbf{x}) = \mathbf{s}_p(\mathbf{x})^\top \phi(\mathbf{x}) + \nabla^\top \phi(\mathbf{x})$ is the *Stein operator* associated to P. The key advantage of KSD is that it quantifies the difference between two measures P and Q where only the score function of P is available, which is the typical situation in Bayesian inference for complex models. The supremum in (3.1) is attained by the function $\phi^*(\cdot) := \mathbb{E}_{\mathbf{X}\sim Q}[\mathcal{A}_p k(\cdot, \mathbf{X})] = \mathbb{E}_{\mathbf{X}\sim Q}[k(\cdot, \mathbf{X})\mathbf{s}_p(\mathbf{X}) + \nabla_2 k(\cdot, \mathbf{X})]$, where $\nabla_2 k(\cdot, \mathbf{X})$ denotes the gradient with respect to the second argument. This leads to the result (Liu and Wang, 2016, Theorem 3.3) $\nabla_\phi \mathrm{KL}(T_{\#}Q, P)|_{\phi=0} = -\phi^*(\mathbf{x})$ where ∇_ϕ is the functional derivative and $T(\mathbf{x})$ is the map parameterized as above. This suggests that ϕ^* is, according to KL divergence, the best choice of function to use in T to map Q to P.

The expression for the optimal ϕ^* forms the basis of the SVGD algorithm. Particles $\mathbf{x}^t = (\mathbf{x}_1^t, \ldots, \mathbf{x}_N^t)$

at step t, are evolved using the following update rule

$$\mathbf{x}_i^{t+1} \;=\; \mathbf{x}_i^t + \epsilon\widehat{\phi}^*(\mathbf{x}_i^t), \tag{3.2}$$

for $i = 1, \ldots, N$ and $t = 0, 1, 2, \ldots$, where

$$\widehat{\phi}^*(\cdot) \;=\; \frac{1}{N}\sum_{j=1}^{N}\left[k(\cdot, \mathbf{x}_j^t)\nabla \log p(\mathbf{x}_j^t) + \nabla_2 k(\cdot, \mathbf{x}_j^t)\right] \tag{3.3}$$

is an estimator of $\phi^*(\cdot)$ obtained by replacing Q with $\frac{1}{N}\sum_{i=1}^N \delta_{\mathbf{x}_i^t}$. Intuitively, (3.2) is an Euler discretization with step-size ϵ of the following system of ODEs

$$\frac{\mathrm{d}\mathbf{x}_i}{\mathrm{d}t}(t) \;=\; \frac{1}{N}\sum_{j=1}^{N}\left[k(\mathbf{x}_i(t), \mathbf{x}_j(t))\nabla \log p(\mathbf{x}_j(t)) + \nabla_2 k(\mathbf{x}_i(t), \mathbf{x}_j(t))\right], \quad i = 1, \ldots, N,$$

for particle positions $\mathbf{x}_1(t), \ldots, \mathbf{x}_N(t)$. Under regularity conditions, SVGD converges to the target P as $N \to \infty$ and $t \to \infty$; see Section 2.3.1 for a discussion.

Tackling the Curse of Dimensionality The SVGD algorithm has shown to be asymptotically exact, in the sense that in the limit of infinite particles and vanishing step-size the particles will converge to the target density (Lu et al., 2019). However, in the finite particle regime, SVGD may fail to adequately explore the state-space, exhibiting mode collapse and variance under-estimation, particularly in high dimensions (Liu et al., 2016; Zhuo et al., 2018; Gong et al., 2021a).

The deterioration in performance in high-dimensions can be understood from the update rule defined in (3.3). The RHS consists of two terms: a kernel-averaged score function $\frac{1}{N}\sum_j k(\cdot, \mathbf{x}_j^t)\nabla \log p(\mathbf{x}_j^t)$ which attracts particles towards modes of the target density, and a repulsion term $\frac{1}{N}\sum_j \nabla_2 k(\cdot, \mathbf{x}_j^t)$ which encourages particle diversity by pushing nearby particles away from each other. As has been observed in Liu and Wang (2016, Section 3.2) and studied analytically in (Zhuo et al., 2018, Section 3), the influence of the repulsion term drops dramatically with increasing dimension, effectively reducing SVGD to a gradient ascent method for $\log p$. As a result, the estimation of SVGD can suffer from a significant under-estimated variance for high-dimensional targets.

Various developments have been proposed to mitigate this issue; we now discuss the most relevant to our work, the sliced approach of Gong et al. (2021a), and defer a review of the others to Section 3.3. Gong et al. (2021a) addressed this problem by projecting both the particles and the score function along one-dimensional directions, known as *slices*, leading to the *max sliced kernel Stein discrepancy* (maxSKSD). This approach hinges on the fact that measuring KSD along finitely many slices is sufficient to capture all geometric information of two distributions, as long as optimal slices are used. Given a user-defined orthonormal basis O in $\mathbb{R}^d$, the *maxSKSD* takes the form

$$\begin{aligned}\mathrm{maxSKSD}(Q, P) \;&=\; \sum_{r\in O}\sup_{g_r\in\mathbb{S}^{d-1}}\sup_{\phi\in\mathcal{B}_k}\mathbb{E}_{\mathbf{X}\sim Q}[\mathbf{s}_p^r(\mathbf{X})\phi(\mathbf{X}^\top g_r) + r^\top g_r \nabla\phi(\mathbf{X}^\top g_r)] &(3.4)\\ &=\; \sum_{r\in O}\sup_{g_r\in\mathbb{S}^{d-1}}\sup_{\phi\in\mathcal{B}_{k_{r,g_r}}}\mathbb{E}_{\mathbf{X}\sim Q}[\mathbf{s}_p(\mathbf{X})^\top\phi(\mathbf{X}) + \nabla\cdot\phi(\mathbf{X})]\,, &(3.5)\end{aligned}$$

where $s_p^r(\mathbf{x}) \coloneqq \mathbf{s}_p(\mathbf{x})^\top r$ is the score projected along a slicing direction r and $\mathcal{B}_k$ is the unit ball of from RKHS $\mathcal{H}_k$ for some kernel $k\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}$. The subscript g_r is used to emphasize that for each r a new optimal g_r must be found. In (3.5), $\mathcal{B}_{k_{r,g_r}}$ is the unit ball of the RKHS corresponding to the matrix valued kernel $k_{r,g_r}(\mathbf{x},\mathbf{y}) = rr^\top k(g_r^\top\mathbf{x}, g_r^\top\mathbf{y})$. This can be seen from standard results regarding matrix valued kernels (Paulsen and Raghupathi, 2016, Chapter 6) and this will be relevant to our later developments.

The idea of the maxSKSD approach is that the $r \in O$ will slice the score function of the target, and for each of those slices the input is sliced according to some other directions g_r. It is known that, under some standard regularity conditions, maxSKSD discriminates measures (Gong et al., 2021a, Corollary 3.1). Replacing KSD with maxSKSD in SVGD results in *sliced-SVGD* (S-SVGD; Gong et al. (2021a)).

An issue with this procedure is that the basis O is a user-defined choice and may result in inefficient performance. Additionally, restricting g_r and r to one dimension might not capture the underlying structure as effectively as higher dimensional projections. Our approach, described in Section 3.5, will use projections onto subspaces that have dimension potentially greater than one. The framework of these subspace projections is described next. This framework also allows us to explore projections in the underlying geometry more efficiently by using Riemannian optimization.

Grassmann Manifold Let $m \in \{1, 2, \dots, d\}$. A crucial ingredient of our proposed method is the *Grassmann manifold* of m-dimensional subspaces in $\mathbb{R}^d$ (Bendokat et al., 2020), defined as

$$\mathrm{Gr}(d, m) \;=\; \{E \subseteq \mathbb{R}^d : \dim(E) = m\}.$$

The intuition behind our proposed method is to define a KSD that seeks the worst possible discrepancy over all distinct subspaces of dimension m, where the worst-case subspace can be found with gradient-type optimization on the Grassmann manifold. We briefly review the key ingredients of optimization on the Grassmann manifold; a more detailed discussion is deferred to Section 3.A.

To make sense of optimization on the Grassmann manifold, we first need to represent each subspace $E \in \mathrm{Gr}(d, m)$ in memory. One way of doing so is by a projection operator that maps every $\mathbf{x} \in \mathbb{R}^d$ to an element in the subspace. To this end, we define a *projector* of rank m to be a $d \times m$ matrix A with orthonormal columns, i.e. $A^\top A = I_m$, where I_m is the $m\times m$ identity matrix. It then follows that $AA^\top$ is a projection matrix since $(AA^\top)^2 = AA^\top$ and $(AA^\top)^\top = AA^\top$. Denote by $[A] = \{A\mathbf{y} : \mathbf{y} \in \mathbb{R}^m\} \subset \mathbb{R}^d$ the image of A.

We represent each subspace $E \in \mathrm{Gr}(d, m)$ by *any* projector A for which $[A] = E$. This is always possible since, for each projector A the subspace $[A]$ is trivially an element of $\mathrm{Gr}(d, m)$, and, conversely, for any $E \in \mathrm{Gr}(d, m)$, we can construct a corresponding projector A by column-wise appending the elements of any orthonormal basis of E. Such A is unique up to an orthogonal transformation: for any projectors A, B, the subspaces $[A] = [B]$ if and only if $A = BC$ for some orthogonal matrix C in $\mathbb{R}^{m\times m}$. As we will show in Section 3.4, the proposed KSD does not depend on which projector is chosen so long as it corresponds to the same subspace E.

Given a representative A of $[A] \in \mathrm{Gr}(d, m)$, the *tangent space* at $[A]$ is defined as $\mathcal{T}_{[A]}\mathrm{Gr}(d, m) = \{\Delta \in$

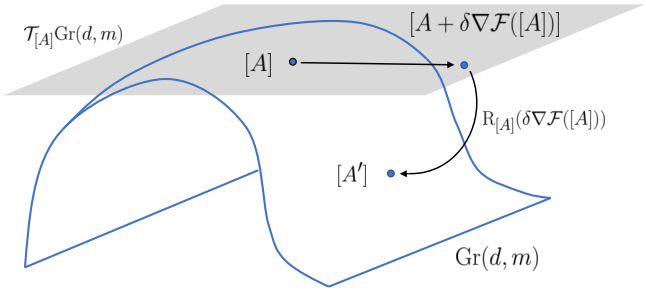


Figure 3.2: An illustration of a gradient ascent step on the Grassmann manifold.

$\mathbb{R}^{d\times m} : A^\top\Delta = 0\}$. We endow a Riemannian metric on $\mathcal{T}_{[A]}\mathrm{Gr}(d,m)$ by restricting the standard matrix inner product $\langle\Delta,\tilde{\Delta}\rangle_0 = \mathrm{Tr}(\Delta^\top\tilde{\Delta})$ to elements $\Delta,\tilde{\Delta}$ in $\mathcal{T}_{[A]}\mathrm{Gr}(d,m)$. It is then clear that $\Pi_A = I_d - AA^\top$ is a projection operator of $\mathbb{R}^{d\times m}$ onto $\mathcal{T}_{[A]}\mathrm{Gr}(d,m)$: $\Pi_A G = (I_d - AA^\top)G \in \mathcal{T}_{[A]}\mathrm{Gr}(d,m)$, for any $G \in \mathbb{R}^{d\times m}$ (Boumal, 2020, Proposition 3.53).

Let $\mathcal{F}$ be a function defined on $\mathrm{Gr}(d,m)$ that is smooth in a proper sense (Boumal, 2020, Section 3). Given a base point $[A] \in \mathrm{Gr}(d,m)$, maximizing $\mathcal{F}$ requires *(i)* finding a vector field of the steepest ascent of $\mathcal{F}$ at $[A]$, and *(ii)* moving from $[A]$ along that vector field without leaving the manifold.

On the Grassmann manifold, the direction of the steepest ascent is no longer represented by the standard (Euclidean) gradient $\nabla\mathcal{F}([A])$. Instead, it is given by the *Riemannian gradient* of $\mathcal{F}$ at $[A]$, given by

$$\mathrm{grad}\,\mathcal{F} \;=\; \Pi_A\nabla\mathcal{F}([A])\;,$$

where $[\nabla\mathcal{F}([A])]_{ij} = \partial\mathcal{F}([A])/\partial A_{ij}$. That is, $\mathrm{grad}\,\mathcal{F}$ is the projection of the Euclidean gradient $\nabla\mathcal{F}([A])$ onto the tangent space at $[A]$. For *(ii)*, one means of moving along a manifold in a given direction is by the *exponential map* (Boumal, 2020, Section 10.2). In our work, we use the *polar retraction* (Boumal, 2020, Definition 3.41), which is an accurate (second order) approximation of the exponential map that is amenable to computation, and is defined as

$$\mathrm{R}_{[A]}(\Delta) \;=\; [UV^\top]\;, \tag{3.6}$$

where $[A] \in \mathrm{Gr}(d,m)$, and $A+\Delta = USV^\top$ is a thin Singular Value Decomposition (see, e.g., Boumal 2020, Section 9.6).

To summarize, with a step-size $\delta > 0$, one step to maximize $\mathcal{F}$ starting from a base point $[A]$ is given by

$$[A'] \;=\; \mathrm{R}_{[A]}(\delta\,\mathrm{grad}\,\mathcal{F})\;.$$

See Figure 3.2 for an illustration. This is a generalization of the standard gradient ascent to the Grassmann manifold (Boumal, 2020, Section 4.3).

3.3 Related Work

Besides the sliced-SVGD of Gong et al. (2021a,b) reviewed in the previous section, many other developments of SVGD involving Rimeannian manifolds or projections have been proposed. Liu and

Zhu (2018) extends SVGD to tasks defined on Riemannian manifolds and Shi et al. (2021) proposes a variant applied to non-Euclidean spaces (such as a simplex) via mirror descent. These are fundamentally different from our work, which concerns problems on the Euclidean space, while evolving the projectors on the Grassmann manifold. Closer to our method is the pSVGD of Chen and Ghattas (2020), where the particles are projected onto a subspace defined by the top eigenvectors of a gradient information matrix. This method however suffers from practical issues, e.g., finding the eigen-decomposition would incur a computational cost that is cubic in the dimension.

Other attempts to alleviate the curse-of-dimensionality problem of SVGD include Message Passing SVGD (Zhuo et al., 2018) and graphical SVGD (Wang et al., 2018), both of which are limited to problems where the target distribution is defined on a *probabilistic graphical model* (PGM) with a known Markov structure. In contrast, GSVGD can be applied to distributions of an arbitrary form.

3.4 Grassmann Kernel Stein Discrepancy

The starting point of our proposed approach is to introduce a novel Stein discrepancy which incorporates dimension reduction features through the use of a projector. To this end, for a projector A of rank m we introduce a matrix-valued kernel k_A arising from an embedding on $\mathbb{R}^d$ into a m-dimensional subspace where $m \leq d$

$$k_A(\mathbf{x}, \mathbf{y}) \;=\; AA^\top k(A^\top \mathbf{x}, A^\top \mathbf{y}) \,, \quad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \,,$$

where $k : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ is a scalar-valued positive definite radial kernel on $\mathbb{R}^m$. That is, $k(\mathbf{u}, \mathbf{v}) = \Psi(\|\mathbf{u} - \mathbf{v}\|_2), \mathbf{u}, \mathbf{v} \in \mathbb{R}^m$, where $\Psi \colon \mathbb{R} \to \mathbb{R}$ and $\|\cdot\|_2$ is the Euclidean 2-norm. Before proceeding, since we shall be performing optimization over $\mathrm{Gr}(d, m)$, we must be sure that for any two projectors A, B that are equivalent in $\mathrm{Gr}(d, m)$, the kernels k_A, k_B are equal. The next lemma assures us of this.

Lemma 3.1. *Let A, B be two projectors of rank m with $[A] = [B]$, then $k_A = k_B$.*

Proof of Lemma 3.1. One can show (e.g., see Bendokat et al. 2020) that $B \in [A]$ if and only if there exists an orthogonal matrix $C \in \mathbb{R}^{m \times m}$ such that $B = AC$. Then, using the orthogonality of C we obtain

$$\begin{aligned} k_B(\mathbf{x}, \mathbf{y}) \;&=\; (AC)(AC)^\top k((AC)^\top \mathbf{x}, (AC)^\top \mathbf{y}) \\ &=\; ACC^\top A^\top k(C^\top A^\top \mathbf{x}, C^\top A^\top \mathbf{y}) \\ &=\; AA^\top \Psi\left(\|C^\top (A^\top \mathbf{x} - A^\top \mathbf{y})\|_2\right) \\ &=\; AA^\top \Psi\left(\|A^\top \mathbf{x} - A^\top \mathbf{y}\|_2\right) \;=\; k_A(\mathbf{x}, \mathbf{y}) \,, \end{aligned}$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. □

Safe in the knowledge that k_A is well-defined over $\mathrm{Gr}(d, m)$, we may define an associated KSD, which

reveals how this choice of k_A induces a projection of the data and score function.

$$\begin{aligned} \mathrm{KSD}_A(Q,P) &= \sup_{\phi\in\mathcal{B}_{k_A}} \mathbb{E}_{\mathbf{X}\sim Q}[\mathbf{s}_p(\mathbf{X})\cdot\phi(\mathbf{X}) + \nabla\cdot\phi(\mathbf{X})] && (3.7)\\ &= \sup_{\phi\in\mathcal{B}_k^m} \mathbb{E}_{\mathbf{X}\sim Q}\big[(A^\top\mathbf{s}_p(\mathbf{X}))\cdot\phi(A^\top\mathbf{X}) + \nabla\cdot\phi(A^\top\mathbf{X})\big]. && (3.8)\end{aligned}$$

where $\mathcal{B}_{k_A}$ is the unit ball of the RKHS corresponding to k_A and $\mathcal{B}_k^m$ is the unit ball of the m-times Cartesian product of $\mathcal{H}_k$, the RKHS corresponding to k. Contrasting this to (3.4), setting $m=1$ and $r, g_r = A$ results in (3.8). The $r^\top g_r$ term in (3.4) becomes I_m in (3.8) since A has orthonormal columns.

Analogous to the approach for KSD (see Liu et al. 2016), we exploit the kernel trick to express KSD_A:

$$\begin{aligned}\mathrm{KSD}_A(Q,P) &= \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\Big[(A^\top\mathbf{s}_p(\mathbf{X}))\cdot A^\top\mathbf{s}_p(\mathbf{X}')k(A^\top\mathbf{X},A^\top\mathbf{X}')\Big]\\ &\quad + 2\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}[(A^\top\mathbf{s}_p(\mathbf{X}))\cdot\nabla_2 k(A^\top\mathbf{X},A^\top\mathbf{X}')] + \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}[\mathrm{Tr}(\nabla_{1,2}k(A^\top\mathbf{X},A^\top\mathbf{X}'))]\,, \end{aligned} \tag{3.9}$$

where $\mathbf{X},\mathbf{X}'\sim Q$ are i.i.d. random variables, $\nabla_i k(A^\top\mathbf{x},A^\top\mathbf{x}')$ denotes the gradient of k with respect to the i-th argument, $\nabla_{1,2}k(\cdot,\cdot)$ is the matrix with ij-th entry $\partial^2 k(\mathbf{x}_1,\mathbf{x}_2)/\partial\mathbf{x}_1^i\partial\mathbf{x}_2^j$, and $\mathrm{Tr}(\cdot)$ denotes the trace.

It is clear that KSD_A does not characterize probability measures for any fixed A. For example if $P=\mathcal{N}(0,C)$ and $Q=\mathcal{N}(0,C')$ are distinct probability measures such that $A^\top C = A^\top C'$, then $\mathrm{KSD}_A(Q,P)=0$. This motivates us to consider the worst-case discrepancy taken over all m-dimensional subspaces. We define the *Grassmann Kernel Stein Discrepancy* (GKSD) as

$$\mathrm{GKSD}(Q,P) = \sup_{[A]\in\mathrm{Gr}(d,m)} \mathrm{KSD}_A(Q,P)\,. \tag{3.10}$$

The following theorem, proved in Section 3.B.1, guarantees that GKSD is able to discriminate distinct distributions. The assumptions on P,Q are standard (Gong et al., 2021a).

Theorem 3.2. *Let P,Q be Borel probability measures with continuously differentiable densities p,q both supported on $\mathbb{R}^d$ with $\lim_{\|\mathbf{x}\|_2\to\infty} q(\mathbf{x})=0$ and suppose the kernel $k(\mathbf{u},\mathbf{v})=\Psi(\|\mathbf{u}-\mathbf{v}\|_2)$ is characteristic and bounded with continuous second order partial derivatives. Then the* GKSD (3.10) *equals* 0 *if and only if $p=q$.*

Remark 3.2.1. *A radial kernel $k(\mathbf{x},\mathbf{y})=\Psi(\|\mathbf{x}-\mathbf{y}\|_2)$ is characteristic if and only if it is integrally strictly positive definite (Sriperumbudur et al., 2011). In particular, the assumptions of Theorem 3.2 hold for Gaussian and Inverse Multi-quadric kernels.*

Intuitively, using projections could be beneficial when the target distribution has a sparse structure. Moreover, projecting on *multivariate* subspaces for $\mathbf{s}_p(\mathbf{x})$ will allow us to better account for the correlation between dimensions. With optimal projectors, correlations can be captured regardless of the projection dimension, since both GKSD and maxSKSD can discriminate distinct distributions. In practice, however, optimal projectors are not available, and hence projecting onto 1 dimension can lead to suboptimal approximation due to weak correlation signals.

Algorithm 4 Grassmann Stein Variational Gradient Descent (GSVGD).

1: **Inputs:** M initialized projectors $[A_{0,l}] \in \mathrm{Gr}(d,m)$, $l = 1,\dots,M$; Initialized particles $\{\mathbf{x}_i^0\}_{i=1}^N$; iteration number n_{epochs}; step sizes $\epsilon, \delta > 0$.
2: **for** $t = 0, 1, \dots, n_{\text{epochs}} - 1$ **do**
3: For $l = 1,\dots,M$, compute the kernel bandwidth using the projected particles $\{A_{t,l}\mathbf{x}_i^t\}_{i=1}^N$ and the median heuristics described in Section 3.6.
4: Update particles $\mathbf{x}_i^{t+1} = \mathbf{x}_i^t + \epsilon \sum_{l=1}^M \widehat{\phi}_{A_{t,l}}(\mathbf{x}_i^t)$ for $i = 1,\dots,N$, where

$$\widehat{\phi}_{A_{t,l}}(\mathbf{x}_i^t) \;=\; \frac{1}{N}\sum_{j=1}^{N}\left[A_{t,l}A_{t,l}^\top \mathbf{s}_p(\mathbf{x}_j^t)k(A_{t,l}^\top\mathbf{x}_j^t, A_{t,l}^\top\mathbf{x}_i^t) + A_{t,l}\nabla_2 k(A_{t,l}^\top\mathbf{x}_j^t, A_{t,l}^\top\mathbf{x}_i^t)\right] , \tag{3.11}$$

and the kernels bandwidths are chosen by median heuristics.
5: Update projectors

$$A_{t+1,l} \;=\; \mathrm{R}_{[A_{t,l}]}\left(\delta\Pi_{A_{t,l}}\nabla\alpha_t([A_{t,l}]) + \sqrt{2T\delta}\Pi_{A_{t,l}}\xi_{t,l}\right) ,$$

for $l = 1,\dots,M$, where $\xi_{t,l} \in \mathbb{R}^{d\times m}$ has i.i.d. $\mathcal{N}(0,1)$ entries, and $\mathrm{R}_{[A_{t,l}]}$ is defined in (3.6).
6: **end for**
7: **Return:** Particles $\{\mathbf{x}_i^{n_{\text{epochs}}}\}_{i=1}^N$.

3.5 Grassmann Stein Variational Gradient Descent

Equipped with this new form of kernel discrepancy we can perform a variant of SVGD which we call *Grassmann Stein Variational Gradient Descent* (GSVGD). More specifically, we seek to find

$$\phi_A^* \;\in\; \arg\sup_{\phi\in\mathcal{B}_{k_A}} -\frac{d}{d\epsilon}\mathrm{KL}(T_{\#}Q, P)\Big|_{\epsilon=0} , \tag{3.12}$$

where $T_{\#} = I + \epsilon\phi$, and I is the identity map. The supremum on the RHS of (3.12) is attained by

$$\phi_A^*(\cdot) \;=\; \mathbb{E}_{\mathbf{X}\sim Q}[AA^\top\mathbf{s}_p(\mathbf{X})k(A^\top\cdot, A^\top\mathbf{X}) + A\nabla_2 k(A^\top\cdot, A^\top\mathbf{X})] ,$$

for which it is clear that

$$-\frac{d}{d\epsilon}\mathrm{KL}(T_{\#}^*Q, P)\Big|_{\epsilon=0} \;=\; \mathrm{KSD}_A(P,Q) , \quad T_{\#}^* \;=\; I + \epsilon\phi_A^* .$$

To transport the particles efficiently, at every step, we seek to identify the subspace $[A] \in \mathrm{Gr}(d,m)$ with the largest discrepancy between P and Q, so that we select

$$[A^*] \in \arg\sup_{[A]\,\in\,\mathrm{Gr}(d,m)} \mathrm{KSD}_A(P,Q), \tag{3.13}$$

for which $\mathrm{KSD}_{A^*} = \mathrm{GKSD}(P,Q)$. In practice, the objective (3.13) can be (approximately) solved by searching for a critical point of $\mathrm{KSD}_A(P,Q)$ at which its Riemannian gradient vanishes. This is justified by the following proposition, which states that, when the probability measures P and Q differ in some low-dimensional subspace $[A_0]$, then $[A_0]$ is indeed a solution sought by (3.13).

Proposition 3.3. *Suppose the conditions in Theorem 3.2 hold for the kernel k and measures P, Q with*

associated densities q, p. Assume further that the kernel satisfies the smoothness condition Assumption 3.1 in Section 3.B.2.1, and that there exists $[A_0] \in Gr(d, m)$ such that q, p admit the following decomposition

$$q(\mathbf{x}) \propto q^m(P_0\mathbf{x})\xi(\Pi_{A_0}\mathbf{x}) \tag{3.14}$$
$$p(\mathbf{x}) \propto p^m(P_0\mathbf{x})\xi(\Pi_{A_0}\mathbf{x}), \tag{3.15}$$

where $P_0 \coloneqq A_0A_0^\top$, $\Pi_{A_0} = I_d - A_0A_0^\top$ as before, and p^m, q^m and ξ are smooth, positive functions that are integrable on $\mathbb{R}^d$ (i.e. they are unnormalized densities). Then $\operatorname{grad} KSD_A(P, Q)|_{A=A_0} = 0$.

Remark 3.3.1. *The assumption on the decomposition (3.14) and (3.15) holds when the density q agrees with the target p except in the subspace $[A_0]$. An example is that q is a prior, and p is the posterior induced from q and a likelihood function that depends on $\mathbf{x}$ only through $A_0\mathbf{x}$. See Section 3.B.2 for more discussions and the proof.*

Sequentially solving the optimization problems (3.12) and (3.13) every time-step yields a particle transport scheme which terminates if and only if the GKSD between the target distribution and the current particle distribution is zero, which implies $P = Q$ under the conditions of Theorem 3.2.

Replacing Q by an empirical distribution of particles $(\mathbf{x}_1^0, \ldots, \mathbf{x}_N^0)$, and denoting by $(\mathbf{x}_1^t, \ldots, \mathbf{x}_N^t)$ the set of particles obtained at step t, the scheme becomes

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t + \frac{\epsilon}{N}\sum_{j=1}^{N}\left[A_tA_t^\top \mathbf{s}_p(\mathbf{x}_j^t)k(A_t^\top\mathbf{x}_i^t, A_t^\top\mathbf{x}_j^t) + A_t\nabla_2 k(A_t^\top\mathbf{x}_i^t, A_t^\top\mathbf{x}_j^t)\right], \quad i = 1, \ldots, N, \tag{3.16}$$

$$A_t \in \arg\sup_{[A]\in \mathrm{Gr}(d,m)} \alpha_t([A]), \tag{3.17}$$

where $\alpha_t([A]) = \mathrm{KSD}_A(Q_t, P)$, for $Q_t(\mathrm{d}\mathbf{x}) = \frac{1}{N}\sum_{i=1}^N \delta_{\mathbf{x}_i^t}(\mathrm{d}\mathbf{x})$.

Optimal Projections via SDEs Proposition 3.3 implies the "optimal" projection A_0 is *one* critical point of α_t, but it does not imply the critical point is unique. Indeed, the problem (3.13) is non-convex and there might be multiple local maximizers, from which finding global maxima is generally NP-hard. A well-known remedy is to add white noise to the gradient flow (Aluffi-Pentini et al., 1985; Geman and Hwang, 1986), allowing the trajectory to escape from local maxima. Mathematically, such methods can be formulated in terms of Stochastic Differential Equations (SDEs). In this particular instance, we explore the use of SDEs taking values on the Grassmann manifold. We consider stochastic dynamics defined by the solution of the following Stratonovich SDE

$$\begin{aligned} dA(t) &= \mu \operatorname{grad}\alpha_t([A])\,dt + \sqrt{2T\mu}\Pi_{A_t} \circ dW(t) \\ &= \mu\Pi_{A(t)}\nabla\alpha_t([A(t)])\,dt + \sqrt{2T\mu}\Pi_{A(t)} \circ dW(t), \end{aligned} \tag{3.18}$$

where $W(t)$ is a $\mathbb{R}^{d\times m}$-valued Brownian motion and $\mu, T > 0$. Combining the particle and projector evolution we can consider the full system as a discretization of the following coupled ODE-SDE

$$\begin{aligned} \dot{\mathbf{x}}_i(t) &= \frac{1}{N}A(t)\sum_{j=1}^{N}\left[A(t)^\top \mathbf{s}_p(\mathbf{x}_j(t))k(A(t)^\top\mathbf{x}_i(t), A(t)^\top\mathbf{x}_j(t)) + \nabla_2 k(A(t)^\top\mathbf{x}_i(t), A(t)^\top\mathbf{x}_j(t))\right], \\ dA(t) &= \mu\Pi_{A(t)}\nabla\alpha_t([A(t)])\,dt + \sqrt{2T\mu}\Pi_{A(t)} \circ dW(t). \end{aligned}$$

We see that the parameter T acts as a temperature, facilitating moving between local minima of the function α by annealing the loss function. The parameter μ controls the relative timescale between the evolution of $A(t)$ and the particles $(\mathbf{x}_1^t \dots, \mathbf{x}_N^t)$. In particular, when $\mu \gg 1$, the distribution of $A(t)$ will converge much faster to equilibrium than the particles. In the other extreme, when $\mu \ll 1$ the projector $A(t)$ will evolve on a slower timescale than the particles, so that the particles will reach their stationary configuration for a fixed projector before it evolves.

To simulate the Stratonovich SDE (3.18), we leverage the formulation of diffusions on a Riemannian manifold, making use of the associated retraction map—see Staneva and Younes (2017); Baxendale et al. (1976); Belopolskaya and Dalecky (2012)—where the $\mathrm{Gr}(d, m)$-valued SDE (3.18) can be expressed in the form of a system of Ito SDEs

$$A(t) = \mathrm{R}_{[A(t)]}\left(B(t)\right), \tag{3.19}$$

$$dB(t) = \mu \Pi_{A(t)} \nabla \alpha_t([A(t)])\, dt + \sqrt{2T\mu}\Pi_{A(t)} dW_t\,. \tag{3.20}$$

With step-size $\delta > 0$, an Euler-Maruyama discretization for (3.20) results in

$$A_{t+1} = \mathrm{R}_{[A_t]}\left(\Pi_{A_t} \nabla \alpha_t([A_t])\, \delta + \sqrt{2T\delta}\Pi_{A_t}\xi_t\right)\,, \tag{3.21}$$

where ξ_t is a $d \times m$ matrix with independent Gaussian entries. Note that we "absorb" the constant μ into the step-size δ for simplicity.

Batch Generalization An important observation is that at each step (3.17), $\mathbf{x}_i^{t+1}$ is different from $\mathbf{x}_i^t$ only in the image of A. This contrasts with both SVGD as well as S-SVGD, both of which update the particles along all dimensions. One extension is to use more than one projector $A_{t,l} \in \mathbb{R}^{m_l \times d}$ for $l = 1, \dots, M$, where $1 \leq M \leq d$ and $\sum_{l=1}^{M} m_l \leq d$. At each optimization step, we could impose orthonormality using QR factorization, so that the $d \times d$ matrix A_t formed by column-wise concatenation of $A_{t,1}, \dots, A_{t,M}$, is orthogonal. However, this requires $\mathcal{O}(d^3)$ operations, which can be prohibitive for large d.

We therefore propose to impose orthonormality every 1000 steps. This means the projectors at each step are no longer guaranteed to be orthonormal and could thus lead to redundancy in the search dimensions. Nevertheless, we still find that using multiple projectors could be beneficial in practice, potentially because it improves the chance of covering the most relevant subspace. The full algorithm is presented in Algorithm 4.

Computational Cost Each iteration of the particles update in Algorithm 4 requires $\mathcal{O}(MNdm(N + d))$ operations. Compared with the $\mathcal{O}(Nd(N + d))$ cost for S-SVGD, the extra factor Mm arises from the fact that (3.11) involves matrix multiplication with $A_{t,l}$, which can be avoided in S-SVGD due to simplification by using the canonical basis. However, this extra computation can be largely reduced by using a fixed number of projectors M and projection dimension m. In our experiments, we find $M = 20$ projectors are sufficient. A reasonable choice of m, however, will depend on the specific problem. Finally, each step of the projector update requires $\mathcal{O}(Mdm^2)$ due to the retraction map, which reduces to the same cost as the slice update in S-SVGD when $M = d$ and $m = 1$.

Connections to S-SVGD The major differences between GSVGD and S-SVGD are that *(i)* GSVGD allows projecting onto a arbitrary low-dimensional subspace, *(ii)* GSVGD updates the projectors via Riemannian optimization on the Grassmann manifold, whereas S-SVGD optimizes the slices in the unit ball of $\mathbb{R}^d$, and *(iii)* the projectors for the score function and the data are set to the same in GSVGD but not in S-SVGD. In particular, S-SVGD slices the score function with a fixed basis O (i.e. the canonical basis), and such arbitrary choice may result in inefficient exploration of the state-space if the basis vectors $r \in O$ do not align with the latent dimensions of the target. On the other hand, optimizing the projectors for the score function and the data separately would lead to a complex joint optimization program that is practically challenging. GSVGD strikes a balance by using the same projector for the score function and the data.

3.6 Experiments

We study the uncertainty quantification property of GSVGD against existing methods such as SVGD and S-SVGD on a number of experiments. We conclude from our experiments that SVGD can underestimate the uncertainty, S-SVGD can overestimate it and GSVGD, with appropriate projection dimensions, yields the best estimates. For GSVGD, we use the batch approach of Algorithm 4, where at most $M = 20$ projectors with the same projection dimension m are used. The temperature T is set to a small value and is gradually incremented to a larger one via an annealing scheme. The learning rates for all methods are tuned for optimal performance on the multimodal mixture example, and the same choices are used for the other experiments.

In all experiments, we use the Gaussian kernel defined by $k(\mathbf{x}, \mathbf{x}') = \exp\left(-\|\mathbf{x} - \mathbf{x}'\|_2^2/(2\sigma^2)\right)$ with bandwidth $\sigma^2 = \texttt{med}^2/(2\log n)$, where $\texttt{med}$ is the median of the pairwise distance of the particles. This follows the heuristic in Liu and Wang (2016). To quantify sample quality, we use the *energy distance* (Székely and Rizzo, 2013) between the approximation and the ground truth. However, we find that the energy distance is not sensitive to differences in correlations between multivariate samples. Therefore, we also evaluate how well each method captures the dependence between multivariate samples by the covariance estimation error $||\hat{\Sigma} - \Sigma||_2$, where $||\cdot||_2$ is the Frobenius norm, Σ is the ground truth sample covariance matrix given by a Hamiltonian Monte Carlo (HMC) (for non-synthetic data), and $\hat{\Sigma}$ is the estimated covariance. For further details including selection of the learning rate ϵ and temperature T, see Section 3.C. All the code is available at github.com/ImperialCollegeLondon/GSVGD.

Multivariate Gaussian The first toy example is a d-dimensional multivariate Gaussian $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; 0, I_d)$. For each method, 500 particles are initialized from $\mathcal{N}(\mathbf{x}; 2\mathbb{1}, 2I_d)$, where $\mathbb{1} \in \mathbb{R}^d$ denotes the vector of ones. This is a standard benchmark used to illustrate the diminishing variance issue of SVGD (Liu and Wang, 2016; Gong et al., 2021a; Zhuo et al., 2018). We show empirically that GSVGD is able to mitigate this problem by plotting the dimension-averaged estimates for the marginal variance in Figure 3.3. We see that GSVGD estimates stay roughly at 1 for all dimensions as expected, whereas the SVGD estimates scale inversely with dimension. S-SVGD, on the other hand, over-estimates the variance. We remark, however, that GSVGD5 appears to be biased with a lower variance. This is due to the variance underestimation issue of SVGD persisting in 5 dimensional subspaces.

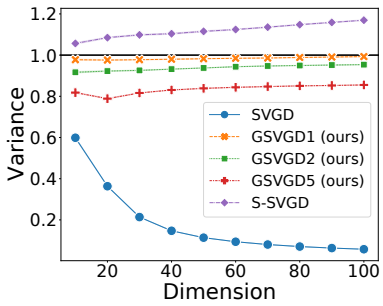
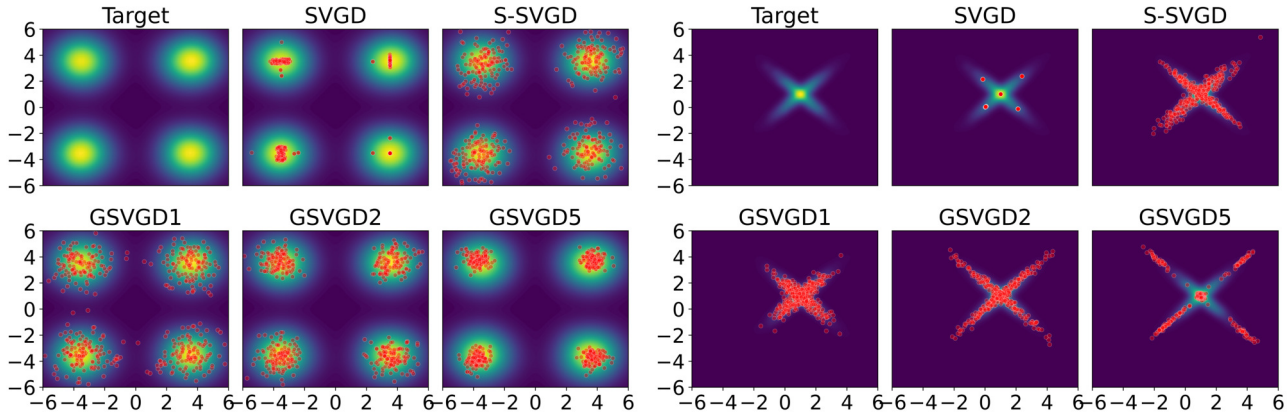


Figure 3.3: Estimating the dimension-averaged marginal variance of $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; 0, I_d)$ across different dimensions d. GSVGD1 means GSVGD with 1-dimensional projections. Black solid line marks the true value. Results are averaged over 20 repetitions. GSVGD1 refers to GSVGD with projection dimension set to $m = 1$, and similarly for GSVGD2 and GSVGD5.



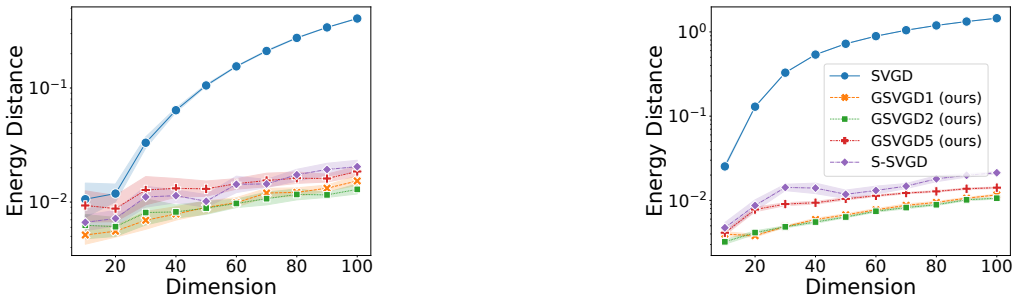
(a) Multimodal mixture.

(b) X-shaped mixture.

Figure 3.4: Marginals of target and estimated densities in the first 2 dimensions. Red dots are the particles after 2000 iterations and the target density is shown by the contour. Both targets have dimension 50.

Multimodal Mixture In the second example, the target distribution is a mixture of 4 d-dimensional Gaussian distributions $p(\mathbf{x}) = \sum_{k=1}^{4} 0.25\mathcal{N}(\mathbf{x}; \mu_k, I_d)$ with uniform mixture ratios. The mean vectors of the multimodal mixture are defined as $\mu_k = (\sqrt{5}\cos(2k\pi/K + \pi/4), \sqrt{5}\sin(2k\pi/k + \pi/4), 0, \ldots, 0)^\top \in \mathbb{R}^d$, for $k = 1, \ldots 4$; see Figure 3.4a. That is, the target in the first two dimensions is a mixture of 4 Gaussian distributions with modes equally spaced on a circle of radius $\sqrt{5}$, whilst in the other coordinates it is a standard Gaussian. Particles are initialized from $\mathcal{N}(0, I_d)$ and only the first two dimensions need to be learned. The primary goal is to investigate whether GSVGD can efficiently recover this low-dimensional latent structure by adapting the projectors. As shown in 3.5a, GSVGD1 and GSVGD2 outperforms S-SVGD in all dimensions in terms of the energy distance, while GSVGD5 is able to achieve a competitive performance. SVGD, on the other hand, fails to capture the uncertainty of the target, as shown in Figure 3.4a.

X-Shaped Mixture The target in this experiment is a d-dimensional mixture of two correlated Gaussian distributions $p(\mathbf{x}) = 0.5\mathcal{N}(\mathbf{x}; \mu_1, \Sigma_1) + 0.5\mathcal{N}(\mathbf{x}; \mu_2, \Sigma_2)$. The means μ_1, μ_2 of each Gaussian have components equal to 1 in the first two coordinates and 0 otherwise, and the covariance matrices



(a) Multimodal mixture.

(b) X-shaped mixture.

Figure 3.5: Energy distance between the target distribution and the particle estimation. Results are averaged over 20 repetitions. 95%-confidence intervals are shown by the shaded regions.

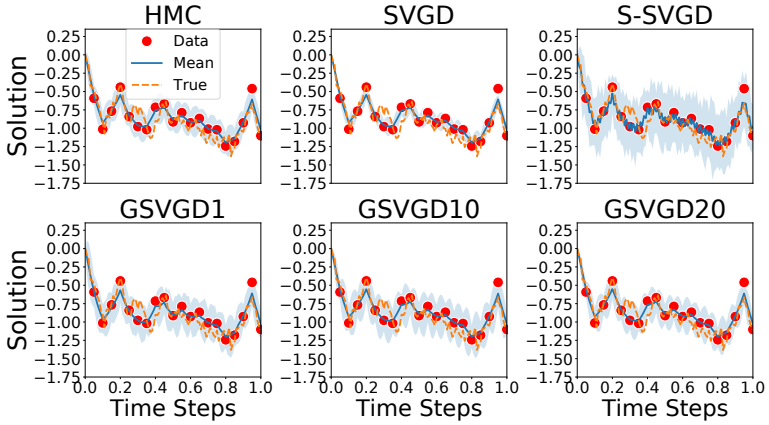


Figure 3.6: HMC, SVGD, S-SVGD and GSVGD solutions, as well as the posterior mean and 95% confidence interval after 1000 iterations.

admit a correlated block diagonal structure as follows:

$$\Sigma_1 = \begin{pmatrix} 1 & \delta & \\ \delta & 1 & \mathbf{0} \\ & \mathbf{0} & I_{d-2} \end{pmatrix}, \qquad \Sigma_2 = \begin{pmatrix} 1 & -\delta & \\ -\delta & 1 & \mathbf{0} \\ & \mathbf{0} & I_{d-2} \end{pmatrix},$$

where $\delta = 0.95$ controls the correlation between the first two dimensions. The mixture hence manifests as an "X-shaped" density marginally in the first two dimensions (see Figure 3.4b). Figure 3.5b shows that GSVGD1 and GSVGD2 achieve a better approximation quality than both SVGD and S-SVGD across all dimensions. This is further verified by plotting the first two coordinates of the final particles in 3.4 (b). On the other hand, GSVGD5 shows a slightly suboptimal performance for dimensions larger than 30. This is potentially because of the inefficiency due to projecting the particles to 5-dimensional subspace in the case we only care about the first 2.

Conditioned Diffusion Process The next example is a benchmark that is often used to test inference methods in high dimensions (Cui et al., 2016; Chen and Ghattas, 2020; Detommaso et al., 2018). We consider a stochastic process $u : [0, 1] \to \mathbb{R}$ governed by

$$du_t = \frac{10u(1-u^2)}{1+u^2}dt + dx_t\,, \quad u_t = 0\,, \tag{3.22}$$

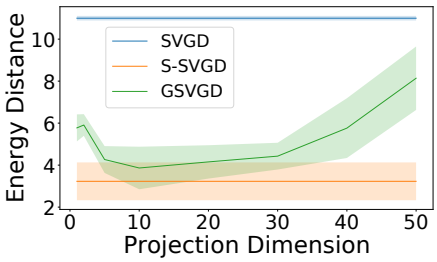
where $t \in (0, 1]$, and the forcing term $x = (x_t)_{t \geq 0}$ follows a Brownian motion so that $x \sim q = \mathcal{N}(0, C)$ with $C(t, t') = \min(t, t')$. We observe noisy data $\mathbf{y} = (y_{t_1}, \ldots, y_{t_{20}})^\top \in \mathbb{R}^{20}$ at 20 equi-spaced time points $t_i = 0.05i$, where $y_{t_i} = u_{t_i} + \epsilon$ for $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.1$, and u_{t_i} is generated by solving (3.22) at a true Brownian path x_{true}. The goal is to use $\mathbf{y}$ to infer the forcing term x and thereby the solution state u.

The prior we use for x is the Brownian motion described above. The dynamic is discretized using an Euler-Maruyama scheme with step size $\Delta t = 10^{-2}$, leading to a 100-dimensional inference problem. Due to the almost sure continuity of the Brownian path, we expect the solution u to have an intrinsic low-dimensional structure. Figure 3.6 shows the mean and 95% credible intervals of the particle estimation of u after 1000 iterations. Compared with a reference HMC, all methods are able to estimate the mean accurately, but SVGD and S-SVGD either underestimate or overestimate the credible intervals. GSVGD projecting on 20 dimensions gives the most accurate uncertainty estimates. Choosing a good projection dimension is non-trivial even in this simple problem; however, by plotting the energy distance between the HMC reference and the results against different projection dimensions (Figure 3.11 in Section 3.D.2), we see that so long as we do not project onto very low or high dimensions, GSVGD achieves a greater degree of agreement with the reference HMC than SVGD and S-SVGD.

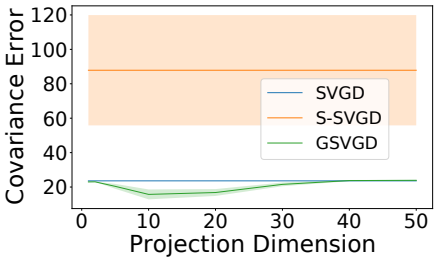
Bayesian Logistic Regression Following Liu and Wang (2016), we consider the Bayesian logistic regression model applied to the Covertype dataset (Asuncion and Newman, 2007). We follow the same setting as Liu and Wang (2016) by placing a Gaussian prior $p_0(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}; 0, \alpha^{-1} I_{d-1})$ on the regression weights $\mathbf{w}$ with $p_0(\alpha) = \text{Gamma}(\alpha; 1, 0.01)$. We are interested in the posterior $p(\mathbf{x}|D)$ of $\mathbf{x} = [\mathbf{w}, \alpha]$. The model was tested on a subset of the Covertype dataset, and we use HMC posterior samples as a gold standard for evaluating the posterior approximations. The original dataset consists of 581,012 data points and 54 features with binary labels, which is too large for the HMC sampler. As a result, a subset of 1,000 randomly selected data points was used, although other methods that are more scalable to data size, such as the stochastic gradient Langevin dynamics (SGLD) of Welling and Teh (2011) or the HMC-ECS of Dang et al. (2019), could have been used to allow inference on the entire dataset.

To evaluate the results, we used both the energy distance and the covariance estimation error $\|\hat{\Sigma} - \Sigma\|_2$, where $\|A\|_2 = \sqrt{\sum_{i,j=1}^{d} A_{ij}^2}$ is the Frobenius norm of a matrix $A \in \mathbb{R}^{d \times d}$, and $\hat{\Sigma}, \Sigma$ are respectively the sample covariance matrices of the particle estimation and of a HMC run treated as the ground truth. The covariance estimation error was used because we found that the energy distance is not sensitive enough to differences in the second moments.

Similar to the synthetic experiments, we see in Figure 3.7 that there exists an optimal projection dimension (around $m = 10$) such that GSVGD more accurate approximates the covariance matrix whilst yielding similar energy distances as S-SVGD. Moreover, S-SVGD severely overestimates the uncertainty, also shown in Figure 3.8, whereas SVGD severely underestimates it. We also observed from Figure 3.7 that the covariance between distinct variables is overestimated compared with the HMC reference, as shown by the large values (in absolute term) of the off-diagonal entries. On the other hand, GSVGD projecting onto $m = 10$ dimensional subspaces showed greater alignment, despite having a similar energy distance as S-SVGD (Figure 3.7a). Again, we emphasize that the performance



(a) Energy distance against projection dimensions.



(b) $||\hat{\Sigma} - \Sigma||_2$ against projection dimensions.

Figure 3.7: Energy distance and covariance errors for SVGD, S-SVGD and GSVGD in the Covertype experiment.

of GSVGD deteriorates when the projection dimension is extreme, e.g. when m equals 1 or the full dimension of the problem ($m = 55$).

3.7 Conclusion and Discussion

We introduced Grassmann Stein Variational Gradient Descent (GSVGD), a novel extension to the SVGD algorithm, to tackle the curse-of-dimensionality problem. At each step, GSVGD seeks the subspace in which the proposal and the target has the largest discrepancy measured by the Grassmann Kernel Stein Discrepancy (GKSD). The evolution of the particles and the projector is determined through a coupled ODE-SDE system, which allows trade-offs between exploration and exploitation. The GSVGD algorithm can be easily extended to a batched version that uses more than one projector. Our experiments demonstrate that GSVGD is able to achieve improved performance over existing methods when the target distribution has an intrinsic low-dimensional structure, especially more accurately quantifying the epistemic uncertainty.

Limitations of GSVGD: One limitation of GSVGD is the choice of projection dimension and number of projectors to use in order to attain superior performance. One possible way is to perform a grid search over a range of feasible values and then assess the performance using metrics that measure particle diversity. We leave this as an open question for future work. In addition, our Grassmann-valued SDE introduces additional hyperparameters, which potentially require tuning on a case-by-case basis for each problem. In Section 3.6 we present some heuristics for tuning GSVGD hyperparameters that we found yielded consistent results, but we would recommend future work to study the sensitivity of the hyperparameters.

3.A Further Background Materials on the Grassmann Manifold

We provide a brief overview of the Grassmann manifold and optimization on the Grassmann manifold. We focus only on the essential components required to understand the proposed method; a thorough treatment of these topics is beyond the scope of the thesis. For more details, we refer the interested

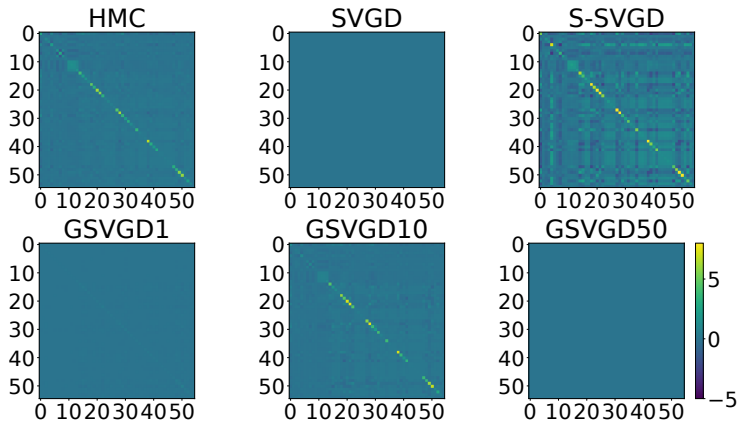


Figure 3.8: Covariance matrices for Covertype.

readers to Bendokat et al. (2020) regarding the theoretical perspectives of the Grassmann manifold, and Boumal (2020); Absil et al. (2008) regarding optimization on Grassmann and other manifolds.

The Stiefel and Grassmann Manifolds The *Grassmann manifold* $\mathrm{Gr}(d, m)$ (Boumal, 2020, Section 9) is the space of all m-dimensional linear subspaces in $\mathbb{R}^d$. Each element of $\mathrm{Gr}(d, m)$ is an equivalence class of $d \times m$ matrices whose columns form an orthonormal basis for the same subspace. To this end, define the *Stiefel manifold* $\mathrm{St}(d, m) \coloneqq \{A \in \mathbb{R}^{d\times m} : A^\top A = I_m\}$, where I_m is the $m \times m$ identity matrix. That is, $\mathrm{St}(d, m)$ is the set of projectors of rank m as defined in Section 3.2. The Grassmann manifold is built from the Stiefel manifold by identifying matrices whose columns span the same subspace. Formally, define an equivalence relationship $\sim$ as $A \sim B \iff A = BC$ for any $C \in O(m)$, the group of orthogonal matrices in $\mathbb{R}^{m\times m}$. $\mathrm{Gr}(d, m)$ is the quotient space

$$\mathrm{Gr}(d, m) \;=\; \mathrm{St}(d, m)/\sim \;=\; \{[A] : A \in \mathrm{St}(d, m)\}\,, \quad \text{where } [A] \;=\; \{B \in \mathrm{St}(d, m) : A \sim B\}\;.$$

The Grassmann manifold is an instance of *quotient manifolds* on which we can define smooth functions; see Boumal (2020, Chapter 9) for a full treatment of quotient manifolds and Boumal (2020, Chapter 3) for the precise definition of smoothness of a manifold-valued function. This allows the use of gradient-type optimization to find the critical points of a smooth function. A function $\mathcal{F}$ defined on $\mathrm{Gr}(d, m)$ differs from one defined on $\mathrm{St}(d, m)$ in that the former must preserve invariance of the chosen projector; that is, $\mathcal{F}([A])$ depends only on the subspace $[A] \in \mathrm{Gr}(d, m)$ but not on A. An example is the matrix-valued kernel defined in Lemma 3.1: $A \mapsto k_A(\mathbf{x}, \mathbf{y}) = AA^\top k(A^\top\mathbf{x}, A^\top\mathbf{y})$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

Tangent Spaces and the Riemannian Gradient To define the Riemannian gradient $\operatorname{grad}\mathcal{F}([A])$ of a smooth function $\mathcal{F}$ on $\mathrm{Gr}(d, m)$, we endow $\mathrm{Gr}(d, m)$ with the standard inner matrix product $g_{[A]}(\Delta, \tilde{\Delta}) \coloneqq \langle\Delta, \tilde{\Delta}\rangle_0 = \mathrm{Tr}(\Delta^\top\tilde{\Delta})$ restricted to $\Delta, \tilde{\Delta} \in \mathrm{Gr}(d, m)$. The Riemannian gradient is then defined by the relationship

$$g_{[A]}(\operatorname{grad}\mathcal{F}([A]), \Delta) \;=\; \frac{\mathrm{d}}{\mathrm{d}\delta}\mathcal{F}(\mathrm{R}_{[A]}(\delta\Delta))\Big|_{\delta=0}\,,$$

for $\Delta \in \mathcal{T}_{[A]}\mathrm{Gr}(d, m) \coloneqq \{\Delta \in \mathbb{R}^{d\times m} : A^\top\Delta = 0\}$, and $\mathrm{R}_{[A]}(\cdot)$ is a *retraction* (Boumal, 2020, Definition 3.41) that sends $[A]$ along a chosen direction while remaining on the Grassmann manifold. In particular,

$\mathcal{T}_{[A]}\mathrm{Gr}(d,m)$ is called the *tangent space* at $[A]$ (Boumal, 2020, Definition 3.10). The Riemannian gradient is simply the orthogonal projection of the standard gradient $\nabla\mathcal{F}([A])$ to the tangent space: $\operatorname{grad}\mathcal{F}([A]) = \Pi_A \nabla\mathcal{F}([A])$, where $\Pi_A = I_d - AA^\top$, and $[\nabla\mathcal{F}([A])]_{ij} = \frac{\partial}{\partial A_{ij}}\mathcal{F}([A])$.

Riemannian Gradient Descent To find a critical point of $\mathcal{F}$ where its Riemannian gradient is zero, *Riemannian gradient descent* (RGD), which generalized the standard gradient descent to a Riemannian manifold, can be used. In RGD, we start from $[A_0] \in \mathrm{Gr}(d,m)$ and iterate $[A_{t+1}] = \mathrm{R}_{[A_t]}(-\delta \operatorname{grad}\mathcal{F}([A_t]))$ for $t = 0, 1, \ldots$ and step size $\delta > 0$. Intuitively, $\operatorname{grad}\mathcal{F}([A_t])$ finds the direction of the steepest descent at $[A_t]$, and the retraction $\mathrm{R}_{[A_t]}(\cdot)$ is the action of moving along $\mathrm{Gr}(d,m)$ in that direction.

3.B Proofs of Theoretical Results

This section contains the proofs for the theoretical results presented in Chapter 3. It is organized as follows: Section 3.B.1 shows Theorem 3.2; Section 3.B.2 includes the proof for Proposition 3.3 and discussions on the assumptions imposed on the kernel and the density functions of the candidate and target distributions.

3.B.1 Proof of Theorem 3.2

The idea of the proof is to rewrite GKSD as a double integral type discrepancy, similar to Liu et al. (2016, Theorem 3.6), and the proof shall largely follow. We will then conclude that the discrepancy distinguishes p, q given the assumptions on the kernel.

The main difference to Liu et al. (2016, Theorem 3.6) is that we shall be using a matrix-valued kernel formed from the projection A, as written in (3.8), rather than a standard scalar-valued kernel. This means we must introduce some extra notation. First, given a density q on $\mathbb{R}^d$, we will be using the operator $\mathcal{A}_q f(\mathbf{x}) = \mathbf{s}_q(\mathbf{x})^\top f(\mathbf{x}) + \nabla^\top f(\mathbf{x})$ for functions $f\colon \mathbb{R}^d \to \mathbb{R}^d$ and abuse notation so that for functions $F\colon \mathbb{R}^d \to \mathbb{R}^{d\times m}$ we have $\mathcal{A}_q F(\mathbf{x}) = (\mathcal{A}_q F_1(\mathbf{x}), \ldots, \mathcal{A}_q F_m(\mathbf{x})) \in \mathbb{R}^{1\times m}$ where $F_i\colon \mathbb{R}^d \to \mathbb{R}^d$ is the i-th column of F.

The next lemma is a vectorized version of Liu et al. (2016, Lemma 2.2, Lemma 2.3) and the proof follows by applying those results elementwise to F.

Lemma 3.4. *Under the assumptions on q in Theorem 3.2, if $F\colon \mathbb{R}^d \to \mathbb{R}^{d\times m}$ has bounded entries and is continuously differentiable, then*

$$\mathbb{E}_{\mathbf{X}\sim Q}\left[(\mathbf{s}_p(\mathbf{X}) - \mathbf{s}_q(\mathbf{X}))^\top F(\mathbf{X})\right] \;=\; \mathbb{E}_{\mathbf{X}\sim Q}\left[\mathcal{A}_p F(\mathbf{X})\right] .$$

Remark 3.4.1. *The boundedness condition on F is to ensure it lies in the* Stein class *of q (Liu et al., 2016, Definition 2.1), since q is assumed to be supported on all of $\mathbb{R}^d$, so that Liu et al. (2016, Lemma2.2, Lemma2.3) can be applied. A similar condition on F can be imposed when the support of q is a compact subset $\mathcal{X}$ of $\mathbb{R}^d$.*

Now we derive the analogy to Liu et al. (2016, Theorem 3.6) using Lemma 3.4, the proof is largely the same as the proof of Liu et al. (2016, Theorem 3.6), which we include for clarity.

Lemma 3.5. *Let $k(\mathbf{u}, \mathbf{v}) = \Psi(\|\mathbf{u} - \mathbf{v}\|_2)$ be a characteristic kernel where Ψ is bounded with continuous second order partial derivatives. Under the assumptions in Theorem 3.2 the projected KSD can be expressed as follows*

$$\mathrm{GKSD}(Q, P) \;=\; \sup_{[A] \in \mathrm{Gr}(d,m)} \mathbb{E}_{\mathbf{X}, \mathbf{X}' \sim Q} \left[(A^\top \delta_{p,q}(\mathbf{X}))^\top A^\top \delta_{p,q}(\mathbf{X}') k(A^\top \mathbf{X}, A^\top \mathbf{X}') \right] , \tag{3.23}$$

where $\delta_{p,q}(\mathbf{x}) = \mathbf{s}_p(\mathbf{x}) - \mathbf{s}_q(\mathbf{x})$.

Proof of Lemma 3.5. Define $\mathbf{v}(\mathbf{x}, \mathbf{x}') = \mathcal{A}_p(Ak(A^\top \cdot, A^\top \mathbf{x}'))(\mathbf{x}) \in \mathbb{R}^{1 \times m}$. Since k is assumed to be bounded, we can apply Lemma 3.4 with $F(\mathbf{x}) = Ak(A^\top \mathbf{x}, A^\top \mathbf{x}')$ to yield

$$\begin{aligned} \mathbb{E}_{\mathbf{X}, \mathbf{X}' \sim Q} \left[(A^\top \delta_{p,q}(\mathbf{X}))^\top A^\top \delta_{p,q}(\mathbf{X}') k(A^\top \mathbf{X}, A^\top \mathbf{X}') \right] &= \mathbb{E}_{\mathbf{X}, \mathbf{X}' \sim Q} \left[\mathbf{v}(\mathbf{X}, \mathbf{X}') A^\top \delta_{p,q}(\mathbf{X}') \right] \\ &= \mathbb{E}_{\mathbf{X}, \mathbf{X}' \sim Q} \left[\delta_{p,q}(\mathbf{X}')^\top A \mathbf{v}(\mathbf{X}, \mathbf{X}')^\top \right] . \end{aligned}$$

The assumption on the second order partial derivatives of the kernel and the proof of Liu et al. (2016, Theorem 3.6) assure us that we can apply Lemma 3.4 again with $F(\mathbf{x}') = A\mathbf{v}(\mathbf{x}, \mathbf{x}')^\top$ to get

$$\mathbb{E}_{\mathbf{X}, \mathbf{X}' \sim Q} \left[(A^\top \delta_{p,q}(\mathbf{X}))^\top A^\top \delta_{p,q}(\mathbf{X}') k(A^\top \mathbf{X}, A^\top \mathbf{X}') \right] \;=\; \mathbb{E}_{\mathbf{X}, \mathbf{X}' \sim Q} \left[\mathcal{A}_p(A\mathbf{v}(\mathbf{X}, \cdot)^\top)(\mathbf{X}') \right] .$$

Straightforward calculations along with $A^\top A = I_m$ show that

$$\begin{aligned} \mathcal{A}_p(A\mathbf{v}(\mathbf{x}, \cdot)^\top)(\mathbf{x}') \;=\; & (A^\top \mathbf{s}_p(\mathbf{x}))^\top (A^\top \mathbf{s}_p(\mathbf{x}')) k(A^\top \mathbf{x}, A^\top \mathbf{x}') + (A^\top \mathbf{s}_p(\mathbf{x}))^\top \nabla_2 k(A^\top \mathbf{x}, A^\top \mathbf{x}') \\ & + \nabla_1 k(A^\top \mathbf{x}, A^\top \mathbf{x}')^\top A^\top \mathbf{s}_p(\mathbf{x}') + \mathrm{Tr}(\nabla_{1,2} k(A^\top \mathbf{x}, A^\top \mathbf{x}')) , \end{aligned}$$

where ∇_i denotes the gradient of k with respect to the i-th argument, and $\nabla_{1,2} k$ is the matrix with ij-th entry $\partial^2 k / \partial \mathbf{x}_1^i \partial \mathbf{x}_2^j$. This completes the proof since it is the integrand of (3.7). □

Proof of Theorem 3.2. First, if $p = q$ then Lemma 3.5 immediately tells us that the discrepancy is zero since $\delta_{p,q} = 0$. Now suppose the discrepancy is zero. Take any projector A, then $\|A^\top \mathbf{x} - A^\top \mathbf{x}'\|_2 \leq \|\mathbf{x} - \mathbf{x}'\|_2$ since $AA^\top$ is a projection matrix. As k is a radial kernel we know that Ψ is monotonically decreasing (Scovel et al., 2010, Theorem 1.1). Therefore, $k(A\mathbf{x}, A\mathbf{x}') \geq k_d(\mathbf{x}, \mathbf{x}')$, where $k_d = \Psi(\|\mathbf{x} - \mathbf{x}'\|_2)$. By Lemma 3.5,

$$0 \;=\; \mathrm{GKSD}(Q, P) \;\geq\; \sup_{[A] \in \mathrm{Gr}(d,m)} \mathbb{E}_{\mathbf{X}, \mathbf{X}' \sim Q} \left[(A^\top \delta_{p,q}(\mathbf{X}))^\top A^\top \delta_{p,q}(\mathbf{X}') k_d(\mathbf{X}, \mathbf{X}') \right] . \tag{3.24}$$

Now choose A as the matrix formed by stacking the first m canonical basis elements. Since k is characteristic and radial, we know that $\Psi(\|\mathbf{x} - \mathbf{x}'\|_2) = L_\nu(\|\mathbf{x} - \mathbf{x}'\|_2)$ where L_ν is the Laplace transform of a measure ν on $[0, \infty)$ with full support (Sriperumbudur et al., 2011, Proposition 5). Therefore, $k_d(\mathbf{x}, \mathbf{x}') = \Psi(\|\mathbf{x} - \mathbf{x}'\|_2) = L_\nu(\|\mathbf{x} - \mathbf{x}'\|_2)$ and k_d is also characteristic and hence integrally strictly positive definite.

Using this we can conclude from (3.24) that $A^\top \delta_{p,q} = 0$, which implies $A^\top \mathbf{s}_q = A^\top \mathbf{s}_q$ so the first m entries of the two score functions are the same. Repeat this argument by setting A to be the next m canonical basis elements until we have checked all canonical basis elements. We then deduce that the score functions are equal in all coordinates, which implies $p = q$ as required. □

3.B.2 Proof of Proposition 3.3

We prove that, provided the densities of the distributions Q, P differ only in a subspace identified by a projector A_0 of rank m, the Riemannian gradient $\operatorname{grad}\alpha([A])$ is zero at $A = A_0$. That is, the optimal subspace $[A_0]$ is indeed a solution sought by the objective

$$\sup_{[A]\in \mathrm{Gr}(d,m)} \mathrm{KSD}_A(Q, P) \ ,$$

where $\mathrm{KSD}_A(Q, P)$ is given by (3.9).

In the rest of this section, we first state an assumption on the form of the kernel in Section 3.B.2.1 that is mild but can greatly simplify the proof. In Section 3.B.2.2, we then provide intuition on the assumption of the decomposibility of the densities of Q and P and give a concrete example. The proof is presented in Section 3.B.2.3.

3.B.2.1 Assumption on the Kernel

In Theorem 3.2, we have assumed the kernel takes the form $k(\mathbf{u}, \mathbf{v}) = \Psi(\|\mathbf{u} - \mathbf{v}\|_2)$, i.e. it is a *radial kernel* defined in Sriperumbudur et al. (2011). To simplify the form of $\operatorname{grad}\alpha([A])$ in our proof, we introduce the following reformulation of Ψ and impose a smoothness condition.

Assumption 3.1 (Smoothness of kernel)**.** *There exists a twice continuously differentiable function $\Phi : \mathbb{R} \to \mathbb{R}$ for which $\Psi(s) = \Phi(s^2)$ for all $s \geq 0$. That is, the kernel has the form $k(\mathbf{u}, \mathbf{v}) = \Phi(\|\mathbf{u} - \mathbf{v}\|_2^2)$.*

Remark 3.5.1. *The continuous differentiability condition on Φ is for convenience only and is mild. In particular, both Gaussian and Inverse Multi-Quadric kernels satisfy Assumption 3.1.*

3.B.2.2 Assumption on the Densities

Intuitively, the decomposition (3.14) and (3.15) states that the the candidate density q and the target p are distinct only in the subspace $[A_0]$. It can be viewed as a type of sparsity assumption. Both the multimodal mixture and the X-shaped mixture examples in the experiments satisfy this assumption, where it is easy to check that such functions q^m, p^m, ξ can be found with $[A_0] = \{ae_1 + be_2 : a, b \in \mathbb{R}\}$ being the subspace spanned by the first two canonical basis vectors e_1, e_2. See Section 3.6 for the precise definition of the target densities in these two examples.

More generally, this assumption holds when $q(\mathbf{x})$ is a prior, and $p(\mathbf{x}) \propto q(\mathbf{x}) f(y|P_0\mathbf{x})$ is the posterior induced by observational data y and a likelihood function $f(y|A_0A_0^\top\mathbf{x})$ that depends on $\mathbf{x}$ only through

its projection onto the subspace $[A_0]$. When the columns of A_0 are canonical basis vectors, this means the data-generating process only depends on a subset of the parameters $\mathbf{x}$. A similar setting is also considered by Chen and Ghattas (2020), where the authors argue that such a likelihood function is a reasonable approximation when the variation of the likelihood is negligible outside some eigenspace of a gradient information matrix.

3.B.2.3 Proof of the Main Result

We begin by deriving an expression for the Euclidean and Riemannian gradients of $\alpha([A])$ with respect to a projector A. The proof for Proposition 3.3 will then be provided later in this

Lemma 3.6. *Suppose that Assumption 3.1 holds, and that p, q satisfy the same assumptions in Theorem 3.2. Then, given a projector A, the Euclidean gradient of $\alpha([A])$ at A is*

$$\begin{aligned}\nabla\alpha([A]) \;&= 2\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}[k(A^\top\mathbf{X}, A^\top\mathbf{X}')\delta_{p,q}(\mathbf{X})\delta_{p,q}(\mathbf{X}')^\top A \\ &\quad + \Phi'\left(\|A^\top\mathbf{X} - A^\top\mathbf{X}'\|_2^2\right)\delta_{p,q}(\mathbf{X})^\top AA^\top\delta_{p,q}(\mathbf{X}')(\mathbf{X}-\mathbf{X}')(\mathbf{X}-\mathbf{X}')^\top A]\;, \quad (3.25)\end{aligned}$$

where $\mathbf{X}, \mathbf{X}' \sim Q$ are i.i.d. copies, Φ' is the first derivative of Φ, and $[\nabla\alpha([A])]_{ij} = \partial\alpha([A])/\partial A_{ij}$. Moreover, letting $\Pi_A \coloneqq I_d - AA^\top$, the Riemannian gradient of $\alpha([A])$ at A is $\operatorname{grad}\alpha([A]) = \Pi_A\nabla\alpha([A])$, where

$$\begin{aligned}\alpha([A]) \;&=\; 2\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\Big[k(A^\top\mathbf{X}, A^\top\mathbf{X}')\delta_{p,q}(\mathbf{X}')\delta_{p,q}(\mathbf{X})^\top A \quad (3.26a)\\ &\quad + \Phi'\left(\|A^\top\mathbf{X} - A^\top\mathbf{X}'\|_2^2\right)\delta_{p,q}(\mathbf{X}')^\top AA^\top\delta_{p,q}(\mathbf{X})(\mathbf{X}-\mathbf{X}')(\mathbf{X}-\mathbf{X}')^\top A\Big]\;. \quad (3.26b)\end{aligned}$$

Proof of Lemma 3.6. Letting G be the right-hand side of (3.25). We will show that, for any $t \in \mathbb{R}$ and projectors A, B of rank m,

$$\alpha([A+tB]) - \alpha([A]) \;=\; t\cdot\operatorname{Tr}(G^\top B) + \mathcal{O}(t^2)\;.$$

Let $\mathbf{w} \coloneqq \mathbf{x} - \mathbf{x}'$ so that $k(A^\top\mathbf{x}, A^\top\mathbf{x}') = \Phi(\|A^\top\mathbf{w}\|_2^2)$ for a projector A. Starting with $\alpha([A])$ in the form of (3.23),

$$\begin{aligned}\alpha([A+tB]) - \alpha([A]) \;&=\; \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[\delta_{p,q}(\mathbf{X})^\top(A+tB)(A+tB)^\top\delta_{p,q}(\mathbf{X}')\Phi\left(\left\|(A+tB)^\top\mathbf{w}\right\|_2^2\right)\right]\\ &\quad - \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[\delta_{p,q}(\mathbf{X})^\top AA^\top\delta_{p,q}(\mathbf{X}')\Phi\left(\left\|A^\top\mathbf{w}\right\|_2^2\right)\right]\;.\end{aligned}$$

With direct computation, the RHS of the above equation gives

$$\begin{aligned}&\alpha([A+tB]) - \alpha([A])\\ &= \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[\delta_{p,q}(\mathbf{X})^\top AA^\top\delta_{p,q}(\mathbf{X}')\left(\Phi\left(\left\|(A+tB)^\top\mathbf{w}\right\|_2^2\right) - \Phi\left(\left\|A^\top\mathbf{w}\right\|_2^2\right)\right)\right] \quad (3.27a)\\ &\quad + 2t\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[\delta_{p,q}(\mathbf{X})^\top AB^\top\delta_{p,q}(\mathbf{X}')\Phi\left(\left\|(A+tB)^\top\mathbf{w}\right\|_2^2\right)\right] \quad (3.27b)\\ &\quad + \mathcal{O}(t^2)\;. \quad (3.27c)\end{aligned}$$

We now seek a linear approximation of the function $A \mapsto \Phi(\|A^\top \mathbf{w}\|_2^2)$, where A is a projector of rank m. Since Φ is continuously differentiable by assumption, it admits a first-order Taylor expansion at $\|A^\top \mathbf{w}\|_2^2$, which gives

$$\begin{aligned}
&\Phi\left(\left\|(A+tB)^\top \mathbf{w}\right\|_2^2\right) - \Phi\left(\left\|A^\top \mathbf{w}\right\|_2^2\right) \\
&= \Phi\left(\left\|A^\top \mathbf{w}\right\|_2^2 + t\left(2w^\top AB^\top \mathbf{w} + t\left\|B^\top \mathbf{w}\right\|_2^2\right)\right) - \Phi\left(\left\|A^\top \mathbf{w}\right\|_2^2\right) \\
&= \Phi\left(\left\|A^\top \mathbf{w}\right\|_2^2\right) + t\Phi'\left(\left\|A^\top \mathbf{w}\right\|_2^2\right)\left(2w^\top AB^\top \mathbf{w} + t\left\|B^\top \mathbf{w}\right\|_2^2\right) - \Phi\left(\left\|A^\top \mathbf{w}\right\|_2^2\right) + \mathcal{O}(t^2) \\
&= t\Phi'\left(\left\|A^\top \mathbf{w}\right\|_2^2\right)\left(2w^\top AB^\top \mathbf{w}\right) + \mathcal{O}(t^2)\,. \qquad (3.28)
\end{aligned}$$

The rest of the proof for the form of the Euclidean gradient follows by substituting (3.28) into (3.27) and using properties of the trace operator:

$$\begin{aligned}
\alpha([A+tB]) - \alpha([A]) &= 2t\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[\Phi'\left(\left\|A^\top \mathbf{w}\right\|_2^2\right)\delta_{p,q}(\mathbf{X})^\top AA^\top \delta_{p,q}(\mathbf{X}')\mathbf{w}^\top AB^\top \mathbf{w}\right] \\
&\quad + 2t\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[\delta_{p,q}(\mathbf{X})^\top AB^\top \delta_{p,q}(\mathbf{X}')\Phi\left(\left\|A^\top \mathbf{w}\right\|_2^2\right)\right] + \mathcal{O}(t^2) \\
&= 2t\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[\Phi'\left(\left\|A^\top \mathbf{w}\right\|_2^2\right)\delta_{p,q}(\mathbf{X})^\top AA^\top \delta_{p,q}(\mathbf{X}')\mathrm{Tr}\left(\mathbf{w}\mathbf{w}^\top AB^\top\right)\right] \\
&\quad + 2t\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[\Phi\left(\left\|A^\top \mathbf{w}\right\|_2^2\right)\mathrm{Tr}\left(\delta_{p,q}(\mathbf{X}')\delta_{p,q}(\mathbf{X})^\top AB^\top\right)\right] + \mathcal{O}(t^2) \\
&= t\cdot\mathrm{Tr}\left(2\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[\Phi'\left(\left\|A^\top \mathbf{w}\right\|_2^2\right)\delta_{p,q}(\mathbf{X})^\top AA^\top \delta_{p,q}(\mathbf{X}')\mathbf{w}\mathbf{w}^\top AB^\top\right]\right) \\
&\quad + t\cdot\mathrm{Tr}\left(2\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[\Phi\left(\left\|A^\top \mathbf{w}\right\|_2^2\right)\delta_{p,q}(\mathbf{X}')\delta_{p,q}(\mathbf{X})^\top AB^\top\right]\right) + \mathcal{O}(t^2) \\
&= t\cdot\mathrm{Tr}(G^\top B) + \mathcal{O}(t^2)\,,
\end{aligned}$$

where in the last equality we have substituted the definition of G and used the fact $\mathrm{Tr}(GB^\top) = \mathrm{Tr}(G^\top B)$.

Finally, it can be shown that $\mathrm{Gr}(d,m)$ is a Riemannian submanifold of the Euclidean manifold, so the Riemannian gradient is the orthogonal projection of the Euclidean gradient to the tangent spaces (see e.g., Boumal 2020, Proposition 3.53). That is, $\mathrm{grad}\,\alpha([A]) = (I_d - AA^\top)\nabla\alpha([A]))$, so (3.26) follows. □

Proof of Proposition 3.3. Decomposing (3.26) into two terms, we have

$$\begin{aligned}
\mathrm{grad}\,\alpha([A]) &= 2\Pi_A\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}[k(A^\top\mathbf{X}, A^\top\mathbf{X}')\delta_{p,q}(\mathbf{X}')\delta_{p,q}(\mathbf{X})^\top A] \\
&\quad + 2\Pi_A\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}[\Phi'\left(\|A^\top\mathbf{X} - A^\top\mathbf{X}'\|_2^2\right)\delta_{p,q}(\mathbf{X}')^\top AA^\top\delta_{p,q}(\mathbf{X})(\mathbf{X}-\mathbf{X}')(\mathbf{X}-\mathbf{X}')^\top A]\,, \qquad (3.29)
\end{aligned}$$

where $\Pi_A = I_d - AA^\top$. We will show that both terms equal to zero at $A = A_0$. Firstly, define

$\tilde{\xi}_{\Pi_{A_0}}(\mathbf{x}) = \xi(\Pi_{A_0}\mathbf{x})$ and $P_0 = A_0 A_0^\top$. The score function of the candidate density is

$$\begin{aligned}\nabla_{\mathbf{x}} \log q(\mathbf{x}) &= \nabla_{\mathbf{x}} \log(q^m \circ P_0)(\mathbf{x}) + \nabla_{\mathbf{x}} \log \tilde{\xi}_{\Pi_{A_0}}(\mathbf{x}) \\ &= P_0 \nabla_{\mathbf{x}} \log q^m(P_0\mathbf{x}) + \nabla_{\mathbf{x}} \log \tilde{\xi}_{\Pi_{A_0}}(\mathbf{x}) \ ,\end{aligned}$$

where $\nabla_{\mathbf{x}} \log q^m(P_0\mathbf{x})$ denotes the gradient of $\log q^m$ evaluated at $P_0\mathbf{x}$, and where the second line follows from the chain rule and the symmetry of P_0. Similarly, the score function of the target density takes the form

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}) = P_0 \nabla_{\mathbf{x}} \log q^m(P_0\mathbf{x}) + \nabla_{\mathbf{x}} \log \xi(\mathbf{x}) \ .$$

Taking the difference,

$$\delta_{p,q}(\mathbf{x}) = P_0(\nabla_{\mathbf{x}} \log p^m(P_0 x) - \nabla_{\mathbf{x}} \log q^m(P_0\mathbf{x})) \ .$$

Since $\Pi_{A_0} P_0 = 0$, we conclude that $\Pi_{A_0}\delta_{p,q}(\mathbf{x}) = 0$, and the first term of (3.29) is zero.

For the second term, we define $a \coloneqq \delta_{p,q}^\top(\mathbf{x}') A_0 A_0^\top \delta_{p,q}(\mathbf{x})$ and $b \coloneqq \Phi'(\|A^\top\mathbf{x} - A^\top\mathbf{x}'\|_2^2)$. Since both the terms $\delta_{p,q}(\mathbf{x})$ and $A_0^\top\mathbf{x} - A_0^\top\mathbf{x}' = A_0^\top(P_0\mathbf{x} - P_0\mathbf{x}')$ are deterministic given $(P_0\mathbf{x}, P_0\mathbf{x}')$, so are a and b. We further introduce the notations $\mathbf{w} \coloneqq P_0\mathbf{x}, \mathbf{w}' \coloneqq P_0\mathbf{x}'$ and $\mathbf{v} \coloneqq \Pi_{A_0}\mathbf{x}, \mathbf{v}' \coloneqq \Pi_{A_0}\mathbf{x}'$. Since q^m, p^m and ξ are unnormalized densities over $\mathbb{R}^d$, we can invoke the tower rule to simplify the second term as

$$\begin{aligned}\Pi_{A_0}\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[ab(\mathbf{X}-\mathbf{X}')(\mathbf{X}-\mathbf{X}')^\top A_0\right] &= \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[ab(\Pi_{A_0}\mathbf{X}-\Pi_{A_0}\mathbf{X}')(P_0\mathbf{X}-P_0\mathbf{X}')^\top A_0\right] \\ &= \mathbb{E}_{\mathbf{W},\mathbf{W}'\sim q^m}\left[\mathbb{E}_{\mathbf{v},\mathbf{v}'\sim q^\perp}\left[ab(\mathbf{v}-\mathbf{v}')(\mathbf{W}-\mathbf{W}')^\top A_0 \mid \mathbf{W},\mathbf{W}'\right]\right] \\ &= \mathbb{E}_{\mathbf{W},\mathbf{W}'\sim q^m}\left[ab\mathbb{E}_{\mathbf{v},\mathbf{v}'\sim q^\perp}\left[\mathbf{v}-\mathbf{v}' \mid \mathbf{W},\mathbf{W}'\right](\mathbf{W}-\mathbf{W}')^\top A_0\right] \ ,\end{aligned}$$

where the first line follows from the fact that $A_0 = A_0 A_0^\top A_0 = P_0 A_0$. Since $\mathbf{X}, \mathbf{X}'$ are i.i.d. copies, we have $\mathbb{E}_{\mathbf{V},\mathbf{V}'\sim q^\perp}[\mathbf{V}-\mathbf{V}'|\mathbf{W},\mathbf{W}'] = \mathbb{E}_{\mathbf{V}\sim q^\perp}[\mathbf{V} \mid \mathbf{W}] - \mathbb{E}_{\mathbf{V}'\sim q^\perp}[\mathbf{V}' \mid \mathbf{W}'] = 0$, so the second term of (3.29) is also zero. This shows that $\operatorname{grad}\alpha([A]) = 0$ at $A = A_0$, as required. □

3.C Experimental Details

In this section, we provide implementation details for the experiments in Section 3.6.

Learning Rates The best step sizes of the gradient method for the particle update in all methods, and for the projector or slice update in GSVGD and S-SVGD, are selected over a grid of values from 10^{-4} to 1 such that they give the minimal energy distance between the particle estimation and the ground truth. This is done for the 50-dimensional multimodal example only, and the same set of learning rates is then used for all experiments. This is done for all methods to ensure a fair comparison.

Temperature T For the temperature parameter T in (3.21), we adopt an annealing scheme where we start with $T = T_0$ and gradually increment it to a large value, T_{large}. Intuitively, setting T small

would lead to faster convergence, but hinders exploration of the maxima of $\alpha_t([A]) = \mathrm{KSD}_A(Q_t, P)$. On the other hand, a large T means the system (3.20) is effectively sampling projectors at random from the Grassmann manifold, thus allowing better exploration of the modes at the expense of a slower convergence. The annealing scheme allows trade-offs between such exploration and exploitation. Starting from T_0, at each iteration t, we multiply T by 10 if the change $|\gamma_t - \gamma_{t-1}|$ in the *particle-averaged magnitude* (Zhuo et al., 2018) of the update function, defined as $\gamma_t \coloneqq \frac{1}{N}\sum_{i=1}^{N} \| \sum_{l=1}^{M} \widehat{\phi}_{A_{t,l}}(\mathbf{x}_i^t)\|_\infty$, is less than a pre-specified threshold, where N is the sample size, M is the number of projectors, and the update function $\widehat{\phi}_{A_{t,l}}$ is defined in (3.11). In the experiments, we used $T_0 = 10^{-4}$ and $T_{\text{large}} = 10^6$, and set the threshold to be $10^{-4}M$. This choice of the threshold is motivated by the dependence of the GSVGD particle update on the number of projectors.

Initialization of Projectors The projectors in GSVGD are initialized with matrices formed by one-hot vectors. This is similar to the setup in S-SVGD, where the slices are initialized to be the canonical basis elements. When more than one projector is used in GSVGD, the projection dimensions, m, are kept fixed for all projectors for simplicity. The number of projectors, M, is chosen such that the projection dimensions sum up to d and is capped at 20, where d is the dimension of the problem. That is, $M = \min(20, \lfloor d/m \rfloor)$, where $\lfloor a \rfloor$ denotes the largest integer smaller than or equal to a.

Other Setups For SVGD, we follow exactly the same setups as in Liu and Wang (2016). For S-SVGD, we follow mostly the same setups in Algorithm 2 and Appendix G of Gong et al. (2021a), except we updated the slices at every step instead of only when the particles have moved a sufficiently far distance from the previous step, as suggested in their paper. As noted in their paper, this trick is to prevent over-fitting of the slices to small samples, which is unlikely an issue in our experiments due to the large sample size we chose. Also, we remark that, for GSVGD, the coupled ODE-SDE system introduced in Section 3.5 suggests a principled way to balance the evolution of the particles and projectors in GSVGD by varying the temperature T.

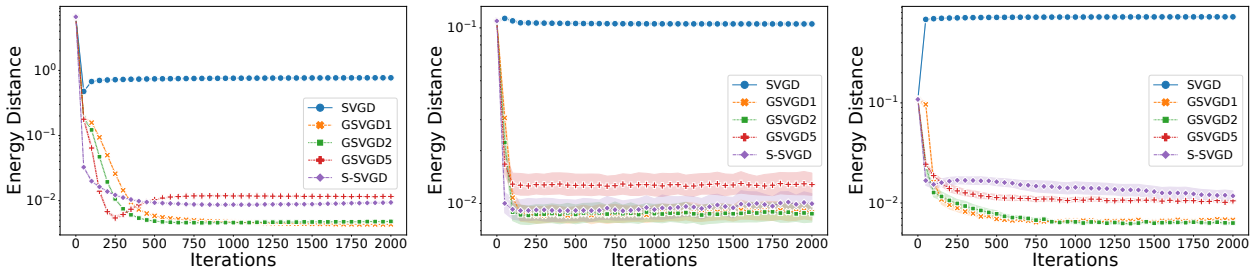
We ran each experiment with 20 random seeds, except Bayesian logistic regression which was repeated 10 times. The mean and 95% confidence intervals are reported in all figures. A No-U-Turn Sampler (NUTS) (Hoffman et al., 2014) is used as a gold standard in the conditioned diffusion and Bayesian logistic regression experiments. Each method ran for 2000 iterations, which we found to be sufficient for convergence (see Figure 3.9).

3.D Supplementary Figures

In this section, we include supplementary figures for the experiments in Section 3.6 as well as some ablation studies.

3.D.1 Synthetic Experiments

Figure 3.9 supplements Figure 3.5, and shows the convergence of SGVD, S-SVGD and GSVGD to the target distribution in the three synthetic experiments in Section 3.6. We can see that the smallest



(a) Multivariate Gaussian. (b) Multimodal mixture. (c) X-shaped mixture.

Figure 3.9: Convergence of particles in the (a) multivariate Gaussian, (b) multimodal mixture and (c) X-shaped experiments. The dimensionality is 50 in all cases.

energy distance between the particle estimation and the ground truth is achieved by GSVGD1 and GSVGD2, followed by S-SVGD and GSVGD5. GSVGD with any projection dimension has comparable per-iteration convergence rate as S-SVGD except for the multivariate Gaussian target. We believe this is because of the extra noise introduced via the coupled ODE-SDE system described in Section 3.5, which added unnecessary complexities for this simple, single-mode target.

3.D.2 Conditioned Diffusion Process

Figure 3.11 plots the energy distance between the particle estimation of each method and a HMC sampler against different projection dimensions, m. We see that there exists a feasible region of projection dimensions where GSVGD achieves a greater degree of agreement with the reference HMC, compared with SVGD and S-SVGD. This is expected given the problem setup: the forcing term is a Brownian motion with covariance $C(t, t') = \min(t, t')$, and hence is correlated between time points. This correlation means the solution state u will admit a low-dimensional structure.

Although in this case an appropriate choice of projection dimension could be inferred from the problem setup, selecting the best projection dimension for a general problem is an open question. Nonetheless, the shape of the curves in Figure 3.11 suggests that a feasible region should not cover the extreme ends.

3.D.3 Ablation Studies

Sample Size To evaluate the influence of sample size on the performance of GSVGD, we show the dimension-averaged variance estimates for the multivariate Gaussian target. We follow the same experimental setup in Section 3.6: for each of $m = 1, 2, 5$, we run each method for 2000 iterations and compute the dimension-averaged marginal variance of the particle estimation of the last iteration. Figure 3.10 shows the estimates for GSVGD1, SVGD and S-SVGD when the target is 50-dimensional. GSVGD with other projection dimensions is not included in the figure for clarity. We see that GSVGD1 remains robust with the sample size and achieves accurate estimation. In contrast, SVGD severely underestimates the variance in high dimensions regardless of the sample size, and S-SVGD shows slight overestimation.

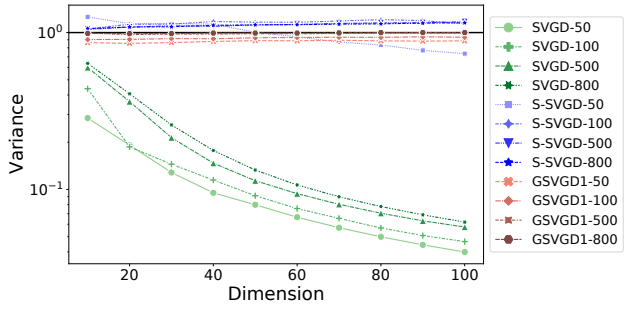


Figure 3.10: Variance estimates for different sample sizes, where the true value is shown by the black solid line. SVGD-50 means 50 particles are used to estimate the variance.

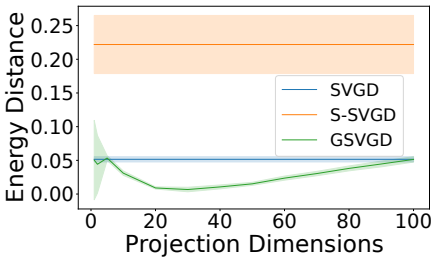


Figure 3.11: Conditioned Diffusion Process: Energy distance between HMC and GSVGD solutions against different projection dimensions. SVGD and S-SVGD are also included for comparison.

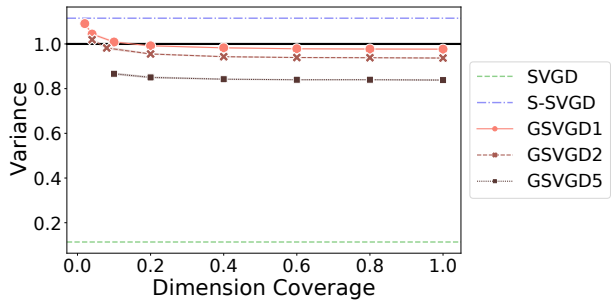


Figure 3.12: Ablation study on the number of projectors M. The target is a 50-dimensional multivariate Gaussian. Black solid line marks the true value.

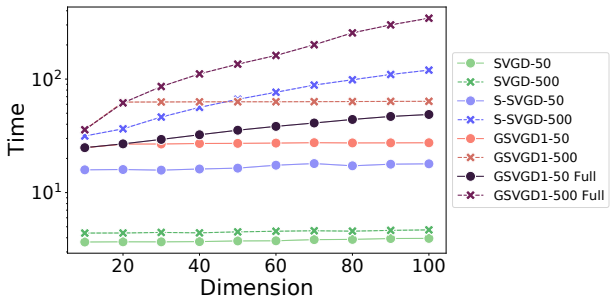


Figure 3.13: Run time (in seconds) against the dimension of the multivariate Gaussian target.

Number of Projectors In Figure 3.12, we perform an ablation study of the performance of GSVGD with respect to the number of projectors M used in Algorithm 4. The target is again the 50-dimensional multivariate Gaussian distribution. The estimate of the dimension-averaged marginal variance is plotted against the *dimension covergae*, which is defined as $(M * m)/d$; that is, it is the total ranks of the M projectors as a proportion of dimension. For a given m, we choose $M \in \{m, 10, 20, \ldots, 50\}$. We also include the result for $M = 2$ and 5 when $m = 1$, and $M = 2$ when $m = 2$.

We observe that the variance estimate of GSVGD improves as M increases until the dimension coverage reaches roughly 0.2, beyond which the performance stabilizes. This suggests that using more projectors does not necessarily improve the estimation quality of GSVGD. It also justifies our choice of $M = \min(20, \lfloor d/m \rfloor)$ (corresponding to a dimension coverage of 0.4 for $m = 1$, 0.8 for $m = 2$ and 1.0 for $m = 5$). We remark that the estimated variance of GSVGD1 and GSVGD2 are consistently better than S-SVGD and SVGD for all choices of M.

Time Complexities We compare the run time of GSVGD against its competitors in Figure 3.13. In GSVGD, we used $M = \min(20, \lfloor d/m \rfloor)$ projectors of rank $m = 1$ (GSVGD1-50, GSVGD1-500, where GSVGD1-50 means 50 particles were used). We have capped M to be at most 20 instead of setting it to the dimensionality d because this was the default setting in all experiments; however, since S-SVGD uses all d slices, we also include the run time of GSVGD1 when using $M = d$ projectors (GSVGD1-50

Full, GSVGD1-500 Full) for a fair comparison.

We see that the run time of GSVGD1-50 Full and GSVGD1-500 Full grows an order-of-magnitude faster than S-SVGD even though they use the same projection dimension and number of projectors/slices. This is due to the extra matrix computation in the particle update of GSVGD, which in turn arises from the fact that the projectors for the score functions are allowed to vary in GSVGD but not in S-SVGD. In particular, the time complexity for a fixed sample size is $\mathcal{O}(d^3)$ for GSVGD1-50 Full and GSVGD1-500 Full, in contrast to $\mathcal{O}(d^2)$ for S-SVGD. However, capping M to be at most 20 can greatly reduce the run time of GSVGD in high dimensions. In fact, with such a choice of M the computational complexity reduces to $\mathcal{O}(d^2)$, as can be seen from the curves GSVGD1-50 and GSVGD1-500.

Finally, the computational burden of the projector update in GSVGD1-50 Full and GSVGD1-500 Full ($\mathcal{O}(d^2)$) is less than that of the particle update. Hence, the main computational bottleneck of GSVGD arises from the particle update (3.11).

Chapter 4

Goodness-of-Fit Testing Part I: Using Perturbation to Improve KSD Tests

The kernel Stein discrepancy (KSD) test is a class of goodness-of-fit tests based on the score function of statistical models. These tests are applicable even when the target distribution has an unknown normalizing factor, such as in Bayesian analysis, for which traditional methods such as likelihood-ratio-tests are no longer feasible. We show theoretically and empirically that the KSD test can suffer from low power when the target distribution is multi-modal with misspecified mixing proportions of the modes. To address this issue, we propose to perturb the observed sample via Markov transition kernels with respect to which the target distribution is invariant. This allows us to then employ the KSD test on the perturbed sample. We provide numerical evidence that with suitably chosen transition kernels, the proposed approach can lead to substantially higher power than the standard KSD test.

4.1 Introduction

Stein discrepancy (SD) (Stein, 1972; Gorham and Mackey, 2015) is a statistical divergence between two probability measures based on Stein's method. More specifically, given two Borel probability measures Q and P supported on $\mathcal{X} \subseteq \mathbb{R}^d$, the Stein discrepancy is defined in (2.4) and recalled below

$$\mathcal{S}(Q, P; \mathcal{F}) \coloneqq \sup_{f \in \mathcal{F}} \mathbb{E}_{\mathbf{X} \sim Q}[\mathcal{A}_P f(\mathbf{X})] \,,$$

where $\mathcal{F}$ is a set of functions on $\mathcal{X}$ and $\mathcal{A}_P$ is an operator acting on $\mathcal{F}$ such that $\mathbb{E}_{\mathbf{X} \sim Q}[\mathcal{A}_P f(\mathbf{X})] = 0$ for all $f \in \mathcal{F}$ if and only if $Q \equiv P$. When $\mathcal{X} = \mathbb{R}^d$ and P admits a positive, continuously differentiable density p with respect to the Lebesgue measure, then the *Langevin Stein operator* is the natural candidate for $\mathcal{A}_P$ which has the crucial property that it only depends on the *score function* $\mathbf{s}_p(\mathbf{x}) \coloneqq \nabla \log p(\mathbf{x})$ of p, which does not require evaluation of the (possibly intractable) normalizing constant of p. When a Reproducing Kernel Hilbert Space (RKHS) (Berlinet and Thomas-Agnan, 2004) is used to construct the function class, this discrepancy is called the *kernel Stein discrepancy* (KSD), and admits a closed-form expression. This has made KSD popular for applications involving an unnormalized density, such as

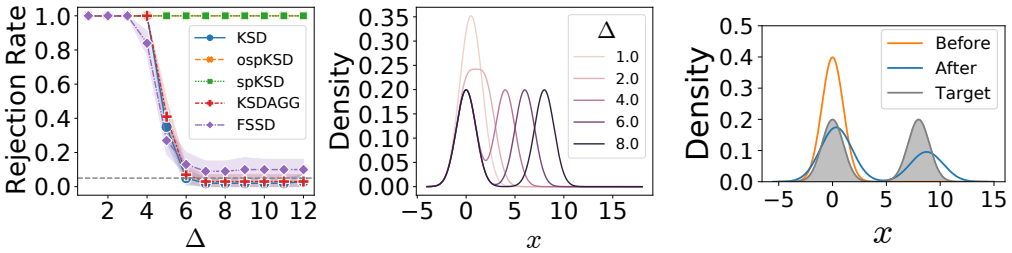


Figure 4.1: Power for a one-dimensional bimodal Gaussian target distribution P with mixing weight 0.5 and mode separation Δ. The candidate distribution Q is only the left component, from which 1000 samples are drawn. Our proposed methods are ospKSD and spKSD; the others are existing benchmarks. *Left and middle.* Rejection rates and target densities for varying Δ; the orange and green lines overlap. *Right.* Density of Q before and after 10 steps of the perturbation described in Section 4.5, as well as the density of the target P.

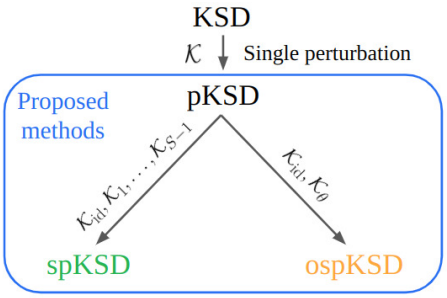


Figure 4.2: Connections between KSD and the proposed divergences.

Bayesian inference (Liu and Wang, 2016), goodness-of-fit testing (Liu et al., 2016; Chwialkowski et al., 2016; Jitkrittum et al., 2017), and sample quality measurement (Gorham and Mackey, 2015, 2017; Gorham et al., 2019); see Anastasiou et al. (2023) for a review.

We focus on goodness-of-fit (GOF) testing, where independent samples from a *candidate distribution* Q are observed and the goal is to test for evidence against the null hypothesis that Q matches a *target distribution* P. When the density p of P is only available in an unnormalized form (i.e. the normalization constant is infeasible to compute) and direct sampling from P is infeasible, classical tests such as the Kolmogorov-Smirnov test (Massey Jr, 1951) or two-sample tests (Gretton et al., 2012; Schrab et al., 2023) cannot be used, as they require either a tractable cumulative distribution function or samples from P. A GOF test based on KSD, on the other hand, does not have these limitations.

However, KSD tests may suffer from low test power when the target probability measure has well-separated modes. For example, when Q and P are mixtures of the same components but differ in the mixing proportions, the power will converge to the test level as the modes of the components become more and more separated (Figure 4.1). This is because the KSD statistic can be numerically close to 0 if the region where the score functions are practically different has low Q-probability. This issue, sometimes known as the *blindness to isolated components* (Wenliang and Kanagawa, 2020), has been noted in a number of works (Gorham et al., 2019; Matsubara et al., 2022; Kanagawa et al., 2022) and addressed in some applications of KSD (Zhang et al., 2022). However, little work has been devoted to tackling this issue in the context of GOF testing.

Contributions Our contribution is twofold. First, we demonstrate theoretically and numerically with a bimodal Gaussian example that the power of KSD tests can converge to the test level when the target distribution has well-separated modes. This is different from the works of Gorham and Mackey (2017) and Wenliang and Kanagawa (2020), which focus on the convergence of the sample KSD but not its limiting null distribution. Second, we address this issue by introducing a *perturbation operator*, giving rise to a family of perturbation-based GOF tests (Figure 4.2) which we call *perturbed kernel Stein discrepancy* (pKSD) tests. The role of the operator is to perturb the candidate and the target distributions simultaneously to create discrepancy that can be more easily detected by KSD. We propose to use Markov transition kernels that are invariant to the target P as the perturbation operator. The P-invariance ensures the resulting GOF tests provably control the Type-I error. The transition kernel is non-irreducible and uses an inter-modal jump proposal, which can increase the test power against multi-modal alternatives, sometimes substantially from the nominal level to almost 1.

Outline Section 4.2 reviews kernel Stein discrepancy. Section 4.4 formalizes the low-power problem of the KSD test. The proposed method is presented in Section 4.5 and Section 4.6. We discuss related work in Section 4.3, followed by experiments in Section 4.7. Section 4.8 concludes.

Notation Throughout this article, we denote by Q, P probability measures on $\mathcal{X} = \mathbb{R}^d$ equipped with the Borel σ-algebra $\mathcal{B}(\mathcal{X})$, and assume P has a continuously differentiable, positive Lebesgue density p. We refer to Q as the *candidate* distribution and P as the *target* distribution. Our interest lies in testing $H_0 : Q = P$ against $H_1 : Q \neq P$ using a finite sample $\{\mathbf{X}_i\}_{i=1}^n$ drawn independently from Q. We assume we can evaluate pointwise the *unnormalized* density $p^*(\mathbf{x}) = p(\mathbf{x})/Z$, where Z is an unknown constant, as well as $\nabla \log p^*(\mathbf{x})$, which is identical to the *score function* of p, namely $\mathbf{s}_p(\mathbf{x}) \coloneqq \nabla \log p(\mathbf{x}) = (\partial_{x^1} \log p(\mathbf{x}), \ldots, \partial_{x^d} \log p(\mathbf{x}))^\top$, where $\partial_{x^j} \log p(\mathbf{x})$ is the partial derivative of $\log p(\mathbf{x})$ with respect to the j-th component of $\mathbf{x}$.

4.2 Background

Choosing the Stein operator $\mathcal{A}_P$ in (2.4) to be the operator mapping continuously differentiable, vector-valued functions $f : \mathbb{R}^d \to \mathbb{R}^d$ to scalar-valued functions via $\mathcal{A}_p f(\mathbf{x}) \coloneqq \langle \nabla \log p(\mathbf{x}), f(\mathbf{x}) \rangle + \langle \nabla, f(\mathbf{x}) \rangle$, one obtains a statistical divergence which only depends on the score function of P. If $f \in \mathcal{F}$ satisfies regularity conditions such as $\lim_{\|\mathbf{x}\|_2 \to \infty} f(\mathbf{x})p(\mathbf{x}) = 0$, then one can show that $\mathbb{E}_{\mathbf{X} \sim P}[\mathcal{A}_P f(\mathbf{X})] = 0$, and f is said to lie in the *Stein class* of P (Liu et al., 2016, Section 2.2). The function class $\mathcal{F}$ is usually chosen to be *(i)* sufficiently broad so that the discrepancy separates distinct probability measures, i.e. $\mathcal{S}(Q, P; \mathcal{F}) = 0 \iff Q = P$, and *(ii)* sufficiently regular so that the right-hand-side of (2.4) can be efficiently solved.

To this end, Liu et al. (2016); Chwialkowski et al. (2016) proposed to construct $\mathcal{F}$ using *reproducing kernel Hilbert spaces* (RKHSs) (Berlinet and Thomas-Agnan, 2004). Specifically, let $\mathcal{H}_k$ be an RKHS associated with positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let $\mathcal{G}_k$ be the unit ball of the d-times Cartesian product $\mathcal{H}_k^d \coloneqq \{f = (f_1, \ldots, f_d) : f_i \in \mathcal{H}_k, \forall i\}$, equipped with the norm $\|f\|_{\mathcal{H}_k^d} = \left(\sum_{i=1}^d \|f\|_{\mathcal{H}_k}^2\right)^{1/2}$.

Choosing $\mathcal{F} = \mathcal{G}_k$ and the operator $\mathcal{A}_P$ yields the *(Langevin) kernel Stein discrepancy* (KSD): $D(Q, P) \coloneqq \mathcal{S}(Q, P; \mathcal{G}_k)$.

Assuming the kernel k has continuous first-order derivatives with respect to both arguments, Chwialkowski et al. (2016, Theorem 2.1) showed that KSD attains a closed form: $D(Q, P) = \mathbb{E}_{\mathbf{X},\mathbf{X}' \sim Q}[u_P(\mathbf{X}, \mathbf{X}')]$, where $\mathbf{X}, \mathbf{X}'$ are independent random variables drawn from Q, and u_P is the *Stein reproducing kernel*, which is defined in (2.10) and recalled below:

$$u_P(\mathbf{x}, \mathbf{x}') \coloneqq \mathbf{s}_p(\mathbf{x})^\top k(\mathbf{x}, \mathbf{x}')\mathbf{s}_p(\mathbf{x}') + \mathbf{s}_p(\mathbf{x})^\top \nabla_{\mathbf{x}'} k(\mathbf{x}, \mathbf{x}') + \nabla_x k(\mathbf{x}, \mathbf{x}')^\top \mathbf{s}_p(\mathbf{x}') + \nabla_1^\top \nabla_2 k(\mathbf{x}, \mathbf{x}') \,,$$

Crucially, u_P (hence also $D(Q, P)$) depends on p only through $\mathbf{s}_p(\mathbf{x}) = \nabla \log p(\mathbf{x})$, so KSD is computable even without the knowledge of the normalizing constant of p.

We will assume $k(\cdot, \mathbf{x})$ lies in the *Stein class* of p for all $\mathbf{x} \in \mathbb{R}^d$, so that $D(P, P) = 0$ (Liu et al., 2016, Theorem 3.6, Proposition 3.3). When Q also admits a density q, and k is *cc*-universal (Sriperumbudur et al., 2011, 2010a) or integrally strictly positive definite (Stewart, 1976, Section 6), KSD is *separating*, meaning that $D(Q, P) = 0 \iff Q = P$, provided that $\mathbb{E}_{\mathbf{x} \sim Q}[\|\mathbf{s}_q(\mathbf{x}) - \mathbf{s}_p(\mathbf{x})\|^2] < \infty$ (Chwialkowski et al., 2016; Liu et al., 2016). The assumption that Q has a density can be relaxed if the target density satisfies additional tail conditions, such as distant dissipativeness (Hodgkinson et al., 2020, Proposition 4); see also Theorem 2.6.

The KSD $D(Q, P)$ can be estimated using a random sample $\{\mathbf{X}_i\}_{i=1}^n$ from Q by the following U-statistic (Serfling, 2009, Section 5.5):

$$\hat{D}_P^2 \coloneqq \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} u_P(\mathbf{X}_i, \mathbf{X}_j) \,. \tag{4.1}$$

We have used a different notation from (2.11) in order to emphasize the dependence on the target distribution P, which will become convenient when introducing our proposed methods later.

The KSD test uses (4.1) as a test statistic. The asymptotic distribution of $\hat{D}_P^2$ under H_0 has no closed form, but can be approximated with a bootstrap procedure (Huskova and Janssen, 1993) using the bootstrap samples

$$\hat{D}_{P,b}^2 \coloneqq \frac{1}{n^2} \sum_{1 \le i \ne j \le n} \left(W_i^b - 1\right)\left(W_j^b - 1\right) u_P(\mathbf{X}_i, \mathbf{X}_j) \,, \tag{4.2}$$

where $(W_1^b, \ldots, W_n^b) \sim \text{Multinomial}\left(n; \frac{1}{n}, \ldots, \frac{1}{n}\right)$ follows a multinomial distribution. The test statistic $\hat{D}_P^2$ is compared against quantiles of $\{\hat{D}_{P,b}^2\}_{b=1}^B$ computed with B i.i.d. draws $(W_1^b, \ldots, W_n^b)$, and H_0 is rejected for large values of $\hat{D}_P^2$. The resulting test achieves the desired level α asymptotically (Huskova and Janssen, 1993; Liu et al., 2016).

Many improvements over the standard KSD test have been proposed, e.g., to reduce the computational cost (Jitkrittum et al., 2017), to address the curse-of-dimensionality (Gong et al., 2021a,b), and to avoid kernel selection by adopting an aggregated testing procedure (Schrab et al., 2022). Moreover, Hagrass et al. (2024) showed that the KSD test is sub-optimal under a certain alternative space, and showed that it can be made minimax optimal by incorporating spectral regularization.

4.3 Related Work

Blindness of score-based divergences The myopia of score-based divergences to mixing proportions is a well-known issue in the literature (Oates et al., 2017; Gorham et al., 2019; Matsubara et al., 2022; Wenliang and Kanagawa, 2020), yet it still lacks a solution. Wenliang and Kanagawa (2020) showed on simple benchmarks that the blindness of KSD-based methods cannot be mitigated by annealing, one approach commonly adopted in Markov chain Monte Carlo to improve their mixing against multi-modal targets. Zhang et al. (2022) addressed the myopia of KSD in the context of parameter learning. However, their method is specific for that application and is not applicable to GOF tests based on KSD. To our knowledge, our work is the first to address this limitation for KSD tests, and our proposed method of improving the test power using Markov transition kernels, which will be articulated in Section 4.5, has not been explored as a solution to the myopia of score-based divergences.

Perturbation with convolution The idea of combining a discrepancy with perturbation has been widely studied in other contexts, where the perturbation is often a convolution operator with Gaussian noise. E.g., the *spread divergence* (Zhang et al., 2020) combines Gaussian convolution with Kullback-Leibler (KL) divergence (more generally, f-divergences) to solve the issue that KL divergence is ill-defined when the distributions have undefined densities or unmatched support. In generative modelling, *denoising score matching* (Vincent, 2011) and *Noise Conditional Score Networks* (Song and Ermon, 2019) combine Gaussian convolution with score matching to improve computational efficiency or estimation quality. Notably, convolution is *not* invariant to the target distribution, thus rendering the score function intractable. This is why we will chose a P-invariance Markov kernel.

Perturbation with convex combination In score matching, Zhang et al. (2022) addresses the blindness of Fisher Divergence by mapping the target and candidate distributions to a convex combination with a Gaussian distribution, thereby "connecting" the well-separated modes. A similar idea cannot be applied to improve the KSD test, as, similarly to convolution, the resulting target distribution no longer has a tractable score function.

Perturbation with annealing Another choice of perturbation is to anneal both distributions by raising the densities to some power, which is studied in Wenliang and Kanagawa (2020). Although the score function remains tractable under this perturbation, annealing alone cannot solve the blindness of score-based discrepancies, as noted by the authors and in Zhang et al. (2020). Moreover, sampling from the annealed candidate distribution is also non-trivial.

4.4 Limitations of KSD Test

The KSD can be blind to certain discrepancies that are strongly visible in other metrics (e.g., in the L_2 norm). One example is mixtures of well-separated components that differ only by the mixing proportions (weights). More generally, KSD will be small when the region where the score difference $\|s_p(x) - s_q(x)\|_2$ takes large values has low Q-probability.

This is known as the "blindness" of KSD (Wenliang and Kanagawa, 2020). This limitation of KSD has been highlighted in a number of works concerning the use of KSD for convergence assessment, model fitting, and generalized Bayesian inference (Gorham et al., 2019; Matsubara et al., 2022; Zhang et al., 2022; Kanagawa et al., 2022); however, its implication to the test power in GOF tests has not yet been formalized.

In Proposition 4.1 (proved in Section 4.A.1), we formally connect the blindness issue with the rates of increase of the sample size and the *Fisher Divergence* (FD) between the two distributions. The FD between two distributions with densities p and q is defined as $F(q,p) \coloneqq \mathbb{E}_{\mathbf{X}\sim Q}[\|\mathbf{s}_p(\mathbf{X}) - \mathbf{s}_q(\mathbf{X})\|_2^2]$ (Liu et al., 2016, Theorem 5.1).

Proposition 4.1. *Let Q and P_ν, $\nu = 1, 2, \ldots$, be probability measures defined on $\mathbb{R}^d$ with positive densities q and p_ν, respectively. Assume $\mathbb{E}_{\mathbf{X}\sim Q}[\|\mathbf{s}_p(\mathbf{X})\|_2^2]$, $\mathbb{E}_{\mathbf{X}\sim Q}[\|\mathbf{s}_q(\mathbf{X})\|_2^2]$, $\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}[u_Q(\mathbf{X},\mathbf{X}')^2] < \infty$, and the kernel k satisfies*

$$\max\left\{\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[|k(\mathbf{X},\mathbf{X}')|\right],\ \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[\|\nabla_{\mathbf{X}'}k(\mathbf{X},\mathbf{X}')\|_2^2\right],\ \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[\|\nabla_{\mathbf{X}}k(\mathbf{X},\mathbf{X}')\|_2^2\right]\right\} < \infty\,. \quad (4.3)$$

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots$ be a sequence of i.i.d. samples from Q. Denote by $F_\nu \coloneqq \mathbb{E}_{\mathbf{X}\sim Q}[\|\mathbf{s}_{p_\nu}(\mathbf{X}) - \mathbf{s}_q(\mathbf{X})\|_2^2]$ the Fisher Divergence between Q and P_ν. If the sequence $n_1, n_2, \ldots \in \mathbb{N}$ satisfies $n_\nu \to \infty$ as $\nu \to \infty$ and $n_\nu = o\big(1/\max\left(F_\nu, F_\nu^{1/2}\right)\big)$, then

$$n_\nu \hat{D}^2_{P_\nu} \xrightarrow{\mathrm{d}} \sum_{j=1}^{\infty} c_j(Z_j^2 - 1) \qquad \text{as } \nu \to \infty\,, \quad (4.4)$$

where $\hat{D}^2_{P_\nu}$ is the sample squared-KSD computed using $\mathbf{X}_1, \ldots, \mathbf{X}_{n_\nu}$, the random variables $Z_j \sim \mathcal{N}(0,1)$ are i.i.d., and $\{c_j\}_{j\geq 1}$ are the eigenvalues of the Stein kernel u_Q under Q, i.e., they are the solutions to $c_j\phi_j(\cdot) = \int_{\mathbb{R}^d} u_Q(\cdot,\mathbf{x}')\phi_j(\mathbf{x}')q(\mathbf{x}')\,\mathrm{d}\mathbf{x}'$ for non-zero ϕ_j.

Remark 4.1.1. *The RHS of (4.4) is the limiting distribution of $n_\nu\hat{D}^2_{P_\nu}$ under $H_0 : Q = P$ (Liu et al., 2016). Hence, this result shows that if the sample size n_ν is $o\big(1/\max\left(F_\nu, F_\nu^{1/2}\right)\big)$, then the test power converges to the probability of rejection under the null hypothesis, which is simply the* nominal level *of the test.*

Remark 4.1.2. *Assumption (4.3) is standard and holds for Inverse Multi-Quadrics (IMQ) and Gaussian kernels. IMQ kernels are preferred as they have desired tail properties to ensure a* convergence determining *KSD for target densities satisfying the distantly dissipative condition (Gorham and Mackey, 2017; Hodgkinson et al., 2020). This includes Gaussian mixtures with common covariance, as well as distributions strongly log-concave outside and on the boundary of a compact set, such as Bayesian linear, logistic, and Huber regression posteriors with Gaussian priors, c.f., Gorham et al. (2019); Gorham and Mackey (2017). Proposition 4.1 does not contradict this result, as it considers a different regime where a* sequence *of target distributions is of interest.*

Proposition 4.1 allows us to study the test power by analyzing the FD. For instance, when P is a mixture of two Gaussian components and Q is one of its component, the FD decreases exponentially fast to 0 with the mode separation. Proposition 4.1 then implies that an unrealistically large sample size would be needed for the test to have a non-trivial power. This is formalized in the following result.

Theorem 4.2. *Let* $Q = \mathcal{N}(0, I_d)$ *and* $P_\nu = \pi\mathcal{N}(0, I_d) + (1-\pi)\mathcal{N}(\Delta_\nu, I_d)$, *where* $\pi \in [0,1]$ *and* $\Delta_\nu \in \mathbb{R}^d$. *With the same notation in Proposition 4.1 and assuming* k *satisfies (4.3), the limit (4.4) holds if* $n_\nu = o\left(e^{\|\Delta_\nu\|_2^2/64}\right)$.

The proof is in Section 4.A.2. Figure 4.1 provides numerical evidence for the blindness phenomenon suggested by Theorem 4.2. It shows the rate of rejection over 100 repetitions at level $\alpha = 0.05$. Using an IMQ kernel whose bandwidth is chosen by median heuristic (Gretton et al., 2012), we observe that the power of the KSD test approaches the prescribed level for $\Delta \geq 6$. A similarly poor performance is observed for KSDAgg (Schrab et al., 2022) and FSSD (Jitkrittum et al., 2017), two variants of KSD. In comparison, our proposed tests, called ospKSD and spKSD, achieve an almost perfect power. Notably, the problem of low test power persists even if the samples are drawn from both components but with a different weight; see Figure 4.4 in Section 4.7.

4.5 KSD Test with Perturbation

We propose to increase the power of KSD test against *multi-modal alternatives* by perturbing both the candidate and the target distributions with a set of *Markov transition kernels* (Robert and Casella, 2004, Chapter 6) and performing KSD tests on the *perturbed* distributions. A Markov transition kernel is a function $\mathcal{K} : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0,1]$ such that *(i)* for all $\mathbf{x} \in \mathcal{X}$, $\mathcal{K}(\mathbf{x}, \cdot)$ is a probability measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, and *(ii)* for all $A \in \mathcal{B}(\mathcal{X})$, $\mathcal{K}(\cdot, A)$ is a measurable function on $\mathcal{X}$. In our example, $\mathcal{K}$ may also be an iterated composition of an underlying kernel, e.g. a Metropolis-Hastings kernel. The perturbed measure of Q is $(\mathcal{K}Q)(\cdot) \coloneqq \int_{\mathcal{X}} \mathcal{K}(\mathbf{x}, \cdot) Q(\mathrm{d}\mathbf{x})$, and similarly for $\mathcal{K}P$.

4.5.1 KSD with a Single Perturbation Kernel

We first consider a *single* transition kernel $\mathcal{K}$. We define the *perturbed kernel Stein discrepancy* (pKSD) as

$$D(Q, P; \mathcal{K}) \coloneqq D(\mathcal{K}Q, \mathcal{K}P) = \sup_{f \in \mathcal{G}_k} \left| \mathbb{E}_{\mathbf{X} \sim \mathcal{K}Q} [\mathcal{A}_{\mathcal{K}P} f(\mathbf{X})] \right|, \tag{4.5}$$

assuming $\mathcal{K}P$ admits a continuously differentiable density so that its score function is well-defined. Notably, $\mathcal{K}Q$ need not have a (Lebesgue) density for (4.5) to exist.

The properties of pKSD are dictated by the operator $\mathcal{K}$. A desirable choice should ensure that *(i)* pKSD is well-defined, and in particular $\mathcal{K}P$ should have a continuously differentiable density whenever P does, *(ii)* pKSD (4.5) can be computed efficiently, and *(iii)* the test can achieve a high power against alternatives with wrong mixing weights.

Given these *desiderata*, we propose to choose a transition kernel $\mathcal{K}$ that is P*-invariant*, i.e. $P(\cdot) = \int_{\mathcal{X}} \mathcal{K}(\mathbf{x}, \cdot) p(\mathbf{x}) \, \mathrm{d}\mathbf{x}$. A P-invariant kernel ensures $\mathcal{K}P = P$, so the score function $\mathbf{s}_{\mathcal{K}p} = \mathbf{s}_p$ is unchanged after perturbation. This means *(i)* and *(ii)* are trivially satisfied. In particular, pKSD will have a

Algorithm 5 Goodness-of-Fit Test with spKSD.

Input: Target P; test level α; observed sample $\{\mathbf{x}_i\}_{i=1}^n$; set of transition kernels $\mathcal{S} = \{\mathcal{K}_s\}_{s=1}^S$ that includes $\mathcal{K}_{\text{id}}$ (i.e., no perturbation); number of transition steps T.
Estimate the mode $\{\mu_1, \dots, \mu_M\}$ and Hessians $\{A_1, \dots, A_M\}$ using Algorithm 6.
Use the estimated modes and Hessians to construct the Markov transition kernels $\mathcal{K}_s$, as described in Section 4.6.1.
for $s = 1, \dots, S$ **do**
  Perturb $\{\mathbf{x}_i\}_{i=1}^n$ with $\mathcal{K}_s$ by T steps to generate perturbed samples $\{\mathbf{x}_i^s\}_{i=1}^n$.
end for
Compute test statistic $\hat{D}^2_{P,\mathcal{S}}$ using (4.7).
Generate bootstrap samples with (4.2) with u_P replaced by $\tilde{u}_P$, and find the $(1-\alpha)$-quantile $\hat{\gamma}_{1-\alpha}$.
Reject H_0 if $\hat{D}^2_{P,\mathcal{S}} \geq \hat{\gamma}_{1-\alpha}$.

closed-form expression

$$D^2(Q, P; \mathcal{K}) \;=\; \mathbb{E}_{\mathbf{X},\mathbf{X}' \sim \mathcal{K}Q}[u_P(\mathbf{X}, \mathbf{X}')] \;,$$

provided that $\mathbb{E}_{\mathbf{X} \sim \mathcal{K}Q}[u_P(\mathbf{X}, \mathbf{X})] < \infty$ (see, e.g., Chwialkowski et al. 2016, Theorem 2.1). Moreover, the P-invariance allows a GOF test similar to the standard KSD test to be constructed, as we will elucidate in Section 4.5.2. To address *(iii)*, we employ a proposal map for $\mathcal{K}$ that "aggregates" densities across the modes of the distribution. As we will demonstrate numerically, such a proposal is sensitive to discrepancies in mixing weights.

Given i.i.d. $\{\mathbf{X}_i\}_i^n \sim Q$, a sample $\{\tilde{\mathbf{X}}_i\}_{i=1}^n$ from $\mathcal{K}Q$ can be constructed by running 1-step transitions under $\mathcal{K}$ starting from each $\mathbf{X}_i$. The pKSD can then be estimated by the U-statistic:

$$\hat{D}^2_{P,\mathcal{K}} \;\coloneqq\; \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} u_P(\tilde{\mathbf{X}}_i, \tilde{\mathbf{X}}_j) \;. \tag{4.6}$$

4.5.2 KSD with Multiple Perturbation Kernels

A single transition kernel can be limited in improving the test power against general multi-modal alternatives. It also does *not* guarantee the *separation* property, as $D(\mathcal{K}Q, \mathcal{K}P) = 0 \;\not\Longrightarrow\; Q = P$, unless $\mathcal{K}$ is injective so that $\mathcal{K}Q = \mathcal{K}P \implies Q = P$ (such as the convolution operator). However, choosing only injective $\mathcal{K}$ would significantly restrict the class of possible options. Instead, we propose to employ a finite *collection* $\mathcal{S} = \{\mathcal{K}_s\}_{s=1}^S$ of P-invariant transition kernels, and require $\mathcal{S}$ to include the identity transition kernel $\mathcal{K}_{\text{id}}$, defined as $\mathcal{K}_{\text{id}}(\mathbf{x}, A) = \delta_x(A)$ for all $\mathbf{x} \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$, where $\delta_x(A) = 1$ if $\mathbf{x} \in A$ and 0 otherwise. In particular, $D(Q, P; \mathcal{K}_{\text{id}})$ reduces to the standard KSD. This gives rise to a separating statistical divergence which we term *sum-pKSD* (spKSD)

$$D(Q, P; \mathcal{S}) \;\coloneqq\; \left(\sum_{\mathcal{K} \in \mathcal{S}} D^2(\mathcal{K}Q, P) \right)^{1/2} ,$$

where we have overloaded the notation $D(Q, P; \mathcal{S})$ with a set $\mathcal{S}$ in place of a single transition kernel to denote spKSD. The next result shows that spKSD indeed separates probability measures so long as $\mathcal{K}_{\text{id}} \in \mathcal{S}$.

Proposition 4.3 (spKSD separation)**.** *Suppose Q, P are probability measures on $\mathcal{X}$ that admit positive (Lebesgue) densities q, p, respectively. Further assume $\mathbb{E}_{\mathbf{X}\sim\mathcal{K}Q}[u_P(\mathbf{X}, \mathbf{X})] < \infty$ for all $\mathcal{K} \in \mathcal{S}$ and $\mathbb{E}_{\mathbf{X}\sim Q}[\|\mathbf{s}_p(\mathbf{X}) - \mathbf{s}_q(\mathbf{X})\|_2^2] < \infty$. If the kernel k is cc-universal and $\mathcal{K}_{id} \in \mathcal{S}$, then $D(Q, P; \mathcal{S}) \geq 0$ with equality if and only if $Q = P$.*

Proof. The stated assumptions ensure that, for all $\mathcal{K} \in \mathcal{S}$, $D(Q, P; \mathcal{K})$ is well-defined, and that $D(Q, P; \mathcal{K}_{\text{id}}) = D(Q, P) = 0 \iff Q = P$ (see, e.g., Chwialkowski et al. (2016, Theorem 2.2)). The desired result then follows since $\mathcal{K}$ is P-invariant for all $\mathcal{K}$ and $D(Q, P; \mathcal{S}) \geq D(Q, P)$. □

The assumption that the alternative distribution Q also admits a density is common in KSD literature when proving separation (e.g., Liu et al. 2016; Chwialkowski et al. 2016; Jitkrittum et al. 2017; Gong et al. 2021b), but it can be relaxed if P is light-tailed or distantly dissipative (Hodgkinson et al., 2020; Gorham and Mackey, 2017).

The spKSD can also be written as a double expectation akin to KSD, provided $\mathbb{E}_{\mathbf{X}\sim\mathcal{K}_s Q}[u_P(\mathbf{X}, \mathbf{X})] < \infty$ for all s. This allows spKSD to be estimated given a random sample $\{\mathbf{X}_i\}_{i=1}^n$ from Q. Formally, for each $\mathcal{K}_s \in \mathcal{S} = \{\mathcal{K}_1, \ldots, \mathcal{K}_S\}$, a sample $\{\mathbf{X}_i^s\}_{i=1}^n$ from $\mathcal{K}_s Q$ can be drawn by running 1-step transitions under $\mathcal{K}_s$ starting from each $\mathbf{x}_i$. Denote by $\mathbf{X}_i^{1:S} \coloneqq \text{concat}(\mathbf{X}_i^1, \ldots, \mathbf{X}_i^S)$ the concatenation of $\mathbf{X}_i^1, \ldots, \mathbf{X}_i^S$ into a single vector. We propose to estimate $D(Q, P; \mathcal{S})$ using the *square-root* of the following U-statistic

$$\hat{D}^2_{P,\mathcal{S}} \coloneqq \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \tilde{u}_P\big(\mathbf{X}_i^{1:S}, \mathbf{X}_j^{1:S}\big) , \tag{4.7}$$

where $\tilde{u}_P\big(\mathbf{X}_i^{1:S}, \mathbf{X}_j^{1:S}\big) \coloneqq \sum_{s=1}^S u_P(\mathbf{X}_i^s, \mathbf{X}_j^s)$.

4.5.3 GOF Testing with spKSD

Having constructed a test statistic for spKSD in the form of a U-statistic, the next result (proved in Section 4.A.3) derives the limiting distribution of spKSD statistic under the null and alternative hypotheses. We denote by R_Q the distribution of $\mathbf{X}_i^{1:S}$ constructed as before and use the same notations as in Proposition 4.3.

Proposition 4.4 (Asymptotic distributions of spKSD)**.** *Suppose the assumptions in Proposition 4.3 hold, and further assume $\mathbb{E}_{\mathbf{W},\mathbf{W}'\sim R_Q}[\tilde{u}_P(\mathbf{W}, \mathbf{W}')^2] < \infty$. Let $\{z_j\}_{j\geq 1}$ be independent draws from $\mathcal{N}(0, 1)$ and denote by $\{c_j\}_{j\geq 1}$ the eigenvalues of $\tilde{u}_P$ under R_Q, i.e., the solutions of $c_j \phi_j(\cdot) = \mathbb{E}_{\mathbf{W}\sim R_Q}[\tilde{u}_P(\cdot, \mathbf{W})\phi_j(\mathbf{W})]$ for non-zero ϕ_j. As $n \to \infty$,*

1. *Under $H_0 : Q = P$, we have $n\hat{D}^2_{P,\mathcal{S}} \to_d \sum_{j=1}^\infty c_j(z_j^2 - 1)$.*
2. *Under $H_1 : Q \neq P$, we have $\sigma_u^2 \coloneqq 4\text{Var}_{\mathbf{w}\sim R_Q}(\mathbb{E}_{\mathbf{W}'\sim R_Q}[\tilde{u}_P(\mathbf{W}, \mathbf{W}')]) > 0$, and $\sqrt{n}\big(\hat{D}^2_{P,\mathcal{S}} - D^2(Q, P; \mathcal{S})\big) \to_d \mathcal{N}(0, \sigma_u^2)$.*

Proposition 4.4 assumes Q also admits a Lebesgue density; when it does not, the stated results still hold true if we additionally assume *i)* the conditions on Q in Proposition 4.3 for KSD to separate probability measures, and *ii)* $R_Q(A) > 0$ whenever $R_P(A) > 0$ for any measurable set $A \subset \mathcal{X}^S$.

Similarly to the case with the standard KSD, the cumulative density function of the limiting distribution under H_0 has no closed-form expression, but the same bootstrap technique can be employed to estimate the p-value using the perturbed samples. The complete algorithm of goodness-of-fit testing with pKSD is given in Algorithm 5.

4.6 A Transition Kernel for Multi-Modal Alternatives

We consider transition kernels of the Metropolis-Hastings (MH) type (Metropolis et al., 1953; Hastings, 1970). At a current state $\mathbf{x}$, a new state $\mathbf{x}'$ is proposed by first generating a d_u-dimensional random vector $\mathbf{u}$ from some known probability mass function g (to be specified later), then mapping to $\mathbf{x}' = \mathbf{h}(\mathbf{x}|\mathbf{u})$, where $\mathbf{h}(\cdot|\mathbf{u})$ is some deterministic, invertible function that is differentiable with differentiable inverse. The proposed state $\mathbf{x}'$ is hence a deterministic function given $\mathbf{x}$ and $\mathbf{u}$.

We choose in this chapter a density g defined on some discrete space $\mathcal{U}$. The transition kernel is

$$\mathcal{K}(\mathbf{x}, A) \;=\; \sum_{\mathbf{u}\in\mathcal{U}} \delta_{\mathbf{x}'}(A) g(\mathbf{u}) \alpha(\mathbf{x}, \mathbf{x}') + \delta_{\mathbf{x}}(A) r(\mathbf{x}) \;,$$

where $\mathbf{x}' = \mathbf{h}(\mathbf{x}|\mathbf{u})$ is the proposed state, $\alpha(\mathbf{x}, \mathbf{x}')$ is an accept-reject rule that guarantees P-invariance, $\delta_x(A) = 1$ if $\mathbf{x} \in A$ and 0 otherwise, and $r(\mathbf{x}) = 1 - \sum_{\mathbf{u}\in\mathcal{U}} g(\mathbf{u})\alpha(\mathbf{x}, \mathbf{x}')$. The accept-reject rule $\alpha(\mathbf{x}, \mathbf{x}')$ is designed to satisfy the *detailed balance condition* (see, e.g., Green and Hastie, 2009, Eq. (1)):

$$\int_{\mathbf{x}\in A} \sum_{\mathbf{u}\in\mathcal{U}} \delta_{\mathbf{h}(\mathbf{x}|\mathbf{u})}(B) p(\mathbf{x}) g(\mathbf{u}) \alpha(\mathbf{x}, \mathbf{h}(\mathbf{x}|\mathbf{u}))\, \mathrm{d}\mathbf{x} \;=\; \int_{\mathbf{x}'\in B} \sum_{\mathbf{u}'\in\mathcal{U}} \delta_{\mathbf{h}(\mathbf{x}'|\mathbf{u}')}(A) p(\mathbf{x}') g(\mathbf{u}') \alpha(\mathbf{x}', \mathbf{h}(\mathbf{x}'|\mathbf{u}'))\, \mathrm{d}\mathbf{x}' \;, \tag{4.8}$$

for all $A, B \in \mathcal{B}(\mathcal{X})$. One valid choice is

$$\alpha(\mathbf{x}, \mathbf{x}') \;=\; \min\left(1, \frac{p(\mathbf{x}')g(\mathbf{u}')}{p(\mathbf{x})g(\mathbf{u})} \left|\frac{\partial \mathbf{h}(\mathbf{x}|\mathbf{u})}{\partial \mathbf{x}}\right|\right) \;, \tag{4.9}$$

if $\mathbf{x}' = \mathbf{h}(\mathbf{x}|\mathbf{u})$ and $\mathbf{x} = \mathbf{h}^{-1}(\mathbf{x}'|\mathbf{u}')$ for some $\mathbf{u}, \mathbf{u}' \in \mathcal{U}$, and zero otherwise. Here, $\partial \mathbf{h}(\mathbf{x}|\mathbf{u})/\partial \mathbf{x}$ denotes the Jacobian of the transformation from $\mathbf{x}$ to $\mathbf{x}'$. Section 4.A.4 proves that $\alpha(\mathbf{x}, \mathbf{x}')$ indeed satisfies (4.8). The accept-reject rule (4.9) resembles those used in Reversible-Jump MCMC (Green, 1995; Green and Hastie, 2009) and recovers the well-known MH rule when the determinant of the Jacobian is 1.

4.6.1 Choosing the Proposal Density

We propose a jump proposal $\mathbf{h}(\mathbf{x}|\mathbf{u})$ that superposes masses at each mode of p. Our choice is motivated by Markov kernels used in the optimization-based MCMC literature, specifically the *deterministic*

jumps proposal in Pompe et al. (2020). New states are proposed by randomly selecting a mapping from a set of candidates that are constructed using the location and geometry of the modes of p. The resulting kernel is *not* irreducible, so the limiting distribution is not necessarily P. Non-irreducibility is essential for the proposed test to work since, under the alternative, the transition kernel should perturb Q to some other distribution for which the KSD between P and the perturbed distribution becomes larger compared with the KSD with the unperturbed one. This is in contrast to MCMC, which requires irreducibility so that asymptotically the chain can sample from the target distribution.

Denote by $\mu_1, \ldots, \mu_M \in \mathbb{R}^d$ the modes of the density p, and $A_1, \ldots, A_M \in \mathbb{R}^{d\times d}$ the *inverse* of the Hessian matrices at those points; how to estimate these quantities will be discussed later. When p is a mixture of elliptic distributions such as Gaussian or multivariate t-distributions, each A_m can be viewed as the covariance matrix of a component. When the Hessians do not exist (e.g., $-\log p$ is not twice differentiable), we can set $A_m = I_d$ and the remaining discussion still follows.

For a current state $\mathbf{x}$, our proposal randomly selects a pair of modes and attempts to map $\mathbf{x}$ from one mode to the "corresponding" point $\mathbf{x}'$ in the other. Formally, let $\mathbf{u} = (u_1, u_2) \sim \mathrm{Unif}(\{(i,j) : 1 \le i \ne j \le M\})$ be a uniform random vector over the index set of all $M(M-1)$ pairs of *distinct* (and ordered) modes, i.e., $g(\mathbf{u}) = 1/(M(M-1))$ for all $\mathbf{u}$. Given a fixed constant $\theta > 0$, the proposal map is

$$\mathbf{h}(\mathbf{x}|\mathbf{u}) \;=\; \mathbf{h}_\theta(\mathbf{x}|\mathbf{u}) \;=\; A_{u_2}^{1/2} A_{u_1}^{-1/2}(\mathbf{x} - \theta\mu_{u_1}) + \theta\mu_{u_2} \;,$$

with the inverse map $\mathbf{h}^{-1}(\mathbf{x}'|\mathbf{u}) = A_{u_1}^{1/2} A_{u_2}^{-1/2}(\mathbf{x} - \theta\mu_{u_2}) + \theta\mu_{u_1}$. Intuitively, $\mathbf{h}$ sends points from mode μ_{u_1} to μ_{u_2} allowing for scaling by local Hessians, and $\mathbf{h}^{-1}$ performs the opposite operation. The constant θ is a hyperparameter introduced to control the scale of the jump, which can increase the ability to detect discrepancies in the mixing weights. Herein, we call θ the *jump scale*.

Given a current state, our proposal chooses two modes randomly, so a proposed state can potentially lie in a low-density region, thus leading to a low acceptance probability. Pompe et al. (2020) address this by recording an auxiliary variable for the mode index and augmenting the state space to $\mathcal{X} \times \{1, 2, \ldots, M(M-1)\}$, so that at every step the new state is *guaranteed* to lie near a mode. However, the same trick cannot be used in our case because the augmented density no longer has a well-defined score function.

4.6.2 Understanding the Source of Test Power

To understand the improvement in test power against multi-modal alternatives, we characterize the limiting distribution of a general distribution Q with a positive density q when we apply the perturbation with infinitely many steps (i.e., $\mathcal{K}^T$ with $T = \infty$). For simplicity, we assume $M = 2$ and $A_1 = A_2 = I_d$ are identity matrices, so that the proposal function is $h_\theta(\mathbf{x}|\mathbf{u}) = \mathbf{x} - \theta(\mu_{u_1} - \mu_{u_2})$ for $\mathbf{x} \in \mathcal{X}$ and $\mathbf{u} = (u_1, u_2) \in \mathcal{U} = \{(i,j) : 1 \le i \ne j \le 2\}$. Thus, given a current state $\mathbf{x}$, the transition kernel proposes moves to $\mathbf{x} + \theta(\mu_1 - \mu_2)$ and $\mathbf{x} - \theta(\mu_1 - \mu_2)$ with equal probability.

Proposition 4.5. *Under the assumptions of Section 4.6.2, the limiting distribution under $\mathcal{K}$ with the*

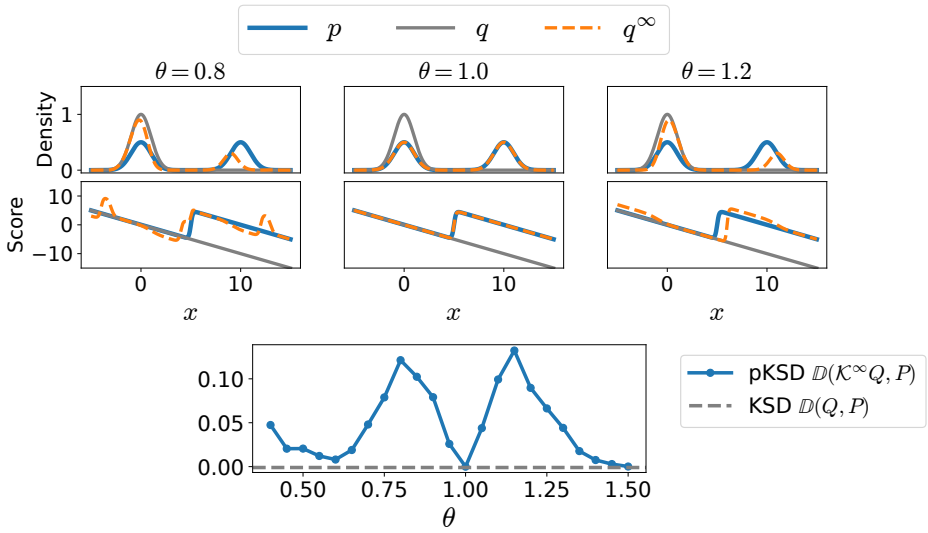


Figure 4.3: *Top.* Densities and score functions of p, q and the limiting distribution q^∞ in (4.10). *Bottom.* pKSD with different jump scales θ, compared with KSD.

initial distribution Q *is* $(\mathcal{K}^\infty Q)(A) = \int_{\mathbf{x}\in A} q^\infty(\mathbf{x})\,\mathrm{d}\mathbf{x}$, $A \in \mathcal{B}(\mathcal{X})$, *where*

$$q^\infty(\mathbf{x}) \;:=\; p(\mathbf{x})\frac{\sum_{s\in\mathbb{Z}} q(\mathbf{x}+s\nu)}{\sum_{k\in\mathbb{Z}} p(\mathbf{x}+k\nu)}\,, \tag{4.10}$$

and $\nu := \theta(\mu_1 - \mu_2)$.

A proof is in Section 4.A.5. Proposition 4.5 shows that the limiting density under $\mathcal{K}$ is the target density p weighted by the ratio between the total masses of q and p over a discrete grid.

To understand why this helps to increase the KSD value, we first rewrite the squared KSD as

$$D^2(Q,P) \;=\; \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}[\delta_{q,p}(\mathbf{X})^\top k(\mathbf{X},\mathbf{X}')\delta_{q,p}(\mathbf{X}')]\,,$$

where $\delta_{q,p}(\mathbf{x}) := \mathbf{s}_q(\mathbf{x}) - \mathbf{s}_p(\mathbf{x})$ is the score difference. This holds whenever k is an integrally strictly positive definite kernel (Liu et al., 2016, Theorem 3.6). The squared pKSD is then

$$D^2(Q,P;\mathcal{K}^\infty) \;=\; \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q^\infty}[\delta_{q^\infty,p}(\mathbf{X})^\top k(\mathbf{X},\mathbf{X}')\delta_{q^\infty,p}(\mathbf{X}')]\,,$$

where, by Proposition 4.5, the score difference becomes $\delta_{q^\infty,p}(\mathbf{x}) = s_{q^\infty}(\mathbf{x}) - \mathbf{s}_p(\mathbf{x}) = \nabla \log \phi_q(\mathbf{x}) - \nabla \log \phi_p(\mathbf{x})$, where $\phi_q(\mathbf{x}) := \sum_{s\in\mathbb{Z}} q(\mathbf{x}+s\nu)$ and similarly for ϕ_p. The operator ϕ superposes densities along a grid, thus allowing to create local discrepancy in the high-probability regions of Q, for example by exchanging masses between modes.

As a concrete example, we consider the setup in Theorem 4.2, where $Q = \mathcal{N}(0, I_d)$, and $P = \pi\mathcal{N}(0, I_d) + (1-\pi)\mathcal{N}(\Delta, I_d)$ for some $\pi \in (0,1)$ and $\Delta \in \mathbb{R}^d$. The operator has created discrepancy in high-probability regions of Q, as demonstrated in Figure 4.3. This also highlights the role of ν: when $\nu = \Delta$, the two components will overlap almost exactly under the perturbation, so $\delta_{q^\infty,p}(\mathbf{x}) \approx 0$ near $\mathbf{x} = 0$, and the KSD will remain small (Figure 4.3). It is hence crucial to tune ν (equivalently, θ). One can in principle select θ by maximizing the (approximate) test power, similarly to the idea in Jitkrittum et al. (2017). However, gradient-based approaches are infeasible as pKSD is *not* differentiable with respect to

θ. An alternative is to use grid-search over some finite set of θ values.

4.6.3 Choosing the Set of Perturbations $\mathcal{S}$

It remains to choose the set of perturbations $\mathcal{S}$ in spKSD. We propose two ways to construct $\mathcal{S}$, one based on grid-search, and the other based on optimization.

For the grid-based approach, we choose a set of values $\{\theta_s\}_{s=1}^{S-1}$ and let $\mathcal{S} = \{\mathcal{K}_{\mathrm{id}}, \mathcal{K}_1, \ldots, \mathcal{K}_{S-1}\}$, where $\mathcal{K}_s$ is the transition kernel described in this section with jump scale θ_s. We propose to choose each θ_s close to 1, following the observations in Section 4.6.2. We still refer to the resulting divergence as spKSD.

For the optimization-based approach, we use only two transition kernels $\mathcal{S} = \{\mathcal{K}_{\mathrm{id}}, \mathcal{K}_\theta\}$, where $\mathcal{K}_\theta$ has jump scale θ that is tuned by maximizing a proxy for the asymptotic test power. Due to the asymptotic normality proved in Proposition 4.4, we can adopt the same approach in Jitkrittum et al. (2017, Proposition 4) to approximate the asymptotic power with the ratio

$$\hat{D}^2_{P,\mathcal{K}_\theta^T} \,/\, \hat{\sigma}_u \,, \tag{4.11}$$

where $\hat{\sigma}_u$ is an estimate of the asymptotic standard deviation σ_u is given by the square root of

$$\hat{\sigma}_u^2 \coloneqq \frac{4}{n^3}\sum_{i=1}^{n}\left(\sum_{j=1}^{n} H_{i,j}\right)^2 - \frac{4}{n^4}\left(\sum_{i,j=1}^{n} H_{i,j}\right)^2 ,$$

with $H_{i,j} \coloneqq u_P(\mathbf{X}_i, \mathbf{X}_j) + u_P(\mathbf{X}_i^\theta, \mathbf{X}_j^\theta)$ and $\mathbf{X}_i^\theta \sim \mathcal{K}_\theta^T Q$; see also Schrab et al. (2022, Eq. 8). Since the objective (4.11) (in particular, $\mathcal{K}_\theta$) is not differentiable with respect to θ, we still choose θ from a pre-specified finite set $\{\theta_s\}_{s=1}^{S}$. The objective is hence $\max_{\theta\in\{\theta_1,\ldots,\theta_S\}} \hat{D}^2_{P,\mathcal{K}_\theta^T}/\hat{\sigma}_u$. We call the resulting discrepancy the *optimized sum-pKSD* (ospKSD).

Whether the grid-based or the optimization-based method should be preferred requires trade-offs and depends on the specific problem at hand. The spKSD requires no held-out sets, but can suffer from a low test power if $\{\theta_s\}_{s=1}^{S-1}$ is poorly chosen in that most of $\mathcal{K}_s$ fail to improve the test power. On the other hand, ospKSD uses a judiciously tuned θ, but the data-splitting can also lead to a drop in test power. In our experiments, we find that spKSD tends to work better for target distributions with a simple geometry, specifically mixtures of *elliptic* distributions (Cambanis et al., 1981) (e.g., the Gaussian mixture examples). However, for distributions whose mixing components have non-elliptic contours (e.g., the mixture of t and banana example, and the sensor network localization example), the benefit of optimization seems to overweigh the negative impact due to data-splitting, and ospKSD outperforms spKSD.

4.6.4 Estimating Mode Vectors and Hessians

We estimate μ_j and A_j by the local minima and Hessians of $-\log p$. To do so, we run in parallel a sequence of BFGS optimizers (Nocedal and Wright, 2006) initiated at different starting points, following

Pompe et al. (2020). BFGS is used because it returns both the local optima and approximated Hessians at those points. The optima are then merged if their weighted Mahalanobis distance is smaller than a pre-specified threshold. In our experiments, we initialize the optimizers from a set of size n_{init}, constructed either by sampling uniformly from a hyper cube $[L_1, U_1] \times \cdots \times [L_d, U_d]$ (for spKSD), or by using both randomly sampled data and some training set (for ospKSD). We now discuss each step in detail.

Finding local modes In practice, the mode locations and Hessians of the density of a non-trivial target distribution are rarely available. Pompe et al. (2020) describes a general framework to estimate these quantities. It proceeds by running in parallel a sequence of optimizers initiated at different starting points. This is done by minimizing the objective $-\log p$ using the BFGS algorithm (Nocedal and Wright, 2006), which returns both the local minima and the approximated Hessian at those points. In our experiments, we use the BFGS algorithm and run for at most 1000 iterations with each initial point.

Mode merging Although the end points of the optimization procedure starting from different initial points may lie close to the local minima, they will still be numerically different from each other. Pompe et al. (2020) proposed to merge two end points m_i and m_j if their Mahalanobis distance weighted by the averaged Hessians at those points is below a given threshold β. The full procedure is stated in Algorithm 6 for completeness.

Choosing the initial points A set of n_{init} initial points for BFGS can be constructed either by sampling randomly from a product of intervals $[L_1, U_1] \times \cdots \times [L_d, U_d]$ in $\mathcal{X}$, or simply from a held-out training set. The first approach will allow modes not covered by the training data to be detected, and the second approach can lead to faster convergence of the optimization algorithm when the training points lie near the modes of P. For spKSD, the first approach is used as we do not assume a held-out set is available. For ospKSD, to combine the best of the two approaches whilst maintaining the same computational budget, half of the n_{init} initial points are drawn randomly from the training set and the other half are initialized uniformly from $[L_1, U_1] \times \cdots \times [L_d, U_d]$.

4.6.5 Choosing the Number of Transitions T

We discuss how to choose the number of transitions T. Clearly, T determines the form of the perturbed distribution, thus impacting the performance of spKSD. Intuitively, T should be set to a large value when the acceptance rate is low to ensure the limiting distribution is achieved. The spKSD could suffer from a low acceptance rate when the estimates of the modes and local Hessians of the target distribution are inaccurate, or when the target distribution cannot be approximated by a mixture of elliptic distributions.

We propose two heuristics to choose this hyperparameter in practice: *(i)* viewing this as another hyperparameter and tuning it using a training set by selecting from a pre-specified set of values, or *(ii)* setting it to a large value (e.g., $T = 1000$) if the computational budget allows.

Algorithm 6 Estimating mode vectors and Hessians; adapted from Pompe et al. (2020, Algorithm 3).

Input: Initial points $\mathbf{s}_1, \ldots, \mathbf{s}_{M_0}$, small positive value β.
Initialize M_0 parallel BFGS optimizers at points $\mathbf{s}_1, \ldots, \mathbf{s}_{M_0}$ and run the algorithm to minimize $-\log p(\mathbf{x})$.
Denote the returned estimates of the local optima by $\mathbf{m}_1, \ldots, \mathbf{m}_{M_0}$ and their corresponding Hessian matrices by $A_1, \ldots, A_{M_0}$.
Set $\mu_1 \coloneqq \mathbf{m}_1$, $A_{\mu_1} \coloneqq A_1$ and $M = 1$.
for $i = 2, \ldots, M_0$ **do**
  if $\min_{j\in\{1,\ldots,M\}} \frac{1}{2}\big((\mu_j - \mathbf{m}_i)^\top A_{\mu_j}(\mu_j - \mathbf{m}_i) + (\mu_j - \mathbf{m}_i)^\top A_i(\mu_j - \mathbf{m}_i)\big) < \beta$ **then**
    $k \coloneqq \arg\min_{j\in\{1,\ldots,M\}} \frac{1}{2}\big((\mu_j - \mathbf{m}_i)^\top A_{\mu_j}(\mu_j - \mathbf{m}_i) + (\mu_j - \mathbf{m}_i)^\top A_i(\mu_j - \mathbf{m}_i)\big)$.
    if $p^*(\mu_k) < p^*(\mathbf{m}_i)$ **then**
      Set $\mu_k \coloneqq \mathbf{m}_i$ and $A_{\mu_k} \coloneqq A_i$.
    end if
  else
    $\mu_{M+1} \coloneqq \mathbf{m}_i$ and $A_{\mu_{M+1}} \coloneqq A_i$.
    $M \coloneqq M + 1$.
  end if
end for
Return $\{\mu_1\}_{i=1}^M$ and $\{A_1\}_{i=1}^M$ as approximates for mode vectors and Hessians, respectively.

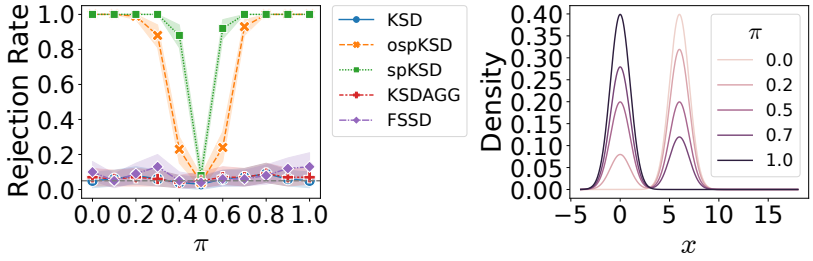


Figure 4.4: One-dimensional Gaussian mixture example. Samples are drawn with a different mixing weight π.

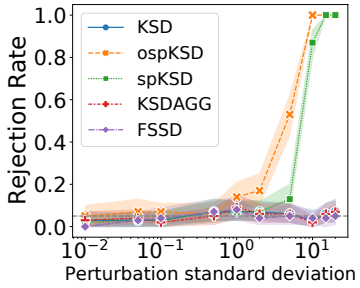


Figure 4.5: Mixture of 10 t and 10 banana distributions.

In particular, we recommend a large T because, with the proposed transition kernel $\mathcal{K}$, the KSD $D(\mathcal{K}^T Q, \mathcal{K}^T P)$ between the perturbed distributions does *not* necessarily decrease as T grows. This is because $\mathcal{K}^T Q$ does *not* necessarily converge to P as $T \to \infty$, since $\mathcal{K}$ is *not* irreducible (see the discussions in Section 4.6.1).

4.7 Experiments

We use 51 jump scales θ equally spaced in $[0.5, 1.5]$, a heuristic that we find works well in practice. All samples have size $n = 1000$. We compare the ospKSD and spKSD tests against benchmarks including KSD test and two variants (KSDAGG and FSSD). All experiments are run with level $\alpha = 0.05$ using the IMQ kernel $k(\mathbf{x}, \mathbf{y}) = (1 + \|\mathbf{x} - \mathbf{y}\|_2^2/\lambda)^{-1/2}$, where λ is chosen to be $\texttt{median}_{i<j}\{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2\}$. KSDAGG follows the setup in Schrab et al. (2022), and FSSD follows Jitkrittum et al. (2017). The probability of rejecting the null hypothesis is estimated by averaging the test output over 100 repetitions, except in the sensors location example, which is repeated 10 times. Translucent shades represent 95%-CIs. The number of transition steps T is selected to be 10 for the Gaussian mixture example, 100 for the

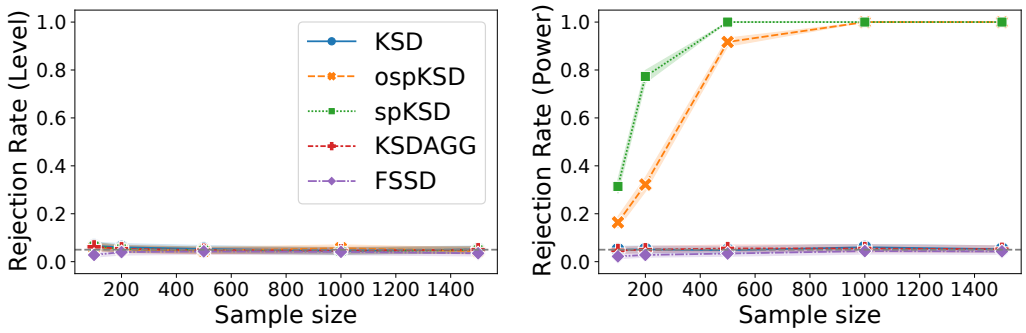


Figure 4.6: Level (left) and power (right) experiments with the multivariate Gaussian mixture example. Gray dashed line shows the test level $\alpha = 0.05$.

mixture of t and banana distributions example, and 1000 for the sensor network localization. Code for reproducing all experiments can be found at github.com/XingLLiu/pksd.

Univariate Gaussian mixture The target has density $p(x) \propto \exp\left(-\frac{1}{2}x^2\right) + 0.5\exp\left(-\frac{1}{2}(x-6)^2\right)$. Samples are drawn with a different mixing weight $\pi \in [0,1]$ of the first component. The results are presented in Figure 4.4. KSD, KSDAGG and FSSD all have a power close to the level 0.05 regardless of the value of π, which is not surprising due to the blindness of KSD. In comparison, ospKSD and spKSD achieve almost perfect power when π deviates from the true value 0.5.

Multivariate Gaussian mixture We then include a level study and a power study for the spKSD and ospKSD tests using a multivariate generalization of the previous Gaussian mixture distribution. The target distribution has dimension 50 and density function $p(\mathbf{x}) \propto \pi_p \exp\left(-\frac{1}{2}\|\mathbf{x}\|_2^2\right) + (1-\pi_p)\exp\left(-\frac{1}{2}\|\mathbf{x}-\Delta e_1\|_2^2\right)$, where $\pi_p = 0.5$, $\Delta = 6$, and $e_1 \in \mathbb{R}^d$ is a vector with 1 in the first coordinate and 0 in others. Samples are drawn either from the target distribution (level experiment), or from only the left component (power experiment). The probability of rejection over 100 repetitions is plotted in Figure 4.6. We can see from the left plot that, under the null hypothesis, all tests have the prescribed test level $\alpha = 0.05$. In the right plot, we investigate the rejection rate under the alternative hypothesis where the samples are drawn from the left mode only. We see that both ospKSD and spKSD achieve a significantly higher power than the benchmarks (KSD, KSDAGG and FSSD), whose power remains close to the level for all sample sizes.

Mixture of t and banana distributions We consider a mixture of 10 multivariate t-distributions and 10 banana-shaped distributions with t-tails in $d = 50$ dimensions, also studied in Pompe et al. (2020). Each component has an equal weight 0.02 and is centered randomly in $[-20, 20]^d$, giving rise to a target with sparsely located, non-elliptic modes. Following Pompe et al. (2020), each t-distribution has 7 degrees of freedom and covariance matrix $0.1\sqrt{d}I_d$, while each banana-shaped distribution has density function $p_{b,\mu} = p_\mu \circ \phi_b$, where $\phi_{b,\mu}(\mathbf{x}) = (x_1, x_2 + bx_1^2 - 100b, x_3, \dots, x_d)^\top$ with $b = 0.003$, and p_μ is the density of a t-distribution with 7 degrees of freedom centered at $\mu \in \mathbb{R}^d$ and with shape matrix $C = \mathrm{diag}(100, 1, \dots, 1) \in \mathbb{R}^{d\times d}$. Samples are drawn with a different set of weights $\{w_j\}_{j=1}^{20}$ formed by sampling $\tilde{w}_j \sim \mathcal{N}(0, \sigma_s^2)$, $\sigma_s > 0$, and normalizing $w_j \propto \exp(\tilde{w}_j)$.

Figure 4.5 (left) shows the results. As σ_s increases, the weights in the two distributions deviate further,

Table 4.1: GOF tests for checking the quality of RAM samples with different scales. Reported values are the number of rejections over 10 repetitions with level 0.05.

Methods		KSD	ospKSD	spKSD	KSDAGG	FSSD
RAM scale σ	0.1	0	8	0	0	0
	0.5	0	0	0	0	1
	1.08	10	8	10	10	7

so ospKSD and spKSD achieve higher power. All the others perform poorly for all values of σ_s, because the components have almost no overlapping high-density regions.

Sensor network localization Tak et al. (2018) use Bayesian methods to infer the locations of sensors from noisy distance data. This is a modification of the example in Ihler et al. (2005) that has been used as a benchmark for MCMC samplers designed for multi-modal distributions (Pompe et al., 2020; Ahn et al., 2013; Lan et al., 2014). Here, six sensors $x_1, \ldots, x_6$ are located in $[0,1]^2$, four of which have unknown locations and the remaining two are known. We observe distance y_{ij} between two sensors $\mathbf{x}_i, \mathbf{x}_j$ with probability $\exp\big(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2/(2\times 0.3^2)\big)$. If observed, the distance follows a Gaussian distribution $y_{ij} \sim \mathcal{N}(\|\mathbf{x}_i - \mathbf{x}_j\|, 0.02^2)$.

Following Tak et al. (2018), we use a diffuse bivariate Gaussian prior distribution $\mathcal{N}(0, 10^2 I_2)$ for each $\mathbf{x}_i \in \mathbb{R}^2$. Let w_{ij} be the binary random variable for which $w_{ij} = 1$ if the distance y_{ij} is observed and 0 otherwise. The full posterior is

$$\pi(x_1, \ldots, x_4 | \mathbf{y}, \mathbf{w}) \propto \exp\left(-\frac{\sum_{k=1}^4 x_k^\top x_k}{2\times 10^2}\right) \prod_{i<j} f_{ij}(\mathbf{x}_i, \mathbf{x}_j | y_{ij}, w_{ij}) ,$$

where $\mathbf{w} = \{w_{ij}\}$, $\mathbf{y} = \{y_{ij}\}$ and

$$\begin{aligned} f_{ij}(\mathbf{x}_i, \mathbf{x}_j | y_{ij}, w_{ij}) = & \left[\exp\left(-\frac{(y_{ij} - \|\mathbf{x}_i - \mathbf{x}_j\|_2)^2}{2\times 0.02^2}\right) \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\times 0.3^2}\right)\right]^{w_{ij}} \\ & \times \left[1 - \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\times 0.3^2}\right)\right]^{1-w_{ij}} . \end{aligned}$$

To draw posterior samples, Tak et al. (2018) propose the *repelling-attracting Metropolis* (RAM), which is an MCMC algorithm designed for efficient learning of multi-modal target distributions. RAM relies on a Gaussian proposal with a fixed covariance matrix $\sigma^2 I_d$ to propose new states in its uphill and downhill steps. The *scale* σ needs to be tuned to facilitate transitions between modes.

We run RAM with different scales σ, each for 420,000 iterations. We discard the first 20,000 particles as burn-in and thin the remaining to obtain a sample of size $n = 4000$. We then evaluate the quality of the samples by applying a GOF test. Figure 4.7 shows the posterior samples generated using each σ. The samples from $\sigma = 0.5$ seem to capture all the modes of the posterior and is consistent with the results reported in Tak et al. (2018, Figure 5), whereas samples from $\sigma = 0.1$ and 1.08 clearly miss some modes. We then compare the test results in Table 4.1, which reports the number of rejections over 10 repetitions. All tests reject most runs for $\sigma = 1.08$ (the scale used in Tak et al. (2018)) and almost no run for $\sigma = 1.08$, which is consistent with the posterior plots. However, for $\sigma = 0.5$, no method except

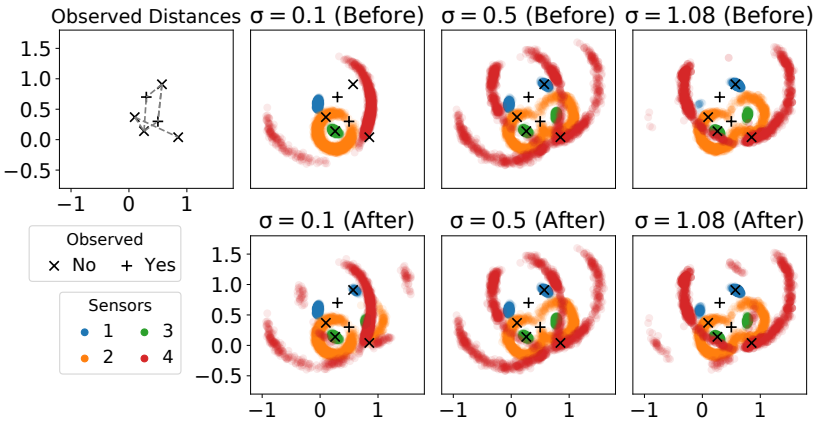


Figure 4.7: True and inferred locations of sensors. Black plus signs mark the location of the unobserved sensors, and black crosses indicate the location of the observed ones.

ospKSD rejected the null hypothesis, demonstrating again the ability of ospKSD to detect missing modes. Figure 4.7 also shows the particles *after* perturbation by the (non-identity) transition kernel used by ospKSD, from which some particles seem to have moved to the missing high-density regions. spKSD in this case also performed poorly with no sample for $\sigma = 0.1$ being rejected, which is potentially because the benefit of not having a held-out set outweighs that of using a tuned transition kernel for this example.

Gaussian-Bernoulli Restricted Boltzmann Machine (GB-RBM) GB-RBMs (Cho et al., 2013) are popular benchmarks for assessing kernel GOF tests (Liu et al., 2016; Jitkrittum et al., 2017; Schrab et al., 2022). It is a latent variable model with joint density $p(\mathbf{x}, \mathbf{h}) \propto \exp\left(\frac{1}{2}\mathbf{x}^\top B\mathbf{h} + \mathbf{b}^\top\mathbf{x} + \mathbf{c}^\top\mathbf{h} - \frac{1}{2}\|\mathbf{x}\|_2^2\right)$, where $\mathbf{h} \in \{-1, 1\}^{d_h}$, and $B, \mathbf{b}, \mathbf{c}$ are fixed hyperparameters. The marginal density $p(\mathbf{x})$ can be rewritten as a mixture of Gaussian distributions:

$$p(\mathbf{x}) = \sum_{\mathbf{h}} \gamma(\mathbf{h}) N\left(\mathbf{x}; \frac{1}{2}B\mathbf{h} + \mathbf{b}, I_d\right), \qquad \text{where } \gamma(\mathbf{h}) \propto \exp\left(\frac{1}{2}\left\|\frac{1}{2}B\mathbf{h} + \mathbf{b}\right\|_2^2 + \mathbf{c}^\top\mathbf{h}\right). \quad (4.12)$$

We consider two parameter settings: *standard* and *multi-modal*. The *standard* setting follows the setups in Liu et al. (2016), in which case the RBM is unimodal. For the target distribution P, we randomly sample the entries of $\mathbf{b}$ and $\mathbf{c}$ from a standard normal, and select the entries of B from $\{-1, 1\}$ with equal probability. We sample from a perturbed version of P where Gaussian noises with standard deviation σ are injected into the entries of B. As σ increases, the problem becomes easier, so all tests are able to reject with a high probability (Figure 4.8).

For the *multi-modal* setting, we set $\mathbf{c} = 0$ and $\mathbf{b} = 0$, and choose B so that the modes are well-separated. Samples are drawn from the same model with a different $\mathbf{c}$, which controls the mixing weights. Specifically, we choose $B = 6E$, where $E \in \mathbb{R}^{d \times d_h}$ is formed by the top d row and d_h columns of the matrix $I_{d_{\max}}$, where $d_{\max} := \max(d, d_h)$. This renders the local modes of p to be located at the corners of a hyper-cube of width 6. Due to the choice of $\mathbf{b}$ and B, changing $\mathbf{c}$ only affects the weights of the components but not their mode locations. We choose $\mathbf{c} = 0$ for the target so that all components have equal weights, and draw samples with from the same model with $\mathbf{c} = (c_0, c_0, 0, \ldots, 0) \in \mathbb{R}^{d_h}$ for

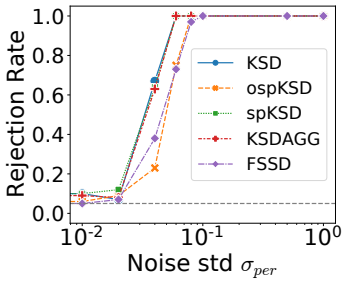


Figure 4.8: GB-RBM with the standard setting.

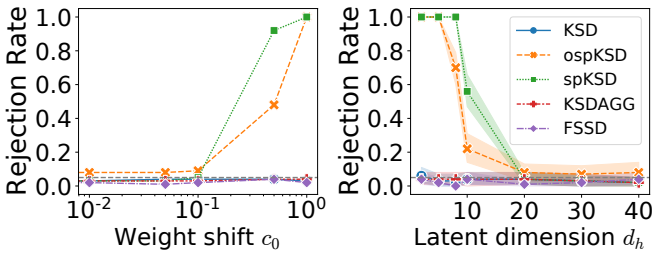


Figure 4.9: GB-RBM with the multi-modal setting.

some c_0 using a Gibbs sampler, which we describe in Section 4.B.

We run ospKSD and spKSD with $T = 50$ steps and report the results in Figure 4.9. The left plot shows the rejection probability with different values of c_0 and $d_h = 5$. As c_0 increases, the weights deviate further from those in the target, which is detected by ospKSD and spKSD as shown by the increasing power. The benchmarks again fail to detect this discrepancy due to the sparsity of the modes. We then fix $c_0 = 5$ and analyze how the performance scales with the latent dimension d_h in the right plot of Figure 4.9. With a moderate d_h, the rejection probabilities of ospKSD and spKSD are significantly larger than the others. This gap vanishes for $d_h \geq 20$, since a larger d_h gives rise to more modes in p and hence more possible jump directions for each point. Therefore, the probability of proposing a "correct" move at each step declines, leading to a small acceptance rate and thus a low power.

4.8 Conclusion and Discussion

We show with a bimodal Gaussian example that GOF tests based on KSD can fail when the target has well-separated modes. To increase its power, we propose to perturb both the candidate and the target distributions using a Markov process before applying the KSD test. Empirical results suggest that our methods (ospKSD and spKSD) are more sensitive to discrepancies in the mixing weights of multimodal distributions, and can achieve remarkably high power particularly when the mixing components are elliptic distributions. We now discuss two limitations of our proposed methods and potential future work.

First, the ospKSD and spKSD rely heavily on accurate estimation of the mode locations and Hessians, which can be extremely challenging and computationally costly for high-dimensional problems. Additionally, the jump proposal of the transition kernel used in the proposed methods is constructed specifically for targets that are mixtures of elliptic distributions, which may be inappropriate for targets with more complicated geometrical structure. Further investigations could aim to find perturbation operators that scale better with dimensionality or that suit a wider family of target distributions.

Second, both spKSD and ospKSD require hyperparameter tuning. The spKSD, as a sum-like statistic, requires a trade-off between the test power and the number of elements in the grid $\mathcal{S}$. Although the heuristic described in Section 4.7 is found to perform decently in our experiments, it is of practical interest to analyze the sensitivity of the test performance to the grid size both empirically and theoretically. The ospKSD, on the other hand, requires a held-out dataset to tune θ, potentially reducing test power

due to data-splitting. One possible approach to mitigate this problem is to combine ospKSD with the aggregated testing framework described in Schrab et al. (2022) to avoid splitting the data.

4.A Proofs of Theoretical Results

This section contains the proofs for the main results in Chapter 4. It is organized as follows:

- Section 4.A.1 proves Proposition 4.1.
- Section 4.A.2 presents and proves an auxiliary lemma, which gives the rate of decay of the Fisher divergence between two distributions: a mixture of Gaussian and one of its components. This result is then used to show Section 4.A.2.
- Section 4.A.3 proves Proposition 4.4, which concerns the asymptotic distributions of spKSD.
- Section 4.A.4 shows that the accept-reject rule (4.9) leads to a transition rule that leaves the target distribution P invariant, as claimed in Section 4.6. It also shows that a class of transition kernels that generalizes (4.9) also leaves P invariant.
- Section 4.A.5 proves Proposition 4.5.

4.A.1 Proof of Proposition 4.1

Fixing positive integer ν, we can write $n_\nu \hat{D}^2_{P_\nu} = n_\nu \hat{D}^2_Q + n_\nu(\hat{D}^2_{P_\nu} - \hat{D}^2_Q)$. Under the stated assumptions on the kernel k and that $\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}[u_Q(\mathbf{X},\mathbf{X}')^2] < \infty$, we can apply Liu et al. (2016, Thm 4.1) to conclude that, as $\nu \to \infty$,

$$n_\nu \hat{D}^2_Q \xrightarrow{\text{d}} \sum_{j=1}^{\infty} c_j(z_j^2 - 1) \ ,$$

where z_j, c_j are as defined in Proposition 4.1. If we could furthermore show that $n_\nu(\hat{D}^2_{P_\nu} - \hat{D}^2_Q) \to 0$ in probability as $\nu \to \infty$, then the desired result would follow from Slutsky's Theorem.

To prove the convergence in probability, we fix $\epsilon > 0$ and denote by $\Pr_{\mathbf{X}_1,\ldots,\mathbf{X}_n\sim Q}$ the probability taken over the i.i.d. random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim Q$. We also omit the dependence of n on ν for brevity. Applying the Markov inequality yields

$$\begin{aligned}
\Pr_{\mathbf{X}_1,\ldots,\mathbf{X}_n\sim Q}(n|\hat{D}^2_{P_\nu} - \hat{D}^2_Q| \geq \epsilon) &\leq \frac{n}{\epsilon}\mathbb{E}_{\mathbf{X}_1,\ldots,\mathbf{X}_n\sim Q}[|\hat{D}^2_{P_\nu} - \hat{D}^2_Q|] \\
&= \frac{n}{\epsilon}\mathbb{E}_{\mathbf{X}_1,\ldots,\mathbf{X}_n\sim Q}\left[\left|\frac{1}{n(n-1)}\sum_{1\leq i\neq j\leq n} u_{P_\nu}(\mathbf{X}_i,\mathbf{X}_j) - u_Q(\mathbf{X}_i,\mathbf{X}_j)\right|\right] \\
&=: T \ .
\end{aligned}$$

Using Jensen's inequality and noting that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d., we have

$$\begin{aligned} T &\leq \frac{n}{\epsilon}\frac{1}{n(n-1)}\sum_{1\leq i\neq j\leq n}\mathbb{E}_{\mathbf{X}_i,\mathbf{X}_j\sim Q}\left[\left|u_{P_\nu}(\mathbf{X}_i,\mathbf{X}_j)-u_Q(\mathbf{X}_i,\mathbf{X}_j)\right|\right] \\ &= \frac{n}{\epsilon}\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[\left|u_{P_\nu}(\mathbf{X},\mathbf{X}')-u_Q(\mathbf{X},\mathbf{X}')\right|\right] \\ &=: T' \,. \end{aligned}$$

To proceed, we substitute the definition of u_{P_ν} and u_Q, and apply a triangle inequality to obtain

$$\begin{aligned} T' &\leq \frac{n}{\epsilon}\Big\{\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[\left|\mathbf{s}_{p_\nu}(\mathbf{X})^\top\mathbf{s}_{p_\nu}(\mathbf{X}')-\mathbf{s}_q(\mathbf{X})^\top\mathbf{s}_q(\mathbf{X}')\right|\left|k(\mathbf{X},\mathbf{X}')\right|\right] \\ &\quad+\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[\left|(\mathbf{s}_{p_\nu}(\mathbf{X})-\mathbf{s}_q(\mathbf{X}))^\top\nabla_{\mathbf{X}'}k(\mathbf{X},\mathbf{X}')\right|\right] \\ &\quad+\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[\left|(\mathbf{s}_{p_\nu}(\mathbf{X}')-\mathbf{s}_q(\mathbf{X}'))^\top\nabla_{x}k(\mathbf{X},\mathbf{X}')\right|\right]\Big\} \\ &\leq \frac{n}{\epsilon}\Big\{\left(\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[(\mathbf{s}_{p_\nu}(\mathbf{X})^\top\mathbf{s}_{p_\nu}(\mathbf{X}')-\mathbf{s}_q(\mathbf{X})^\top\mathbf{s}_q(\mathbf{X}'))^2\right]\right)^{1/2}\left(\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[k(\mathbf{X},\mathbf{X}')^2\right]\right)^{1/2} \\ &\quad+\left(\mathbb{E}_{\mathbf{X}\sim Q}\left[\|\mathbf{s}_{p_\nu}(\mathbf{X})-\mathbf{s}_q(\mathbf{X})\|_2^2\right]\right)^{1/2}\left(\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[\|\nabla_{\mathbf{X}'}k(\mathbf{X},\mathbf{X}')\|_2^2\right]\right)^{1/2} \\ &\quad+\left(\mathbb{E}_{\mathbf{X}\sim Q}\left[\|\mathbf{s}_{p_\nu}(\mathbf{X})-\mathbf{s}_q(\mathbf{X})\|_2^2\right]\right)^{1/2}\left(\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[\|\nabla_{\mathbf{X}}k(\mathbf{X},\mathbf{X}')\|_2^2\right]\right)^{1/2}\Big\}\,, \end{aligned} \tag{4.13}$$

where the last inequality follows from Cauchy-Schwarz inequality applied to each expectation. We bound each of the three terms individually. For the first term, we can bound as

$$\begin{aligned} &\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[\left(\mathbf{s}_{p_\nu}(\mathbf{X})^\top\mathbf{s}_{p_\nu}(\mathbf{X}')-\mathbf{s}_q(\mathbf{X})^\top\mathbf{s}_q(\mathbf{X}')\right)^2\right] \\ &= \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[\left(\mathbf{s}_{p_\nu}(\mathbf{X})^\top(\mathbf{s}_{p_\nu}(\mathbf{X}')-\mathbf{s}_q(\mathbf{X}'))+(\mathbf{s}_{p_\nu}(\mathbf{X})-\mathbf{s}_q(\mathbf{X}))^\top\mathbf{s}_q(\mathbf{X}')\right)^2\right] \\ &\leq \underbrace{2\,\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[\left(\mathbf{s}_{p_\nu}(\mathbf{X})^\top(\mathbf{s}_{p_\nu}(\mathbf{X}')-\mathbf{s}_q(\mathbf{X}'))\right)^2\right]}_{=:T_1}+\underbrace{2\,\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}\left[\left((\mathbf{s}_{p_\nu}(\mathbf{X})-\mathbf{s}_q(\mathbf{X}))^\top\mathbf{s}_q(\mathbf{X}')\right)^2\right]}_{=:T_2}, \end{aligned}$$

where the last line follows from the fact that $(a+b)^2 \leq 2a^2+2b^2$ for any $a, b \in \mathbb{R}$. Now, applying the Cauchy-Schwarz inequality gives

$$\begin{aligned} T_1 &\leq 2\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}[\|\mathbf{s}_{p_\nu}(\mathbf{X})\|_2^2\|\mathbf{s}_{p_\nu}(\mathbf{X}')-\mathbf{s}_q(\mathbf{X}')\|_2^2] \\ &\leq 2\left(2\mathbb{E}_{\mathbf{X}\sim Q}[\|\mathbf{s}_{p_\nu}(\mathbf{X})-\mathbf{s}_q(\mathbf{X})\|_2^2]+2\mathbb{E}_{\mathbf{X}\sim Q}[\|\mathbf{s}_q(\mathbf{X})\|_2^2]\right)\mathbb{E}_{\mathbf{X}'\sim Q}[\|\mathbf{s}_{p_\nu}(\mathbf{X}')-\mathbf{s}_q(\mathbf{X}')\|_2^2] \\ &\leq 4F_\nu^2+4\mathbb{E}_{\mathbf{X}\sim Q}[\|\mathbf{s}_q(\mathbf{X})\|_2^2]F_\nu\,, \end{aligned}$$

where $F_\nu := \mathbb{E}_{\mathbf{X}\sim Q}[\|\mathbf{s}_{p_\nu}(\mathbf{X})-\mathbf{s}_q(\mathbf{X})\|_2^2]$ is the Fisher Divergence between P_ν and Q, and where the second line holds because

$$\|\mathbf{s}_{p_\nu}(\mathbf{x})\|_2^2 = \|\mathbf{s}_{p_\nu}(\mathbf{x})-\mathbf{s}_q(\mathbf{x})+\mathbf{s}_q(\mathbf{x})\|_2^2 \leq 2\|\mathbf{s}_{p_\nu}(\mathbf{x})-\mathbf{s}_q(\mathbf{x})\|_2^2+2\|\mathbf{s}_q(\mathbf{x})\|_2^2\,.$$

A similar argument by Cauchy-Schwarz inequality shows that

$$T_2 \leq \mathbb{E}_{\mathbf{X}\sim Q}[\|\mathbf{s}_q(\mathbf{X})\|_2^2]F_\nu\,.$$

Combining the bounds for T_1 and T_2 yields

$$\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}[(\mathbf{s}_{p_\nu}(\mathbf{X})^\top \mathbf{s}_{p_\nu}(\mathbf{X}') - \mathbf{s}_q(\mathbf{X})^\top \mathbf{s}_q(\mathbf{X}'))^2] \;\leq\; 8F_\nu^2 + 10\mathbb{E}_{\mathbf{X}\sim Q}[\|\mathbf{s}_q(\mathbf{X})\|_2^2]F_\nu \,.$$

By the assumed boundedness of the kernel and its gradients, there exists a positive constant $M < \infty$ depending only on k and Q such that

$$\max\left\{\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}[|k(\mathbf{X},\mathbf{X}')|],\; \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}[\|\nabla_{\mathbf{X}'}k(\mathbf{X},\mathbf{X}')\|_2^2],\; \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}[\|\nabla_{\mathbf{X}}k(\mathbf{X},\mathbf{X}')\|_2^2]\right\} \;\leq\; M \,.$$

We hence conclude from (4.13) that

$$\begin{aligned}\mathrm{Pr}_{\mathbf{X}_1,\ldots,\mathbf{X}_n\sim Q}(n|\hat{D}_{P_\nu}^2 - \hat{D}_Q^2| \geq \epsilon) \;&\leq\; \frac{n}{\epsilon}\left[M^{1/2}\left(8F_\nu^2 + 10\mathbb{E}_{\mathbf{X}\sim Q}[\|\mathbf{s}_q(\mathbf{X})\|_2^2]F_\nu\right)^{1/2} + 2M^{1/2}F_\nu^{1/2}\right] \\ &\leq\; \frac{n}{\epsilon}M^{1/2}\underbrace{\left[F_\nu^{1/2}\left(8F_\nu + 10\mathbb{E}_{\mathbf{X}\sim Q}[\|\mathbf{s}_q(\mathbf{X})\|_2^2]\right)^{1/2} + 2F_\nu^{1/2}\right]}_{=:T_3} \,. \quad (4.14)\end{aligned}$$

The term T_3 is $O\big(\max(F_\nu, F_\nu^{1/2})\big)$. Therefore, if $n = n_\nu = o\big(1/\max(F_\nu, F_\nu^{1/2})\big)$, then the right hand side of (4.14) converges to 0 as $\nu \to \infty$, thus $n_\nu|\hat{D}_{P_\nu}^2 - \hat{D}_Q^2| \to 0$ in probability. This completes the proof.

4.A.2 Proof of Theorem 4.2

Theorem 4.2 follows directly from Proposition 4.1 and the next lemma, which states that the Fisher Divergence between Q and P_Δ decays with a rate at least exponentially fast in the inter-modal distance $\|\Delta\|_2$.

Lemma 4.6. *Under the same assumptions in Proposition 4.1, we have* $\mathbb{E}_{\mathbf{X}\sim Q}[\|\mathbf{s}_{p_\Delta}(\mathbf{X}) - \mathbf{s}_q(\mathbf{X})\|_2^2] = o\left(e^{-\|\Delta\|_2^2/32}\right)$.

Proof of Lemma 4.6. For any $\delta > 0$, define $B_\delta := \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \leq \delta\}$. We have the following decomposition

$$\begin{aligned}&\mathbb{E}_{\mathbf{X}\sim Q}\left[\|\mathbf{s}_{p_\Delta}(\mathbf{X}) - \mathbf{s}_q(\mathbf{X})\|_2^2\right] \\ &= \mathbb{E}_{\mathbf{X}\sim Q}\left[\delta_{\mathbf{X}}(B_\delta)\|\mathbf{s}_{p_\Delta}(\mathbf{X}) - \mathbf{s}_q(\mathbf{X})\|_2^2\right] + \mathbb{E}_{\mathbf{X}\sim Q}\left[\delta_{\mathbf{X}}(\mathbb{R}^d\backslash B_\delta)\|\mathbf{s}_{p_\Delta}(\mathbf{X}) - \mathbf{s}_q(\mathbf{X})\|_2^2\right] \,. \quad (4.15)\end{aligned}$$

The rest of the proof proceeds with bounding the two terms separately. We first note that standard computation gives

$$\frac{p_\Delta(\mathbf{x})}{q(\mathbf{x})} \;=\; \frac{\pi\exp\left(-\frac{1}{2}\|\mathbf{x}\|^2\right) + (1-\pi)\exp\left(-\frac{1}{2}\|\mathbf{x}-\Delta\|^2\right)}{\exp\left(-\frac{1}{2}\|\mathbf{x}\|^2\right)} \;=\; \pi + (1-\pi)\exp\left(\Delta^\top\mathbf{x} - \frac{1}{2}\|\Delta\|^2\right) \,,$$

and

$$\|\mathbf{s}_{p_\Delta}(\mathbf{x}) - \mathbf{s}_q(\mathbf{x})\|_2^2 \;=\; \left\|\frac{(1-\pi)\Delta\exp\left(\Delta^\top\mathbf{x} - \frac{1}{2}\|\Delta\|_2^2\right)}{\pi + (1-\pi)\exp\left(\Delta^\top\mathbf{x} - \frac{1}{2}\|\Delta\|_2^2\right)}\right\|_2^2 \;=\; \frac{(1-\pi)^2\|\Delta\|_2^2}{\left(1 - \pi + \pi\exp\left(-\Delta^\top\mathbf{x} + \frac{1}{2}\|\Delta\|_2^2\right)\right)^2} \,.$$

For $\mathbf{x} \in B_\delta$, Cauchy-Schwarz inequality implies $\Delta^\top \mathbf{x} \leq \|\Delta\|_2 \|\mathbf{x}\|_2 \leq \delta \|\Delta\|_2$. Hence,

$$\|\mathbf{s}_{p_\Delta}(\mathbf{x}) - \mathbf{s}_q(\mathbf{x})\|_2^2 \;\leq\; \frac{(1-\pi)^2\|\Delta\|_2^2}{\pi^2 \exp\left(-2\Delta^\top \mathbf{x} + \|\Delta\|_2^2\right)} \;\leq\; \frac{(1-\pi)^2\|\Delta\|_2^2}{\pi^2 \exp\left(-2\delta\|\Delta\|_2 + \|\Delta\|_2^2\right)} \,,$$

and the first term of (4.15) can be bounded as

$$\begin{aligned}\mathbb{E}_{\mathbf{X}\sim Q}\left[\delta_{\mathbf{X}}(B_\delta)\|\mathbf{s}_{p_\Delta}(\mathbf{X}) - \mathbf{s}_q(\mathbf{X})\|_2^2\right] \;&\leq\; \mathbb{E}_{\mathbf{X}\sim Q}\left[\delta_{\mathbf{X}}(B_\delta)\frac{(1-\pi)^2\|\Delta\|_2^2}{\pi^2 \exp\left(-2\delta\|\Delta\|_2 + \|\Delta\|_2^2\right)}\right] \\ &\leq\; \frac{(1-\pi)^2\|\Delta\|_2^2}{\pi^2 \exp\left(-2\delta\|\Delta\|_2 + \|\Delta\|_2^2\right)} \,,\end{aligned}$$

where the last inequality follows from the fact that $\mathbb{E}_{\mathbf{X}\sim Q}[\delta_{\mathbf{X}}(B_\delta)] \leq 1$.

To bound the second term of (4.15), we note that for $\mathbf{x} \in \mathbb{R}^d \backslash B_\delta$,

$$\|\mathbf{s}_{p_\Delta}(\mathbf{x}) - \mathbf{s}_q(\mathbf{x})\|_2^2 \leq \frac{(1-\pi)^2\|\Delta\|_2^2}{(1-\pi)^2} \;=\; \|\Delta\|_2^2 \,.$$

Therefore,

$$\mathbb{E}_{\mathbf{X}\sim Q}\left[\delta_{\mathbf{X}}(B_\delta)\|\mathbf{s}_{p_\Delta}(\mathbf{X}) - \mathbf{s}_q(\mathbf{X})\|_2^2\right] \;\leq\; \|\Delta\|_2^2 \mathbb{E}_{\mathbf{X}\sim Q}\left[\delta_{\mathbf{X}}(B_\delta)\right] \;\leq\; 5^d\|\Delta\|_2^2 \exp\left(-\frac{\delta^2}{8}\right) \,,$$

by the tail probability of the norm of centered Gaussian random vectors (see, e.g., Wainwright 2019, Proposition 2.5). Combining these results we have

$$\begin{aligned}\mathbb{E}_{\mathbf{X}\in Q}[\|\mathbf{s}_{p_\Delta}(\mathbf{X}) - \mathbf{s}_q(\mathbf{X})\|_2^2] \;&\leq\; \frac{(1-\pi)^2\|\Delta\|_2^2}{\pi^2 \exp\left(-2\delta\|\Delta\|_2 + \|\Delta\|_2^2\right)} + 5^d\|\Delta\|_2^2 \exp\left(-\frac{\delta^2}{8}\right) \\ &=\; (1-\pi)^2\pi^{-2}\|\Delta\|_2^2 \exp\left(-\frac{1}{17}\|\Delta\|_2^2\right) + 5^d\|\Delta\|_2^2 \exp\left(-\frac{1}{17}\|\Delta\|_2^2\right) \,,\end{aligned}$$

where the last line follows by choosing $\delta = 8\|\Delta\|_2/17$. Noting the RHS of the last inequality is $o\left(-\frac{1}{32}\|\Delta\|_2^2\right)$ completes the proof. □

4.A.3 Proof of Proposition 4.4

By Serfling (2009, Sections 5.5.1, 5.5.2), sufficient conditions for the stated results are

C1. $\mathbb{E}_{\mathbf{W},\mathbf{W}'\sim R_Q}[\tilde{u}_P(\mathbf{W},\mathbf{W}')^2] < \infty$.

C2. Under H_0: $\xi_1 \coloneqq \mathrm{Var}_{\mathbf{W}\sim R_Q}(\mathbb{E}_{\mathbf{W}'\sim R_Q}[\tilde{u}_P(\mathbf{W},\mathbf{W}')]) = 0$ and $\xi_2 \coloneqq \mathrm{Var}_{\mathbf{W},\mathbf{W}'\sim R_Q}(\tilde{u}_P(\mathbf{W},\mathbf{W}')) > 0$.

C3. Under H_1: $\xi_1 > 0$.

Now C1. holds by assumption. To show C2., we start with the following decomposition. For $\mathbf{W}' \sim R_Q$, we can write $\mathbf{W}' = (\mathbf{y}^1, \ldots, \mathbf{y}^S)$, where the marginal distribution of each $\mathbf{y}^s$ is $\mathcal{K}_s Q$. Therefore, for any

deterministic $\mathbf{w} = (\mathbf{x}^1, \ldots, \mathbf{x}^S) \in \mathcal{X}^S$,

$$\mathbb{E}_{\mathbf{W}' \sim R_Q}[\tilde{u}_P(\mathbf{W}, \mathbf{W}')] \;=\; \sum_{s=1}^{S} \mathbb{E}_{\mathbf{W}' \sim R_Q}[u_P(bX^s, \mathbf{Y}^s)] \;=\; \sum_{s=1}^{S} \mathbb{E}_{\mathbf{Y}^s \sim \mathcal{K}_s Q}[u_P(\mathbf{X}^s, \mathbf{Y}^s)] \,, \tag{4.16}$$

where the first equality holds by the additive form of $\tilde{u}_P(\mathbf{w}, \mathbf{w}')$, and the second equality follows since each expectation is only taken over the distribution of a single $\mathbf{y}^s$. When $Q = P$, each term in (4.16) equals to $\mathbb{E}_{\mathbf{Y}^s \sim P}[u_P(\mathbf{x}^s, \mathbf{Y}^s)]$ by P-invariance. Now, under the assumed conditions in Proposition 4.3, the same argument in the proof of Liu et al. (2016, Theorem 4.1) shows that $\mathbb{E}_{\mathbf{Y}^s \sim P}[u_P(\mathbf{x}^s, \mathbf{Y}^s)] = 0$. Hence, $\xi_1 = 0$.

To prove $\xi_2 > 0$ when $Q = P$, we suppose for a contradiction that $\xi_2 = 0$. We then must have $\tilde{u}_P \equiv c'$ P-almost surely for some fixed constant c'. Since P admits a positive density on $\mathcal{X}$ by assumption, this implies that $c' = 0$. On the other hand, for any probability measure Q' on $\mathcal{X}$, taking expectation with respect to $R_{Q'}$ yields

$$\begin{aligned} 0 \;=\; c' \;=\; \mathbb{E}_{\mathbf{W}, \mathbf{W}' \sim R_{Q'}}[\tilde{u}_P(\mathbf{W}, \mathbf{W}')] \;&=\; \sum_{s=1}^{S} \mathbb{E}_{(\mathbf{X}^1, \ldots, \mathbf{X}^S), (\mathbf{Y}^1, \ldots, \mathbf{Y}^m) \sim R_{Q'}}[u_P(\mathbf{X}^s, \mathbf{Y}^s)] \\ &= \sum_{s=1}^{S} \mathbb{E}_{\mathbf{X}^s, \mathbf{Y}^s \sim \mathcal{K}_s Q'}[u_P(\mathbf{X}^s, \mathbf{Y}^s)] \\ &= \sum_{s=1}^{S} D(\mathcal{K}_s Q', P) \;\geq\; D(Q', P) \,, \end{aligned}$$

where the second equality follows from a similar argument in (4.16), and the inequality holds because $\mathcal{K}_{\mathrm{id}} \in \mathcal{S}$. This is a contradiction, as it would imply $D(Q', P) = 0$ for any $Q' \neq P$.

To show C3, we prove the contrapositive by supposing $\xi_1 = 0$ and aiming to show that $Q = P$. If $\xi_1 = 0$ then there must exist a constant c for which

$$c \;=\; \mathbb{E}_{\mathbf{W}' \sim R_Q}[\tilde{u}_P(\mathbf{w}, \mathbf{W}')] \,,$$

for R_Q-almost-all $\mathbf{w}$. With the stated choice of $\mathcal{K}$ and the assumption that Q admits a Lebesgue density, the above identity also holds for R_P-almost-all $\mathbf{w}$, where R_P is constructed in the same way as R_Q by replacing Q with P. Taking expectation of both sides with respect to $\mathbf{W} \sim R_P$ then yields

$$c \;=\; \mathbb{E}_{\mathbf{W} \sim R_P} \mathbb{E}_{\mathbf{W}' \sim R_Q}[\tilde{u}_P(\mathbf{W}, \mathbf{W}')] \;=\; \mathbb{E}_{\mathbf{W}' \sim R_Q} \mathbb{E}_{\mathbf{W} \sim R_P}[\tilde{u}_P(\mathbf{W}, \mathbf{W}')] \,,$$

where the last equality follows from the Fubini-Toneli Theorem. Following the same argument above for $Q = P$, we conclude that $\mathbb{E}_{\mathbf{W} \sim R_P}[\tilde{u}_P(\mathbf{W}, \mathbf{w}')] = 0$ for R_Q-almost-all $\mathbf{w}'$, and hence

$$0 \;=\; \mathbb{E}_{\mathbf{W}' \sim R_Q} \mathbb{E}_{\mathbf{W} \sim R_Q}[\tilde{u}_P(\mathbf{W}, \mathbf{W}')] \;=\; \sum_{s=1}^{S} D(\mathcal{K}_s Q, P) \;\geq\; D(Q, P) \,.$$

It follows that $D(Q, P) = 0$, thus $Q = P$.

4.A.4 Validity of the Accept-Reject Rule

In this section, we first show in Section 4.A.4.1 that the accept-reject rule $\alpha(\mathbf{x}, \mathbf{x}')$ defined in (4.9) leaves the target distribution P invariant by proving that it fulfills the detailed balance equation. We then present in Section 4.A.4.2 a more general class of accept-reject rules, which contains (4.9) as a special case, and show that any member of this class also satisfies the detailed balance equation. The results in Section 4.A.4.2 could be of independent interest.

4.A.4.1 A Sufficient Condition

Let $\mathcal{K}$ be the Markov transition kernel studied in Section 4.6.1. We first present a sufficient condition for the detailed balance equation (4.8):

$$\int_{\mathbf{x}\in A} \sum_{\mathbf{u}\in\mathcal{U}} \delta_{\mathbf{h}(\mathbf{x}|\mathbf{u})}(B) p(\mathbf{x}) g(\mathbf{u}) \alpha(\mathbf{x}, \mathbf{h}(\mathbf{x}|\mathbf{u}))\, \mathrm{d}\mathbf{x} \;=\; \int_{\mathbf{x}'\in B} \sum_{\mathbf{u}'\in\mathcal{U}} \delta_{\mathbf{h}(\mathbf{x}'|\mathbf{u}')}(A) p(\mathbf{x}') g(\mathbf{u}') \alpha(\mathbf{x}', \mathbf{h}(\mathbf{x}'|\mathbf{u}'))\, \mathrm{d}\mathbf{x}' \tag{4.17}$$

for all $A, B \in \mathcal{B}(\mathcal{X})$. For simplicity, we will write $\mathbf{x}' = \mathbf{h}(\mathbf{x}|\mathbf{u})$ and $\mathbf{x} = \mathbf{h}^{-1}(\mathbf{x}'|\mathbf{u}')$, so that the dependence of $\mathbf{x}'$ on $\mathbf{u}$ and of $\mathbf{x}$ on $\mathbf{x}'$ is implicit.

Proposition 4.7. *Let p be a probability density function on $\mathcal{X} \subset \mathbb{R}^d$. Suppose that $\mathbf{h}$ is a deterministic, invertible function that is differentiable with differentiable inverse. Furthermore, let g be a known density defined on some discrete space $\mathcal{U}$. Consider a Markov transition kernel of the form*

$$\mathcal{K}(\mathbf{x}, A) \;=\; \sum_{\mathbf{u}\in\mathcal{U}} \delta_{\mathbf{x}'}(A) g(\mathbf{u}) \alpha(\mathbf{x}, \mathbf{x}') + \delta_x(A) r(\mathbf{x})\;, \tag{4.18}$$

where $\mathbf{x}' \coloneqq \mathbf{h}(\mathbf{x}|\mathbf{u})$, $\delta_x(A) = 1$ if $\mathbf{x} \in A$ and 0 otherwise, and $r(\mathbf{x}) = 1 - \sum_{\mathbf{u}\in\mathcal{U}} g(\mathbf{u}(\mathbf{x}, \mathbf{x}'))\alpha(\mathbf{x}, \mathbf{x}')$. Then an accept-reject rule $\alpha(\mathbf{x}, \mathbf{x}')$ satisfies the detailed balance condition (4.17) if

$$p(\mathbf{x}) g(\mathbf{u}) \alpha(\mathbf{x}, \mathbf{x}') \;=\; p(\mathbf{x}') g(\mathbf{u}') \alpha(\mathbf{x}', \mathbf{x}) \left| \frac{\partial \mathbf{h}(\mathbf{x}|\mathbf{u})}{\partial \mathbf{x}} \right| . \tag{4.19}$$

Proof of Proposition 4.7. The proof largely imitates Green and Hastie (2009, Section 2.1), which shows the claim when the density g is defined on a continuous space. Defining $\mathcal{U}_B \coloneqq \{\mathbf{u} : \mathbf{x}' = \mathbf{h}(\mathbf{x}|\mathbf{u}) \in B \text{ for some } \mathbf{x} \in \mathcal{X}\}$ and $\mathcal{U}_A \coloneqq \{\mathbf{u} : \mathbf{x} = \mathbf{h}^{-1}(\mathbf{x}'|\mathbf{u}) \in B \text{ for some } \mathbf{x}' \in \mathcal{X}\}$, we can rewrite (4.17) as

$$\int_{\mathbf{x}\in A} \sum_{\mathbf{u}\in\mathcal{U}_B} p(\mathbf{x}) g(\mathbf{u}) \alpha(\mathbf{x}, \mathbf{x}')\, \mathrm{d}\mathbf{x} \;=\; \int_{\mathbf{x}'\in B} \sum_{\mathbf{u}'\in\mathcal{U}_A} p(\mathbf{x}') g(\mathbf{u}') \alpha(\mathbf{x}', \mathbf{x})\, \mathrm{d}\mathbf{x}' .$$

Noting that $(\mathbf{x}, \mathbf{u}) \in A \times \mathcal{U}_B \iff (\mathbf{x}', \mathbf{u}') = (\mathbf{h}(\mathbf{x}|\mathbf{u}), \mathbf{u}) \in B \times \mathcal{U}_A$, and by the invertibility of the transformation $\mathbf{h}$, a change-of-variable formula can be applied to the right-hand-side of (4.17) to yield

$$\int_{\mathbf{x}\in A} \sum_{\mathbf{u}\in\mathcal{U}_A} p(\mathbf{x}) g(\mathbf{u}) \alpha(\mathbf{x}, \mathbf{x}')\, \mathrm{d}\mathbf{x} \;=\; \int_{\mathbf{x}\in A} \sum_{\mathbf{u}\in\mathcal{U}_A} p(\mathbf{x}') g(\mathbf{u}) \alpha(\mathbf{x}', \mathbf{x}) \left| \frac{\partial \mathbf{h}(\mathbf{x}|\mathbf{u})}{\partial \mathbf{x}} \right| \mathrm{d}\mathbf{x} .$$

We therefore conclude that a sufficient condition is

$$p(\mathbf{x})g(\mathbf{u})\alpha(\mathbf{x},\mathbf{x}') \;=\; p(\mathbf{x}')g(\mathbf{u}')\alpha(\mathbf{x}',\mathbf{x})\left|\frac{\partial \mathbf{h}(\mathbf{x}|\mathbf{u})}{\partial \mathbf{x}}\right| .$$

□

From Proposition 4.7, it follows that, for any probability mass function g on the discrete space $\mathcal{U}$ with $g(\mathbf{u}) > 0$ for all $\mathbf{u} \in \mathcal{U}$, the detailed balance condition holds with

$$\alpha(\mathbf{x},\mathbf{x}') \;=\; \min\left(1, \frac{p(\mathbf{x}')g(\mathbf{u}')}{p(\mathbf{x})g(\mathbf{u})}\left|\frac{\partial \mathbf{h}(\mathbf{x}|\mathbf{u})}{\partial \mathbf{x}}\right|\right) , \tag{4.20}$$

by verifying that it indeed satisfies (4.19). This can be viewed as a generalization of the Metropolis-Hastings (MH) rule $\alpha(\mathbf{x},\mathbf{x}') = \min\left(1, \frac{p(\mathbf{x}')g(\mathbf{u}')}{p(\mathbf{x})g(\mathbf{u})}\right)$.

4.A.4.2 A Class of Valid Accept-Reject Rules

Accept-reject rules of the form (4.20) is not the only choice that satisfies the detailed balance condition. For the standard Metropolis-Hastings transition kernel, alternative accept-reject rules have been studied (Barker, 1965; Peskun, 1973; Hird et al., 2020). We follow Hird et al. (2020) to propose a class of accept-reject rules that are valid for proposed kernels of the form (4.18).

Similarly as in Proposition 4.7, in this subsection, we will assume g is a probability mass function defined on the discrete space $\mathcal{U}$ with $g(\mathbf{u}) > 0$ for all $\mathbf{u} \in \mathcal{U}$.

Lemma 4.8. *Using the same notations in Proposition 4.7, define* $t(\mathbf{x},\mathbf{x}') \coloneqq \frac{p(\mathbf{x}')g(\mathbf{u}')}{p(\mathbf{x})g(\mathbf{u})}$ *when* $p(\mathbf{x})g(\mathbf{u}) > 0$, *and* $t(\mathbf{x},\mathbf{x}') = 0$ *otherwise, where* $\mathbf{u},\mathbf{u}' \in \mathcal{U}$ *such that* $\mathbf{x}' = \mathbf{h}(\mathbf{x}|\mathbf{u})$ *and* $\mathbf{x} = \mathbf{h}(\mathbf{x}'|\mathbf{u}')$. *Then the equality (4.19) holds for*

$$\alpha(\mathbf{x},\mathbf{x}') \;=\; \rho\left(\left|\frac{\partial \mathbf{h}(\mathbf{x}|\mathbf{u})}{\partial \mathbf{x}}\right| t(\mathbf{x},\mathbf{x}')\right) ,$$

where ρ *is any function that satisfies* $\rho(s) = s\rho(1/s)$, *for all* $s > 0$, *and* $\rho(0) \coloneqq 0$.

Proof of Lemma 4.8. We follow the derivation in Hird et al. (2020, Eq. 4). By the definition of t, it is obvious that $t(\mathbf{x},\mathbf{x}') = 1/t(\mathbf{x}',\mathbf{x})$ and $p(\mathbf{x})g(\mathbf{u})t(\mathbf{x},\mathbf{x}') = p(\mathbf{x}')g(\mathbf{u}')$. The assumption on ρ then implies

$$\begin{aligned}
p(\mathbf{x})g(\mathbf{u})\alpha(\mathbf{x},\mathbf{x}') \;&=\; p(\mathbf{x})g(\mathbf{u})\rho\left(\left|\frac{\partial \mathbf{h}(\mathbf{x}|\mathbf{u})}{\partial \mathbf{x}}\right| t(\mathbf{x},\mathbf{x}')\right) \\
&=\; p(\mathbf{x})g(\mathbf{u})t(\mathbf{x},\mathbf{x}')\left|\frac{\partial \mathbf{h}(\mathbf{x}|\mathbf{u})}{\partial \mathbf{x}}\right| \rho\left(\frac{1}{\left|\frac{\partial \mathbf{h}(\mathbf{x}|\mathbf{u})}{\partial \mathbf{x}}\right| t(\mathbf{x},\mathbf{x}')}\right) \\
&=\; p(\mathbf{x}')g(\mathbf{u}')\left|\frac{\partial \mathbf{h}(\mathbf{x}|\mathbf{u})}{\partial \mathbf{x}}\right| \rho\left(\left|\frac{\partial \mathbf{h}(\mathbf{x}'|\mathbf{u}')}{\partial \mathbf{x}}\right| t(\mathbf{x}',\mathbf{x})\right) \\
&=\; p(\mathbf{x}')\alpha(\mathbf{x}',\mathbf{x})\left|\frac{\partial \mathbf{h}(\mathbf{x}|\mathbf{u})}{\partial \mathbf{x}}\right| ,
\end{aligned} \tag{4.21}$$

where (4.21) follows from the fact that $\left|\frac{\partial \mathbf{h}(\mathbf{x}'|\mathbf{u}')}{\partial \mathbf{x}'}\right| = \left|\frac{\partial \mathbf{h}(\mathbf{x}|\mathbf{u})}{\partial \mathbf{x}}\right|^{-1}$ by the invertibility of $\mathbf{h}$. □

In particular, choosing $\rho(t) = \min(1, t)$ gives the generalized MH accept-reject rule (4.20). Another feasible choice is $g(t) = t/(1+t)$, which leads to a generalized version of the Barker's rule (Barker, 1965; Peskun, 1973; Livingstone and Zanella, 2021)

$$\alpha(\mathbf{x}', \mathbf{x}) \;=\; \frac{p(\mathbf{x}')g(\mathbf{u}')\left|\frac{\partial \mathbf{h}(\mathbf{x}|\mathbf{u})}{\partial \mathbf{x}}\right|}{p(\mathbf{x})g(\mathbf{u}) + p(\mathbf{x}')g(\mathbf{u}')\left|\frac{\partial \mathbf{h}(\mathbf{x}|\mathbf{u})}{\partial \mathbf{x}}\right|} \;,$$

whenever $p(\mathbf{x})g(\mathbf{u}) > 0$, and 0 otherwise.

4.A.5 Proof of Proposition 4.5

We first characterize the limiting distribution when the initial distribution is a point mass δ_{x_0} for any $x_0 \in \mathbb{R}^d$, then generalize the result to an arbitrary probability measure Q.

4.A.5.1 Limiting Distribution with a Point Mass Initial Distribution

Fixing $\mathbf{x}_0 \in \mathbb{R}^d$, we define $\mathcal{I}_{\mathbf{x}_0} := \{\mathbf{x}_0 + k\nu : k \in \mathbb{Z}\}$. We first identify a stationary distribution and aim to show that it is also the limiting distribution.

Lemma 4.9. *The following probability mass function defines a stationary distribution under $\mathcal{K}$:*

$$r_{\mathbf{x}_0}(\mathbf{x}) \;=\; \frac{p(\mathbf{x})}{\sum_{k\in\mathbb{Z}} p(\mathbf{x}_0 + k\nu)} \;,$$

if $\mathbf{x} \in \mathcal{I}_{\mathbf{x}_0}$, and $r_{\mathbf{x}_0}(\mathbf{x}) = 0$, otherwise.

Proof of Lemma 4.9. A sufficient condition for the detailed-balance condition in this case is

$$r_{\mathbf{x}_0}(\mathbf{x})g(\mathbf{u})\alpha(\mathbf{x}, \mathbf{x}') \;=\; r_{\mathbf{x}_0}(\mathbf{x}')g'(\mathbf{u}')\alpha(\mathbf{x}', \mathbf{x}) \;, \tag{4.22}$$

for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$. Fix $\mathbf{x}$. Since $\alpha(\mathbf{x}, \mathbf{x}') = 0$ unless $\mathbf{x}' \in \{\mathbf{x} - \nu, \mathbf{x} + \nu\}$, it is sufficient to check whether the above equation holds for $\mathbf{x}' \in \{\mathbf{x} - \nu, \mathbf{x} + \nu\}$. For example, for $\mathbf{x}' = \mathbf{x} + \nu$,

$$\begin{aligned}
r_{\mathbf{x}_0}(\mathbf{x})g(\mathbf{u})\alpha(\mathbf{x}, \mathbf{x}') \;&=\; \frac{p(\mathbf{x})}{\sum_{k\in\mathbb{Z}} p(\mathbf{x}_0 + k\nu)} g(\mathbf{u}) \min\left(1, \frac{g(\mathbf{u}')p(\mathbf{x}')}{g(\mathbf{u})p(\mathbf{x})}\right) \\
&\overset{(a)}{=}\; \frac{1}{\sum_{k\in\mathbb{Z}} p(\mathbf{x}_0 + k\nu)} g(\mathbf{u}') \min\left(p(\mathbf{x}), p(\mathbf{x}')\right) \\
&\overset{(b)}{=}\; \frac{p(\mathbf{x}')}{\sum_{k\in\mathbb{Z}} p(\mathbf{x}_0 + k\nu)} g(\mathbf{u}') \min\left(\frac{g(\mathbf{u})p(\mathbf{x})}{g(\mathbf{u}')p(\mathbf{x}')}, 1\right) \\
&=\; r_{\mathbf{x}_0}(\mathbf{x}')g'(\mathbf{u}')\alpha(\mathbf{x}', \mathbf{x}) \;,
\end{aligned}$$

where (a) holds as $g(\mathbf{u}) = g(\mathbf{u}')$ by definition, and (b) follows again from the identity $g(\mathbf{u}) = g(\mathbf{u}')$. This verifies ((4.22)). A similar derivation for $\mathbf{x}' = \mathbf{x} - \nu$ completes the proof. □

The next result shows that the Markov chain defined on $\mathcal{I}_{x_0}$ is irreducible and aperiodic under mild conditions.

Lemma 4.10. *Given $x_0 \in \mathcal{X}$ and consider the Markov chain with initial distribution δ_{x_0}. Then*

1. *All state in $\mathcal{I}_{x_0}$ are irreducible.*
2. *A state $\mathbf{x} \in \mathcal{I}_{x_0}$ is aperiodic if $p(\mathbf{x}+\nu) < p(\mathbf{x})$ or $p(\mathbf{x}-\nu) < p(\mathbf{x})$.*

Proof of Lemma 4.10. To prove the first claim, it is sufficient to show that any state $\mathbf{x} \in \mathcal{I}_{\mathbf{x}_0}$ can reach any other state $\mathbf{y} \in \mathcal{I}_{\mathbf{x}_0}$ with positive probability, i.e., for any singleton $A = \{\mathbf{y}\}$ where $\mathbf{y} \in \mathcal{I}_{\mathbf{x}_0}$, there exists $T \in \mathbb{N}$ so that $\mathcal{K}^T(\mathbf{x}, A) > 0$, where

$$\mathcal{K}(\mathbf{x}, A) \;=\; \sum_{\mathbf{u}\in\mathcal{U}} \delta_{\mathbf{x}'}(A) g(\mathbf{u}) \alpha(\mathbf{x}, \mathbf{x}') + \delta_x(A) r(\mathbf{x}) \;,$$

where $\mathcal{U} = \{(1,2),(2,1)\}$, $g(\mathbf{u}) = 1/2$ for all $\mathbf{u} \in \mathcal{U}$, and $\mathbf{x}' = \mathbf{x}'(\mathbf{x}, \mathbf{u}) = \mathbf{x} + \theta(\mu_{u_1} - \mu_{u_2})$ (see Section 4.6.2). We fix $\mathbf{x} \in \mathcal{I}_{d_0}$ and pick $\mathbf{y} \in \{\mathbf{x}-\nu, \mathbf{x}+\nu\}$, i.e., $\mathbf{y}$ is the point immediately to the left or right of $\mathbf{x}$ in $\mathcal{I}_{x_0}$. The transition probability in this case reduces to

$$\mathcal{K}(\mathbf{x}, \{\mathbf{y}\}) \;=\; \frac{1}{2}\alpha(\mathbf{x}, \mathbf{y}) \;=\; \frac{1}{2}\min\left(1, \frac{p(\mathbf{y})}{p(\mathbf{x})}\right) \;,$$

which is positive as p is positive on $\mathbb{R}^d$, i.e., $\mathbf{x}$ can move to its left or right state in one step with positive probability. An inductive argument directly shows that $\mathbf{x}$ can move to *any* $\mathbf{y} \in \mathcal{I}_{x_0}$ with positive probability in finitely many steps. This shows *1.*

To prove the second claim of this lemma, we note that if $p(\mathbf{x}+\nu) < p(\mathbf{x})$ or $p(\mathbf{x}-\nu) < p(\mathbf{x})$, then the 1-step transition probability of starting from $\mathbf{x}$ and staying is non-zero. Indeed,

$$\mathcal{K}(\mathbf{x}, \{\mathbf{x}\}) \;=\; 1 - \mathcal{K}(\mathbf{x}, \{\mathbf{x}+\nu\}) - \mathcal{K}(\mathbf{x}, \{\mathbf{x}-\nu\}) \;=\; 1 - \frac{1}{2}\alpha(\mathbf{x}, \mathbf{x}+\nu) - \frac{1}{2}\alpha(\mathbf{x}, \mathbf{x}-\nu) \;.$$

Since $p(\mathbf{x}+\nu) < p(\mathbf{x})$, we have $\alpha(\mathbf{x}, \mathbf{x}+\nu) = \min(1, p(\mathbf{x}+\nu)/p(\mathbf{x})) < 1$. Similarly, $\alpha(\mathbf{x}, \mathbf{x}-\nu) < 1$. Hence, $\mathcal{K}(\mathbf{x}, \{\mathbf{x}\}) > 0$, thus $\mathbf{x}$ is aperiodic. □

Combining Lemma 4.9 and 4.10, we can identify the limiting distribution when the initial distribution is a point mass at $\mathbf{x} \in \mathcal{I}_{\mathbf{x}_0}$.

Proposition 4.11. *If $p(\mathbf{x}+\nu) < p(\mathbf{x})$ or $p(\mathbf{x}-\nu) < p(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{I}_0$, then $r_{\mathbf{x}_0}$ is also the unique limiting distribution, i.e., $\mathcal{K}^T(\mathbf{x}, A) \to \sum_{\mathbf{x}'\in A} r_{\mathbf{x}_0}(\mathbf{x}')$, for all $\mathbf{x} \in \mathcal{I}_{\mathbf{x}_0}$ and $A \subset \mathcal{I}_{\mathbf{x}_0}$. Furthermore, for a Lebesgue-measurable set $A \subset \mathcal{X}$ and a state $\mathbf{x} \in \mathcal{X}$,*

$$\lim_{n\to\infty} \mathcal{K}^n(\mathbf{x}, A) \;=\; \sum_{\mathbf{x}'\in\mathcal{I}_{\mathbf{x}}} \delta_{\mathbf{x}'}(A) \frac{p(\mathbf{x}')}{\sum_{k\in\mathbb{Z}} p(\mathbf{x}+k\nu)} \;. \tag{4.23}$$

Proof of Proposition 4.11. Under the stated assumption, Lemma 4.10 shows that the Markov chain is irreducible and aperiodic on $\mathcal{I}_{x_0}$. Since the stationary distribution of an irreducible and aperiodic

Markov chain defined on a countable space is also the unique limiting distribution (e.g., Meyn and Tweedie 1993), the first part follows. (4.23) holds because, for any $\mathbf{x}$, the function $r_{\mathbf{x}}$ is a probability mass function taking zero values outside of $\mathcal{I}_{\mathbf{x}}$. □

4.A.5.2 Limiting Distribution with a General Initial Distribution

We now prove Proposition 4.5, which characterizes the limiting distribution with a general initial distribution Q.

Proof of Proposition 4.5. Let $A \subset \mathcal{X}$ be Lebesgue-measurable. For any fixed $n \in \mathbb{Z}_+$, the probability of A under the n-step perturbed distribution is

$$(\mathcal{K}^n Q)(A) \;=\; \int_{\mathbf{x}\in\mathcal{X}} \mathcal{K}^n(\mathbf{x}, A) Q(\mathrm{d}\mathbf{x}) \;=\; \int_{\mathbf{x}\in\mathcal{X}} \mathcal{K}^n(\mathbf{x}, A) q(\mathbf{x})\,\mathrm{d}\mathbf{x}\,.$$

Since $|\mathcal{K}^n(\mathbf{x}, A)| \le |\mathcal{K}^n(\mathbf{x}, \mathcal{I}_{\mathbf{x}})| \le 1$ and $\int_{\mathbf{x}\in\mathcal{X}} q(\mathbf{x})\,\mathrm{d}\mathbf{x} = 1 < \infty$, we can apply Dominated Convergence Theorem (Bartle, 1995, Section 5.6) to conclude

$$\lim_{n\to\infty} (\mathcal{K}^n Q)(A) \;=\; \int_{\mathbf{x}\in\mathcal{X}} \mathcal{K}^\infty(\mathbf{x}, A) q(\mathbf{x})\,\mathrm{d}\mathbf{x} \;=\; \int_{\mathbf{x}\in\mathcal{X}} \sum_{\mathbf{x}'\in\mathcal{I}_{\mathbf{x}}} \delta_{\mathbf{x}'}(A) \frac{p(\mathbf{x}')}{\sum_{k\in\mathbb{Z}} p(\mathbf{x}+k\nu)} q(\mathbf{x})\,\mathrm{d}\mathbf{x}\,,$$

where the second equality holds by applying Proposition 4.11. Substituting the definition of $\mathcal{I}_{\mathbf{x}}$, the RHS of the above equality can be written as

$$\begin{aligned}
\int_{\mathbf{x}\in\mathcal{X}} \sum_{\mathbf{x}'\in\mathcal{I}_{\mathbf{x}}} \delta_{\mathbf{x}'}(A) \frac{p(\mathbf{x}')}{\sum_{k\in\mathbb{Z}} p(\mathbf{x}+k\nu)} q(\mathbf{x})\,\mathrm{d}\mathbf{x} \;&=\; \int_{\mathbf{x}\in\mathcal{X}} \sum_{s\in\mathbb{Z}} \delta_{\mathbf{x}+s\nu}(A) \frac{p(\mathbf{x}+s\nu)}{\sum_{k\in\mathbb{Z}} p(\mathbf{x}+k\nu)} q(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&\overset{(a)}{=} \sum_{s\in\mathbb{Z}} \int_{\mathbf{x}\in\mathcal{X}} \delta_{\mathbf{x}+s\nu}(A) \frac{p(\mathbf{x}+s\nu)}{\sum_{k\in\mathbb{Z}} p(\mathbf{x}+k\nu)} q(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&= \sum_{s\in\mathbb{Z}} \int_{\mathbf{u}\in\mathcal{X}} \delta_{\mathbf{u}}(A) \frac{p(\mathbf{u})}{\sum_{k\in\mathbb{Z}} p(\mathbf{w}+(k-s)\nu)} q(\mathbf{u}-s\nu)\,\mathrm{d}\mathbf{u} \\
&=: T\,, \qquad (4.24)
\end{aligned}$$

where (a) holds by applying Fubini-Toneli Theorem (Bartle, 1995, Section 10.9, 10.10) to exchange the integral and the infinite sum, and the last line follows from a change of variable $\mathbf{u} := \mathbf{x} + s\nu$ for a given s. In particular, we can apply Fubini-Toneli Theorem as $\int_{\mathbf{x}\in\mathcal{X}} \sum_{s\in\mathbb{Z}} \left|\delta_{\mathbf{x}+s\nu}(A) \frac{p(\mathbf{x}+s\nu)}{\sum_{k\in\mathbb{Z}} p(\mathbf{x}+k\nu)}\right| q(\mathbf{x})\,\mathrm{d}\mathbf{x} \le \int_{\mathbf{x}\in\mathcal{X}} \sum_{s\in\mathbb{Z}} \frac{p(\mathbf{x}+s\nu)}{\sum_{k\in\mathbb{Z}} p(\mathbf{x}+k\nu)} q(\mathbf{x})\,\mathrm{d}\mathbf{x} < \infty$. Applying Fubini-Toneli Theorem again, the term T defined in (4.24) can be written as

$$\begin{aligned}
T \;&=\; \int_{\mathbf{u}\in\mathcal{X}} \sum_{s\in\mathbb{Z}} \delta_{\mathbf{u}}(A) \frac{p(\mathbf{u})}{\sum_{k\in\mathbb{Z}} p(\mathbf{w}+(k-s)\nu)} q(\mathbf{u}-s\nu)\,\mathrm{d}\mathbf{u} \\
&= \int_{\mathbf{u}\in\mathcal{X}} \delta_{\mathbf{u}}(A) \frac{p(\mathbf{u})}{\sum_{k\in\mathbb{Z}} p(\mathbf{w}+(k-s)\nu)} \sum_{s\in\mathbb{Z}} q(\mathbf{u}-s\nu)\,\mathrm{d}\mathbf{u} \\
&= \int_{\mathbf{u}\in A} p(\mathbf{u}) \frac{\sum_{s\in\mathbb{Z}} q(\mathbf{u}+s\nu)}{\sum_{k\in\mathbb{Z}} p(\mathbf{u}+k\nu)}\,\mathrm{d}\mathbf{u}\,,
\end{aligned}$$

where the last line holds by a re-indexing of the sums on the numerator and on the denominator. □

4.B Details on Sampling from Gaussian-Bernoulli RBMs

We discuss practical considerations when sampling from the Gaussian-Bernoulli RBM with our particular choice of B (the *multi-modal* setting). Denote by GB-RBM$(B, \mathbf{b}, \mathbf{c})$ the joint density of the Gaussian-Bernoulli RBM with parameters $B, \mathbf{b}, \mathbf{c}$ (see the previous section for definition).

In the *multi-modal* setting, we set $\mathbf{c} = (6, 6, 0, \ldots, 0) \in \mathbb{R}^{d_h}$, and $B = \lambda E$, where λ is a large, positive constant, and $E \in \mathbb{R}^{d \times d_h}$ is the top $d \times d_h$ sub-matrix of $I_{d_{\max}}$, where $d_{\max} \coloneqq \max(d, d_h)$. These lead to a model that is a mixture of Gaussian distributions with *distantly* located modes. More specifically, the modes are located at the corners of the hyper-cube in $\mathbb{R}^d$ with length 6.

The standard way to sample from a GB-RBM is to use a Gibbs sampler (see, e.g., Melchior et al. 2017; Jitkrittum et al. 2017). However, the Gibbs sampling can suffer from poor mixing when λ is large. This is because the modes will become more disconnected as λ increases, thus making it more challenging for the sampler to learn the mixing weights correctly. This means an impractically long burn-in period would be required for the Gibbs sampler to produce faithful samples from the ground truth.

We propose a practical method to generate faithful samples when λ is large. It is based on the observation that, with B of the above form, varying the value of λ only affects the mean locations of the Gaussian components in the GB-RBM, but *not* the mixing ratios. Indeed, for any $\mathbf{h} \in \{-1, 1\}^{d_h}$, we have $\|Bh\|_2^2 = \lambda^2 \|Eh\|_2^2 = \lambda^2 d_h$, which is constant for all $\mathbf{h}$. Substituting this into $\gamma(\mathbf{h})$ of (4.12),

$$\gamma(\mathbf{h}) \;=\; \frac{\exp\left(\frac{1}{8}\lambda^2 d_h + \mathbf{c}^\top \mathbf{h}\right)}{\sum_{\mathbf{h}'} \exp\left(\frac{1}{8}\lambda^2 d_h + \mathbf{c}^\top \mathbf{h}'\right)} \;=\; \frac{\exp\left(\mathbf{c}^\top \mathbf{h}\right)}{\sum_{\mathbf{h}'} \exp\left(\mathbf{c}^\top \mathbf{h}'\right)} \,,$$

which does not depend on λ. It follows that, given any $\lambda, \lambda' \geq 0$, if $(\mathbf{x}, \mathbf{h}) \sim$ GB-RBM$(\lambda E, \mathbf{b}, \mathbf{c})$, then $(\mathbf{y}, \mathbf{h}) \sim$ GB-RBM$(\lambda' E, \mathbf{b}, \mathbf{c})$, where $\mathbf{y} \coloneqq \mathbf{x} - \frac{1}{2}(\lambda - \lambda')E\mathbf{h}$. This implies that we can sample from GB-RBM$(\lambda E, \mathbf{b}, \mathbf{c})$ with a large λ by the following procedure:

1. Use a Gibbs sampler to sample $(\mathbf{x}, \mathbf{h}) \sim$ GB-RBM$(\lambda' E, \mathbf{b}, \mathbf{c})$, where λ' is small.
2. Set $\mathbf{y} = \mathbf{x} - \frac{1}{2}(\lambda' - \lambda)E\mathbf{h}$.

Therefore, assuming the Gibbs sampler is capable of generating faithful samples from GB-RBM$(\lambda' E, \mathbf{b}, \mathbf{c})$ for some $\lambda' \geq 0$, this will produce faithful samples from GB-RBM$(\lambda E, \mathbf{b}, \mathbf{c})$ for any $\lambda \geq 0$ large. In our experiments, we used $\lambda' = 0$.

Chapter 5

Goodness-of-Fit Testing Part II: On the Robustness of KSD Tests

Goodness-of-fit testing is often criticized for its lack of practical relevance; since "all models are wrong", the null hypothesis that the data conform to our model is ultimately always rejected when the sample size is large enough. Despite this, probabilistic models are still used extensively, raising the more pertinent question of whether the model is good enough for a specific task. This question can be formalized as a robust goodness-of-fit testing problem by asking whether the data were generated by a distribution corresponding to our model up to some mild perturbation. In this chapter, we show that kernel goodness-of-fit tests based on KSD are not robust according to common notions of robustness including qualitative and quantitative robustness. We also show that robust techniques based on tilted kernels from the parameter estimation literature are not sufficient for ensuring both types of robustness in the context of goodness-of-fit testing. We therefore propose the first robust kernel goodness-of-fit test which resolves this open problem using kernel Stein discrepancy balls, which encompass perturbation models such as Huber contamination models and density uncertainty bands.

5.1 Introduction

Goodness-of-fit (GOF) testing (D'Agostino, Ralph B, 2017; Lehmann and Romano, 2022, Chapter 16) tackles the question of how well a given probabilistic model describes some observed data. More formally, given a model P and observations drawn independently from some distribution Q, GOF testing compares the null hypothesis $H_0 : Q = P$ against the alternative hypothesis $H_1 : Q \neq P$. This process is fundamental to detecting clear violation of model assumptions before a model is used for predictions, decision-making, or providing probabilistic guarantees on important quantities of interest.

This chapter continues our discussion on kernel GOF tests, namely the KSD tests (Chwialkowski et al., 2016; Liu et al., 2016). Unlike most classical alternatives such as likelihood-ratio tests (Hogg et al., 1977, Chapter 8.7) and Kolmogorov-Smirnov tests (Kolmogorov, 1933), KSD tests can be used for models for which we do not have a closed-form expression of the density or distribution function. This

is a significant advantage since it enables their use for models which are beyond the reach of other tests, including many modern flexible density estimation models, energy-based models, large graphical models, and Bayesian posteriors. Kernel-based GOF tests achieve this through a tractable test statistic based on the *score function* (i.e., the gradient of the log-density), which is widely available since it can be evaluated without knowledge of the normalization constant of the density. This property has made KSD tests popular and has led to numerous extensions specializing the approach to problems from survival analysis (Fernandez et al., 2020) to inference with manifold data (Xu and Matsuda, 2020), graphs (Xu and Reinert, 2021) and protein data (Amin et al., 2023).

A significant drawback of GOF testing is the following conundrum: in almost all real-world applications, models are wrong, leading GOF tests to ultimately reject the null hypothesis as the sample size grows. However, there is often a difference between statistical significance from a GOF perspective, and practical relevance of the model for downstream tasks. For instance, *data corruption* may arise in signal processing due to sensor failures (Rizzoni and Min, 1991; Sharma et al., 2010), in classification tasks due to mislabelled entities (Frénay and Verleysen, 2013), and in radio systems due to impulsive noise (Blackard et al., 1993). In these cases, H_0 may be rejected even though P is *almost* correct up to these mild perturbations, and as such could still be a useful model for predictions or providing probabilistic statements on quantities of interest. This conundrum has led to the development of *robust tests* (Huber, 1965; LeCam, 1973; Dabak and Johnson, 1994; Fauß et al., 2021), which aim to control the Type-I error rate (the probability of rejecting the null hypothesis) under a *composite* null hypothesis $H_0^{\mathrm{C}} : Q \in \mathcal{P}_0$, where $\mathcal{P}_0$ is a family of probability distributions called the uncertainty set (Fauß et al., 2021). This uncertainty set is constructed to include P and distributions close to P in some appropriate notion of similarity. By designing $\mathcal{P}_0$ to include potential contaminations, such as adversarial contaminations or outliers, the composite null H_0^{C} can now hold in realistic scenarios, thus avoiding the aforementioned conundrum and thereby aligning statistical and practical significance.

Unfortunately, robust kernel GOF tests are absent from the existing literature on kernel tests. The closest work is that of Sun and Zou (2023); Schrab and Kim (2024), who proposed robust tests using uncertainty sets defined by the Maximum Mean Discrepancy (MMD; Gretton et al. 2012). These tests require samples from *both* Q and P and are thus applicable when simulating from P is feasible (see, e.g., Lloyd and Ghahramani 2015), but this would likely lead to significant additional computational cost and a lack of statistical efficiency due to the extra error induced by simulations.

Instead, we propose a novel class of robust kernel GOF tests based on the kernel Stein discrepancy (KSD). These tests are appropriate in the *one-sample* setting where only samples from Q are available, and where only an unnormalized density of P is available. To guarantee robustness to mild perturbations, we make use of tilted kernels proposed in a related construction for robust parameter estimation in Barp et al. (2019); Matsubara et al. (2022). We then study qualitative and quantitative robustness, two notions of robustness which concern the insensitivity of a test to perturbations in a neighborhood of P and which are formally introduced in Section 5.2.3.

Contributions Our contributions are summarized in Table 5.1 and described as follows:

1. In Section 5.3, we study the robustness properties of existing KSD tests. We show that KSD tests

Table 5.1: Summary of theoretical contributions to the robustness of kernel GOF tests provided in this chapter.

Goodness-of-Fit Test	Qualitative robustness	Quantitative Robustness	Result
Existing KSD test + stationary kernels	✗	✗	Theorem 5.1
Existing KSD test + tilted kernels	✓	✗	Theorem 5.3
Novel KSD tests proposed in this chapter	✓	✓	Theorem 5.4

with stationary kernels are *not* robust both under infinitesimally small model deviation (which we call *qualitative robustness*) and against a fixed class of misspecified models (called *quantitative robustness*). We then show that qualitative robustness can be guaranteed by using appropriately tilted kernels, but this is *not* sufficient to guarantee quantitative robustness.

2. In Section 5.4, we propose a kernel GOF test that controls the Type-I error rate against all distributions in a KSD-ball of radius $\theta > 0$ centered at P. This test is straightforward to implement since it is identical to existing KSD tests up to a shift of the test statistic by θ. We then make recommendation on how θ should be selected to ensure common perturbations such as Huber contaminations (which includes outliers) or density-band contaminations are contained within the KSD-ball.

Outline Section 5.2 reviews the necessary background, including kernel GOF testing based on KSD and robust GOF testing. Section 5.3 contains our main theoretical contributions concerning the (lack of) robustness of existing kernel GOF tests. A novel robust kernel GOF test is proposed in Section 5.4. Numerical experiments are provided in Section 5.5. Section 5.6 concludes and discusses open questions.

5.2 Background

We now review the background that is essential for understanding the contributions of this work, including the kernel Stein discrepancy and robust GOF testing. We begin by briefly summarizing the notation for this chapter.

Notation Let $\mathcal{P}(\mathbb{R}^d)$ denote the set of probability measures on $\mathbb{R}^d$. Given an integer $r > 0$, $\mathcal{C}^r$ is the set of functions $f : \mathbb{R}^d \to \mathbb{R}$ that are r-times continuously differentiable. The gradient of $f \in \mathcal{C}^1$ is denoted $\nabla f(\mathbf{x}) = (\partial_1 f\mathbf{x}), \ldots, \partial_d f(\mathbf{x}))^\top$. The set $\mathcal{C}_b^1$ contains functions in $\mathcal{C}^1$ that are bounded and with bounded gradient, and the set $\mathcal{C}_b^2$ contains $f \in \mathcal{C}_b^1$ for which $\sup_{\mathbf{x}\in\mathbb{R}^d} |\nabla^\top \nabla f(\mathbf{x})| < \infty$, where $\nabla^\top \nabla f(\mathbf{x}) = \sum_{j=1}^d \partial_j \partial_j f(\mathbf{x})$. For a function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with two arguments, $\nabla_i k(\mathbf{x}, \mathbf{x}')$ will be used to denote its gradient with respect to the i-th argument respectively for $i = 1, 2$, and we define $\nabla_1^\top \nabla_2 k(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^d \partial_{1j} \partial_{2j} k(\mathbf{x}, \mathbf{x}')$, where $\partial_{ij} k(\mathbf{x}, \mathbf{x}') = [\nabla_i k(\mathbf{x}, \mathbf{x}')]_j$ for $i = 1, 2$ and $j = 1, \ldots, d$. The set $\mathcal{C}^{(1,1)}$ (resp., $\mathcal{C}_b^{(1,1)}$) denotes all functions $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with $(\mathbf{x}, \mathbf{x}') \mapsto \partial_{1j} \partial_{2j} k(\mathbf{x}, \mathbf{x}')$ continuous (resp., continuous and bounded) for all $j = 1, \ldots, d$.

5.2.1 The Kernel Stein Discrepancy

A key ingredient for kernel GOF tests is the kernel Stein discrepancy (KSD, Chwialkowski et al. 2016; Liu et al. 2016; Gorham and Mackey 2017; Oates et al. 2017). A detailed introduction to KSD was given in Section 2.2; here, we provide a brief summary.

Throughout, we assume $P, Q \in \mathcal{P}(\mathbb{R}^d)$, and P admits a Lebesgue density function $p \in \mathcal{C}^1$ such that $p > 0$ on $\mathbb{R}^d$. Let $k \in \mathcal{C}^{(1,1)}$ be a scalar-valued reproducing kernel with associated reproducing kernel Hilbert space (RKHS, Berlinet and Thomas-Agnan, 2004) denoted by $\mathcal{H}_k$. The *Langevin KSD* gives a notion of discrepancy between Q and P and is defined as

$$D(Q, P) \coloneqq \sup_{f \in \mathcal{B}} \left| \mathbb{E}_{\mathbf{X} \sim Q}[(\mathcal{A}_p f)(\mathbf{X})] \right| = \sup_{f \in \mathcal{B}} \left| \int_{\mathbb{R}^d} (\mathcal{A}_p f)(\mathbf{x}) Q(\mathrm{d}\mathbf{x}) \right| ,$$

where $\mathcal{B} \coloneqq \{h = (h_1, \ldots, h_d) : \ h_j \in \mathcal{H}_k \text{ and } \sum_{j=1}^d \|h_j\|_{\mathcal{H}_k}^2 \leq 1\}$ is the unit ball in the d-times Cartesian product $\mathcal{H}_k^d$ of the RKHS $\mathcal{H}_k$. The operator $\mathcal{A}_p$ is a linear operator called the *Langevin Stein operator* and it maps (suitably regular) vector-valued functions f to scalar-valued ones via $(\mathcal{A}_p f)(\mathbf{x}) \coloneqq \mathbf{s}_p(\mathbf{x})^\top f(\mathbf{x}) + \nabla^\top f(\mathbf{x})$, where $\mathbf{s}_p(\mathbf{x}) \coloneqq \nabla_{\mathbf{x}} \log p(\mathbf{x})$ is known as the *(Hyvärinen) score function* of P (Hyvärinen, 2005).

When $\mathbb{E}_{\mathbf{X} \sim P}[\|\mathbf{s}_p\|_2] < \infty$, the *squared* KSD admits the following double-integral form (Barp et al., 2022, Corollary 2.10; Gorham and Mackey, 2017)

$$\begin{aligned} D^2(Q, P) &= \mathbb{E}_{\mathbf{X}, \mathbf{X}' \sim Q}[u_P(\mathbf{X}, \mathbf{X}')] = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} u_P(\mathbf{x}, \mathbf{x}') Q(\mathrm{d}\mathbf{x}) Q(\mathrm{d}\mathbf{x}') , \\ u_P(\mathbf{x}, \mathbf{x}') &\coloneqq \mathbf{s}_p(\mathbf{x})^\top \mathbf{s}_p(\mathbf{x}') k(\mathbf{x}, \mathbf{x}') + \mathbf{s}_p(\mathbf{x})^\top \nabla_2 k(\mathbf{x}, \mathbf{x}') + \nabla_1 k(\mathbf{x}, \mathbf{x}')^\top \mathbf{s}_p(\mathbf{x}') + \nabla_1^\top \nabla_2 k(\mathbf{x}, \mathbf{x}') , \end{aligned} \tag{5.1}$$

whenever $D^2(Q, P)$ is well-defined. The function u_P is also a reproducing kernel, known as the *Stein (reproducing) kernel*. The main advantage of the KSD is that it is computable even if p is unnormalized. Indeed, u_P depends on p only through $\mathbf{s}_p$, which does not depend on the normalizing constant of p (as the constant is cancelled due to the differentiation). The KSD can be straightforwardly estimated at $\mathcal{O}(n^2)$ cost using a V-statistic estimator (5.2), which is non-negative and biased but consistent. It is formed by replacing Q with its empirical version Q_n based on independent observations $\mathbb{X}_n \coloneqq \{\mathbf{X}_i\}_{i=1}^n \sim Q$:

$$D^2(\mathbb{X}_n) \coloneqq D^2(Q_n, P) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n u_P(\mathbf{X}_i, \mathbf{X}_j) , \tag{5.2}$$

where we have overloaded the notation by writing $D^2 : (\mathbb{R}^d)^n \to [0, \infty)$ as a test statistic. An alternative estimator is the U-statistic (2.11), which differs from (5.2) by summing only over *disjoint* pairs $i \neq j$. Although the U-statistic is unbiased, it can take negative values. Therefore, in this chapter, we focus on V-statistics, leaving a discussion on how to extend our results to U-statistics to Section 5.C.

When the kernel k is *characteristic*, the KSD is *P-separating*, meaning that $D(Q, P) \geq 0$ with equality if and only if $Q = P$ for all Q that finitely integrates $\|\mathbf{s}_p\|_2$; see, e.g., Barp et al. (2022, Theorem 3.2). Most characteristic kernels used in kernel GOF tests are sufficiently smooth stationary kernels of the form $k(\mathbf{x}, \mathbf{x}') = h(\mathbf{x} - \mathbf{x}')$, where $h \in \mathcal{C}_b^2$ and $h(0) > 0$. These include the squared-

exponential kernel $k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2/(2\lambda^2))$ and the inverse multi-quadric (IMQ) kernel $k(\mathbf{x}, \mathbf{x}') = (1 + \|\mathbf{x} - \mathbf{x}'\|_2^2/\lambda^2)^{-b}$, where $b > 0$, and $\lambda > 0$ is a hyperparameter known as the *bandwidth*. The IMQ kernel is often preferred in practice since the resulting KSD has the desirable property of *P-convergence control* (Gorham and Mackey, 2017; Barp et al., 2022). The bandwidth λ also plays a crucial role in the performance of kernel-based tests (Reddi et al., 2015; Ramdas et al., 2015; Huang et al., 2023), and a *median-heuristic* (Fukumizu et al., 2009, Section 5) is often used to select its value in practice (Liu et al., 2016; Chwialkowski et al., 2016).

5.2.2 Kernel Goodness-of-Fit Testing

The KSD provides a natural quantity to construct a test statistic since, by the P-separation property of KSD, testing $H_0 : Q = P$ is equivalent to testing whether $D(Q, P) = 0$. Since the KSD is non-negative, A GOF test can therefore be constructed whereby H_0 is rejected for large values of the KSD estimate (Chwialkowski et al., 2016; Liu et al., 2016). KSD tests have been used for GOF testing with unnormalized models in a wide range of applications and data structures (Yang et al., 2018; Fernandez et al., 2020; Xu and Reinert, 2021; Xu and Matsuda, 2021; Amin et al., 2023). Extensions have also been developed to address its limitations, such as high computational cost (Jitkrittum et al., 2017; Huggins and Mackey, 2018), difficulty in bandwidth selection (Schrab et al., 2022) and lack of test power against certain alternatives (Liu et al., 2023).

A significant practical challenge with KSD tests is determining an appropriate decision threshold. A valid threshold can be obtained by considering the quantiles of the null distribution of $D^2(\mathbb{X}_n)$, but since this distribution is not usually tractable, a bootstrap procedure is often employed to estimate it. This is done by computing the empirical quantiles based on bootstrap samples of the form

$$D^2_{\mathbf{W}}(\mathbb{X}_n) \;=\; \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(W_i - 1)(W_j - 1)u_P(\mathbf{X}_i, \mathbf{X}_j)\;, \tag{5.3}$$

where $\mathbf{W} \coloneqq (W_1, \ldots, W_n) \sim \text{Multinomial}(n; 1/n, \ldots, 1/n)$. This procedure, called *Efron's bootstrap* or *weighted bootstrap* (Arcones and Gine, 1992; Janssen, 1994), assigns a multinomial weight to each observation, which mimics recomputing the test statistic using data re-sampled from $\mathbb{X}_n$ without replacement. Other bootstrap methods such as wild bootstraps (Leucht and Neumann, 2013; Shao, 2010) are also viable, particularly when the samples are potentially correlated (Chwialkowski et al., 2016), but we focus on the weighted bootstrap since they perform similarly for independent samples, as shown in Section 5.B.7. The test threshold is selected as the $(1 - \alpha)$-quantile of the distribution of $D^2_{\mathbf{W}}(\mathbb{X}_n)$ conditional on $\mathbb{X}_n$, i.e.,

$$q^2_{\infty,1-\alpha}(\mathbb{X}_n) \;\coloneqq\; \inf\left\{u \in \mathbb{R} :\; 1 - \alpha \leq \Pr_{\mathbf{W}}\left(D^2_{\mathbf{W}}(\mathbb{X}_n) \leq u \mid \mathbb{X}_n\right)\right\}. \tag{5.4}$$

In practice, this quantile is approximated with Monte-Carlo estimation by first drawing B independent copies $\{\mathbf{W}^b\}_{b=1}^B$ of $\mathbf{W}$, and then computing

$$q^2_{B,1-\alpha}(\mathbb{X}_n) \;\coloneqq\; \inf\left\{u \in \mathbb{R} :\; 1 - \alpha \leq \frac{1}{B+1}\left(\mathbb{1}\{D^2(\mathbb{X}_n) \leq u\} + \sum_{b=1}^{B}\mathbb{1}\{D^2_{\mathbf{W}^b}(\mathbb{X}_n) \leq u\}\right)\right\}, \tag{5.5}$$

where $\mathbb{1}\{\mathcal{A}\}$ denotes the indicator function for the event $\mathcal{A}$. The KSD test then rejects H_0 if $D^2(\mathbb{X}_n) > q^2_{B,1-\alpha}(\mathbb{X}_n)$. This test is asymptotically well-calibrated for $H_0 : Q = P$ (Liu et al., 2016, Theorem 4.3). When KSD is P-separating, this test is also *consistent*, meaning that, whenever $Q \neq P$, the probability of rejection approaches one as the sample size $n \to \infty$ (Chwialkowski et al., 2016; Liu et al., 2016).

5.2.3 Robustness for Goodness-of-Fit Testing and Related Work

Robustness of GOF tests refers to the lack of sensitivity of the test outcome to small model deviations (Rieder, 1982; Lambert, 1982). Model deviations can be formalized as a neighborhood $\mathcal{P}_0$ around a nominal model P. The neighborhood $\mathcal{P}_0$ is often called an *uncertainty set* and encodes the practitioner's uncertainty on the model P. One popular example for $\mathcal{P}_0$ is *Huber's contamination model* (Huber, 1964, 1965):

$$\mathcal{P}(P;\epsilon) \;:=\; \{(1-\epsilon')P + \epsilon' R :\; 0 \le \epsilon' \le \epsilon,\; R \in \mathcal{P}(\mathbb{R}^d)\}\;, \tag{5.6}$$

where $\epsilon \in [0,1]$ is the maximal contamination ratio, and the probability measure R acts as arbitrary contamination. This model is appropriate when practitioners believe that P accurately describes all but a small proportion of the data. Huber's models have been studied in the context of robust GOF testing including Huber (1965); Qin and Priebe (2017), as well as in robust estimation (Hampel, 1974; Huber and Ronchetti, 2011). Special interests lie in the case when R is restricted to point masses, i.e., $R = \delta_{\mathbf{z}}$ for some $\mathbf{z} \in \mathbb{R}^d$, where robustness is often called *bias-robustness* (Huber and Ronchetti, 2011). Beyond Huber's models, $\mathcal{P}_0$ can also be chosen as *density-band models* (Kassam, 1981), which assume the density of the data-generating distribution is close to the model density up to a small error (see Section 5.4.2). Moreover, $\mathcal{P}_0$ can be set to a ball defined via a statistical divergence or metric, such as Hellinger distance (LeCam, 1973), Wasserstein metric (Gao et al., 2018), and Maximum Mean Discrepancy (Sun and Zou, 2023).

Uncertainty sets $\mathcal{P}_0$ provide a framework for assessing the robustness of a GOF test targeting the classic null $H_0 : Q = P$. By studying the probability of rejection when the data-generating distribution Q deviates from the model P but remains within $\mathcal{P}_0$, we can gauge the test's robustness. In the robust testing literature, a common approach is to consider a sequence $\{\mathcal{P}_0^n\}_{n=1}^\infty$ of uncertainty sets with *shrinking* size as the sample size n increases. This is formalized in the following notion of *qualitative robustness*, inspired by Rieder (1982).

Definition 5.1 (Qualitative robustness to a sequence of neighborhood)**.** *Let $\mathcal{P}_0^1 \supseteq \mathcal{P}_0^2 \supseteq \ldots$ be a sequence of subsets of $\mathcal{P}(\mathbb{R}^d)$ that contain P and such that $\cap_{n=1}^\infty \mathcal{P}_0^n = \{P\}$. For any positive integer n, let $T_n : (\mathbb{R}^d)^n \to \mathbb{R}$ be a test statistic where a large value of T_n suggests deviation from the null $H_0 : Q = P$, and let $\gamma_n : (\mathbb{R}^d)^n \to \mathbb{R}$ be a function that computes the decision threshold. A sequence of hypothesis tests (indexed by n) that reject H_0 when $T_n > \gamma_n$ is* qualitatively robust *to $\{\mathcal{P}_0^n\}_{n=1}^\infty$ if, as $n \to \infty$,*

$$\sup_{Q \in \mathcal{P}_0^n} \left| \Pr_{\mathbb{X}_n \sim Q}\left(T_n(\mathbb{X}_n) > \gamma_n(\mathbb{X}_n)\right) - \Pr_{\mathbb{X}_n^* \sim P}\left(T_n(\mathbb{X}_n^*) > \gamma_n(\mathbb{X}_n^*)\right)\right| \;\to\; 0\;. \tag{5.7}$$

Intuitively, this notion of robustness asks how sensitive the test outcome is under infinitesimally small

model deviation. The shrinking size condition ensures that the probability of rejection under any $Q \neq P$ does not trivially approach one due to the consistency of the test. By choosing $\mathcal{P}_0^n$ to be Prokhorov balls (Prokhorov, 1956) in Definition 5.1, we recover the qualitative robustness introduced in Rieder (1982, Definition 2.1), which parallels the conventional notion of qualitative robustness for estimators as argued in (Rieder, 1982, Remark 2). However, distributions in a Prokhorov ball do not have a simple form, posing challenges to the analysis. Our definition extends it to a general sequence of neighborhood. In Section 5.3, we will choose $\mathcal{P}_0^n$ to be Huber's contamination models, which both significantly simplifies the analysis and encompasses a wide range of relevant scenarios. We will show that, within Huber's neighborhood, the standard KSD test is *not* qualitatively robust with stationary kernels, but is qualitatively robust with appropriately *tilted* kernels.

However, qualitative robustness on its own has limitations. It only concerns the insensitivity of the test to sufficiently small neighborhoods, but does not provide guarantees for neighborhoods of a *fixed* size. The latter scenario is more pertinent, as practitioners often need to specify a certain form of model misspecification and require the test to remain well-calibrated against such misspecified models. This can be formalized by relaxing the point null hypothesis H_0 to a composite hypothesis $H_0^{\mathrm{C}} : Q \in \mathcal{P}_0$ for some uncertainty set $\mathcal{P}_0$, and requiring calibration against H_0^{C}. This leads to the following notion of *quantitative robustness*.

Definition 5.2 (Quantitative robustness to a neighborhood)**.** *Given $\alpha \in (0,1)$ and $\mathcal{P}_0 \subset \mathcal{P}(\mathbb{R}^d)$, a test is* quantitatively robust to $\mathcal{P}_0$ *at level α if its rejection probability under* any $Q \in \mathcal{P}_0$ *does not exceeds α.*

Quantitatively robust GOF tests have been developed for various types of uncertainty sets $\mathcal{P}_0$, including Huber's contamination model (Huber, 1965) and neighborhoods defined by Kullback-Leibler divergence (Levy, 2008; Yang and Chen, 2018) or α-divergence (Gül and Zoubir, 2016). These tests enjoy minimax optimality but require the normalizing constant of P to be known, thus not applicable to unnormalized models. Two-sample tests that are quantitatively robust to Hellinger distance (LeCam, 1973), Wasserstein distance (Gao et al., 2018), and Maximum Mean Discrepancy (Gao et al., 2018; Sun and Zou, 2023) have also been proposed. However, to be used for GOF testing, they require approximation of P by finite samples, which incurs extra error and can be computationally demanding due to the non-trivial task of sampling from P. In the remainder of this chapter, we will choose $\mathcal{P}_0$ as a *KSD ball* centered at P. We will show how this choice naturally lends itself to a GOF test that is both easy to implement for unnormalized models and provably robust to various types of contamination.

5.3 The (Lack of) Robustness of Existing Kernel Goodness-of-fit Tests

We will now study the robustness of standard KSD tests using stationary kernels in Section 5.3.1, then study their tilted counterparts which are popular in the parameter estimation literature in Section 5.3.2.

5.3.1 Existing KSD Tests with Stationary Kernels are not Qualitatively Robust

Our first result states that contamination of Huber's type can considerably affect the probability of the standard KSD test rejecting the null hypothesis $H_0 : Q = P$. Our result holds with the bootstrap

threshold defined in (5.4), and all probabilities are taken over the randomness of both the sample and the bootstrap weights $\mathbf{W} \sim \text{Multinomial}(n; 1/n, \dots, 1/n)$. The proof is in Section 5.A.1.

Theorem 5.1. *Assume* $\mathbb{E}_{\mathbf{X}\sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2^4] < \infty$, *the function* $\mathbf{x} \mapsto \|\mathbf{s}_p(\mathbf{x})\|_2$ *is unbounded,* $k(\mathbf{x}, \mathbf{x}') = h(\mathbf{x} - \mathbf{x}')$ *with* $h \in \mathcal{C}_b^2$ *and* $h(0) > 0$, *and assume the integrability conditions:*

$$\sup_{\mathbf{z}\in\mathbb{R}^d} \|\mathbf{s}_p(\mathbf{z})\|_2^4 \mathbb{E}_{\mathbf{X}\sim P}\left[h(\mathbf{X} - \mathbf{z})^4\right] < \infty\,, \quad \textit{and} \quad \sup_{\mathbf{z}\in\mathbb{R}^d} \|\mathbf{s}_p(\mathbf{z})\|_2^2 \mathbb{E}_{\mathbf{X}\sim P}\left[\|\nabla h(\mathbf{x} - \mathbf{z})\|_2^2\right] < \infty\,.$$

Then, for any test level $\alpha \in (0, 1)$ *and any sequence* $\{\epsilon_n\}_{n=1}^{\infty}$ *with* $\epsilon_n^{-1} = o(n)$, *the following holds as* $n \to \infty$,

$$\sup_{Q\in\mathcal{P}(P;\epsilon_n)} \left| \Pr_{\mathbb{X}_n\sim Q,\mathbf{W}}\left(D^2(\mathbb{X}_n) > q^2_{\infty,1-\alpha}(\mathbb{X}_n)\right) - \Pr_{\mathbb{X}_n^*\sim P,\mathbf{W}}\left(D^2(\mathbb{X}_n^*) > q^2_{\infty,1-\alpha}(\mathbb{X}_n^*)\right)\right| \to 1 - \alpha\,.$$

Theorem 5.1 immediately implies that the standard KSD test is *not* qualitatively robust to any sequence of Huber's models for which $\epsilon_n n$ grows with n, where $\epsilon_n n$ is the expected number of contamination in a sample of size n. At first glance, the lack of qualitative robustness might seem trivial given that the KSD test is consistent, meaning that the rejection probability under any $Q \neq P$ approaches one as $n \to \infty$. However, one subtlety is that the ball $\mathcal{P}(P; \epsilon_n)$ can also shrink in size, since ϵ_n is allowed to decay. The rate condition on ϵ_n imposes a lower bound on this decay rate. Intuitively, it implies that $\epsilon_n n$, the expected number of contamination in the sample, grows with n. In particular, this excludes the case where $\epsilon_n n$ is a constant. Moreover, since Huber's models $\mathcal{P}(P; \epsilon_n)$ are contained within Prokhorov balls, Theorem 5.1 immediately implies non-qualitative robustness to Prokhorov neighborhoods.

The moment conditions in Theorem 5.1 are mild and are satisfied by many thin-tail distributions and standard kernel functions. One example where Theorem 5.1 applies is when $P = \mathcal{N}(\mu, \Sigma)$ is a Gaussian with mean $\mu \in \mathbb{R}^d$, covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$, and unbounded score function $\mathbf{s}_p(\mathbf{x}) = -\Sigma^{-1}(\mathbf{x} - \mu)$, and k is a squared-exponential or IMQ kernel. In this case, it is straightforward to verify that the moment assumptions in Theorem 5.1 hold. We provide numerical evidence in Section 5.5.1.

Remark 5.1.1 (Unbounded Stein kernel)**.** *Theorem 5.1 is a consequence of the fact that the Stein kernel evaluated at* $\mathbf{x}$, *i.e.,* $\mathbf{x} \mapsto u_P(\mathbf{x}, \mathbf{x})$, *is not bounded when the score function* $\mathbf{s}_p$ *is not bounded. This situation arises in light-tailed models such as Gaussian models, or, more generally,* strongly sub-Gaussian *models (Palmer et al., 2010) which, when* $d = 1$, *are defined as those with density* p *such that* $p(\sqrt{x})$ *is log-concave on* $(0, \infty)$. *On the other hand, this exploding-score condition excludes heavy-tailed models with bounded score functions, such as super-Laplacian or t-distributions (Gorham and Mackey, 2017; Barp et al., 2022).*

5.3.2 Existing KSD Tests with Tilted Kernels are Qualitatively Robust

Since the lack of robustness of existing KSD tests with stationary kernels arises because the Stein kernel is unbounded, a natural approach is to choose another kernel k which prevents this from happening. We now show that this can be achieved using a KSD with tilted kernel, a common approach to induce robustness for parameter estimation (Barp et al., 2019; Matsubara et al., 2022). The proof is deferred to Section 5.A.2.

Lemma 5.2 (Bounded Stein kernel). *Suppose* $k(\mathbf{x}, \mathbf{x}') = w(\mathbf{x})h(\mathbf{x} - \mathbf{x}')w(\mathbf{x}')$, *where* $h \in \mathcal{C}_b^2$ *is a stationary reproducing kernel, and the weighting function* $w \in \mathcal{C}_b^1$ *satisfies* $\sup_{\mathbf{x}\in\mathbb{R}^d} \|w(\mathbf{x})\mathbf{s}_p(\mathbf{x})\|_2 < \infty$. *Then*

$$\sup_{\mathbf{x},\mathbf{x}'\in\mathbb{R}^d} |u_P(\mathbf{x}, \mathbf{x}')| \leq \sup_{\mathbf{x}\in\mathbb{R}^d} u_P(\mathbf{x}, \mathbf{x}) = \tau_\infty < \infty.$$

Notably, the function $\mathbf{x} \mapsto u_P(\mathbf{x}, \mathbf{x})$ is non-negative since u_P is a reproducing kernel. Moreover, Barp et al. (2022, Theorem 3.11) shows that KSD with such tilted kernels is P-separating whenever $\|\mathbf{s}_p\|_2$ grows at most root-exponentially. Most stationary kernels used in KSD tests (such as the IMQ or squared-exponential kernels) satisfy the continuity condition required of h. In this construction, the weighting function w is used to counteract the growth of the score function $\mathbf{s}_p$. When the score grows polynomially like $\|\mathbf{s}_p(\mathbf{x})\|_2 = \mathcal{O}(\|\mathbf{x}\|_2^r)$ for some $r > 0$, it suffices to choose $w(\mathbf{x}) = (1 + a^2\|\mathbf{x}\|_2^2)^{-b}$ for any $a > 0$ and $b \geq 1/(2r)$. This weight function is common in the frequentist parameter estimation literature (Barp et al., 2019) and the generalized Bayesian inference literature (Matsubara et al., 2022; Altamirano et al., 2023, 2024; Duran-Martin et al., 2024).

Intuitively, w reduces KSD's capacity to detect details by down-weighting the tail regions of the density p. As we will show, this is beneficial for robustifying the KSD test, although this comes at the expense of reducing its power if detecting discrepancy in the tails is desired. This trade-off is not unique to KSD and is common in the robust estimation literature (Basu and Sarkar, 1994; Barp et al., 2019).

The following result states that, with a tilted kernel, a small proportion ϵ_n of contamination in the data has negligible influence on the outcome of the standard KSD test. The proof is provided in Section 5.A.3.

Theorem 5.3. *Assume* $\mathbb{E}_{\mathbf{X}\sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2] < \infty$ *and that* k *is a tilted kernel satisfying the conditions in Lemma 5.2. Then, for any test level* $\alpha \in (0, 1)$ *and any sequence* $\epsilon_n = o(n^{-1/2})$, *the following holds as* $n \to \infty$,

$$\sup_{Q\in\mathcal{P}(P;\epsilon_n)} \left| \Pr_{\mathbb{X}_n\sim Q,\mathbf{W}} \left(D^2(\mathbb{X}_n) > q^2_{\infty,1-\alpha}(\mathbb{X}_n)\right) - \Pr_{\mathbb{X}_n^*\sim P,\mathbf{W}} \left(D^2(\mathbb{X}_n^*) > q^2_{\infty,1-\alpha}(\mathbb{X}_n^*)\right)\right| \to 0 .$$

The LHS of the above convergence is the worst-case difference between the rejection probabilities of the standard KSD test with and without data corruptions, and $\mathcal{P}(P; \epsilon_n)$ is the Huber's contamination model defined in (5.6). This result thus immediately implies that the standard KSD test is qualitative robust to any sequence of Huber's models so long as the contamination ratio decays sufficiently fast as $\epsilon_n = o(n^{-1/2})$. The condition on ϵ_n allows the expected number of contaminated data $n\epsilon_n$ to grow with n but requires the growth rate to be slow. In particular, it is met when the expected number of contaminated data is bounded, i.e., $\epsilon_n = \mathcal{O}(n^{-1})$. This condition is not an artifact of the proof: in Section 5.B.1, we show empirically that the tilted-KSD test is no longer qualitatively robust when $\epsilon_n = n^{-r}$ for any $r \leq 1/2$.

The boundedness condition on $\|w(\mathbf{x})\mathbf{s}_p(\mathbf{x})\|_2$ bears similarity with the ones imposed for robust estimation methods based on KSD (Barp et al., 2022; Matsubara et al., 2022). For example, Barp et al. (2019, Proposition 7) shows that their KSD-based estimator is *globally bias-robust* assuming

$\mathbf{x} \mapsto \|\mathbf{s}_p(\mathbf{x})\|_2 \int_{\mathbb{R}^d} \|k(\mathbf{x}, \mathbf{x}')\mathbf{s}_p(\mathbf{x}')\|_2 Q(\mathrm{d}\mathbf{x}')$ is bounded. This condition is slightly weaker, but often hard to verify in practice. Moreover, although tilted kernels enforce *qualitative* robustness, they are *not* enough to guarantee *quantitative* robustness. Indeed, the consistency of the standard test implies that $\Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}(D^2(\mathbb{X}_n) > q^2_{\infty,1-\alpha}(\mathbb{X}_n)) \to 1$ whenever $Q \neq P$ (Liu et al. 2016, Proposition 4.2; Schrab et al. 2022, Theorem 3.3), so KSD tests with either stationary or tilted kernels are not quantitatively robust to any neighborhoods *strictly* larger than $\{P\}$. We provide numerical evidence in Section 5.5.1.

5.4 Robust Kernel Goodness-of-fit Tests for KSD-Ball Uncertainty Sets

We now propose a novel robust KSD test for the setting where the uncertainty set is a KSD-ball; i.e.,

$$H_0^{\mathrm{C}} : \; Q \in \mathcal{B}^{\mathrm{KSD}}(P; \theta) \,, \qquad H_1^{\mathrm{C}} : \; Q \notin \mathcal{B}^{\mathrm{KSD}}(P; \theta) \,, \tag{5.8}$$

where $\theta \geq 0$ and for $P \in \mathcal{P}(\mathbb{R}^d)$ with density $p \in \mathcal{C}^1$, we define $\mathcal{B}^{\mathrm{KSD}}(P; \theta) \coloneqq \{Q : \; D(Q, P) \leq \theta\}$. The advantage of using a KSD-ball as the uncertainty set is that it lends naturally to a simple, tractable test that is robust to the contamination models we reviewed in Section 5.2.3. We first describe our novel test, which we call *robust-KSD* test, and show that it is quantitatively robust to KSD balls in the sense of Definition 5.2. We then discuss in Section 5.4.2 how to choose the uncertainty radius θ to incorporate other contamination models, such as Huber's contamination models and density-band models.

5.4.1 A Robust KSD Test

Given a prescribed test level $\alpha \in (0, 1)$, our robust KSD test uses the following test statistic

$$\Delta_\theta(\mathbb{X}_n) \;\coloneqq\; \max\left(0, D(\mathbb{X}_n) - \theta\right) \,, \tag{5.9}$$

where $D(\mathbb{X}_n)$ is the square-root of the V-statistic (5.2). The test rejects H_0^{C} for large values of $\Delta_\theta(\mathbb{X}_n)$. The decision threshold should be chosen to control the Type-I error rate for all possible $Q \in \mathcal{B}^{\mathrm{KSD}}(P; \theta)$, i.e., we require $\gamma = \gamma(\mathbb{X}_n)$ so that $\sup_{Q \in \mathcal{B}^{\mathrm{KSD}}(P;\theta)} \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}(\Delta_n(\mathbb{X}_n) > \gamma) \leq \alpha$. One approach to construct γ is to use deviation inequalities for bounded functions as done in Gretton et al. (2012, Corollary 11). However, such approach tends to be overly conservative (Gretton et al., 2012). Instead, we propose a bootstrap procedure to construct a decision threshold. We will show that choosing $\gamma = q_{\infty,1-\alpha}(\mathbb{X}_n)$, the square-root of the (population) bootstrap quantile defined in (5.4), gives the desired Type-I error control asymptotically. Our robust-KSD test therefore rejects $H_0^{\mathrm{C}} : Q \in \mathcal{B}^{\mathrm{KSD}}(P; \theta)$ if $\Delta_\theta(\mathbb{X}_n) > q_{\infty,1-\alpha}(\mathbb{X}_n)$. The validity and consistency of this test is stated in the following result, proved in Section 5.A.5. In particular, it implies that our robust-KSD test is quantitatively robust to KSD balls in the infinite-sample limit.

Theorem 5.4. *Suppose* $\mathbb{E}_{\mathbf{X} \sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2] < \infty$ *and* $k(\mathbf{x}, \mathbf{x}') = w(\mathbf{x})h(\mathbf{x} - \mathbf{x}')w(\mathbf{x}')$*, with stationary reproducing kernel* $h \in \mathcal{C}_b^2$ *and weighting function* $w \in \mathcal{C}_b^1$*. Assume* $Q \in \mathcal{P}(\mathbb{R}^d)$ *with* $\mathbb{E}_{\mathbf{X} \sim Q}[\|w(\mathbf{X})\mathbf{s}_p(\mathbf{X})\|_2^4] <$

∞. *Then*

1. *Under* $H_0^{\mathrm{C}} : Q \in \mathcal{B}^{\mathrm{KSD}}(P;\theta)$, *we have* $\limsup_{n\to\infty} \Pr_{\mathbb{X}_n\sim Q,\mathbf{W}}(\Delta_\theta(\mathbb{X}_n) > q_{\infty,1-\alpha}(\mathbb{X}_n)) \le \alpha$.
2. *Under* $H_1^{\mathrm{C}} : Q \notin \mathcal{B}^{\mathrm{KSD}}(P;\theta)$, *we have* $\lim_{n\to\infty} \Pr_{\mathbb{X}_n\sim Q,\mathbf{W}}\left(\Delta_\theta(\mathbb{X}_n) > q_{\infty,1-\alpha}(\mathbb{X}_n)\right) \to 1$.

Theorem 5.4 assumes the weighting function w in the kernel is bounded with bounded gradient. This covers the case $w(x) = 1$, in which case $k(\mathbf{x},\mathbf{x}') = h(\mathbf{x}-\mathbf{x}')$ is an *untilted* stationary kernel such as IMQ or squared-exponential kernels. Theorem 5.4 also assumes the data-generating distribution Q finitely integrates the fourth moment of the weighted score, i.e., $\|w(\mathbf{x})\mathbf{s}_p(\mathbf{x})\|_2^4$. For tilted kernels satisfying the conditions in Lemma 5.2, $\|w(\mathbf{x})\mathbf{s}_p(\mathbf{x})\|_2$ is bounded, so this moment condition automatically holds for any $Q \in \mathcal{P}(\mathbb{R}^d)$.

In practice, the decision threshold $q_{\infty,1-\alpha}$ is intractable, and we again use the Monte Carlo estimate $q_{B,1-\alpha}$ defined in (5.5) as an approximation. Compared with the standard KSD test targeting the point null $H_0 : Q = P$, our robust-KSD test uses the same bootstrap procedure to compute the decision threshold, but has a slightly different test statistic $\Delta_\theta(\mathbb{X}_n)$ instead of $D^2(\mathbb{X}_n)$ (given in (5.2)) to account for the composite null. Moreover, as $\theta \to 0$, the test statistic Δ_θ approaches $D(\mathbb{X}_n)$, the square-root of the test statistic of the standard test, and the KSD-ball $\mathcal{B}^{\mathrm{KSD}}(P;\theta)$ falls to the singleton $\{P\}$, so the robust-KSD test reduces to the standard test. Our robust KSD test can hence be viewed as a generalization of the standard KSD test to the composite hypotheses in (5.8). In particular, the robust KSD test is also qualitatively robust whenever the conditions on k and $\mathbf{s}_p$ from Theorem 5.3 hold; see Proposition 5.14 in Section 5.A.3 for a formal statement and proof.

Another advantage of our robust KSD test being a direct generalization of the standard test is that no extra computation is required to guarantee robustness. For a given $\theta \ge 0$, our robust test has the same computational cost as the standard one, which scales as $\mathcal{O}(n^2 d)$, and only requires a minor transformation of the test statistic. However, extra computation is potentially needed in determining θ, which might require optimizing the Stein kernel. This is discussed in detail in Section 5.4.2.

5.4.2 Choosing the Uncertainty Radius

A crucial design choice in our robust tests is the uncertainty radius θ. This should be guided by the types of contamination that the practitioner is willing to tolerate. In this section, we discuss how to choose θ for Huber's contamination models and density-band models when using KSD balls based on tilted kernels.

Firstly, suppose we are considering Huber's contamination model $\mathcal{P}(P;\epsilon_0)$ for some $\epsilon_0 \in [0,1]$. Then θ should be chosen such that the KSD-ball $\mathcal{B}^{\mathrm{KSD}}(P;\theta)$ contains $\mathcal{P}(P;\epsilon_0)$. The following result, proved in Section 5.A.6, shows how to achieve this given an upper bound on the Stein kernel $\tau_\infty = \sup_{\mathbf{x}\in\mathbb{R}^d} u_P(\mathbf{x},\mathbf{x})$.

Proposition 5.5. *Suppose* k *satisfies the conditions in Lemma 5.2 and* $\mathbb{E}_{\mathbf{X}\sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2] < \infty$. *Let* $\epsilon_0 \in [0,1]$. *Then* $\mathcal{P}(P;\epsilon_0) \subseteq \mathcal{B}^{\mathrm{KSD}}(P;\theta)$ *if* $\theta = \epsilon_0 \tau_\infty^{1/2}$ *and this bound is tight, i.e.,* $\sup_{Q\in\mathcal{P}(P;\epsilon_0)} D(Q,P) = \epsilon_0\tau_\infty^{1/2}$.

This result suggests that, when at most a proportion ϵ_0 of data is corrupted, setting $\theta = \epsilon_0 \tau_\infty^{1/2}$ will ensure quantitative robustness to $\mathcal{P}(P; \epsilon_0)$. Note that using a similar argument as in the proof of Proposition 5.5, Huber's contamination models can also be related to the MMD balls; see Section 5.B.2.

One practical issue is that τ_∞, the supremum of the Stein kernel, does not always admit a closed-form expression. One way to compute it is by numerical optimization of $\mathbf{x} \mapsto u_P(\mathbf{x}, \mathbf{x})$, but this assumes the optimizer converges and requires extra computation, thus not suitable in high dimension or when evaluation of the score function is costly. We propose an alternative approach, where τ_∞ is approximated by the maximum of the Stein kernel evaluated at the observed data, i.e., $\max_{i=1,\dots,n} u_P(\mathbf{X}_i, \mathbf{X}_i)$. This approach requires no extra computation and gives a reasonable estimate with moderate or large n. In Section 5.5, we use this approach and show empirically that the resulting test is still well-calibrated despite this approximation. Further discussions on how this approximation affects the test performance can be found in Section 5.B.3.

Another uncertainty model commonly studied in the literature is the *density-band* model (Kassam, 1981; Hafner, 1993), defined as distributions whose density function lies within an error band of a nominal model with density p, i.e., $\{Q \in \mathcal{P}(\mathbb{R}^d) : \ Q \text{ has density } q \text{ such that } |q(\mathbf{x}) - p(\mathbf{x})| \leq \delta(\mathbf{x}) \text{ for all } \mathbf{x}\}$, for some function $\delta : \mathbb{R}^d \to [0, \infty)$. It can be shown that density-band models can be rewritten as Huber's contamination models (5.6) with additional constraints on the outlier distribution (Fauß and Zoubir, 2016). However, compared with Huber's models, density-band models have the advantage of being more interpretable and more natural for certain forms of model disparity such as heavy tails. The following result suggests how the uncertainty radius θ should be chosen for such models and is proved in Section 5.A.6.3.

Proposition 5.6. *Suppose k satisfies the conditions in Lemma 5.2 and $\mathbb{E}_{\mathbf{X}\sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2] < \infty$. Furthermore, assume $Q, P \in \mathcal{P}(\mathbb{R}^d)$ admit positive densities q, p on $\mathbb{R}^d$ and $p \in \mathcal{C}^1$. If $|q(\mathbf{x}) - p(\mathbf{x})| \leq \delta(\mathbf{x})$ for some function $\delta : \mathbb{R}^d \to [0, \infty)$ such that $\delta_0 := \int_{\mathbb{R}^d} \delta(\mathbf{x}) \, \mathrm{d}\mathbf{x} < \infty$, then $D(Q, P) \leq \delta_0 \tau_\infty^{1/2}$.*

5.5 Experiments

We will now evaluate numerically the proposed GOF tests using both synthetic and real data. Unless otherwise mentioned, all standard KSD tests are based on an IMQ kernel $k(\mathbf{x}, \mathbf{x}') = h_{\mathrm{IMQ}}(\mathbf{x} - \mathbf{x}')$ where $h_{\mathrm{IMQ}}(\mathbf{u}) = (1 + \|\mathbf{u}\|_2^2/\lambda^2)^{-1/2}$ with a bandwidth $\lambda^2 > 0$ selected via the median heuristic, i.e., $\lambda_{\mathrm{med}} = \mathrm{Median}\big\{\|\mathbf{X}_i - \mathbf{X}_j\|_2 : \ 1 \leq i < j \leq n\big\}$. All tilted-KSD and robust-KSD tests are based on a *tilted* IMQ kernel with the weight function $w(\mathbf{x}) = (1 + \|\mathbf{x} - \mathbf{a}\|_2^2/c)^{-b}$, where $\mathbf{a} \in \mathbb{R}^d$ and $c > 0$. Intuitively, $\mathbf{a}$ and c respectively centers and scales the input. In our experiments, we choose $\mathbf{a} = 0$ and $c = 1$, as all data will always be centered and on a suitable scale. Moreover, since we will focus on sub-Gaussian models, we choose $b = 1/2$ in all experiments. This ensures the Stein kernel is bounded. All tests have nominal level $\alpha = 0.05$.

The probability of rejection is computed by averaging over 100 repetitions, and the 95% confidence intervals are reported. Our robust-KSD test will be shortened as *R-KSD*. Code for reproducing all experiments can be found at github.com/XingLLiu/robust-kernel-test.

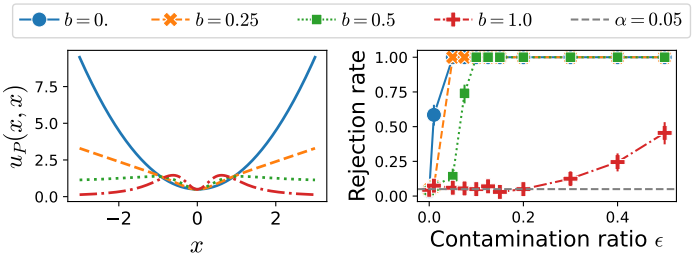


Figure 5.1: *Left.* Stein kernel for $P = \mathcal{N}(0, 1)$ and an IMQ kernel tilted by $w(x) = (1 + x^2)^{-b}$. The larger b is, the more the tails of the function $x \mapsto u_P(x, x)$ are down-weighted. The choice $b = 0$ corresponds to no weighting, reducing to an IMQ kernel. *Right.* The rejection probability under contamination by $R = \delta_z$ with $z = 10$.

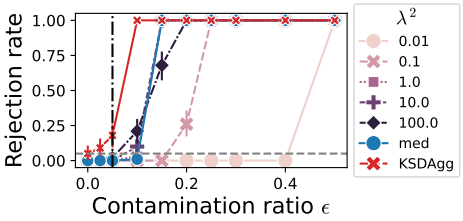


Figure 5.2: Rejection probability of robust-KSD with different bandwidths λ. "med" is the median heuristic. "KSDAgg" is the test of Schrab et al. (2022). The dashed line is $\alpha = 0.05$. The vertical line is the maximal proportion of contamination $\epsilon_0 = 0.05$ controlled by robust-KSD.

5.5.1 Toy Gaussian Model

We first consider a Gaussian model $P = \mathcal{N}(0, 1)$, in which case the score function of the model is $s_p(x) = -x$ and is hence unbounded. This toy example is simplistic and not representative of the unnormalized models our test is most suited for, but it will nonetheless be helpful to study our algorithmic choices, to verify Theorem 5.1 and Theorem 5.3 numerically, and to compare against alternative robust tests.

Decay rate of weighting function We first draw random samples $\mathbb{X}_n$ of size $n = 500$ from $Q = (1 - \epsilon)P + \epsilon\delta_z$, for different values of $\epsilon \in [0, 1]$ and $z \in \mathbb{R}$. This setting mimics the presence of outliers at z. Given a stationary kernel h, a natural question is "how does the choice of weight w affect the Stein kernel and the test power?". On the left-hand side of Figure 5.1, we plot the Stein kernel $x \mapsto u_P(x, x)$ for different values of b. As b grows, the tails of the Stein kernel are progressively down-weighted. We then plot the rejection probability of the standard KSD tests using these Stein kernels when the outlier is $z = 10$ on the right-hand side of Figure 5.1. As b increases, the rejection probability decreases, suggesting that the test power is lower when the tails of the Stein kernel are overly down-weighted. This is not surprising since the Stein kernel decays faster for larger values of b, making the test insensitive to model deviations at the tails. Ideally, w should decay just enough so that u_P remains bounded, but not too fast as it would lose power. This highlights the trade-off between robustness and power in the choice of w. A similar trade-off has been observed in robust estimation, where an overly decaying w can impede the estimation accuracy (Barp et al., 2019, Appendix F.2). Henceforth, we choose $b = 1/2$ to balance this trade-off, as it is the smallest value that ensures u_P is bounded for sub-Gaussian models.

Kernel bandwidth We now use the same dataset to investigate how the choice of the kernel bandwidth λ affects the performance of the robust-KSD test. It is well-known that the bandwidth plays a crucial role in kernel-based tests and can considerably affect the ability of the test to detect model disparities (Gretton et al., 2012; Ramdas et al., 2015; Reddi et al., 2015; Schrab et al., 2022, 2023). In general, a smaller bandwidth is better at detecting local differences, while a larger bandwidth is

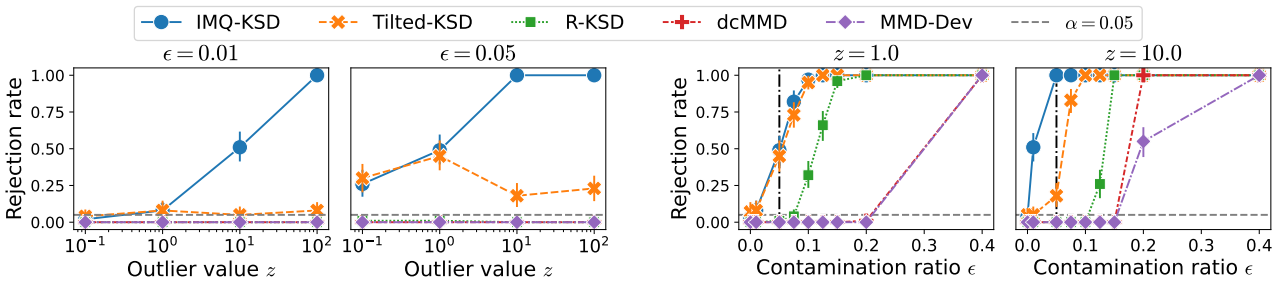


Figure 5.3: Rejection probability under an outlier-contaminated Gaussian model with different outlier values z and contamination ratios ϵ. The grey dotted horizontal line is the test level $\alpha = 0.05$, and the black dash-dot vertical line corresponds to $\epsilon_0 = 0.05$. The KSD tests with IMQ kernel lack both qualitative and quantitative robustness since they reject even for small z or ϵ. The tilted-KSD test is more robust in cases where z or ϵ are larger, but ultimately still reject the null due to their lack of quantitative robustness.

more suited for global deviations (Schrab et al., 2023). A common strategy for bandwidth selection in kernel-based testing is the median heuristic. Schrab et al. (2022, 2023) also proposed testing frameworks that aggregate the test result with multiple bandwidths to give the final decision, thereby avoiding bandwidth selection and achieving higher test power. A natural question is therefore: "how does the bandwidth affects the robustness of the proposed test?".

We run the robust-KSD test with various choices of bandwidths $\lambda^2 \in \Lambda \cup \{\lambda^2_{\text{med}}\}$, where $\Lambda = \{0.01, 0.1, 1, 10, 100\}$ and the uncertainty radius θ is set to control at most $\epsilon = 0.05$ contaminations in the sample. We also compare it with the KSDAgg test of Schrab et al. (2022) using the same tilted kernel and aggregating over the collection of bandwidths Λ. Other configurations of KSDAgg follow the recommendations in Schrab et al. (2022, Section 4.2). This experiment is repeated 200 times to reduce randomness, and the results are reported in Figure 5.2. For any choice of λ, the robust-KSD test is able to control the Type-I error when $\epsilon \leq \epsilon_0$, where ϵ is the proportion of contamination in the data, suggesting that robust-KSD remains well-calibrated for all bandwidths. If instead $\epsilon > \epsilon_0$, all robust-KSD tests eventually reject H_0^{C}, with λ chosen by median heuristic achieving the highest power. The KSDAgg test is not robust and more sensitive to contamination than all robust-KSD tests at all contamination levels ϵ. This is not surprising given that KSDAgg is designed to *maximize* the test power over multiple bandwidths, rendering this test more sensitive to a wide range of alternatives, including contaminated models. As a result, we henceforth do not include KSDAgg in our experiments. A similar study for the *standard* KSD test is given in Section 5.B.4.

Standard vs. robust KSD tests Using the same dataset, we then compare KSD tests with a stationary kernel to tilted-KSD tests and our proposed robust-KSD test in Figure 5.3. As shown in the left two plots, with a contamination ratio ϵ of 0.01 or 0.05, the standard IMQ-KSD test rejects the point null with high probability for large outlier values z. This aligns with the lack of qualitative robustness of tests based on stationary kernels proved in Theorem 5.1. In contrast, the standard Tilted-KSD test rejects with lower probability for *all* z values, aligning with the qualitative robustness result with tilted kernels in Theorem 5.3. Notably, our theoretical results in both Section 5.3 and Section 5.4 apply only to *fixed* kernels, thus excluding kernels based on observed data such as those selected via median heuristic. However, our numerical results suggest that the conclusion of our theory may hold more

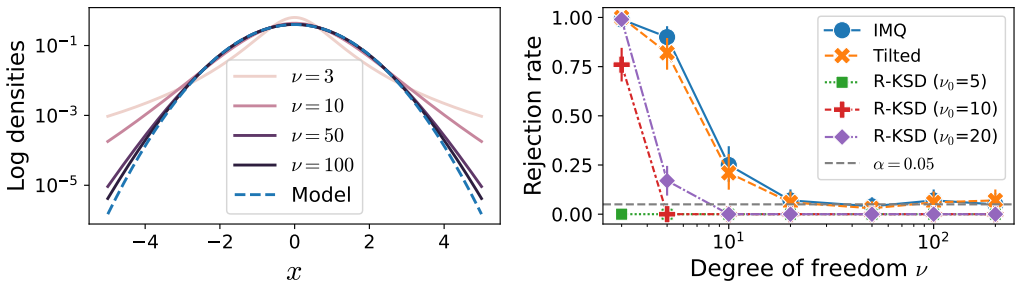


Figure 5.4: Heavy-tailed experiment. *Left.* Log densities of a standard Gaussian model and scaled t-distributions with different degree-of-freedom (dof) and moment-matched to Gaussian. *Right.* Rejection probability of the standard and robust tests with different θ set to control the cases $\nu \geq \nu_0$ for different values of ν_0.

broadly.

The Type-I error of the tilted-KSD test still exceeds the nominal level $\alpha = 0.05$, suggesting that a tilted kernel alone is not enough to enforce *quantitative* robustness. Notably, the rejection probability of Tilted-KSD is highest when $z = 1$ but declines for larger z. This is because $u_P(z, z)$ with the tilted kernel peaks at around $|z| \approx 1$ and then decreases with z (see Figure 5.1), making the test more sensitive when $|z| \approx 1$ but less sensitive to large outliers. In contrast, running our robust-KSD test with a pre-set contamination control $\epsilon_0 = 0.05$, we can see from the left two plots of Figure 5.3 that it remains well-calibrated for all values of outlier z. The results also suggest that R-KSD is conservative since its rejection probability is close to 0; we attribute this to the fact that the robust tests are designed to control *all* types of contamination, of which the injected outlier is only one specific type. To demonstrate the non-trivial power of the robust tests, we show in the right two plots the rejection probability against different contamination ratios ϵ. As ϵ grows, R-KSD eventually is capable of rejecting the null hypothesis $\mathcal{H}_0 : Q \in \mathcal{B}^{\mathrm{KSD}}(P; \theta)$ with test power approaching one.

We also compared our tests with two existing robust kernel tests, namely the robust MMD tests of Schrab and Kim (2024) (denoted *dcMMD*) and of Sun and Zou (2023, Eq. (38)) (denoted *MMD-Dev*). Both tests are provably quantitatively robust to Huber's contamination models, but they are two-sample tests that require extra samples from the model P. When the same number of samples from Q and from P are used and the cost of simulation is negligible, the cost of a single evaluation of MMD comparable to that of KSD. However, this cost could be prohibitively expensive when P is a more complex model. Implementation details of these tests are deferred to Section 5.B.2. As shown in Figure 5.3, both tests are more conservative than our proposed KSD tests. The conservativeness of dcMMD is because it is designed for a contamination model that is *stronger* than Huber's model, while the conservativeness of MMD-Dev is because it uses a deviation inequality to construct the decision threshold, which is known to be conservative; see e.g., Gretton et al. (2012).

Misspecified tails Finally, we now change the data-generating mechanism to study the impact of misspecified tails. We sample from $Q_\nu = t_\nu \sqrt{(\nu - 2)/\nu}$, where t_ν is the Student's t-distribution with degree-of-freedom (dof) $\nu > 2$, and the scaling factor $\sqrt{(\nu - 2)/\nu}$ ensures its second moment matches that of P. We set the uncertainty radius so that the R-KSD test is robust to Q_ν with $\nu \geq \nu_0$ for different values of ν_0. This is achieved using Proposition 5.6 to derive a bound on $D(Q_\nu, P)$, which is

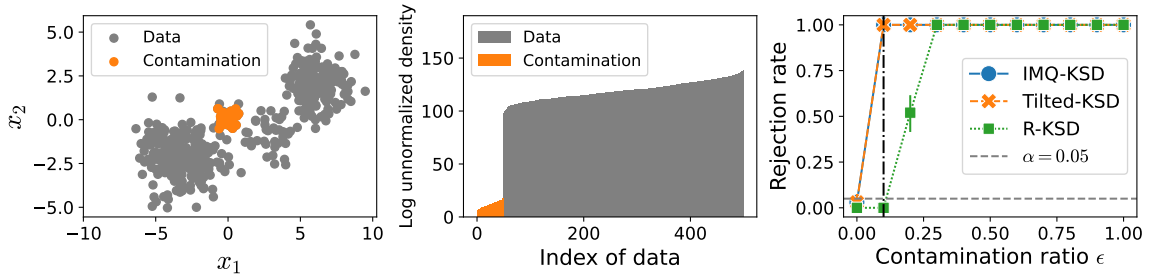


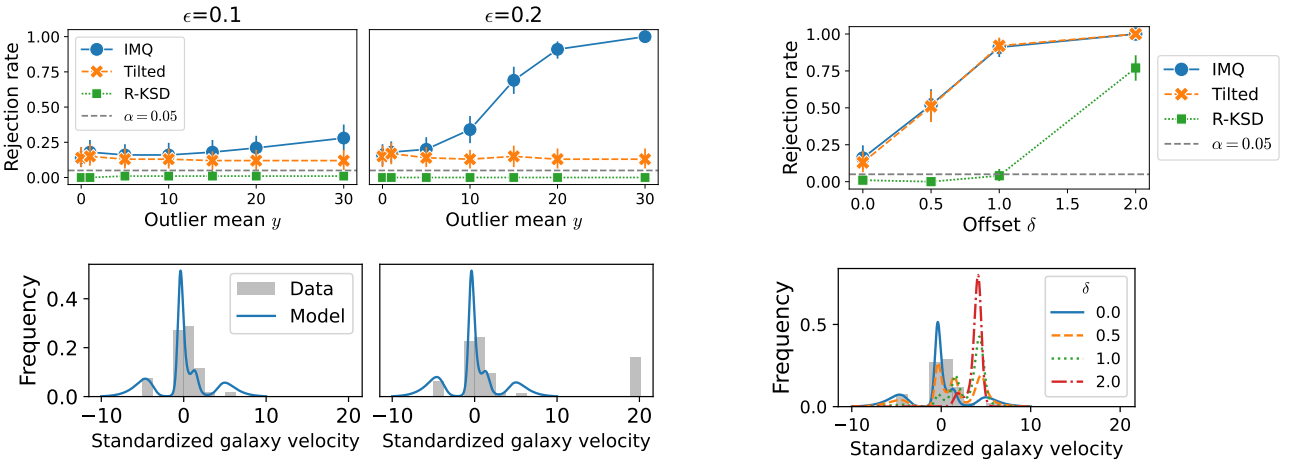
Figure 5.5: Gaussian-Bernoulli RBM experiment. *Left.* Data generated from P and injected contamination in the first two dimensions. *Middle.* Log unnormalized densities of the data and contamination ordered from small to large; the injected contamination is indeed abnormal since they have much lower densities. *Right.* Probability of rejection with a Gaussian-Bernoulli RBM against the contamination ratio ϵ. The robust tests are calibrated to control the Type-I error with no more than $\epsilon_0 = 0.1$ proportion of contamination.

discussed in detail in Proposition 5.22 of Section 5.A.6.4. The results are reported in Figure 5.4. For small values of ν, the tail of the model is misspecified, leading the robust tests with $\nu_0 = 10$ or 20 to reject with high probability because the KSD between Q_ν and P is large. As ν grows, Q_ν converges weakly to the standard Gaussian, making the model well-specified in the limit $\nu \to \infty$. Consequently, neither the standard tests or the robust tests reject the null hypothesis for large ν. The robust test with $\nu_0 = 5$ shows no power because, to be robust to tails that are so misspecified, the uncertainty radius θ computed using Proposition 5.6 must be very large, thus resulting in a small test statistic (cf. (5.9)) and low test power. Overall, the robust tests are more conservative than the standard one. This is because the robust tests control *all* possible contaminations within a KSD-ball, while misspecified tails are only one of the many possible such forms of contamination.

5.5.2 Gaussian-Bernoulli Restricted Boltzmann Machines

Our next experiment is an example of energy-based model called a Gaussian-Bernoulli Restricted Boltzmann Machine (RBM) model (Cho et al., 2013). RBMs have been used in a wide range of scientific applications (see Fischer and Igel, 2014; Zhang et al., 2019, for a review), and they are a common benchmarks for assessing kernel GOF tests (Liu et al., 2016; Jitkrittum et al., 2017; Schrab et al., 2022). RBMs are latent-variable models with joint density $p(\mathbf{x}, \mathbf{h}) \propto \exp(\frac{1}{2}\mathbf{x}^\top B\mathbf{h} + \mathbf{b}^\top\mathbf{x} + \mathbf{c}^\top\mathbf{h} - \frac{1}{2}\|\mathbf{x}\|_2^2)$, where $\mathbf{x} \in \mathbb{R}^d$ is an observable variable, $\mathbf{h} \in \{\pm 1\}^{d'}$ is a binary hidden variable with latent dimension d', and $B \in \mathbb{R}^{d\times d'}, \mathbf{b} \in \mathbb{R}^d$ and $\mathbf{c} \in \mathbb{R}^{d'}$ are model parameters. Computing the normalizing constant of the marginal $p(\mathbf{x})$ requires summing over $2^{d'}$ terms, so it becomes intractable when d' is large. However, its score function has a closed-form expression $\mathbf{s}_p(\mathbf{x}) = b - \mathbf{x} + B\tanh(B^\top\mathbf{x} + \mathbf{c})$, where tanh is applied entry-wise. We set $d = 50$ and $d' = 10$, and randomly initialize $B, \mathbf{b}, \mathbf{c}$ by sampling each entry independently from a $\mathcal{N}(0, 1)$. We then generate $n = 500$ samples from the model by block Gibbs sampling following Cho et al. (2013); Jitkrittum et al. (2017), and randomly replace $\epsilon \in [0, 1]$ proportions of the data with outliers drawn from $R = \mathcal{N}(0, 0.1^2 I_d)$; see the left and middle of Figure 5.5. As shown by the plots, the outliers are clearly abnormal since they have low density values.

We plot the rejection probabilities of the tests on the right-hand side of Figure 5.5. The standard



(a) Level (b) Power

Figure 5.6: Probability of rejection (top) and data and fitted models (bottom) for the KEF experiment. *(a)* Extra outliers are added to the data; data and fitted models with no contamination (bottom left) and contamination ratio $\epsilon = 0.2$ (bottom right). *(b)* Different offset values δ are added to the fitted parameter to create model deviation. The radius θ is chosen to control contamination of up to a proportion of $\epsilon_0 = 0.2$.

IMQ-KSD and Tilted-KSD tests are not well-calibrated under the composite null $H_0^{\mathrm{C}} : Q \in \mathcal{B}^{\mathrm{KSD}}(P; \theta)$, while robust-KSD with θ chosen by setting $\epsilon_0 = 0.1$ controls the rejection probability below the level $\alpha = 0.05$ when the contamination ratio ϵ does not exceed ϵ_0. On the other hand, when ϵ exceeds the maximal allowed proportion ϵ_0, the robust test rejects H_0^{C} with probability approaching 1, thus showing its power.

When generating the synthetic data from P, we run a Gibbs sampler for 7000 steps and discard the first 2000 as burn-in and keep every 10-th datum to reduce correlation. This takes 58.40 seconds on average, compared with 0.84 seconds required to perform the robust-KSD test. As a result, it will clearly not be reasonable to use the robust MMD tests from Sun and Zou (2023); Schrab et al. (2023) here as simulating a large enough number of samples from P would cost orders of magnitude more than the cost of the GOF test. Moreover, the samples generated through Gibbs sampling are no longer i.i.d., thus violating the assumptions in our theoretical results. Extension to the non-independent case could potentially be made following the approach in Chérief-Abdellatif and Alquier (2022), which studied the robustness of estimators (instead of GOF tests) under correlated data.

5.5.3 A Kernel Exponential Family Model for Density Estimation

Our next example concerns density estimation using a kernel exponential family (KEF) model (Canu and Smola, 2006). Given a reproducing kernel $k_0 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and a reference density p_0 on $\mathbb{R}^d$, we have $p_\eta(\mathbf{x}) \propto p_0(\mathbf{x}) \exp(-f_\eta(\mathbf{x}))$, where f_η lies in the RKHS $\mathcal{H}_{k_0}$ associated with k_0. We follow the setting in Matsubara et al. (2022) which uses a finite-basis approximation of the RKHS function so that $f_\eta(\mathbf{x}) = \sum_{l=1}^{25} \eta_l \phi_l(\mathbf{x})$, where $\eta = (\eta_1, \ldots, \eta_{25})^\top \in \mathbb{R}^{25}$ and $\phi_l(\mathbf{x}) = (x^l/\sqrt{l!}) \exp(-x^2/2)$ for $l = 1, 2, \ldots, 25$.

We are interested in testing the goodness-of-fit of the KEF model under data contamination. Suppose that the practitioners believe the observed data are subject to contamination, so that a *robust* estimator for η is used. Robust parameter estimation for exponential family models was studied in Barp et al. (2019), which proposed a minimum-distance estimator using KSD with tilted kernels. With a suitable weighting function, the resulting estimator is *bias-robust* (Barp et al., 2019, Proposition 7) and thus not susceptible to contamination. However, how to test the goodness-of-fit of the estimated model in this setting remains an open question. The standard IMQ-KSD test is not suitable since it is not robust and can thus falsely reject a good model due to contamination. Using a tilted kernel is also not enough since it is not quantitatively robust and hence can suffer from uncontrolled Type-I error. We therefore propose to use our robust-KSD test for this task. One challenge with the KEF model is that its density does not have a closed form, and although sampling from P is feasible in this low-dimensional example ($d = 1$), it would become challenging in high dimensions, making MMD-based tests not well-suited for this model. Our robust-KSD test, however, does not suffer from this limitation.

We use the dataset as Matsubara et al. (2022); Key et al. (2021), which is a 1-dimensional dataset of 82 galaxy velocities (Postman et al., 1986; Roeder, 1990). To avoid using the same data for model training and testing, we randomly split the data into equal halves, each containing $n_{\text{data}} = 41$ data points. To each half, we then add n_{ol} independent realisations from $R = \mathcal{N}(z, 0.1^2)$ for some value of z for the mean contamination, so that the resulting dataset has a size of $n = n_{\text{data}} + n_{\text{ol}}$ and a contamination ratio of $\epsilon = n_{\text{ol}}/(n_{\text{data}} + n_{\text{ol}})$. We then fit the KEF model to one half of the data, and perform the tests on the other half. Both the robust minimum KSD estimation and the robust-KSD test use the same tilted kernel $k(x, y) = w(x)h_\Lambda(x - y)w(y)$ with $w(x) = (1 + x^2)^{-1/2}$ and a sum of IMQ kernels $h_\Lambda(x - y) = \sum_{\lambda^2 \in \Lambda}(1 + |x - y|^2/(2\lambda^2))^{-1/2}$ where $\Lambda = \{0.6, 1, 1.2\}$. The choice of the weighting function follows Matsubara et al. (2022) and the use of a sum kernel is recommended in Key et al. (2021). Direct computation shows that the Stein kernel u_P is bounded in this case. We choose $\epsilon_0 = 0.2$ when computing θ for R-KSD.

The results are shown in Figure 5.6. We first study the level of our tests in Figure 5.6a. As the outlier mean z increases, the IMQ-KSD test rejects the null $H_0 : Q = P$ with increasing probability, suggesting that this is due to the presence of outliers. Notably, this happens even though the parameter estimator is not susceptible to contamination; see the bottom row of Figure 5.6a for the fitted model with and without outliers. When the tilted kernel was used, the standard KSD test has a lower rejection probability, although it is still above the nominal level. On the other hand, the robust-KSD test is well-calibrated for all values of z.

We then study the power of KSD tests in Figure 5.6b. To show the robust test has non-trivial power when no outliers are present, we fit the model parameters $\hat{\eta}$ on half of the data (without contamination) and replace the first component by $\hat{\eta}_1 + \delta$, where $\delta \in \{0, 0.5, 1, 2\}$ is an error offset, so that the resulting model has a poor fit. Other experimental setups including the choice of θ remain the same as before. As expected, the robust test has lower power than the standard tests; however, as the offset value δ increases, the robust test eventually rejects with high probability, thus showing its power under no contamination.

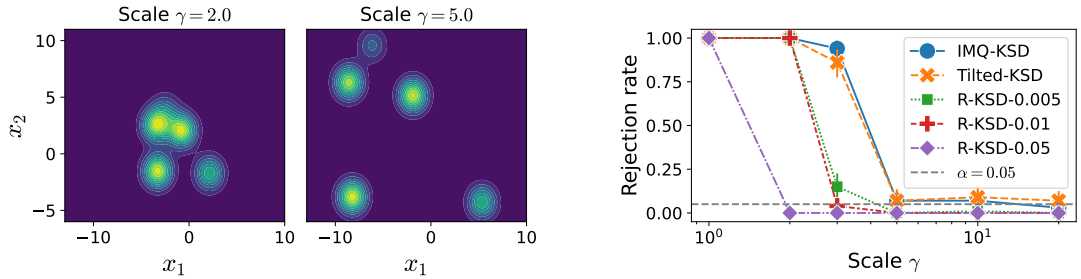


Figure 5.7: Mixture-of-Gaussian experiment. *Left.* Contour plots of model densities. *Right.* Rejection probability of the standard and robust tests. "R-KSD-0.01" refers to the R-KSD test with uncertainty radius θ chosen by setting the maximal allowed contamination proportion to $\epsilon_0 = 0.01$, and similarly for the others.

5.5.4 Limitations for Multimodal Models

We investigate the performance of the robust test when the model has multimodality and the data are drawn from the same model with misspecified mixing ratios, thus revealing a limitation of our robust-KSD test. It is well-known that the standard KSD test performs poorly with mixture models with well-separated modes (Liu et al., 2023). This is intuitively due to the myopia of the score function, making the test blind to disparities in the mixing ratios, and it is a general issue for KSD and other score-based discrepancies (Wenliang and Kanagawa, 2020; Zhang et al., 2022). We demonstrate that our robust test can also suffer from this limitation, which is made even more prominent due to the relaxation of the point null to a KSD-ball. We consider a 2-dimensional mixture of Gaussian model $P = \sum_{j=1}^{5} \pi_j \mathcal{N}(\gamma\mu_j, I_2)$, where $\pi_j \propto u_j$ with $u_1, \ldots, u_5$ drawn independently from Uniform$(0, 1)$, each mean $\mu_j \in \mathbb{R}^d$ has entries drawn independently from Uniform$(-2, 2)$, and $\gamma > 0$ is a scaling factor controlling the mode separation. The data-generating distribution Q has a different set of mixture ratios that are randomly generated. For R-KSD, the uncertainty radius is chosen to be $\theta = \epsilon_0 \tau_\infty^{1/2}$ with different values of ϵ_0, and we label the corresponding result as R-KSD-ϵ_0. The results are reported in Figure 5.7. As the scale γ increases, the modes become more separated, so $D(Q, P)$ decreases and eventually Q lies in a small neighbourhood of P, and hence all the KSD tests fail to reject with high probability. This is a consequence of the aforementioned myopia of KSD tests, which affects both the standard and the robust tests. The opposite is observed when γ is small, in which case the modes overlap, so all tests can correctly reject the null with high probability. Compared with the standard tests, the robust one is more conservative because it controls all possible contamination; nevertheless, it still achieves non-trivial power when ϵ_0, thus also the radius θ, is set to a small value.

5.6 Conclusion and Discussion

This chapter studied the robustness of the kernel GOF test of Liu et al. (2016); Chwialkowski et al. (2016). We showed that existing KSD tests are not qualitatively robust when stationary kernels are used, but achieves qualitative robustness using the tilted kernels from the robust estimation literature (Barp et al., 2019; Matsubara et al., 2022). Since qualitative robustness is not enough to ensure calibration when the proportion of contaminations does not vanish asymptotically, we proposed a novel robust KSD

test that provably controls the Type-I error when the data-generating distribution lies in a KSD-ball around the reference model, and consistently rejects otherwise. We then discussed how to choose the radius of the ball under different contamination models, namely Huber's contamination model and the density-band model. Through various experiments including synthetic and real datasets, we showed that the robust GOF test is well-calibrated and maintains non-trivial power. We now discuss several interesting directions for future work.

Firstly, we showed consistency of the proposed robust KSD test against *all* alternative distributions not lying in the null set, i.e., $H_1^{\mathrm{C}} : Q \notin \mathcal{B}^{\mathrm{KSD}}(P;\theta)$. In the robust testing literature, it is however also common to restrict the alternative set to a neighborhood centered on an alternative model $P' \neq P$ and study minimaxity against such alternative hypothesis (Huber, 1965; Levy, 2008; Gül and Zoubir, 2016; Sun and Zou, 2023). Understanding whether our robust KSD test achieves such minimax property could be an interesting future work.

Secondly, as discussed in Section 5.2.3, our proposed robust KSD test does not require the Stein kernel u_P to be bounded. A bounded Stein kernel is convenient as the resulting KSD-ball $\mathcal{B}^{\mathrm{KSD}}(P;\theta)$ contains many contamination models of interest, such as Huber's model with an *arbitrary* noise distribution. However, this could be too strict in some applications where outliers cannot take an arbitrary value, e.g., when the domain is bounded or the data acquisition procedure is highly regulated to prevent extreme outliers. In these cases, even with an unbounded kernel, the KSD-ball could still capture all contamination models of interest. Unbounded Stein kernels possess desirable theoretical properties such as controlling the convergence of expectations of certain function classes (Barp et al., 2022; Kanagawa et al., 2022); exploring whether they can lead to more powerful robust-KSD tests could therefore also be a worthwhile future research direction.

5.A Proof of Theoretical Results

This section contains the proofs of all main theoretical results in Chapter 5. In particular,

- Section 5.A.1 proves the non-qualitative robustness result Theorem 5.1.
- Section 5.A.2 proves the qualitative robustness result Lemma 5.2.
- Section 5.A.3 proves Theorem 5.3.
- Section 5.A.4 presents auxiliary tools needed to prove the validity and consistency of the robust-KSD test Theorem 5.4.
- Section 5.A.5 proves Theorem 5.4.
- Section 5.A.6 proves the results concerning the connection between KSD balls and different contamination models, namely Proposition 5.5 and Proposition 5.6.

We will use the following notation throughout this section. Given i.i.d. random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim Q \in \mathcal{P}(\mathbb{R}^d)$ and an integrable function $f : (\mathbb{R}^d)^n \to \mathbb{R}$, we will write for brevity $\mathbb{E}[f(\mathbf{X}_1, \ldots, \mathbf{X}_n)] = \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} f(\mathbf{x}_1, \ldots, \mathbf{x}_n) Q(\mathrm{d}\mathbf{x}_1) \cdots Q(\mathrm{d}\mathbf{x}_n)$.

5.A.1 Proof of Theorem 5.1 and Related Preliminary Results

Sketch of proof For any $\epsilon \in [0,1]$ and $R \in \mathcal{P}(\mathbb{R}^d)$, a random variable $\mathbf{X}_i$ drawn from $Q = (1-\epsilon)P + \epsilon R$ can be written as $\mathbf{X}_i = (1-\xi_i)\mathbf{X}_i^* + \xi_i \mathbf{Z}_i$, where $\xi_i \sim \text{Bernoulli}(\epsilon)$, $\mathbf{X}_i^* \sim P$ and $\mathbf{Z}_i \sim R$ are independent, and the equality is understood as equality in distribution. Defining the random variable $M' = \sum_{i=1}^n \xi_i$, then $M' \sim \text{Binomial}(n, \epsilon)$ and represents the number of contaminated data in a random sample $\mathbb{X}_n$ drawn from Q. We will first show that the conclusion in Theorem 5.1 holds *conditional* on the event $\{M' \geq m_0'\}$ for some integer $m_0' \in [0, n]$, i.e., when there are at least m' contaminated data in the sample. This will be sufficient to show Theorem 5.1 by proving that this event occurs with high probability.

We will use the following notation in the rest of this section: Let $\mathbb{X}_n^*$ be a random sample drawn from Q. For any integer $m' \in [0, n]$ giving the number of corrupted data and any $\mathbf{z} \in \mathbb{R}^d$, we define $\mathbb{X}_{n,\mathbf{z}} = \mathbb{X}_{n-m'}^* \cup \{\mathbf{z}\}^{m'}$.

The rest of this section is organized as follows:

- We first present three preliminary results. Lemma 5.7 and Lemma 5.8 respectively bound the test statistic $D^2(\mathbb{X}_{n,\mathbf{z}})$ and $q^2_{\infty,1-\alpha}(\mathbb{X}_{n,\mathbf{z}})$ for a given number of corrupted data, m', by quantities of the form $a_n u_P(\mathbf{z},\mathbf{z}) + b_n$ for some a_n, b_n. These results allow us to compare the relative rate at which $D^2(\mathbb{X}_{n,\mathbf{z}})$ and $q^2_{\infty,1-\alpha}(\mathbb{X}_{n,\mathbf{z}})$ scale with $\mathbf{z}$. Both results will be used to prove Proposition 5.9, which states that the claim in Theorem 5.1 holds conditionally on the number of corrupted data.
- Section 5.A.1.1, 5.A.1.2, 5.A.1.3 and 5.A.1.4 provide the proofs for Lemma 5.7, Lemma 5.8, Proposition 5.9 and Theorem 5.1, respectively.

Assumption 5.1. *Assume* $\mathbb{E}_{\mathbf{X}\sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2^4] < \infty$ *and* $k(\mathbf{x},\mathbf{x}') = h(\mathbf{x}-\mathbf{x}')$ *for some* $h \in \mathcal{C}_b^2$ *with* $h(0) > 0$.

Lemma 5.7. *Assume* $\sigma_\infty^2 := \sup_{\mathbf{z}\in\mathbb{R}^d} \mathbb{E}_{\mathbf{X}\sim P}[u_P(\mathbf{X},\mathbf{z})^2] < \infty$ *and suppose Assumption 5.1 holds. Then, for any* $\delta > 0$ *and integers* $0 < m' < n$, *the following event holds with probability at least* $1-\delta$:

$$D^2(\mathbb{X}_{n,\mathbf{z}}) \;\geq\; \epsilon_{m,n}^2 u_P(\mathbf{z},\mathbf{z}) - \frac{2\sigma_\infty \epsilon_{m,n}}{\sqrt{\delta m}}\,,$$

where $\epsilon_{m,n} = m'/n$ *and* $m = n - m'$.

Lemma 5.8. *Suppose Assumption 5.1 holds. Assume* (i) $\sigma_\infty^2 := \sup_{\mathbf{z}\in\mathbb{R}^d} \mathbb{E}_{\mathbf{X}\sim P}[u_P(\mathbf{X},\mathbf{z})^2] < \infty$, (ii) $\xi_P^2 := \mathbb{E}_{\mathbf{X}\sim P}[u_P(\mathbf{X},\mathbf{X})^2] < \infty$, *and* (iii) $\Sigma_P^2 := \text{Var}_{\mathbf{X},\mathbf{X}'\sim P}(u_P(\mathbf{X},\mathbf{X}')) < \infty$. *Then there exist constants* $C, C_{P,1}$ *and* $C_{P,2}$ *depending only on* P *such that, for any* $\delta > 0$ *and integers* $0 < m' < n$, *the following event holds with probability at least* $1-\delta$:

$$q^2_{\infty,1-\alpha}(\mathbb{X}_{n,\mathbf{z}}) \;\leq\; \frac{C\log(1/\alpha)}{n\sqrt{\delta}}\epsilon_{m,n} u_P(\mathbf{z},\mathbf{z}) + \frac{C_{P,1}}{n\delta} + \frac{C_{P,2}\log(1/\alpha)}{n\sqrt{\delta}}\,,$$

where $\epsilon_{m,n} = m'/n$ *and* $m = n - m'$, *and the probability is over the joint distribution of* $\mathbb{X}_{n,\mathbf{z}}$ *and* $\mathbf{W} \sim \text{Multinomial}(n; 1/n, \ldots, 1/n)$.

Proposition 5.9. *There exists a constant $C > 0$ such that the following holds. Suppose Assumption 5.1 holds. Assume the following integrability conditions hold:*

$$\sup_{\mathbf{z}\in\mathbb{R}^d} \|\mathbf{s}_p(\mathbf{z})\|_2^4 \mathbb{E}_{\mathbf{X}\sim P}\left[h(\mathbf{X}-\mathbf{z})^4\right] \;<\; \infty\,, \qquad \textit{and} \qquad \sup_{\mathbf{z}\in\mathbb{R}^d} \|\mathbf{s}_p(\mathbf{z})\|_2^2 \mathbb{E}_{\mathbf{X}\sim P}\left[\|\nabla h(\mathbf{x}-\mathbf{z})\|_2^2\right] \;<\; \infty\,.$$

Further assume that $\mathbf{z} \mapsto \|\mathbf{s}_p(\mathbf{z})\|_2$ is unbounded. Then for any $s \in [0,1)$, there exists $\mathbf{z} \in \mathbb{R}^d$ such that, for any sample size n and integer $m' \in [n^{1-s}/2, n)$ giving the number of corrupted data, the condition

$$n^{\frac{(1-s)}{2}} \;>\; 2\sqrt{2}C \log\left(\frac{4}{\alpha}\right)$$

implies

$$\Pr_{\mathbb{X}^*_{n-m'}\sim P, \mathbf{W}}\left(D^2(\mathbb{X}_{n,\mathbf{z}}) > q^2_{\infty,1-\alpha}(\mathbb{X}_{n,\mathbf{z}})\right) \;\geq\; 1 - n^{-(1-s)}.$$

The assumption $m' \in [n^{1-s}/2, n)$ ensures there is considerable amount of corrupted data in the sample, while excluding the case where all data are contaminated ($m' = n$). The constant factor $1/2$ is arbitrary.

5.A.1.1 Proof of Lemma 5.7

The test statistic can be decomposed in three terms: a term with uncorrupted samples, a term with corrupted samples, and term for interactions between these two groups:

$$\begin{aligned}
D^2(\mathbb{X}_{n,\mathbf{z}}) \;&=\; \frac{1}{n}\sum_{1\leq i,j\leq n} u_P(\mathbf{X}_i, \mathbf{X}_j) \\
&=\; \frac{1}{n^2}\sum_{1\leq i,j\leq m} u_P(\mathbf{X}^*_i, \mathbf{X}^*_j) + \frac{2}{n^2}\sum_{1\leq i\leq m<j\leq n} u_P(\mathbf{X}^*_i, \mathbf{z}) + \frac{1}{n^2}\sum_{m<i,j\leq n} u_P(\mathbf{z},\mathbf{z}) \\
&=\; \frac{m^2}{n^2} D^2(\mathbb{X}^*_m) + \frac{2(n-m)}{n^2}\sum_{i=1}^{m} u_P(\mathbf{X}^*_i, \mathbf{z}) + \frac{(n-m)^2}{n^2} u_P(\mathbf{z},\mathbf{z}) \\
&=\; (1-\epsilon_{m,n})^2 D^2(\mathbb{X}^*_m) + 2\epsilon_{m,n}(1-\epsilon_{m,n}) S_{m,\mathbf{z}} + \epsilon^2_{m,n} u_P(\mathbf{z},\mathbf{z})\,, \qquad (5.10)
\end{aligned}$$

where in the last line we have defined $S_{m,\mathbf{z}} := m^{-1}\sum_{i=1}^m u_P(\mathbf{X}_i, \mathbf{z})$ and $\epsilon_{m,n} := (n-m)/n$.

The remainder of the proof proceeds by lower-bounding the first two terms with high probability. The first term can be lower-bounded by 0, since $D^2(\mathbb{X}^*_m)$ is a V-statistic and is hence non-negative. This lower bound will be relatively good as we would expect $D^2((\mathbb{X}^*_m)$ to approach zero at a root-m rate.

Bounding the second term in (5.10) requires bounding $S_{m,\mathbf{z}}$. Because by assumption $\mathbf{X}_i = \mathbf{X}^*_i \sim P$ are i.i.d. for $i = 1, \ldots, m$, the term $S_{m,\mathbf{z}}$ is a sum of i.i.d. random variables $u_P(\mathbf{X}^*_i, \mathbf{z})$. Moreover, $u_P(\mathbf{X}^*_i, \mathbf{z})$ are zero-mean, because the proof of Gorham and Mackey (2017, Proposition 1) shows that $\mathbb{E}_{\mathbf{X}^*\sim P}[u_P(\mathbf{X}^*, \cdot)] = 0$ whenever $\mathbb{E}_{\mathbf{X}^*\sim P}[\|\mathbf{s}_p(\mathbf{X}^*)\|_2] < \infty$, which holds under Assumption 5.1. We can therefore compute the variance of $S_{m,\mathbf{z}}$ as

$$\mathrm{Var}(S_{m,\mathbf{z}}) \;=\; \mathbb{E}\left[S^2_{m,\mathbf{z}}\right] \;=\; \frac{1}{m}\mathbb{E}_{\mathbf{X}^*_1\sim P}\left[u_P(\mathbf{X}^*_1, \mathbf{z})^2\right] \;\leq\; \frac{\sigma^2_\infty}{m}\,, \qquad (5.11)$$

which is finite since $\sigma_\infty^2 < \infty$ by assumption. We can hence apply Chebyshev's inequality. Choosing

$$t \coloneqq \left(\frac{\sigma_\infty^2}{\delta m}\right)^{\frac{1}{2}} = \frac{\sigma_\infty}{\sqrt{\delta m}}\,, \tag{5.12}$$

the set $\mathcal{A}_1 \coloneqq \left\{|S_{m,\mathbf{z}}| \leq t\right\}$ holds with probability at least $1-\delta$. Furthermore, on $\mathcal{A}_1$, we have by (5.10) that

$$D^2(\mathbb{X}_n) \;\geq\; (1-\epsilon_{m,n})^2 D^2(\mathbb{X}_m^*) - 2\epsilon_{m,n}(1-\epsilon_{m,n})t + \epsilon_{m,n}^2 u_P(\mathbf{z},\mathbf{z}) \;\geq\; -2\epsilon_{m,n}t + \epsilon_{m,n}^2 u_P(\mathbf{z},\mathbf{z})\,,$$

where in the last step we have used $\epsilon_{m,n} \in [0,1]$. This shows the claim.

5.A.1.2 Proof of Lemma 5.8

To prove this lemma, we will need the following concentration bound for quadratic forms of multinomial random vectors. This is a special instance of the Hanson-Wright inequality due to Adamczak (2015) when the dependent random vectors follow multinomial distributions.

Lemma 5.10 (Hanson-Wright inequality; Adamczak 2015, Theorem 2.5)**.** *Let* $\mathbf{W} = (W_1, \ldots, W_n) \sim$ *Multinomial*$(n; 1/n, \ldots, 1/n)$*. Then there exists constant* $C > 0$ *such that, for any* $n \times n$ *matrix* A *with entries* $a_{ij} \in \mathbb{R}$ *and any* $\delta > 0$*, we have*

$$\Pr_{\mathbf{W}}\left(\left|\mathbf{W}^\top A \mathbf{W} - \mu\right| \geq C\|A\|_{\mathrm{F}} \log\left(\frac{1}{\delta}\right)\right) \;\leq\; \delta\,,$$

where $\mu \coloneqq -n^{-1}\sum_{1\leq i,j\leq n} a_{ij} + \sum_{1\leq i\leq n} a_{ii}$*, and* $\|A\|_{\mathrm{F}} = (\sum_{1\leq i,j\leq n} a_{ij}^2)^{1/2}$ *is the Frobenius norm.*

Proof of Lemma 5.10. Adamczak (2015, Remark 2.2) shows that multinomial random vectors satisfy the convex concentration inequality (Adamczak, 2015, Definition 2.2). Therefore, we can apply Adamczak (2015, Theorem 2.5) to conclude that there exists constants C_1, C_2 such that, for any $n \times n$ matrix A and any $t > 0$,

$$\begin{aligned}\Pr_{\mathbf{W}}\left(\left|\mathbf{W}^\top A \mathbf{W} - \mu\right| \;\geq\; t\right) \;&\leq\; 2\exp\left(-\frac{1}{C_1}\min\left(\frac{t^2}{C_2\|A\|_{\mathrm{F}}^2}, \frac{t}{\|A\|_{\mathrm{op}}}\right)\right)\\ &\leq\; 2\exp\left(-\frac{1}{C_1}\min\left(\frac{t^2}{C_2\|A\|_{\mathrm{F}}^2}, \frac{t}{\|A\|_{\mathrm{F}}}\right)\right)\,,\end{aligned}$$

where $\|A\|_{\mathrm{op}} = \sup_{\|\mathbf{x}\|_2\leq 1}\|A\mathbf{x}\|_2$ is the operator norm, $\mu = \mathbb{E}_{\mathbf{W}}[\sum_{1\leq i,j\leq n}(W_i - 1)(W_j - 1)a_{ij}]$, and the last step holds due to the well-known relationship between operator norm and Frobenius norm $\|A\|_{\mathrm{op}} \leq \|A\|_{\mathrm{F}}$ (see, e.g., Golub and van Loan, 2013, Eq. 2.3.7). Moreover, we can compute μ using explicit expressions for the joint central moments of multinomial random variables (Ouimet, 2020,

Theorem 2):

$$\begin{aligned}
\mu &= \sum_{1\le i,j\le n} a_{ij}\mathbb{E}_{\mathbf{W}}\left[(W_i-1)(W_j-1)\right] \\
&= \sum_{1\le i,j\le n} a_{ij}\mathbb{E}_{\mathbf{W}}\left[(\mathbf{W}_i-1)(\mathbf{W}_j-1)\right] + \sum_{1\le i\le n} a_{ii}\mathbb{E}_{\mathbf{W}}\left[(\mathbf{W}_i)^2\right] \\
&= -\frac{1}{n}\sum_{1\le i\ne j\le n} a_{ij} + \left(1-\frac{1}{n}\right)\sum_{1\le i\le n} a_{ii} \\
&= -\frac{1}{n}\sum_{1\le i,j\le n} a_{ij} + \sum_{1\le i\le n} a_{ii}\,.
\end{aligned}$$

For any $\delta > 0$, we claim that the choice

$$t \;=\; \max\left(C_1^{\frac{1}{2}}C_2^{\frac{1}{2}},\, C_1\right)\|A\|_{\mathrm{F}}\log\left(\frac{1}{\delta}\right) \tag{5.13}$$

implies that $\Pr_{\mathbf{W}}(|\mathbf{W}^\top A\mathbf{W}-\mu| \ge t) \le \delta$, which would complete the proof by setting $C = \max(C_1^{1/2}C_2^{1/2}, C_1)$. To see this, we note that the the following are equivalent

$$\begin{aligned}
&\delta \;\ge\; 2\exp\left(-\frac{1}{C_1}\min\left(\frac{n^4t^2}{C_2\|A\|_{\mathrm{F}}^2},\, \frac{n^2t}{\|A\|_{\mathrm{F}}}\right)\right) \\
&\iff\; C_1\log\left(\frac{2}{\delta}\right) \;\le\; \min\left(\frac{n^4t^2}{C_2\|A\|_{\mathrm{F}}^2},\, \frac{n^2t}{\|A\|_{\mathrm{F}}}\right) \\
&\iff\; C_1\log\left(\frac{2}{\delta}\right) \;\le\; \frac{n^4t^2}{C_2\|A\|_{\mathrm{F}}^2} \qquad \text{and} \quad C_1\log\left(\frac{2}{\delta}\right) \;\le\; \frac{n^2t}{\|A\|_{\mathrm{F}}} \\
&\iff\; t \;\ge\; C_1^{1/2}C_2^{1/2}\frac{\|A\|_{\mathrm{F}}}{n^2}\sqrt{\log\left(\frac{2}{\delta}\right)} \qquad \text{and} \quad t \;\ge\; C_1\frac{\|A\|_{\mathrm{F}}}{n^2}\log\left(\frac{2}{\delta}\right) \\
&\iff\; t \;\ge\; \max\left(C_1^{1/2}C_2^{1/2},\, C_1\sqrt{\log\left(\frac{2}{\delta}\right)}\right)\frac{\|A\|_{\mathrm{F}}}{n^2}\sqrt{\log\left(\frac{2}{\delta}\right)}\,.
\end{aligned}$$

Since $\log(2/\delta) > 1$ for $\delta \in (0,1)$, the last inequality is met with t defined in (5.13). This shows that $\Pr_{\mathbf{W}}\left(\left|\mathbf{W}^\top A\mathbf{W} - \mu\right| \;\ge\; t\right) \le \delta$, thus finishing the proof. □

We are now ready to prove Lemma 5.8.

Proof of Lemma 5.8. Denote by F_∞ the cumulative distribution function for the conditional distribution of the bootstrap sample $D^2_{\mathbf{W}}(\mathbb{X}_{n,\mathbf{z}})$ (defined in (5.3)) given $\mathbb{X}_{n,\mathbf{z}}$, and denote its $(1-\alpha)$-quantile as

$$q^2_{\infty,1-\alpha}(\mathbb{X}_{n,\mathbf{z}}) \;\coloneqq\; \inf\{t\in\mathbb{R}:\; 1-\alpha \le F_\infty(t)\}\,.$$

To bound $q^2_{\infty,1-\alpha/2}(\mathbb{X}_{n,\mathbf{z}})$, we use a similar argument as in Schrab et al. (2023, Appendix E4) by bounding the exceedance probability of $D^2_{\mathbf{W}}(\mathbb{X}_{n,\mathbf{z}})$. Notably, their proof cannot be applied directly here, because they use a wild bootstrap (Leucht and Neumann, 2013) that corresponds to replacing $\mathbf{W}_i - 1$ with independent Rademacher weights, whereas in our case W_i are correlated and multinomially distributed. To proceed, we first note that we can write $D^2_{\mathbf{W}}(\mathbb{X}_{n,\mathbf{z}})$ as a quadratic form as $D^2_{\mathbf{W}}(\mathbb{X}_{n,\mathbf{z}}) = n^{-2}\mathbf{W}^\top A\mathbf{W}$,

where $A = (a_{ij})_{ij}$ with $a_{ij} \coloneqq u_P(\mathbf{X}_i, \mathbf{X}_j)$. We can then use the Hanson-Wright inequality stated in Lemma 5.10 applied with $\delta = \alpha/2$ to show that there exist positive constant C such that, for any almost sure realizations $\mathbb{X}_{n,\mathbf{z}}$,

$$\begin{aligned}
&\Pr_{\mathbf{W}}\left(D^2_{\mathbf{W}}(\mathbb{X}_{n,\mathbf{z}}) \;\geq\; \mu/n^2 + \frac{C\|A\|_{\mathrm{F}}}{n^2}\log\left(\frac{1}{\alpha}\right) \;\middle|\; \mathbb{X}_{n,\mathbf{z}}\right) \\
&= \Pr_{\mathbf{W}}\left(\mathbf{W}^\top A\mathbf{W} \;\geq\; \mu + C\|A\|_{\mathrm{F}}\log\left(\frac{1}{\alpha}\right) \;\middle|\; \mathbb{X}_{n,\mathbf{z}}\right) \\
&\leq \Pr_{\mathbf{W}}\left(\left|\mathbf{W}^\top A\mathbf{W} - \mu\right| \;\geq\; C\|A\|_{\mathrm{F}}\log\left(\frac{1}{\alpha}\right) \;\middle|\; \mathbb{X}_{n,\mathbf{z}}\right) \\
&\leq \alpha\,,
\end{aligned}$$

where $\mu \coloneqq -n^{-1}\sum_{1\leq i,j\leq n} a_{ij} + \sum_{1\leq i\leq n} a_{ii}$. This implies that the $(1-\alpha)$-quantile of $D^2_{\mathbf{W}}(\mathbb{X}_n)$ can be bounded as

$$q^2_{\infty,1-\alpha} \;\leq\; \frac{\mu}{n^2} + t \;=\; \frac{\mu}{n^2} + \frac{\rho_\alpha\|A\|_{\mathrm{F}}}{n^2}\,, \tag{5.14}$$

where we have defined $\rho_\alpha \coloneqq C\log(1/\alpha)$. We now further simplifies this bound by bounding μ and $\|A\|_{\mathrm{F}}$. The matrix A is the Gram matrix of the Stein kernel u_P, thus positive semi-definite. This implies $\sum_{1\leq i,j\leq n} a_{ij} \geq 0$ and $a_{ii} \geq 0$ for all i, so we have

$$\begin{aligned}
\mu \;=\; -\frac{1}{n}\sum_{1\leq i,j\leq n} a_{ij} + \sum_{1\leq i\leq n} a_{ii} \;\leq\; \sum_{1\leq i\leq n} a_{ii} \;&=\; \sum_{1\leq i\leq m} a_{ii} + \sum_{m<i\leq n} a_{ii} \\
&= \sum_{1\leq i\leq m} a_{ii} + (n-m)u_P(\mathbf{z},\mathbf{z})\,,
\end{aligned} \tag{5.15}$$

where the last quality holds since by assumption $\mathbf{X}_i = \mathbf{z}$ for $i > m$. Moreover, since $a_{ii} = u_P(\mathbf{X}_i, \mathbf{X}_i) \geq 0$ for any i, we can apply Markov inequality to bound the first term on the RHS. This gives that the following holds with probability at least $1 - \delta/4$,

$$\begin{aligned}
\sum_{1\leq i\leq m} a_{ii} \;\leq\; \frac{4}{\delta}\sum_{1\leq i\leq m}\mathbb{E}[a_{ii}] \;=\; \frac{4m}{\delta}\mathbb{E}_{\mathbf{X}^*\sim P}[u_P(\mathbf{X}^*,\mathbf{X}^*)] \;&\leq\; \frac{4m}{\delta}\left(\mathbb{E}_{\mathbf{X}^*\sim P}[u_P(\mathbf{X}^*,\mathbf{X}^*)^2]\right)^{\frac{1}{2}} \\
&= \frac{4m}{\delta}\xi_P\,,
\end{aligned}$$

where the second step holds since by assumption $\mathbf{X}_i = \mathbf{X}^*_i \sim P$ are i.i.d. for $i = 1, \ldots, m$, and where we have substituted $\xi^2_P = \mathbb{E}_{\mathbf{X}_1\sim P}[u_P(\mathbf{X}_1,\mathbf{X}_1)^2]$. Combining with (5.15) implies that, with probability at least $1 - \delta/4$,

$$\mu \;\leq\; \frac{4m}{\delta}\xi_P + (n-m)u_P(\mathbf{z},\mathbf{z})\,. \tag{5.16}$$

We now bound $\|A\|^2_{\mathrm{F}}$ by again applying Markov's inequality to show that, with probability at least

$1-\delta/4$, where the probability is taken over the randomness of $\mathbb{X}_m^*$, we have

$$\begin{aligned}
\|A\|_{\mathrm{F}}^2 &= \sum_{1\le i,j\le n} u_P(\mathbf{X}_i,\mathbf{X}_j)^2 \\
&\le \frac{4}{\delta}\sum_{1\le i,j\le n}\mathbb{E}[u_P(\mathbf{X}_i,\mathbf{X}_j)^2] \\
&= \frac{4}{\delta}\Big(\sum_{1\le i,j\le m}\mathbb{E}[u_P(\mathbf{X}_i^*,\mathbf{X}_j^*)^2] + 2(n-m)\sum_{1\le i\le m}\mathbb{E}[u_P(\mathbf{X}_i^*,\mathbf{z})^2] + (n-m)^2\mathbb{E}[u_P(\mathbf{z},\mathbf{z})^2]\Big) \\
&= \frac{4}{\delta}\Big(\big(m(m-1)\Sigma_P^2 + m\xi_P^2\big) + 2m(n-m)\mathbb{E}_{\mathbf{X}_1^*\sim P}[u_P(\mathbf{X}_1^*,\mathbf{z})^2] + (n-m)^2 u_P(\mathbf{z},\mathbf{z})^2\Big) \\
&\le \frac{4}{\delta}\Big(\big(m(m-1)\Sigma_P^2 + m\xi_P^2\big) + 2m(n-m)\sigma_\infty^2 + (n-m)^2 u_P(\mathbf{z},\mathbf{z})^2\Big)\,, \qquad (5.17)
\end{aligned}$$

where the second last line holds since

$$\begin{aligned}
\sum_{1\le i,j\le m}\mathbb{E}[u_P(\mathbf{X}_i^*,\mathbf{X}_j^*)^2] &= \sum_{1\le i\ne j\le m}\mathbb{E}[u_P(\mathbf{X}_i^*,\mathbf{X}_j^*)^2] + \sum_{1\le i\le m}\mathbb{E}[u_P(\mathbf{X}_i^*,\mathbf{X}_i^*)^2] \\
&= m(m-1)\mathbb{E}[u_P(\mathbf{X}_i^*,\mathbf{X}_j^*)^2] + m\mathbb{E}[u_P(\mathbf{X}_i^*,\mathbf{X}_i^*)^2] \\
&= m(m-1)\Sigma_P^2 + m\xi_P^2\,,
\end{aligned}$$

and where we have substituted $\Sigma_P^2 = \mathbb{E}_{\mathbf{X}_1,\mathbf{X}_2\sim P}[u_P(\mathbf{X}_1,\mathbf{X}_2)^2]$ and $\sigma_\infty^2 = \sup_{\mathbf{z}\in\mathbb{R}^d}\mathbb{E}_{\mathbf{X}\sim P}[u_P(\mathbf{X},\mathbf{z})^2]$. Using the above inequality and (5.16) to bound (5.14), we conclude that the following event holds with probability at least $1-\delta/2-\delta/4-\delta/4 = 1-\delta$, where the randomness is taken over the joint distribution of $\mathbb{X}_m^*$ and $\mathbf{W}$,

$$\begin{aligned}
&q_{\infty,1-\alpha}^2(\mathbb{X}_{n,\mathbf{z}}) \\
&\le \frac{4}{n\delta}\xi_P + \frac{n-m}{n^2}u_P(\mathbf{z},\mathbf{z}) \\
&\quad + \frac{\rho_\alpha}{n^2}\sqrt{\frac{4}{\delta}\big(m(m-1)\Sigma_P^2 + m\xi_P^2 + 2m(n-m)\sigma_\infty^2 + (n-m)^2u_P(\mathbf{z},\mathbf{z})^2\big)} \\
&\le \frac{4}{n\delta}\xi_P + \frac{n-m}{n^2}u_P(\mathbf{z},\mathbf{z}) \\
&\quad + \frac{\rho_\alpha}{n^2\sqrt{\delta}}\Big(2\sqrt{m(m-1)}\Sigma_P + \sqrt{m}\xi_P + \sqrt{2m(n-m)}\sigma_\infty + (n-m)u_P(\mathbf{z},\mathbf{z})\Big)\,, \\
&\le \frac{4}{n\delta}\xi_P + \frac{\epsilon_{m,n}}{n}u_P(\mathbf{z},\mathbf{z}) + \frac{\rho_\alpha}{n\sqrt{\delta}}\Big(2\Sigma_P + \xi_P + \sqrt{2\epsilon_{m,n}}\sigma_\infty + \epsilon_{m,n}u_P(\mathbf{z},\mathbf{z})\Big) \\
&\le \frac{4}{n\delta}\xi_P + \frac{2\rho_\alpha\epsilon_{m,n}}{n\sqrt{\delta}}u_P(\mathbf{z},\mathbf{z}) + \frac{\rho_\alpha}{n\sqrt{\delta}}\big(2\Sigma_P + \xi_P + \sqrt{2}\sigma_\infty\big)\,, \qquad (5.18)
\end{aligned}$$

where the first inequality holds by (5.16) and (5.17) (and using $m\le n$), the second inequality follows from using twice the inequality $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$ for any $a,b\ge 0$, the third inequality holds since $m\le n$ and $\epsilon_{m,n} = (n-m)/n$, and (5.18) holds since $\epsilon_{m,n}\le 1$ and since the conditions $\delta\in(0,1)$ and $\rho_\alpha\ge 1$ imply

$$\frac{\epsilon_{m,n}}{n}u_P(\mathbf{z},\mathbf{z}) \;\le\; \frac{\epsilon_{m,n}}{n\sqrt{\delta}}u_P(\mathbf{z},\mathbf{z}) \;\le\; \frac{\rho_\alpha\epsilon_{m,n}}{n\sqrt{\delta}}u_P(\mathbf{z},\mathbf{z})\,.$$

Substituting the definition of $\rho_\alpha = C\log(1/\alpha)$, and defining the constants $C' \coloneqq 2C$ and $C_{P,1} \coloneqq 4\xi_P$

and $C_{P,2} \coloneqq C(2\Sigma_P + \xi_P + \sqrt{2}\sigma_\infty)$, we have from (5.18) that

$$q^2_{\infty,1-\alpha}(\mathbb{X}_{n,\mathbf{z}}) \;\le\; \frac{C' \log(1/\alpha)\epsilon_{m,n}}{n\sqrt{\delta}} u_P(\mathbf{z},\mathbf{z}) + \frac{C_{P,1}}{n\delta} + \frac{C_{P,2}\log(1/\alpha)}{n\sqrt{\delta}} \,.$$

□

5.A.1.3 Proof of Proposition 5.9

We first show that the moment conditions (i)-(iii) in Lemma 5.7 and Lemma 5.8 are met, and apply these lemmas to conclude the proof.

Step 1: Verifying moment conditions

For any $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$, we can bound the squared Stein kernel as

$$\begin{aligned}
&u_P(\mathbf{x},\mathbf{z})^2 \\
&= \left[\mathbf{s}_p(\mathbf{x})^\top \mathbf{s}_p(\mathbf{z}) h(\mathbf{x}-\mathbf{z}) - \mathbf{s}_p(\mathbf{x})^\top \nabla h(\mathbf{x}-\mathbf{z}) + \mathbf{s}_p(\mathbf{z})^\top \nabla h(\mathbf{x}-\mathbf{z}) + \nabla^\top \nabla h(\mathbf{x}-\mathbf{z})\right]^2 \\
&\le 4\Big[\big(\mathbf{s}_p(\mathbf{x})^\top \mathbf{s}_p(\mathbf{z}) h(\mathbf{x}-\mathbf{z})\big)^2 + \big(\mathbf{s}_p(\mathbf{x})^\top \nabla h(\mathbf{x}-\mathbf{z})\big)^2 + \big(\mathbf{s}_p(\mathbf{z})^\top \nabla h(\mathbf{x}-\mathbf{z})\big)^2 \\
&\qquad + \big(\nabla^\top \nabla h(\mathbf{x}-\mathbf{z})\big)^2\Big] \\
&\le 4\Big[\|\mathbf{s}_p(\mathbf{x})\|_2^2 \|\mathbf{s}_p(\mathbf{z})\|_2^2 h(\mathbf{x}-\mathbf{z})^2 + \|\mathbf{s}_p(\mathbf{x})\|_2^2 \|\nabla h(\mathbf{x}-\mathbf{z})\|_2^2 + \|\mathbf{s}_p(\mathbf{z})\|_2^2 \|\nabla h(\mathbf{x}-\mathbf{z})\|_2^2 \qquad (5.19a) \\
&\qquad + \big(\nabla^\top \nabla h(\mathbf{x}-\mathbf{z})\big)^2\Big] \,, \qquad (5.19b)
\end{aligned}$$

where the first inequality holds since $(a+b+c+d)^2 \le 4(a^2+b^2+c^2+d^2)$ for any $a,b,c,d \in \mathbb{R}$, and the second inequality is due to Cauchy-Schwarz inequality. The assumption $h \in \mathcal{C}_b^2$ immediately implies that the last term in (5.19) is uniformly bounded over $\mathbf{x}, \mathbf{z}$. For the first term, taking expectation over P and supremum over $\mathbf{z}$ gives

$$\begin{aligned}
&\sup_{\mathbf{z}\in\mathbb{R}^d} \mathbb{E}_{\mathbf{X}\sim P}\big[\|\mathbf{s}_p(\mathbf{X})\|_2^2 \|\mathbf{s}_p(\mathbf{z})\|_2^2 h(\mathbf{X}-\mathbf{z})^2\big] \\
&\le \sup_{\mathbf{z}\in\mathbb{R}^d} \|\mathbf{s}_p(\mathbf{z})\|_2^2 \big(\mathbb{E}_{\mathbf{X}\sim P}\big[h(\mathbf{X}-\mathbf{z})^4\big]\big)^{1/2} \big(\mathbb{E}_{\mathbf{X}\sim P}\big[\|\mathbf{s}_p(\mathbf{X})\|_2^4\big]\big)^{1/2} \\
&= C_P^{1/2} \sup_{\mathbf{z}\in\mathbb{R}^d} \|\mathbf{s}_p(\mathbf{z})\|_2^2 \big(\mathbb{E}_{\mathbf{X}\sim P}\big[h(\mathbf{X}-\mathbf{z})^4\big]\big)^{1/2} \,,
\end{aligned}$$

where the first step holds by Hölder's inequality, and $C_P \coloneqq \mathbb{E}_{\mathbf{X}\sim P}\big[\|\mathbf{s}_p(\mathbf{X})\|_2^4\big]$. The RHS of the last line is finite by assumption. Similarly, the second term in (5.19) can also be bounded by Hölder's inequality

$$\sup_{\mathbf{z}\in\mathbb{R}^d} \mathbb{E}_{\mathbf{X}\sim P}\big[\|\mathbf{s}_p(\mathbf{x})\|_2^2 \|\nabla h(\mathbf{x}-\mathbf{z})\|_2^2\big] \;\le\; C_P^{1/2} \sup_{\mathbf{z}\in\mathbb{R}^d} \mathbb{E}_{\mathbf{X}\sim P}\big(\big[\|\nabla h(\mathbf{x}-\mathbf{z})\|_2^4\big]\big)^{1/2} \,,$$

which is finite since $h \in \mathcal{C}_b^2$. For the third term in (5.19),

$$\sup_{\mathbf{z}\in\mathbb{R}^d} \mathbb{E}_{\mathbf{X}\sim P}\big[\|\mathbf{s}_p(\mathbf{z})\|_2^2 \|\nabla h(\mathbf{x}-\mathbf{z})\|_2^2\big] \;=\; \sup_{\mathbf{z}\in\mathbb{R}^d} \|\mathbf{s}_p(\mathbf{z})\|_2^2 \mathbb{E}_{\mathbf{X}\sim P}\big[\|\nabla h(\mathbf{x}-\mathbf{z})\|_2^2\big] \,,$$

which is again finite by assumption. Substituting these bounds into (5.19), we have shown that $\sigma_\infty^2 = \sup_{\mathbf{z}\in\mathbb{R}^d} \mathbb{E}_{\mathbf{X}\sim P}[u_P(\mathbf{X},\mathbf{z})^2] < \infty$, which verifies condition (i).

To proceed, we note that a translation-invariant kernel corresponds to choosing a constant weighting function $w(\mathbf{x}) \equiv 1$ in the tilted kernel of Lemma 5.2, so that we can leverage the derivations in the proof of Lemma 5.2. To show condition (ii), setting $w(\mathbf{x}) \equiv 1$ in (5.28) to simplify $u_P(\mathbf{x},\mathbf{x})$ yields

$$\begin{aligned}\xi_P^2 \coloneqq \mathbb{E}_{\mathbf{X}\sim P}\big[u_P(\mathbf{X},\mathbf{X})^2\big] &= \mathbb{E}_{\mathbf{X}\sim P}\Big[\big(\|\mathbf{s}_p(\mathbf{X})\|_2^2 h(0) - \nabla^\top\nabla h(0)\big)^2\Big] \\ &\le 2h(0)^2\mathbb{E}_{\mathbf{X}\sim P}\big[\|\mathbf{s}_p(\mathbf{X})\|_2^4\big] + 2\big|\nabla^\top\nabla h(0)\big|^2 \\ &< \infty\,,\end{aligned}$$

where the first inequality holds since $(a+b)^2 \le 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$. Finally, (5.30) shows that $u_P(\mathbf{x},\mathbf{x}')^2 \le u_P(\mathbf{x},\mathbf{x})u_P(\mathbf{x}',\mathbf{x}')$ for any $\mathbf{x},\mathbf{x}'$, so

$$\begin{aligned}\Sigma_P^2 \coloneqq \mathrm{Var}_{\mathbf{X},\mathbf{X}'\sim P}(u_P(\mathbf{X},\mathbf{X}')) \le \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim P}[u_P(\mathbf{X},\mathbf{X}')^2] &\le \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim P}[u_P(\mathbf{X},\mathbf{X})u_P(\mathbf{X}',\mathbf{X}')] \\ &\overset{(a)}{=} \big(\mathbb{E}_{\mathbf{X}\sim P}[u_P(\mathbf{X},\mathbf{X})]\big)^2 \\ &\overset{(b)}{\le} \mathbb{E}_{\mathbf{X}\sim P}[u_P(\mathbf{X},\mathbf{X})^2] \\ &= \xi_P^2 < \infty\,,\end{aligned}$$

where (a) holds since $\mathbf{X},\mathbf{X}'$ are i.i.d. and (b) follows from Jensen's inequality. This shows condition (iii).

Step 2: Proving the main claim

Pick any $s \in [0,1)$ and any integer $m' \in [n^{1-s}/2, n)$. Define $\epsilon_{m,n} \coloneqq m'/n$ and $m = n - m'$. Let $C, C_{P,1}, C_{P,2}$ be the constants defined in Lemma 5.8, and define the events

$$\mathcal{A}_1 \coloneqq \big\{D^2(\mathbb{X}_n) \ge \epsilon_{m,n}^2 u_P(\mathbf{z},\mathbf{z}) - t_1(m')\big\}\,, \tag{5.20a}$$

$$\mathcal{A}_2 \coloneqq \left\{q_{B,1-\alpha}^2(\mathbb{X}_n) \le \frac{C\log(1/\alpha)}{n\sqrt{\delta/2}}\epsilon_{m,n}u_P(\mathbf{z},\mathbf{z}) + t_2(m')\right\}, \tag{5.20b}$$

where

$$t_1(m') \coloneqq \frac{2\sigma_\infty\epsilon_{m,n}}{\sqrt{\delta m/2}}\,, \qquad t_2(m') \coloneqq \frac{2C_{P,1}}{n\delta} + \frac{C_{P,2}\log(1/\alpha)}{n\sqrt{\delta/2}}\,,$$

and the dependence on m' in these quantities is through $m = n - m'$. It follows from Lemma 5.7 and Lemma 5.8 applied to $\delta/2$ that the event $\mathcal{A} \coloneqq \mathcal{A}_1 \cap \mathcal{A}_2$ occurs with probability $\Pr(\mathcal{A}) \ge 1 - \delta/2 - \delta/2 = 1 - \delta$. These events allow us to bound $D^2(\mathbb{X}_{n,\mathbf{z}})$ and $q_{B,1-\alpha}^2(\mathbb{X}_{n,\mathbf{z}})$, which then allows us to show that $D^2(\mathbb{X}_{n,\mathbf{z}}) > q_{B,1-\alpha}^2(\mathbb{X}_{n,\mathbf{z}})$ for sufficiently large $\mathbf{z}$. On $\mathcal{A}$, we have

$$\begin{aligned}D^2(\mathbb{X}_n) - q_{B,1-\alpha}^2(\mathbb{X}_{n,\mathbf{z}}) &\ge \epsilon_{m,n}^2 u_P(\mathbf{z},\mathbf{z}) - t_1(m') - \frac{C\log(1/\alpha)}{n\sqrt{\delta/2}}\epsilon_{m,n}u_P(\mathbf{z},\mathbf{z}) - t_2(m') \\ &= \epsilon_{m,n}u_P(\mathbf{z},\mathbf{z})\left(\epsilon_{m,n} - \frac{C'}{n\sqrt{\delta}}\right) - T(m')\,,\end{aligned}$$

where in the last line we have defined $C' \coloneqq \sqrt{2}C\log(1/\alpha)$ and $T(m') \coloneqq t_1(m') + t_2(m')$. This implies

$$\begin{aligned}
&\Pr_{\mathbb{X}^*_m \sim P, \mathbf{W}}\left(D^2(\mathbb{X}_{n,\mathbf{z}}) > \hat{q}^B_{1-\alpha}(\mathbb{X}_{n,\mathbf{z}})\right) \\
&\geq \Pr_{\mathbb{X}^*_m \sim P, \mathbf{W}}\left(D^2(\mathbb{X}_{n,\mathbf{z}}) > \hat{q}^B_{1-\alpha}(\mathbb{X}_{n,\mathbf{z}}) \mid \mathcal{A}\right) \Pr_{\mathbb{X}^*_m \sim P, \mathbf{W}}(\mathcal{A}) \\
&\geq (1-\delta)\Pr_{\mathbb{X}^*_m \sim P, \mathbf{W}}\left(D^2(\mathbb{X}_{n,\mathbf{z}}) > \hat{q}^B_{1-\alpha}(\mathbb{X}_{n,\mathbf{z}}) \mid \mathcal{A}\right) \\
&\geq (1-\delta)\mathbb{1}\left\{\epsilon_{m,n} u_P(\mathbf{z},\mathbf{z})\left(\epsilon_{m,n} - \frac{C'}{n\sqrt{\delta}}\right) - T(m') \;>\; 0\right\}, \qquad (5.21)
\end{aligned}$$

where the first step holds due to the law of total probability. We claim that the indicator function in (5.21) takes value 1 for all $m' \in [n^{1-s}/2, n)$ when the following inequalities hold:

$$n^{\frac{(1-s)}{2}} \;>\; 4C', \qquad \|\mathbf{s}_p(\mathbf{z})\|_2^2 \;>\; \frac{1}{h(0)}\left(4T(n-1)n^{2s} + \nabla_1^\top \nabla_2 h(0)\right). \qquad (5.22)$$

Indeed, since $\delta > 0$ was arbitrary, we can set $\delta = n^{-(1-s)}$, so we have

$$\epsilon_{m,n} - \frac{C'}{n\sqrt{\delta}} \;=\; \epsilon_{m,n} - \frac{C'}{n^{\frac{1}{2}+\frac{s}{2}}} \overset{(a)}{>} \epsilon_{m,n} - \frac{C'}{4C'n^s} \;=\; \epsilon_{m,n} - \frac{1}{4n^s} \overset{(b)}{\geq} \epsilon_{m,n} - \frac{\epsilon_{m,n}}{2} \;=\; \frac{\epsilon_{m,n}}{2}, \qquad (5.23)$$

where (a) holds since the first condition in (5.22) implies $n^{\frac{1}{2}+\frac{s}{2}} = n^{s+\frac{(1-s)}{2}} \geq 4C'n^s$, and (b) holds since the assumption $m' \geq n^{1-s}/2$ implies $n^s \leq n/(2m') = \epsilon_{m,n}/2$. Using this and the second inequality in (5.22),

$$\begin{aligned}
\epsilon_{m,n} u_P(\mathbf{z},\mathbf{z})\left(\epsilon_{m,n} - \frac{C'}{n\sqrt{\delta}}\right) &= \epsilon_{m,n}\left(\|\mathbf{s}_p(\mathbf{z})\|_2^2 h(0) - \nabla_1^\top \nabla_2 h(0)\right)\left(\epsilon_{m,n} - \frac{C'}{n\sqrt{\delta}}\right) \\
&> \epsilon_{m,n} \times 4T(n-1)n^{2s} \times \frac{\epsilon_{m,n}}{2} \\
&\geq \epsilon_{m,n} \times \frac{2T(n-1)}{\epsilon_{m,n}^2} \times \frac{\epsilon_{m,n}}{2} \\
&= T(n-1) \\
&\geq T(m'),
\end{aligned}$$

where the first equality holds by setting $w(\mathbf{x}) \equiv 1$ in (5.28) to compute $u_P(\mathbf{z},\mathbf{z})$, the second line follows from the second inequality in (5.22) and (5.23), the third line holds since $n^{2s} \geq n^{2s}/(m')^2 = \epsilon_{m,n}^{-2}$, and the last line holds as direct computation shows $T(n-1) \geq T(m')$ since $m' < n$ by assumption. To summarize, we have shown that for any n and $\mathbf{z}$ satisfying (5.22), the following holds

$$\Pr_{\mathbb{X}^*_m \sim P, \mathbf{W}}\left(D^2(\mathbb{X}_{n,\mathbf{z}}) > q^2_{\infty,1-\alpha}(\mathbb{X}_{n,\mathbf{z}})\right) \;\geq\; 1-\delta \;=\; 1 - n^{-(1-s)}.$$

It remains to show that there exists $\mathbf{z}$ not dependent on m' that satisfies the second inequality in (5.22), which would imply the claimed result. Such $\mathbf{z}$ always exists, because, for the second inequality in (5.22), the RHS does not depend on $\mathbf{z}$ nor m' and is finite since $\epsilon_{m,n} = m'/n \in (n^{-s}/2, 1)$ by assumption, while the LHS explodes with $\mathbf{z}$ as we have assumed $\mathbf{z} \mapsto \|\mathbf{s}_p(\mathbf{z})\|_2$ is unbounded. This concludes the proof.

5.A.1.4 Proof of Theorem 5.1

It suffices to show the claim for any sequence of the form $\epsilon_n = n^{-s}$, where $s \in [0,1)$ is arbitrary. Let the random variable M' be defined as at the start of Section 5.A.1. For any random sample $\mathbb{X}_n$, define the event $\mathcal{A}(\mathbb{X}_n) = \{D^2(\mathbb{X}_n) > q^2_{\infty,1-\alpha}(\mathbb{X}_n)\}$. For n sufficiently large so that $n^{(1-s)/2} > 2\sqrt{2}C\log(1/\alpha)$, we can apply Proposition 5.9 to conclude that, for all n, there exists $\mathbf{z}_n \in \mathbb{R}^d$ such that the following holds for all $m' \in [n^{1-s}/2, n) = [\epsilon_n n/2, n)$, we have

$$\begin{aligned} \Pr_{\mathbb{X}_n\sim Q,\mathbf{W}}\left(\mathcal{A}(\mathbb{X}_n) \mid M' = m'\right) &= \Pr_{\mathbb{X}^*_{n-m'}\sim P,\mathbf{W}}\left(D^2(\mathbb{X}_{n,\mathbf{z}}) > q^2_{\infty,1-\alpha}(\mathbb{X}_{n,\mathbf{z}})\right) \\ &\geq 1 - n^{-(1-s)}\,, \end{aligned} \tag{5.24}$$

where we have used the shorthand notation $\mathbb{X}_{n,\mathbf{z}_n} = \mathbb{X}^*_{n-m'} \cup \{\mathbf{z}_n\}^{m'}$. We will show that the sequence of probability measures $Q := (1-\epsilon_n)P + \epsilon_n\delta_{\mathbf{z}_n}$ satisfies

$$\Pr_{\mathbb{X}_n\sim Q,\mathbf{W}}\left(D^2(\mathbb{X}_n) > q^2_{\infty,1-\alpha}(\mathbb{X}_n)\right) - \Pr_{\mathbb{X}^*_n\sim P,\mathbf{W}}\left(D^2(\mathbb{X}^*_n) > q^2_{\infty,1-\alpha}(\mathbb{X}^*_n)\right) \to 1-\alpha\,, \tag{5.25}$$

which will imply the claimed result of our theorem. Using a concentration inequality for Binomial distributions (e.g., Chung and Lu, 2002, Lemma 2.1) applied to the case of a Binomial(n,ϵ_n), we can show that the event $\mathcal{B} := \{\epsilon_n n/2 \leq M' \leq 3\epsilon_n n/2\}$ holds with high probability:

$$\begin{aligned} \Pr_{M'}(\mathcal{B}) &= 1 - \Pr_{M'}\left(\left|M' - \epsilon_n n\right| < \frac{\epsilon_n n}{2}\right) \\ &\geq 1 - \exp\left(-\frac{(\epsilon_n n)^2}{8\epsilon_n n}\right) - \exp\left(-\frac{(\epsilon_n n)^2}{2(\epsilon_n n + \epsilon_n n/6)}\right) \\ &= 1 - \exp\left(-\frac{\epsilon_n n}{8}\right) - \exp\left(-\frac{3\epsilon_n n}{7}\right) \\ &= 1 - o\big(n^{-(1-s)}\big)\,, \end{aligned} \tag{5.26}$$

where the last line holds since $r \in [0,1)$ implies $\exp(-C\epsilon_n n) = \exp(-Cn^{1-r}) = o\big(n^{-(1-r)/2}\big)$ for any constant $C > 0$. Define the index set I_n to be the set of integers in the interval $[\epsilon_n n/2, 3\epsilon_n n/3]$. Then

$$\begin{aligned} &\Pr_{\mathbb{X}_n\sim Q,\mathbf{W}}\left(D^2(\mathbb{X}_n) > q^2_{\infty,1-\alpha}(\mathbb{X}_n)\right) \\ &\geq \Pr_{\mathbb{X}_n\sim Q,\mathbf{W}}\left(\mathcal{A}(\mathbb{X}_n) \cap \mathcal{B}\right) \\ &\overset{(a)}{=} \sum_{m'\in I_n} \Pr_{\mathbb{X}_n\sim Q,\mathbf{W}}\left(\mathcal{A}(\mathbb{X}_n) \mid M' = m'\right)\Pr(M' = m') \\ &\geq \Big(\min_{m'\in I_n} \Pr_{\mathbb{X}_n\sim Q,\mathbf{W}}\left(\mathcal{A}(\mathbb{X}_n) \mid M' = m'\right)\Big) \sum_{m'=m'_0}^{n} \Pr(M' = m') \\ &\overset{(b)}{=} \min_{m'\in I_n} \Pr_{\mathbb{X}_n\sim Q,\mathbf{W}}\left(\mathcal{A}(\mathbb{X}_n) \mid M' = m'\right)\Pr(\mathcal{B}) \\ &\overset{(c)}{\geq} \left(1 - n^{-(1-s)}\right)\left(1 - o\big(n^{-(1-s)}\big)\right) = 1 - \mathcal{O}\big(n^{-(1-s)}\big)\,, \end{aligned} \tag{5.27}$$

where the first inequality follows from the law of total probability, (a) and (b) hold by the partition $\mathcal{B} = \cup_{m'\in I_n}\{M' = m'\}$ and again the law of total probability, and (c) follows from (5.24) and (5.26). On the other hand, the assumed moment conditions imply $\mathbb{E}_{\mathbf{X}^*\sim P}[u_P(\mathbf{X}^*,\cdot)] = 0$ as argued in the paragraph

before (5.11) and $\mathbb{E}_{\mathbf{X}_1^*,\mathbf{X}_2^*\sim P}[u_P(\mathbf{X}_1^*,\mathbf{X}_2^*)^2] \leq \mathbb{E}_{\mathbf{X}^*\sim P}[u_P(\mathbf{X}^*,\mathbf{X}^*)^2] < \infty$ as argued in Section 5.A.1.3, so we can apply Arcones and Gine (1992, Theorem 3.5) with $r = 2$ to yield the asymptotic validity of the KSD test when $Q = P$, namely

$$\Pr_{\mathbb{X}_n^*\sim P,\mathbf{W}}\left(D^2(\mathbb{X}_n^*) > q^2_{\infty,1-\alpha}(\mathbb{X}_n^*)\right) \;\to\; \alpha\;.$$

Combining this with (5.27) shows the desired convergence (5.25).

5.A.1.5 Implications for KSDs on Non-Euclidean Spaces

The core argument underpinning the proofs in Theorem 5.1 is that the function $\mathbf{z} \mapsto u_P(\mathbf{z},\mathbf{z})$ is unbounded on the support of the model, which in our case is $\mathbb{R}^d$. The proof per se does not rely on the support being an Euclidean space. Consequently, this argument can potentially be extended to KSDs constructed for models defined on other spaces, such as models for Riemannian data (Xu and Matsuda, 2021), sequence data with varying dimensionality (Baum et al., 2023), functional data (Wynne et al., 2022), or censored time-to-event data (Fernandez et al., 2020).

5.A.2 Proof of Lemma 5.2

Direct expansion of the Stein kernel in (5.1) with a tilted kernel k gives

$$\begin{aligned}
u_P(\mathbf{x},\mathbf{x}') \;=\; & \langle w(\mathbf{x})\mathbf{s}_p(\mathbf{x}),\; w(\mathbf{x}')\mathbf{s}_p(\mathbf{x}')\rangle h(\mathbf{x},\mathbf{x}') \\
& + \langle w(\mathbf{x}')\mathbf{s}_p(\mathbf{x}'),\; \nabla w(\mathbf{x})\rangle h(\mathbf{x}-\mathbf{x}') + \langle w(\mathbf{x}')\mathbf{s}_p(\mathbf{x}'), \nabla h(\mathbf{x}-\mathbf{x}')\rangle w(\mathbf{x}) \\
& + \langle w(\mathbf{x})\mathbf{s}_p(\mathbf{x}),\; \nabla w(\mathbf{x}')\rangle h(\mathbf{x}-\mathbf{x}') - \langle w(\mathbf{x})\mathbf{s}_p(\mathbf{x}), \nabla h(\mathbf{x}-\mathbf{x}')\rangle w(\mathbf{x}') \\
& + \langle \nabla w(\mathbf{x}),\; \nabla w(\mathbf{x}')\rangle h(\mathbf{x}-\mathbf{x}') + \langle w(\mathbf{x})\nabla h(\mathbf{x}-\mathbf{x}'),\; \nabla w(\mathbf{x}')\rangle \\
& - \langle \nabla w(\mathbf{x}),\; w(\mathbf{x}')\nabla h(\mathbf{x}-\mathbf{x}')\rangle - w(\mathbf{x})w(\mathbf{x}')\nabla^\top \nabla h(\mathbf{x}-\mathbf{x}') \\
=\; & \langle \mathbf{s}_{p,w}(\mathbf{x}),\; \mathbf{s}_{p,w}(\mathbf{x}')\rangle h(\mathbf{x}-\mathbf{x}') \\
& + \langle \mathbf{s}_{p,w}(\mathbf{x}'),\; \nabla w(\mathbf{x})\rangle h(\mathbf{x}-\mathbf{x}') + \langle \mathbf{s}_{p,w}(\mathbf{x}'),\; \nabla h(\mathbf{x}-\mathbf{x}')\rangle w(\mathbf{x}) \\
& + \langle \mathbf{s}_{p,w}(\mathbf{x}),\; \nabla w(\mathbf{x}')\rangle h(\mathbf{x}-\mathbf{x}') - \langle \mathbf{s}_{p,w}(\mathbf{x}),\; \nabla h(\mathbf{x}-\mathbf{x}')\rangle w(\mathbf{x}') \\
& + \langle \nabla w(\mathbf{x}),\; \nabla w(\mathbf{x}')\rangle h(\mathbf{x}-\mathbf{x}') + \langle \nabla h(\mathbf{x}-\mathbf{x}'),\; \nabla w(\mathbf{x}')\rangle w(\mathbf{x}) \\
& - \langle \nabla w(\mathbf{x}),\; \nabla h(\mathbf{x}-\mathbf{x}')\rangle w(\mathbf{x}') - w(\mathbf{x})w(\mathbf{x}')\nabla^\top \nabla h(\mathbf{x}-\mathbf{x}')\;,
\end{aligned}$$

where $\mathbf{s}_{p,w}(\cdot) \coloneqq w(\cdot)\mathbf{s}_p(\cdot)$ and $\nabla h(\mathbf{x}-\mathbf{x}')$ denotes ∇h evaluated at $\mathbf{x}-\mathbf{x}'$. By setting $\mathbf{x} = \mathbf{x}'$ and eliminating identical terms,

$$\begin{aligned}
& u_P(\mathbf{x},\mathbf{x}) \\
=\; & \|\mathbf{s}_{p,w}(\mathbf{x})\|_2^2 h(0) + 2\langle \mathbf{s}_{p,w}(\mathbf{x}),\; \nabla w(\mathbf{x})\rangle h(0) + \|\nabla w(\mathbf{x})\|_2^2 h(0) - w(\mathbf{x})^2\nabla^\top\nabla h(0) && (5.28)\\
\leq\; & \|\mathbf{s}_{p,w}(\mathbf{x})\|_2^2 h(0) + 2\|\mathbf{s}_{p,w}(\mathbf{x})\|_2\|\nabla w(\mathbf{x})\|_2 h(0) + \|\nabla w(\mathbf{x})\|_2^2 h(0) + w(\mathbf{x})^2|\nabla^\top\nabla h(0)|\;, && (5.29)
\end{aligned}$$

where the last line follows from Cauchy-Schwarz inequality. The RHS of (5.29) is bounded over $\mathbf{x}$ assuming the stated conditions on h, w and the supremum of $\|\mathbf{s}_p(\mathbf{x})w(\mathbf{x})\|_2 = \|s_{p,w}(\mathbf{x})\|_2$. To prove the

bound on $u_P(\mathbf{x}, \mathbf{x}')$, we denote by $\mathcal{H}_u$ the RKHS associated with u_P, which is a reproducing kernel (see, for example, Barp et al., 2022, Theorem 2.8). We us the reproducing property of the Stein reproducing kernel u_P and the Cauchy-Schwarz inequality for the corresponding RKHS norm $\|\cdot\|_{\mathcal{H}_u}$ yields

$$u_P(\mathbf{x}, \mathbf{x}') = \langle u_P(\cdot, \mathbf{x}), u_P(\cdot, \mathbf{x}')\rangle_{\mathcal{H}_u} \leq \|u_P(\cdot, \mathbf{x})\|_{\mathcal{H}_u}\|u_P(\cdot, \mathbf{x}')\|_{\mathcal{H}_u} = u_P(\mathbf{x}, \mathbf{x})^{\frac{1}{2}} u_P(\mathbf{x}', \mathbf{x}')^{\frac{1}{2}} . \quad (5.30)$$

Hence, $\sup_{\mathbf{x},\mathbf{x}'\in\mathbb{R}^d} u_P(\mathbf{x}, \mathbf{x}') \leq \infty$. In particular, $u_P(\mathbf{x}, \mathbf{x})^{1/2}$ is well-defined since $u_P(\mathbf{x}, \mathbf{x}) \geq 0$, which is because the Stein kernel u_P is positive definite.

5.A.3 Proof of Theorem 5.3 and Related Preliminary Results

Sketch of proof We will show a general result in Proposition 5.14, which states that the robust-KSD test that rejects $H_0^{\mathrm{C}} : \mathcal{B}^{\mathrm{KSD}}(P; \theta)$ if $\Delta_\theta(\mathbb{X}_n) := \max(0, D(\mathbb{X}_n) - \theta) > q_{B,1-\alpha}(\mathbb{X}_n)$ is qualitatively robust for any $\theta \geq 0$. This immediately shows Theorem 5.3 by setting $\theta = 0$. To show Proposition 5.14, we follow a similar approach in the proof of Theorem 5.1, where we will first show that the result holds conditionally on the number of contaminated data, and use a high-probability argument to complete the proof.

The rest of this section is organized as follows:

- Section 5.A.3.1 shows Lemma 5.11, which states that the bootstrap quantiles $q_{B,1-\alpha}(\mathbb{X}_n)$ and $q_{B,1-\alpha}(\mathbb{X}_n^*)$ computed using any two samples $\mathbb{X}_n$ and $\mathbb{X}_n^*$ that differ by at most m' elements are close to each other with high probability.
- Section 5.A.3.2 shows Lemma 5.12, which says that the difference in the exceedance probabilities of $\Delta_\theta(\mathbb{X}_n)$ and of $\Delta_\theta(\mathbb{X}_n^*)$ is small.
- Section 5.A.3.3 shows Lemma 5.13, which says that the rejection probability of a robust-KSD test using $\mathbb{X}_n$ is close to that of a test using $\mathbb{X}_n^*$, assuming $\mathbb{X}_n$ and $\mathbb{X}_n^*$ differ by no more than $o(n^{1/2})$ elements. Its proof relies on Lemma 5.11 and Lemma 5.12.
- Section 5.A.3.4 uses Lemma 5.13 to show Proposition 5.14, which says that the robust-KSD test is qualitatively robust for any $\theta \geq 0$. This result immediately shows Theorem 5.3, the proof of which is also contained in this section.

Lemma 5.11. *Assume k is a tilted kernel satisfying the conditions in Lemma 5.2. Then there exists absolute constants $C_1, C_2 > 0$ such that, for any $\delta > 0$ and any (deterministic) realisations $\mathbb{X}_n = \{\mathbf{x}_i\}_{i=1}^n$ and $\mathbb{X}_n^* = \{\mathbf{x}_i^*\}_{i=1}^n$ that differ by at most m' elements, we have*

$$\Pr_{\mathbf{W}}\left(\left|D^2_{\mathbf{W}}(\mathbb{X}_n) - D^2_{\mathbf{W}}(\mathbb{X}_n^*)\right| \leq \frac{\tau_\infty \epsilon_{m,n}^{\frac{1}{2}}}{n}\left(C_1 + C_2 \log\left(\frac{8}{\delta}\right)\right)\right) \geq 1 - \delta ,$$

where $D^2_{\mathbf{W}}(\mathbb{X}_n)$, defined in (5.3), is the bootstrap sample computed using $\mathbb{X}_n$, and $\epsilon_{m,n} = m'/n$. Moreover,

the above inequality implies

$$\Pr_{\mathbf{W}}\left(|q_{\infty,1-\alpha}(\mathbb{X}_n) - q_{\infty,1-\alpha}(\mathbb{X}_n^*)| \;\leq\; \frac{\tau_\infty^{\frac{1}{2}}\epsilon_{m,n}^{\frac{1}{4}}}{n^{\frac{1}{2}}}\sqrt{C_1 + C_2\log\left(\frac{16}{\delta}\right)}\right) \geq\; 1-\delta\,.$$

Lemma 5.12. *Assume $\mathbb{E}_{\mathbf{X}\sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2] < \infty$ and that k is a tilted kernel satisfying the conditions in Lemma 5.2. For any $\delta, \gamma > 0$ and any integer $m' \in [0, n]$, it holds that*

$$\begin{aligned}&\sup_{\mathbf{z}_1,\ldots,\mathbf{z}_{m'}\in\mathbb{R}^d}\left|\Pr_{\mathbb{X}_m^*\sim P}\left(D\left(\mathbb{X}_{n-m'}^* \cup \{\mathbf{z}_i\}_{i=1}^{m'}\right) \;>\; \gamma\right) - \Pr_{\mathbb{X}_n^*\sim P}\left(D(\mathbb{X}_n^*) \;>\; \gamma\right)\right| \\ &\leq\; \Pr_{\mathbb{X}_n^*\sim P}\left(|D(\mathbb{X}_n^*) - \gamma| \;\leq\; t_{m,n,\delta}\right) + \delta\,,\end{aligned}$$

where $t_{m,n,\delta} \coloneqq \left(4\sqrt{2}\tau_\infty\epsilon_{m,n}/\sqrt{\delta n} + 2\tau_\infty\epsilon_{m,n}^2\right)^{1/2}$.

Lemma 5.13. *Assume $\mathbb{E}_{\mathbf{X}\sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2] < \infty$ and that k is a tilted kernel satisfying the conditions in Lemma 5.2. For any $\theta \geq 0$, any $f = f(n) = o(n^{1/2})$ and any $\delta > 0$, there exists n_0 such that for any $n \geq n_0$, we have*

$$\max_{m'\leq f(n)} \omega(m') \;<\; \delta\,,$$

where the max *is over all integers non-negative $m' \leq f(n)$, and*

$$\begin{aligned}\omega(m') \;\coloneqq\; \big|\Pr_{\mathbb{X}_m^*\sim P,\mathbf{W}}\left(\Delta_\theta(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) > q_{\infty,1-\alpha}(\mathbb{X}_m^* \cup \mathbb{Z}_{m'})\right) \\ - \Pr_{\mathbb{X}_n^*\sim P,\mathbf{W}}\left(\Delta_\theta(\mathbb{X}_n^*) > q_{\infty,1-\alpha}(\mathbb{X}_n^*)\right)\big|\,.\end{aligned}$$

Proposition 5.14. *Assume $\mathbb{E}_{\mathbf{X}\sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2] < \infty$ and that k is a tilted kernel satisfying the conditions in Lemma 5.2. If $\theta \geq 0$ and $\epsilon_n = o(n^{-1/2})$, then as $n \to \infty$,*

$$\sup_{Q\in\mathcal{P}(P;\epsilon_n)}\left|\Pr_{\mathbb{X}_n\sim Q,\mathbf{W}}\left(\Delta_\theta(\mathbb{X}_n) > q_{\infty,1-\alpha}(\mathbb{X}_n)\right) - \Pr_{\mathbb{X}_n^*\sim P,\mathbf{W}}\left(\Delta_\theta(\mathbb{X}_n^*) > q_{\infty,1-\alpha}(\mathbb{X}_n^*)\right)\right| \;\to\; 0\,,$$

where $\mathcal{P}(P;\epsilon_n)$ is the Huber's contamination model defined in (5.6).

5.A.3.1 Proof of Lemma 5.11

Without loss of generality, we assume $\mathbb{X}_n$ and $\mathbb{X}_n^*$ are ordered so that $\mathbf{x}_i = \mathbf{x}_i^*$ for $i = 1, \ldots, m$. We also define the $n \times n$ matrices $U \coloneqq (u_{ij})_{1\leq i,j\leq n}$ and $U' \coloneqq (u'_{ij})_{1\leq i,j\leq n}$, where $u_{ij} \coloneqq u_P(\mathbf{x}_i, \mathbf{x}_j)$ and $u'_{ij} \coloneqq u_P(\mathbf{x}_i^*, \mathbf{x}_j^*)$. Since u_P is a reproducing kernel, U and U' are Gram matrices, thus symmetric and positive semi-definite. Furthermore, we define $W_i^0 \coloneqq W_i - 1$ for any $i = 1, \ldots, n$, and define the vectors $\mathbf{V}_1 \coloneqq (W_1^0, \ldots, W_m^0, 0, \ldots, 0)^\top$ and $\mathbf{V}_2 \coloneqq (0, \ldots, 0, W_{m+1}^0, \ldots, W_n^0)^\top$, so that $\mathbf{W}^0 = \mathbf{W} - 1 = \mathbf{V}_1 + \mathbf{V}_2$. It follows that the bootstrap sample $D_{\mathbf{W}}^2(\mathbb{X}_n)$ defined in (5.3) can be decomposed as

$$D_{\mathbf{W}}^2(\mathbb{X}_n) \;=\; \frac{1}{n^2}\sum_{1\leq i,j\leq n} W_i^0 W_j^0 u_{ij} \;=\; \frac{1}{n^2}(\mathbf{W}^0)^\top U\mathbf{W}^0 \;=\; \frac{1}{n^2}\left(\mathbf{V}_1^\top U\mathbf{V}_1 + \mathbf{V}_2^\top U\mathbf{V}_2 + 2\mathbf{V}_1^\top U\mathbf{V}_2\right)\,.$$

Similarly, we can decompose $D^2_{\mathbf{W}}(\mathbb{X}^*_n)$ as

$$D^2_{\mathbf{W}}(\mathbb{X}^*_n) \;=\; \frac{1}{n^2}\Big(\mathbf{V}_1^\top U'\mathbf{V}_1 + \mathbf{V}_2^\top U'\mathbf{V}_2 + 2\mathbf{V}_1^\top U'\mathbf{V}_2\Big)\,.$$

Since $\mathbf{x}_i = \mathbf{x}^*_i$ for all $i \le m$ by construction, we have $u_{ij} = u'_{ij}$ for $1 \le i,j \le m$, so $\mathbf{V}_1^\top U\mathbf{V}_1 = \sum_{1\le i,j\le m} W^0_i W^0_j u_{ij} = \sum_{1\le i,j\le m} W^0_i W^0_j u'_{ij} = \mathbf{V}_1^\top U'\mathbf{V}_1$, and we can bound the following difference as

$$\begin{aligned}
&n^2\big|D^2_{\mathbf{W}}(\mathbb{X}_n) - D^2_{\mathbf{W}}(\mathbb{X}^*_n)\big| \\
&= \big|\mathbf{V}_2^\top U\mathbf{V}_2 + 2\mathbf{V}_1^\top U\mathbf{V}_2 - \mathbf{V}_2^\top U'\mathbf{V}_2 - 2\mathbf{V}_1^\top U'\mathbf{V}_2\big| \\
&\le |\mathbf{V}_2^\top U\mathbf{V}_2| + 2|\mathbf{V}_1^\top U\mathbf{V}_2| + |\mathbf{V}_2^\top U'\mathbf{V}_2| + 2|\mathbf{V}_1^\top U'\mathbf{V}_2| \\
&\le |\mathbf{V}_2^\top U\mathbf{V}_2| + 2|\mathbf{V}_1^\top U\mathbf{V}_1|^{\frac12}|\mathbf{V}_2^\top U\mathbf{V}_2|^{\frac12} + |\mathbf{V}_2^\top U'\mathbf{V}_2| + 2|\mathbf{V}_1^\top U'\mathbf{V}_1|^{\frac12}|\mathbf{V}_2^\top U'\mathbf{V}_2|^{\frac12}\,, \qquad (5.31)
\end{aligned}$$

where the last line follows from Cauchy-Schwarz inequality applied to the U-weighted inner product $(\mathbf{x},\mathbf{x}') \mapsto \mathbf{x}^\top U\mathbf{x}'$ and the U'-weighted inner product $(\mathbf{x},\mathbf{x}') \mapsto \mathbf{x}^\top U'\mathbf{x}'$, which are well-defined since U, U' are positive semi-definite. The proof proceeds by bounding each term separately. We discuss how to bound $|\mathbf{V}_1 U\mathbf{V}_1|$ and $|\mathbf{V}_2 U\mathbf{V}_2|$, and the argument for U' is identical. We first define the $n\times n$ matrix with $(A_1)_{ij} = u_{ij}$ for $i,j \le m$ and $(A_1)_{ij} = 0$, so that we can write $\mathbf{V}_1^\top U\mathbf{V}_1 = \mathbf{W}^\top A_1\mathbf{W}$. We can apply the Hanson-Wright lemma (Lemma 5.10) to conclude that there exists positive constants C such that, for any $\delta > 0$ and any almost-sure sequence $\mathbb{X}_n$, the following event occurs with probability at least $1 - \delta/4$,

$$\begin{aligned}
|\mathbf{V}_1^\top A\mathbf{V}_1| \;=\; |\mathbf{W}^\top A_1\mathbf{W}| \;&\le\; \left|-\frac1n\sum_{1\le i,j\le m} u_{ij} + \sum_{1\le i\le m} u_{ii}\right| + C\log\left(\frac{8}{\delta}\right)\left(\sum_{1\le i,j\le m} u_{ij}^2\right)^{\frac12} \\
&\le\; \frac1n\sum_{1\le i,j\le m}|u_{ij}| + \sum_{1\le i\le m}|u_{ii}| + C\log\left(\frac{8}{\delta}\right)\left(\sum_{1\le i,j\le m} u_{ij}^2\right)^{\frac12} \\
&\le\; \frac{m^2}{n}\tau_\infty + m\tau_\infty + Cm\tau_\infty\log\left(\frac{8}{\delta}\right) \\
&\le\; 2m\tau_\infty + Cm\tau_\infty\log\left(\frac{8}{\delta}\right)\,, \qquad (5.32)
\end{aligned}$$

where the second last inequality holds since $|u_{ij}| \le \sup_{\mathbf{x},\mathbf{x}'\in\mathbb{R}^d}|u_P(\mathbf{x},\mathbf{x}')| \le \tau_\infty$, and the last line holds since $m \le n$. Similarly, defining the matrix A_2 with $(A_2)_{ij} = 0$ for $i,j \le m$ and $(A_2)_{ij} = u_{ij}$ otherwise, we can write $\mathbf{V}_2^\top U\mathbf{V}_2 = \mathbf{W}^\top A_2\mathbf{W}$, and the same argument as before shows that the following holds with probability at least $1 - \delta/4$,

$$\begin{aligned}
|\mathbf{v}_2^\top A\mathbf{v}_2| \;=\; |\mathbf{W}^\top A_2\mathbf{W}| \;&\le\; \left|-\frac1n\sum_{m<i,j\le n} u_{ij} + \sum_{m<i\le n} u_{ii}\right| + C\log\left(\frac{8}{\delta}\right)\left(\sum_{m<i,j\le n} u_{ij}^2\right)^{\frac12} \\
&\le\; \frac{(n-m)^2\tau_\infty}{n} + (n-m)\tau_\infty + C(n-m)\tau_\infty\log\left(\frac{8}{\delta}\right) \\
&\le\; 2(n-m)\tau_\infty + C(n-m)\tau_\infty\log\left(\frac{8}{\delta}\right)\,, \qquad (5.33)
\end{aligned}$$

Combining (5.32) and (5.33), we conclude that the following holds with probability at least $1-\delta/2$,

$$\begin{aligned}
&|\mathbf{V}_2^\top U \mathbf{V}_2| + 2|\mathbf{V}_1^\top U \mathbf{V}_1|^{\frac{1}{2}} |\mathbf{V}_2^\top U \mathbf{V}_2|^{\frac{1}{2}} \\
&\leq \left(2(n-m)\tau_\infty + C(n-m)\tau_\infty \log\left(\frac{8}{\delta}\right)\right) \\
&\quad + 2\left(2m\tau_\infty + Cm\tau_\infty \log\left(\frac{8}{\delta}\right)\right)^{\frac{1}{2}} \left(2(n-m)\tau_\infty + C(n-m)\tau_\infty \log\left(\frac{8}{\delta}\right)\right)^{\frac{1}{2}} \\
&= \left((n-m) + 2m^{\frac{1}{2}}(n-m)^{\frac{1}{2}}\right) \times \tau_\infty \left(2 + C\log\left(\frac{8}{\delta}\right)\right) .
\end{aligned}$$

The same argument shows that the last two terms of (5.31) can be bounded by the same quantity on an event with probability at least $1-\delta/2$. To conclude, we have shown that with probability at least $1-\delta$,

$$\begin{aligned}
\left|D^2_{\mathbf{W}}(\mathbb{X}_n) - D^2_{\mathbf{W}}(\mathbb{X}_n^*)\right| &\leq \frac{2}{n^2} \times \left((n-m) + 2(n-m)^{\frac{1}{2}} m^{\frac{1}{2}}\right) \times \tau_\infty \left(2 + C\log\left(\frac{8}{\delta}\right)\right) \\
&= \frac{2}{n} \times \frac{(n-m)^{\frac{1}{2}}}{n^{\frac{1}{2}}} \left(\frac{(n-m)^{\frac{1}{2}}}{n^{\frac{1}{2}}} + \frac{2m^{\frac{1}{2}}}{n^{\frac{1}{2}}}\right) \times \tau_\infty \left(2 + C\log\left(\frac{8}{\delta}\right)\right) \\
&\leq \frac{6\tau_\infty \epsilon_{m,n}^{\frac{1}{2}}}{n} \left(2 + C\log\left(\frac{8}{\delta}\right)\right) = \frac{\tau_\infty \epsilon_{m,n}^{\frac{1}{2}}}{n} \left(12 + 6C\log\left(\frac{8}{\delta}\right)\right) ,
\end{aligned}$$

where in the second last line we have defined $\epsilon_{m,n} = (n-m)/n$ and used $n-m \leq n$ and $m \leq n$. Defining $C_1 = 12$ and $C_2 = 6C$ shows the first claim. The second claim can be shown by first noting that the above inequality implies that their $(1-\alpha)$-quantiles must satisfy

$$\left|q^2_{\infty,1-\alpha}(\mathbb{X}_n) - q^2_{\infty,1-\alpha}(\mathbb{X}_n^*)\right| \leq \frac{\tau_\infty \epsilon_{m,n}^{\frac{1}{2}}}{n} \left(C_1 + C_2 \log\left(\frac{8}{\delta}\right)\right) =: \rho_{m,n,\delta} .$$

We will argue separately for the two cases $(q_{\infty,1-\alpha}(\mathbb{X}_n) + q_{\infty,1-\alpha}(\mathbb{X}_n^*)) \geq \rho_{m,n,\delta}^{1/2}$ and $(q_{\infty,1-\alpha}(\mathbb{X}_n) + q_{\infty,1-\alpha}(\mathbb{X}_n^*)) < \rho_{m,n,\delta}^{1/2}$. In the former case, the above inequality implies

$$\begin{aligned}
|q_{\infty,1-\alpha}(\mathbb{X}_n) - q_{\infty,1-\alpha}(\mathbb{X}_n^*)| &= \frac{\left|q_{\infty,1-\alpha}(\mathbb{X}_n) - q_{\infty,1-\alpha}(\mathbb{X}_n^*)\right| \times \left|q_{\infty,1-\alpha}(\mathbb{X}_n) + q_{\infty,1-\alpha}(\mathbb{X}_n^*)\right|}{q_{\infty,1-\alpha}(\mathbb{X}_n) + q_{\infty,1-\alpha}(\mathbb{X}_n^*)} \\
&= \frac{\left|q^2_{\infty,1-\alpha}(\mathbb{X}_n) - q^2_{\infty,1-\alpha}(\mathbb{X}_n^*)\right|}{q_{\infty,1-\alpha}(\mathbb{X}_n) + q_{\infty,1-\alpha}(\mathbb{X}_n^*)} \\
&\leq \frac{\rho_{m,n,\delta}}{\rho_{m,n,\delta}^{\frac{1}{2}}} = \rho_{m,n,\delta}^{\frac{1}{2}} .
\end{aligned}$$

When instead $(q_{\infty,1-\alpha}(\mathbb{X}_n) + q_{\infty,1-\alpha}(\mathbb{X}_n^*)) < \rho_{m,n,\delta}^{1/2}$, we have

$$\begin{aligned}
|q_{\infty,1-\alpha}(\mathbb{X}_n) - q_{\infty,1-\alpha}(\mathbb{X}_n^*)| \leq \max(q_{\infty,1-\alpha}(\mathbb{X}_n),\, q_{\infty,1-\alpha}(\mathbb{X}_n^*)) &\leq q_{\infty,1-\alpha}(\mathbb{X}_n) + q_{\infty,1-\alpha}(\mathbb{X}_n^*) \\
&\leq \rho_{m,n,\delta}^{\frac{1}{2}} ,
\end{aligned}$$

which combined with the previous inequality shows the second claim.

5.A.3.2 Proof of Lemma 5.12

Pick $\{\mathbf{z}_i\}_{i=1}^{m'} \subset \mathbb{R}^d$, and define $\mathbb{Z}_{m'} \coloneqq \{\mathbf{z}_i\}_{i=1}^{m'}$. Decomposing the test statistic $D^2(\mathbb{X}_m^* \cup \mathbb{Z}_{m'})$ using a similar approach as in (5.10), we have

$$D^2(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) \;=\; \frac{m^2}{n^2} D^2(\mathbb{X}_m^*) + \frac{2}{n^2} S_m + \frac{1}{n^2} \sum_{m<i,j\le n} u_P(\mathbf{z}_i, \mathbf{z}_j)\;, \tag{5.34}$$

where $S_m \coloneqq \sum_{i=1}^m T_i$ and $T_i \coloneqq \sum_{j=m+1}^n u_P(\mathbf{X}_i^*, \mathbf{z}_j)$. Similarly, the test statistic computed using the pure sample $\mathbb{X}_n^*$ can be written as

$$D^2(\mathbb{X}_n^*) \;=\; \frac{m^2}{n^2} D^2(\mathbb{X}_m^*) + \frac{2}{n^2} S_m^* + \frac{1}{n^2} \sum_{m<i,j\le n} u_P(\mathbf{X}_i^*, \mathbf{X}_j^*)\;,$$

where $S_m^* \coloneqq \sum_{i=1}^m T_i^*$ and $T_i^* \coloneqq \sum_{j=m+1}^n u_P(\mathbf{X}_i^*, \mathbf{X}_j^*)$. Using a triangle inequality to bound their difference,

$$\begin{aligned} n^2\left|D^2(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) - D^2(\mathbb{X}_n^*)\right| \;&=\; \left|2S_m + \sum_{m<i,j\le n} u_P(\mathbf{z}_i, \mathbf{z}_j) - 2S_m^* - \sum_{m<i,j\le n} u_P(\mathbf{X}_i, \mathbf{X}_j)\right| \\ &\le\; 2|S_m| + 2|S_m^*| + \sum_{m<i,j\le n} |u_P(\mathbf{z}_i, \mathbf{z}_j)| + |u_P(\mathbf{X}_i, \mathbf{X}_j)| \\ &\le\; 2|S_m| + 2|S_m^*| + 2(n-m)^2\tau_\infty\;, \end{aligned} \tag{5.35}$$

where the last line holds since $\sup_{\mathbf{x},\mathbf{x}'\in\mathbb{R}^d} u_P(\mathbf{x},\mathbf{x}') \le \tau_\infty$ under the assumed kernel conditions due to Lemma 5.2. We then bound $|S_m|$ and $|S_m^*|$ with high probability. When $\mathbb{E}_{\mathbf{X}^*\sim P}[\|\mathbf{s}_p(\mathbf{X}^*)\|_2] < \infty$, we have $\mathbb{E}_{\mathbf{X}^*\sim P}[u_P(\mathbf{X}^*, \cdot)] = 0$ as argued in the paragraph before (5.11), so it is straightforward to see that $\mathbb{E}_{\mathbf{X}_i^*\sim P}[T_i] = 0$ for all i, and thus S_m is a sum of zero-mean i.i.d. random variables T_i. Therefore,

$$\mathrm{Var}(S_m^2) \;=\; m\mathbb{E}_{\mathbf{X}_1^*\sim P}\left[T_1^2\right] \;=\; m \sum_{m<j,l\le n} \mathbb{E}_{\mathbf{X}_1^*\sim P}[u_P(\mathbf{X}_1^*, \mathbf{z}_j) u_P(\mathbf{X}_1^*, \mathbf{z}_l)] \;\le\; m(n-m)^2\tau_\infty^2\;,$$

where the last step holds again since $\sup_{\mathbf{x},\mathbf{x}'\in\mathbb{R}^d} u_P(\mathbf{x},\mathbf{x}') \le \tau_\infty$. For any $\delta > 0$, Chebyshev's inequality implies that the event $\mathcal{A}_1 \coloneqq \{|S_m| \le \sqrt{m}(n-m)\tau_\infty/\sqrt{\delta/2}\}$ occurs with probability at least $1-\delta/2$. In other words, $|S_m|$ can be upper bounded on the high-probability event $\mathcal{A}_1$. A similar argument applied to S_m^* shows that the event $\mathcal{A}_2 \coloneqq \{|S_m^*| \le \sqrt{m}(n-m)\tau_\infty/\sqrt{\delta/2}\}$ also occurs with probability at least $1-\delta/2$. On $\mathcal{A} \coloneqq \mathcal{A}_1 \cap \mathcal{A}_2$, which occurs with probability at least $1-\delta$, we have by (5.35) that

$$\begin{aligned} \left|D^2(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) - D^2(\mathbb{X}_n^*)\right| \;&\le\; \frac{4\sqrt{m}(n-m)\tau_\infty}{n^2\sqrt{\delta/2}} + \frac{2(n-m)^2\tau_\infty}{n^2} \\ &\le\; \frac{4\sqrt{2}(n-m)\tau_\infty}{n^{3/2}\sqrt{\delta}} + \frac{2(n-m)^2\tau_\infty}{n^2} \\ &=\; 4\sqrt{2}\epsilon_{m,n}\frac{\tau_\infty}{\sqrt{\delta n}} + 2\tau_\infty\epsilon_{m,n}^2 \;=:\; t_{m,n,\delta}^2\;. \end{aligned} \tag{5.36}$$

where the second step holds since $\sqrt{m} \le \sqrt{n}$, and in the last line we have substituted $\epsilon_{m,n} = (n-m)/n$. We claim that this implies that the following holds on $\mathcal{A}$:

$$\left|D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) - D(\mathbb{X}_n^*)\right| \;\le\; t_{m,n,\delta}\;. \tag{5.37}$$

To see this, we first assume $D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) + D(\mathbb{X}_n^*) \geq t_{m,n,\delta}$. In this case, (5.36) implies

$$\begin{aligned}\left|D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) - D(\mathbb{X}_n^*)\right| &= \frac{\left|D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) - D(\mathbb{X}_n^*)\right| \times \left|D(\mathbb{X}_n) + D(\mathbb{X}_n^*)\right|}{D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) + D(\mathbb{X}_n^*)} \\ &= \frac{D^2(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) - D^2(\mathbb{X}_n^*)}{D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) + D(\mathbb{X}_n^*)} \\ &\leq t_{m,n,\delta} \,.\end{aligned}$$

On the other hand, when $D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) + D(\mathbb{X}_n^*) < t_{m,n,\delta}$, we have

$$\left|D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) - D(\mathbb{X}_n^*)\right| \leq \max\left(D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}),\, D(\mathbb{X}_n^*)\right) \leq D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) + D(\mathbb{X}_n^*) < t_{m,n,\delta} \,,$$

where the first inequality holds since $D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}), D(\mathbb{X}_n^*)$ are non-negative. Combining these two cases shows (5.37). It then follows that, for any $\gamma > 0$,

$$\begin{aligned}&\Pr_{\mathbb{X}_n^* \sim P}\left(D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) > \gamma\right) \\ &= \Pr_{\mathbb{X}_n^* \sim P}\left(\{D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) > \gamma\} \cap \mathcal{A}\right) + \Pr_{\mathbb{X}_n^* \sim P}\left(\{D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) > \gamma\} \cap \mathcal{A}^{\mathsf{c}}\right) \\ &\leq \Pr_{\mathbb{X}_n^* \sim P}\left(D(\mathbb{X}_n^*) + t_{m,n,\delta} > \gamma\right) + \delta \\ &= \Pr_{\mathbb{X}_n^* \sim P}\left(D(\mathbb{X}_n^*) > \gamma - t_{m,n,\delta}\right) + \delta \,,\end{aligned}$$

where the first step follows from the law of total probability, and the second line holds since $\Pr_{\mathbb{X}_n^* \sim P}\left(\{D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) > \gamma\} \cap \mathcal{A}^{\mathsf{c}}\right) \leq \Pr(\mathcal{A}^{\mathsf{c}}) \leq \delta$ This implies one side of the desired inequality:

$$\begin{aligned}&\Pr_{\mathbb{X}_n^* \sim P}\left(D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) > \gamma\right) - \Pr_{\mathbb{X}_n^* \sim P}\left(D(\mathbb{X}_n^*) > \gamma\right) \\ &\leq \Pr_{\mathbb{X}_n^* \sim P}\left(D(\mathbb{X}_n^*) > \gamma - t_{m,n,\delta}\right) - \Pr_{\mathbb{X}_n^* \sim P}\left(D(\mathbb{X}_n^*) > \gamma\right) + \delta \\ &= \Pr_{\mathbb{X}_n^* \sim P}\left(\gamma - t_{m,n,\delta} \leq D(\mathbb{X}_n^*) \leq \gamma\right) + \delta \\ &\leq \Pr_{\mathbb{X}_n^* \sim P}\left(\left|D(\mathbb{X}_n^*) - \gamma\right| \leq t_{m,n,\delta}\right) + \delta \,,\end{aligned}$$

where in the last line we have used

$$\begin{aligned}\Pr_{\mathbb{X}_n^* \sim P}\left(\gamma - t_{m,n,\delta} \leq D(\mathbb{X}_n^*) \leq \gamma\right) &\leq \Pr_{\mathbb{X}_n^* \sim P}\left(\gamma - t_{m,n,\delta} \leq D(\mathbb{X}_n^*) \leq \gamma + t_{m,n,\delta}\right) \\ &= \Pr_{\mathbb{X}_n^* \sim P}\left(\left|D(\mathbb{X}_n^*) - \gamma\right| \leq t_{m,n,\delta}\right) .\end{aligned}$$

A similar argument shows the other direction:

$$\begin{aligned}&\Pr_{\mathbb{X}_n^* \sim P}\left(D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) > \gamma\right) - \Pr_{\mathbb{X}_n^* \sim P}\left(D(\mathbb{X}_n^*) > \gamma\right) \\ &\geq \Pr_{\mathbb{X}_n^* \sim P}\left(\{D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) > \gamma\} \cap \mathcal{A}\right) - \Pr_{\mathbb{X}_n^* \sim P}\left(D(\mathbb{X}_n^*) > \gamma\right) \\ &\geq \Pr_{\mathbb{X}_n^* \sim P}\left(D(\mathbb{X}_n^*) > \gamma + t_{m,n,\delta}\right) - \Pr_{\mathbb{X}_n^* \sim P}\left(D(\mathbb{X}_n^*) > \gamma\right) \\ &= -\Pr_{\mathbb{X}_n^* \sim P}\left(\gamma \leq D(\mathbb{X}_n^*) \leq \gamma + t_{m,n,\delta}\right) \\ &\geq -\Pr_{\mathbb{X}_n^* \sim P}\left(\gamma - t_{m,n,\delta} \leq D(\mathbb{X}_n^*) \leq \gamma + t_{m,n,\delta}\right) \\ &= -\Pr_{\mathbb{X}_n^* \sim P}\left(\left|D(\mathbb{X}_n^*) - \gamma\right| \leq t_{m,n,\delta}\right) \,,\end{aligned}$$

where the first step holds by the law of total probability, and the second step follows from (5.37).

5.A.3.3 Proof of Lemma 5.13

Fix any $\theta, \gamma \geq 0$ and any integers m', n with $m' \leq f(n)$, where $f \in o(n^{1/2})$. Define $m = n - m'$. Pick $Z_{m'} = \{\mathbf{z}_i\}_{i=1}^{m'} \subset \mathbb{R}^d$. The LHS of the inequality can be bounded as

$$\begin{aligned}
&\left| \Pr_{\mathbb{X}_m^* \sim P, \mathbf{W}} \left(\Delta_\theta(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) > q_{\infty,1-\alpha}(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) \right) - \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \left(\Delta_\theta(\mathbb{X}_n^*) > q_{\infty,1-\alpha}(\mathbb{X}_n^*) \right) \right| \\
&= \left| \Pr_{\mathbb{X}_m^* \sim P, \mathbf{W}} \left(D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) > q_{\infty,1-\alpha}(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) + \theta \right) \right. \\
&\qquad \left. - \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \left(D(\mathbb{X}_n^*) > q_{\infty,1-\alpha}(\mathbb{X}_n^*) + \theta \right) \right| \\
&\leq \left| \Pr_{\mathbb{X}_m^* \sim P, \mathbf{W}} \left(D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) > q_{\infty,1-\alpha}(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) + \theta \right) \right. \\
&\qquad \left. - \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \left(D(\mathbb{X}_n^*) > q_{\infty,1-\alpha}(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) + \theta \right) \right| \\
&\quad + \left| \Pr_{\mathbb{X}_m^* \sim P, \mathbf{W}} \left(D(\mathbb{X}_n^*) > q_{\infty,1-\alpha}(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) + \theta \right) \right. \\
&\qquad \left. - \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \left(D(\mathbb{X}_n^*) > q_{\infty,1-\alpha}(\mathbb{X}_n^*) + \theta \right) \right| \\
&=: T_1 + T_2 \,, \qquad (5.38)
\end{aligned}$$

where the first equality holds since, for any $\gamma \geq 0$, it can be checked that the inequality $\Delta_\theta = \max(0, D(\mathbb{X}_n) - \theta) > \gamma$ holds if and only if $D(\mathbb{X}_n) < \gamma + \theta$. We will now bound the terms T_1 and T_2, respectively. Denote for brevity $\gamma_{\mathbf{z}} := q_{\infty,1-\alpha}(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) + \theta$ and $\gamma_* := q_{\infty,1-\alpha}(\mathbb{X}_n^*) + \theta$. Fix any $\delta > 0$. Applying Lemma 5.12 with $\gamma = \gamma_{\mathbf{z}}$ yields

$$\begin{aligned}
T_1 &\leq \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \left(\left| D(\mathbb{X}_n^*) - \gamma_{\mathbf{z}} \right| \leq t_{m,n,\delta} \right) + \delta \\
&\leq \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \left(\left| D(\mathbb{X}_n^*) - \gamma_* \right| \leq t_{m,n,\delta} + |\gamma_{\mathbf{z}} - \gamma_*| \right) + \delta \,, \qquad (5.39)
\end{aligned}$$

where $t_{m,n,\delta} := \left(4\sqrt{2}\tau_\infty \epsilon_{m,n} / \sqrt{\delta n} + 2\tau_\infty \epsilon_{m,n}^2 \right)^{1/2}$ is defined in Lemma 5.12, and (5.39) follows from a triangle inequality. Moreover, by Lemma 5.11, there exists an event, say $\mathcal{A}$, with probability at least $1 - \delta$ such that, on $\mathcal{A}$,

$$|\gamma_{\mathbf{z}} - \gamma_*| = \left| q_{\infty,1-\alpha}(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) - q_{\infty,1-\alpha}(\mathbb{X}_n^*) \right| \leq \frac{\tau_\infty^{\frac{1}{2}} \epsilon_{m,n}^{\frac{1}{4}}}{n^{\frac{1}{2}}} \sqrt{C_1 + C_2 \log\left(\frac{8}{\delta}\right)} =: \rho_{m,n,\delta} \,.$$

Combining this with (5.39) implies that T_1 can be bounded as

$$\begin{aligned}
T_1 &\leq \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \left(\left\{ \left| D(\mathbb{X}_n^*) - \gamma_* \right| \leq t_{m,n,\delta} + |\gamma_{\mathbf{z}} - \gamma_*| \right\} \cap \mathcal{A} \right) + \Pr(\mathcal{A}^\mathsf{c}) + \delta \\
&\leq \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \left(\left| D(\mathbb{X}_n^*) - \gamma_* \right| \leq t_{m,n,\delta} + \rho_{m,n,\delta} \right) + 2\delta \,, \qquad (5.40)
\end{aligned}$$

where the second line holds since $\Pr(\mathcal{A}^\mathsf{c}) \leq \delta$. To bound T_2, we first note that

$$\begin{aligned}
T_2 &= \left| \Pr_{\mathbf{X}_n^* \sim P} \left(D(\mathbb{X}_n^*) > \gamma_{\mathbf{z}} \right) - \Pr_{\mathbf{X}_n^* \sim P} \left(D(\mathbb{X}_n^*) > \gamma_* \right) \right| \\
&= \Pr_{\mathbf{X}_n^* \sim P} \left(\min(\gamma_{\mathbf{z}}, \gamma_*) < D(\mathbb{X}_n^*) \leq \max(\gamma_{\mathbf{z}}, \gamma_*) \right) \\
&\leq \Pr_{\mathbf{X}_n^* \sim P} \left(\left\{ \min(\gamma_{\mathbf{z}}, \gamma_*) < D(\mathbb{X}_n^*) \leq \max(\gamma_{\mathbf{z}}, \gamma_*) \right\} \cap \mathcal{A} \right) + \Pr(\mathcal{A}^\mathsf{c}) \\
&\leq \Pr_{\mathbf{X}_n^* \sim P} \left(\left| D(\mathbb{X}_n^*) - \gamma_* \right| \leq \rho_{m,n,\delta} \right) + \delta \,, \qquad (5.41)
\end{aligned}$$

where the second line holds since $|\Pr(Y > a) - \Pr(Y > b)| = \Pr(\min(a,b) < Y \leq \max(a,b))$ for any constants a, b and random variable Y, and the last line holds since $\Pr(\mathcal{A}^\mathsf{c}) \leq \delta$ and since on $\mathcal{A}$ the

following holds

$$\gamma_* - \rho_{m,n,\delta} \;\le\; \gamma_* - \left|\gamma_{\mathbf{z}} - \gamma_*\right| \;\le\; \min(\gamma_{\mathbf{z}},\, \gamma_*) \;\le\; \max(\gamma_{\mathbf{z}},\, \gamma_*) \;\le\; \gamma_* + \left|\gamma_{\mathbf{z}} - \gamma_*\right| \;\le\; \gamma_* + \rho_{m,n,\delta}\,.$$

Substituting the bounds (5.40) and (5.41) into (5.38) gives

$$\begin{aligned}
&\left|\Pr_{\mathbb{X}_m^*\sim P,\mathbf{W}}\left(\Delta_\theta(\mathbb{X}_m^*\cup\mathbb{Z}_{m'}) > q_{\infty,1-\alpha}(\mathbb{X}_m^*\cup\mathbb{Z}_{m'})\right) - \Pr_{\mathbb{X}_n^*\sim P,\mathbf{W}}\left(\Delta_\theta(\mathbb{X}_n^*) > q_{\infty,1-\alpha}(\mathbb{X}_n^*)\right)\right| \\
&\le \Pr_{\mathbb{X}_n^*\sim P,\mathbf{W}}\left(\left|D(\mathbb{X}_n^*) - \gamma_*\right| \;\le\; t_{m,n,\delta} + \rho_{m,n,\delta}\right) + \Pr_{\mathbb{X}_n^*\sim P,\mathbf{W}}\left(\left|D(\mathbb{X}_n^*) - \gamma_*\right| \;\le\; \rho_{m,n,\delta}\right) + 3\delta \\
&\le 2\Pr_{\mathbb{X}_n^*\sim P,\mathbf{W}}\left(\left|D(\mathbb{X}_n^*) - \gamma_*\right| \;\le\; t_{m,n,\delta} + \rho_{m,n,\delta}\right) + 3\delta\,, \qquad (5.42)
\end{aligned}$$

where the last inequality holds since $t_{m,n,\delta} \ge 0$. Since $m' \le f(n)$ and $f \in o(n^{1/2})$ by assumption, we have $\epsilon_{m,n} = m'/n \le f(n)/n = o(n^{-1/2})$. This implies

$$\begin{aligned}
t_{m,n,\delta} \;&=\; \left(4\sqrt{2}\tau_\infty\delta^{-\frac{1}{2}}\epsilon_{m,n}n^{-\frac{1}{2}} + 2\tau_\infty\epsilon_{m,n}^2\right)^{\frac{1}{2}} \;\le\; \left(4\sqrt{2}\tau_\infty\delta\right)^{\frac{1}{2}}\epsilon_{m,n}^{\frac{1}{2}}n^{-\frac{1}{4}} + \sqrt{2}\tau_\infty^{\frac{1}{2}}\epsilon_{m,n} \\
&=\; o\left(n^{-\frac{1}{2}}\right)\,,
\end{aligned}$$

which holds since τ_∞, δ are constants, and

$$\begin{aligned}
\rho_{m,n,\delta} \;&=\; \frac{\tau_\infty^{\frac{1}{2}}\epsilon_{m,n}^{\frac{1}{4}}}{n^{\frac{1}{2}}}\sqrt{C_1 + C_2\log\left(8n^{2(1-2s)}\right)} \\
&=\; \frac{\tau_\infty^{\frac{1}{2}}\epsilon_{m,n}^{\frac{1}{4}}}{n^{\frac{1}{2}}}\sqrt{C_1' + C_2'\log(n)} \\
&=\; o\left(n^{-\frac{1}{2}}\right)\,, \qquad (5.43)
\end{aligned}$$

where we have defined $C_1' \coloneqq C_1 + C_2\log 8$ and $C_2' \coloneqq 2C_2(1-2s)$, and the last step holds since $\epsilon_{m,n}^{1/2}\sqrt{\log(n)} = o(n^{1/4}\sqrt{\log(n)}) = o(1)$.

Define $\eta(m') \coloneqq t_{m,n,\delta} + \rho_{m,n,\delta}$, where the dependence on m' is through $m = n - m'$. Then the above derivations show that $\eta(m') \in o(n^{-1/2})$ for all $m' \le f(n)$. In particular, $\eta\left(f(n)\right) \in o(n^{-1/2})$. Moreover, substituting these into (5.42) gives

$$\begin{aligned}
&\left|\Pr_{\mathbb{X}_m^*\sim P,\mathbf{W}}\left(\Delta_\theta(\mathbb{X}_m^*\cup\mathbb{Z}_{m'}) > q_{\infty,1-\alpha}(\mathbb{X}_m^*\cup\mathbb{Z}_{m'})\right) - \Pr_{\mathbb{X}_n^*\sim P,\mathbf{W}}\left(\Delta_\theta(\mathbb{X}_n^*) > q_{\infty,1-\alpha}(\mathbb{X}_n^*)\right)\right| \\
&\le 2\Pr_{\mathbb{X}_n^*\sim P,\mathbf{W}}\left(\left|D(\mathbb{X}_n^*) - \gamma_*\right| \;\le\; \eta(m')\right) + 3\delta \\
&\le 2\Pr_{\mathbb{X}_n^*\sim P,\mathbf{W}}\left(\left|D(\mathbb{X}_n^*) - \gamma_*\right| \;\le\; \eta(f(n))\right) + 3\delta\,,
\end{aligned}$$

where the last line holds since $m' \le n^s$ by assumption and $m' \mapsto \eta(m')$ is monotone increasing by direct computation. Taking supremum over $\mathbb{Z}_{m'} \subset \mathbb{R}^d$ and $m' \le f(n)$ gives

$$\begin{aligned}
&\max_{m'\le f(n)}\sup_{\mathbf{z}_1,\ldots,\mathbf{z}_{m'}\in\mathbb{R}^d}\left|\Pr_{\mathbb{X}_m^*\sim P,\mathbf{W}}\left(\Delta_\theta(\mathbb{X}_m^*\cup\mathbb{Z}_{m'}) > q_{\infty,1-\alpha}(\mathbb{X}_m^*\cup\mathbb{Z}_{m'})\right)\right. \\
&\qquad\qquad \left. - \Pr_{\mathbb{X}_n^*\sim P,\mathbf{W}}\left(\Delta_\theta(\mathbb{X}_n^*) > q_{\infty,1-\alpha}(\mathbb{X}_n^*)\right)\right| \\
&\le 2\Pr_{\mathbb{X}_n^*\sim P,\mathbf{W}}\left(\left|D(\mathbb{X}_n^*) - \gamma_*\right| \;\le\; \eta(f(n))\right) + 3\delta
\end{aligned}$$

To complete our proof, it remains to show that the first term can be bounded by δ, from which the

claimed result in Lemma 5.13 would follow by redefining δ. To proceed, we write

$$\begin{aligned}
&\Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \left(|\Delta_\theta(\mathbb{X}_n^*) - q_{\infty,1-\alpha}(\mathbb{X}_n^*)| \;\leq\; \eta\left(f(n)\right)\right) \\
&= \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \left(\Delta_\theta(\mathbb{X}_n^*) \;\leq\; q_{\infty,1-\alpha}(\mathbb{X}_n^*) + \eta\left(f(n)\right)\right) \\
&\quad - \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \left(\Delta_\theta(\mathbb{X}_n^*) \;\leq\; q_{\infty,1-\alpha}(\mathbb{X}_n) - \eta\left(f(n)\right)\right) \\
&= \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \left(\sqrt{n} D(\mathbb{X}_n^*) \;\leq\; \sqrt{n}\zeta_{n,\theta} + \sqrt{n}\eta\left(f(n)\right)\right) \qquad (5.44a) \\
&\quad - \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \left(\sqrt{n} D(\mathbb{X}_n^*) \;\leq\; \sqrt{n}\zeta_{n,\theta} - \sqrt{n}\eta\left(f(n)\right)\right) \qquad (5.44b)
\end{aligned}$$

where the first equality holds again because $|\Pr(Y > a) - \Pr(Y > b)| = \Pr(\min(a,b) < Y \leq \max(a,b))$ for any constants a, b and random variable Y, the second equality holds by noting that $\Delta_\theta(\mathbb{X}_n^*) = \max(0, D(\mathbb{X}_n^*) - \theta) \leq t$ for some $t > 0$ if and only if $D(\mathbb{X}_n^*) - \theta \leq t$, and (5.44) holds by multiplying $\sqrt{n}$ on both sides of the inequalities within the probabilities and defining $\zeta_{n,\theta} \coloneqq q_{\infty,1-\alpha}(\mathbb{X}_n^*) + \theta$.

Under the assumed conditions on k and assuming $\mathbb{E}_{\mathbf{X}^* P}[\|\mathbf{s}_p(\mathbf{X})\|_2] < \infty$, Gorham and Mackey (2017, Proposition 1) shows that $\mathbb{E}_{\mathbf{X}^* \sim P}[u_P(\mathbf{X}^*, \cdot)] = 0$, so the V-statistic $D^2(\mathbb{X}_n^*)$ is *degenerate* (see, e.g., Serfling, 2009, Chapter 6), and classic results on the asymptotics of degenerate V-statistics (Serfling, 2009, Theorem 6.4.1 A) shows that $nD^2(\mathbb{X}_n^*)$ converges weakly to a non-negative distribution. Since the square-root function $x \mapsto \sqrt{x}$ is continuous on $[0, \infty)$ and continuous functions preserve weak limits by the Continuous Mapping Theorem (van der Vaart, 2000, Theorem 2.3), the squared-root statistic, $\sqrt{n}D(\mathbb{X}_n^*)$, also converges weakly to a non-negative distribution. In particular, $\sqrt{n}\zeta_{n,\theta}$ does not vanish as $n \to \infty$. Also since $\eta\left(f(n)\right) = o(n^{-1/2})$ as argued below (5.43), we have $\sqrt{n}\eta\left(f(n)\right) \to 0$. In summary, we have shown that

$$\begin{aligned}
&\lim_{n\to\infty} \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \left(|\Delta_\theta(\mathbb{X}_n^*) - q_{\infty,1-\alpha}(\mathbb{X}_n^*)| \;\leq\; \eta\left(f(n)\right)\right) \\
&= \lim_{n\to\infty} \Big(\Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \left(\sqrt{n} D(\mathbb{X}_n^*) \;\leq\; \sqrt{n}\zeta_{n,\theta} + \sqrt{n}\eta\left(f(n)\right)\right) \\
&\qquad\qquad - \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \left(\sqrt{n} D(\mathbb{X}_n^*) \;\leq\; \sqrt{n}\zeta_{n,\theta} - \sqrt{n}\eta\left(f(n)\right)\right) \Big) \\
&= \lim_{n\to\infty} \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \left(\sqrt{n} D(\mathbb{X}_n^*) \;\leq\; \sqrt{n}\zeta_{n,\theta}\right) - \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \left(\sqrt{n} D(\mathbb{X}_n^*) \;\leq\; \sqrt{n}\zeta_{n,\theta}\right) \\
&= 0 \,,
\end{aligned}$$

which completes the proof.

5.A.3.4 Proofs for Proposition 5.14 and Theorem 5.3

Proof of Proposition 5.14. For each n and $R \in \mathcal{P}(\mathbb{R}^d)$, define $Q = (1 - \epsilon_n)P + \epsilon_n R$. Let $\mathbb{X}_n$ and $\mathbb{X}_n^*$ be random samples drawn from Q and P, respectively. As argued in Section 5.A.1, each random variable in $\mathbb{X}_n$ can be written as $\mathbf{X}_i \overset{d}{=} (1 - \xi_i)\mathbf{X}_i^* + \xi_i \mathbf{Z}_i$, where $\mathbf{Z}_i \sim R$ and $\xi_i \sim \text{Bernoulli}(n, \epsilon_n)$ are independent, and $\overset{d}{=}$ denotes equality in distribution. Define the random variable $M' = \sum_{i=1}^n \xi_i$, then $M' \sim \text{Binomial}(n, \epsilon_n)$.

Pick $\delta > 0$. Given any sequence $\epsilon_n = o(n^{-1/2})$, there must exists $f = f(n) \in o(n^{1/2})$ such that $\epsilon_n \leq f(n)/n$ and $f(n) \to \infty$ as $n \to \infty$. Defining the event $\mathcal{B} \coloneqq \{M' \leq \epsilon_n n + f(n)\}$, we will first show that $\mathcal{B}$ occurs with high probability. Using a concentration inequality for Binomial distributions (e.g.,

Chung and Lu, 2002, Lemma 2.1, Eq. (2.2)) applied to Binomial(n, ϵ_n), we have

$$\Pr_{M'}(\mathcal{B}^{\mathsf{c}}) \;=\; \Pr_{M'}\left(M' - \epsilon_n n > f(n)\right) \;\leq\; \exp\left(-\frac{f^2(n)}{2(\epsilon_n n + f(n)/3)}\right) \;=:\; t_n \,. \tag{5.45}$$

Define the event $\mathcal{A}(\mathbb{X}_n) = \{\Delta_\theta(\mathbb{X}_n) > q_{\infty,1-\alpha}(\mathbb{X}_n)\}$ and similarly for $\mathcal{A}(\mathbb{X}_n^*)$. Using the above inequality to decompose $\Pr_{\mathbb{X}_n\sim Q,\mathbf{W}}(\mathcal{A}(\mathbb{X}_n))$ yields

$$\begin{aligned}
&\Pr_{\mathbb{X}_n\sim Q,\mathbf{W}}(\mathcal{A}(\mathbb{X}_n)) \\
&\leq \Pr_{\mathbb{X}_n\sim Q,\mathbf{W}}(\mathcal{A}(\mathbb{X}_n)\cap\mathcal{B}) + \Pr(\mathcal{B}^{\mathsf{c}}) \\
&= \sum_{m'\leq \epsilon_n n + f(n)} \Pr_{\mathbb{X}_n\sim Q,\mathbf{W}}(\mathcal{A}(\mathbb{X}_n) \mid M' = m')\Pr(M' = m') + \Pr(\mathcal{B}^{\mathsf{c}}) \\
&= \sum_{m'\leq \epsilon_n n + f(n)} \Pr_{\mathbb{X}^*_{n-m'}\sim P,\, \mathbb{Z}_{m'}\sim R,\mathbf{W}}(\mathcal{A}(\mathbb{X}^*_{n-m'}\cup\mathbb{Z}_{m'}))\Pr(M' = m') + \Pr(\mathcal{B}^{\mathsf{c}}) \\
&\leq \max_{m'\leq \epsilon_n n + f(n)} \Pr_{\mathbb{X}^*_{n-m'}\sim P,\, \mathbb{Z}_{m'}\sim R,\mathbf{W}}(\mathcal{A}(\mathbb{X}^*_{n-m'}\cup\mathbb{Z}_{m'})) \sum_{m'\leq \epsilon_n n + f(n)} \Pr(M' = m') + \Pr(\mathcal{B}^{\mathsf{c}}) \\
&= \max_{m'\leq \epsilon_n n + f(n)} \Pr_{\mathbb{X}^*_{n-m'}\sim P,\, \mathbb{Z}_{m'}\sim R,\mathbf{W}}(\mathcal{A}(\mathbb{X}^*_{n-m'}\cup\mathbb{Z}_{m'}))\Pr(\mathcal{B}) + \Pr(\mathcal{B}^{\mathsf{c}}) \\
&\leq \max_{m'\leq \epsilon_n n + f(n)} \Pr_{\mathbb{X}^*_{n-m'}\sim P,\, \mathbb{Z}_{m'}\sim R,\mathbf{W}}(\mathcal{A}(\mathbb{X}^*_{n-m'}\cup\mathbb{Z}_{m'})) + t_n \,,
\end{aligned}$$

where in the last line we have used (5.45) and that $\Pr(\mathcal{B}) \leq 1$. This implies

$$\begin{aligned}
&\Pr_{\mathbb{X}_n\sim Q,\mathbf{W}}(\mathcal{A}(\mathbb{X}_n)) - \Pr_{\mathbb{X}_n^*\sim P}(\mathcal{A}(\mathbb{X}_n^*)) \\
&\leq \max_{m'\leq \epsilon_n n + f(n)} \left(\Pr_{\mathbb{X}^*_{n-m'}\sim P,\, \mathbb{Z}_{m'}\sim R,\mathbf{W}}(\mathcal{A}(\mathbb{X}^*_{n-m'}\cup\mathbb{Z}_{m'})) - \Pr_{\mathbb{X}_n^*\sim P}(\mathcal{A}(\mathbb{X}_n^*))\right) + t_n \\
&\leq \max_{m'\leq \epsilon_n n + f(n)} \sup_{\mathbf{z}_1,\ldots,\mathbf{z}_{m'}\in\mathbb{R}^d} \left|\Pr_{\mathbb{X}_m^*\sim P}(\mathcal{A}(\mathbb{X}_m^*\cup\{\mathbf{z}_i\}_{i=1}^{m'})) - \Pr_{\mathbb{X}_n^*\sim P}(\mathcal{A}(\mathbb{X}_n^*))\right| + t_n \\
&=: \max_{m'\leq \epsilon_n n + f(n)} \omega(m') + t_n \,,
\end{aligned}$$

where in the second inequality we have taken supremum over all possible values of $\mathbb{Z}_{m'}$, and in the last line we have defined

$$\omega(m') \;:=\; \sup_{\mathbf{z}_1,\ldots,\mathbf{z}_{m'}\in\mathbb{R}^d} \left|\Pr_{\mathbb{X}^*_{n-m'}\sim P}\left(\mathcal{A}\left(\mathbb{X}^*_{n-m'}\cup\{\mathbf{z}_i\}_{i=1}^{m'}\right)\right) - \Pr_{\mathbb{X}_n^*\sim P}\left(\mathcal{A}(\mathbb{X}_n^*)\right)\right| .$$

Similarly, we have the lower bound

$$\begin{aligned}
\Pr_{\mathbb{X}_n\sim Q,\mathbf{W}}(\mathcal{A}(\mathbb{X}_n)) \;&\geq\; \Pr_{\mathbb{X}_n\sim Q,\mathbf{W}}(\mathcal{A}(\mathbb{X}_n)\cap\mathcal{B}) \\
&= \sum_{m'\leq \epsilon_n n + f(n)} \Pr_{\mathbb{X}_n\sim Q,\mathbf{W}}(\mathcal{A}(\mathbb{X}_n) \mid M' = m')\Pr(M' = m') \\
&= \sum_{m'\leq \epsilon_n n + f(n)} \Pr_{\mathbb{X}^*_{n-m'}\sim P,\, \mathbb{Z}_{m'}\sim R,\mathbf{W}}(\mathcal{A}(\mathbb{X}^*_{n-m'}\cup\mathbb{Z}_{m'}))\Pr(M' = m') \\
&\geq \min_{m'\leq \epsilon_n n + f(n)} \Pr_{\mathbb{X}^*_{n-m'}\sim P,\, \mathbb{Z}_{m'}\sim R,\mathbf{W}}(\mathcal{A}(\mathbb{X}^*_{n-m'}\cup\mathbb{Z}_{m'})) \sum_{m'\leq \epsilon_n n + f(n)} \Pr(M' = m') \\
&= \min_{m'\leq \epsilon_n n + f(n)} \Pr_{\mathbb{X}^*_{n-m'}\sim P,\, \mathbb{Z}_{m'}\sim R,\mathbf{W}}(\mathcal{A}(\mathbb{X}^*_{n-m'}\cup\mathbb{Z}_{m'}))\Pr(\mathcal{B}) \,,
\end{aligned}$$

where in the second last line we have taken infimum over the values of $\mathbb{Z}_{m'}$. Taking infimum over $\mathbf{z}_1, \ldots, \mathbf{z}_{m'}$, the last line is lower bounded by

$$\begin{aligned}
&\min_{m' \leq \epsilon_n n + f(n)} \Pr_{\mathbb{X}^*_{n-m'} \sim P,\, \mathbb{Z}_{m'} \sim R, \mathbf{W}}(\mathcal{A}(\mathbb{X}^*_{n-m'} \cup \mathbb{Z}_{m'}))\, \Pr(\mathcal{B}) \\
\geq\ & \min_{m' \leq \epsilon_n n + f(n)} \inf_{\mathbf{z}_1, \ldots, \mathbf{z}_{m'}} \Pr_{\mathbb{X}^*_{n-m'} \sim P}(\mathcal{A}(\mathbb{X}^*_{n-m'} \cup \{\mathbf{z}_i\}_{i=1}^{m'}))\, \Pr(\mathcal{B}) \\
\geq\ & \min_{m' \leq \epsilon_n n + f(n)} \inf_{\mathbf{z}_1, \ldots, \mathbf{z}_{m'}} \Pr_{\mathbb{X}^*_{n-m'} \sim P}(\mathcal{A}(\mathbb{X}^*_{n-m'} \cup \{\mathbf{z}_i\}_{i=1}^{m'})) \times (1 - t_n) \\
=\ & \min_{m' \leq \epsilon_n n + f(n)} \inf_{\mathbf{z}_1, \ldots, \mathbf{z}_{m'}} \Big(\Pr_{\mathbb{X}^*_{n-m'} \sim P}(\mathcal{A}(\mathbb{X}^*_{n-m'} \cup \{\mathbf{z}_i\}_{i=1}^{m'})) \\
&\qquad\qquad - t_n \Pr_{\mathbb{X}^*_{n-m'} \sim P}(\mathcal{A}(\mathbb{X}^*_{n-m'} \cup \{\mathbf{z}_i\}_{i=1}^{m'})) \Big) \\
\geq\ & \min_{m' \leq \epsilon_n n + f(n)} \inf_{\mathbf{z}_1, \ldots, \mathbf{z}_{m'}} \Pr_{\mathbb{X}^*_{n-m'} \sim P}(\mathcal{A}(\mathbb{X}^*_{n-m'} \cup \{\mathbf{z}_i\}_{i=1}^{m'})) - t_n\ , \qquad (5.46)
\end{aligned}$$

where the second inequality holds since (5.45) implies $\Pr(\mathcal{B}) \geq 1 - t_n$, and the last line holds as a probability is always smaller than or equal to 1. This implies

$$\begin{aligned}
&\Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}(\mathcal{A}(\mathbb{X}_n)) - \Pr_{\mathbb{X}^*_n \sim P}(\mathcal{A}(\mathbb{X}^*_n)) \\
\geq\ & \min_{m' \leq \epsilon_n n + f(n)} \inf_{\mathbf{z}_1, \ldots, \mathbf{z}_{m'}} \left(\Pr_{\mathbb{X}^*_{n-m'} \sim P}\left(\mathcal{A}\left(\mathbb{X}^*_{n-m'} \cup \{\mathbf{z}_i\}_{i=1}^{m'}\right)\right) - \Pr_{\mathbb{X}^*_n \sim P}\left(\mathcal{A}(\mathbb{X}^*_n)\right)\right) - t_n \\
\geq\ & -\max_{m' \leq \epsilon_n n + f(n)} \sup_{\mathbf{z}_1, \ldots, \mathbf{z}_{m'} \in \mathbb{R}^d} \left| \Pr_{\mathbb{X}^*_{n-m'} \sim P}\left(\mathcal{A}\left(\mathbb{X}^*_{n-m'} \cup \{\mathbf{z}_i\}_{i=1}^{m'}\right)\right) - \Pr_{\mathbb{X}^*_n \sim P}\left(\mathcal{A}(\mathbb{X}^*_n)\right) \right| - t_n \\
=\ & -\max_{m' \leq \epsilon_n n + f(n)} \omega(m') - t_n\ .
\end{aligned}$$

We have therefore shown that

$$\left| \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}(\mathcal{A}(\mathbb{X}_n)) - \Pr_{\mathbb{X}^*_n \sim P}(\mathcal{A}(\mathbb{X}^*_n)) \right| \ \leq\ \max_{m' \leq \epsilon_n n + f(n)} \omega(m') + t_n\ . \qquad (5.47)$$

It remains to bound the two terms on the RHS of the above inequality. Since $\epsilon_n \leq f(n)/n$ and $f(n) \in o(n^{1/2})$, it is clear that t_n, defined in (5.45), converges to 0 as $n \to \infty$, so $t_n \leq \delta/2$ for sufficiently large n. To bound the first term, we can apply Lemma 5.13, since $\epsilon_n n + f(n) \leq 2f(n) \in o(n^{1/2})$ and thus the assumptions in Lemma 5.13 holds. We can therefore conclude that there exists n_0 such that for any $n \geq n_0$, we have

$$\max_{m' \leq \epsilon_n n + f(n)} \omega(m') \ <\ \frac{\delta}{2}\ .$$

Using these arguments to bound (5.47), we have shown that, for sufficiently large n,

$$\left| \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}(\mathcal{A}(\mathbb{X}_n)) - \Pr_{\mathbb{X}^*_n \sim P}(\mathcal{A}(\mathbb{X}^*_n)) \right| \ \leq\ \frac{\delta}{2} + \frac{\delta}{2} \ =\ \delta\ .$$

Taking the supremum over $Q \in \mathcal{P}(P; \epsilon_n)$ then implies, for sufficiently large n,

$$\sup_{Q \in \mathcal{P}(P; \epsilon_n)} \left| \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\left(\Delta_\theta(\mathbb{X}_n) > q_{\infty, 1-\alpha}(\mathbb{X}_n)\right) - \Pr_{\mathbb{X}^*_n \sim P, \mathbf{W}}\left(\Delta_\theta(\mathbb{X}^*_n) > q_{\infty, 1-\alpha}(\mathbb{X}^*_n)\right) \right| \ \leq\ \delta\ .$$

Since $\delta > 0$ was arbitrary, this shows the claim. □

Proof of Theorem 5.3. This result immediately follows from Proposition 5.14 by setting $\theta = 0$, in which case

$$\begin{aligned}
&\sup_{Q\in\mathcal{P}(P;\epsilon_n)} \left| \Pr_{\mathbb{X}_n\sim Q,\mathbf{W}} \left(D^2(\mathbb{X}_n) > q^2_{\infty,1-\alpha}(\mathbb{X}_n)\right) - \Pr_{\mathbb{X}_n^*\sim P,\mathbf{W}} \left(D^2(\mathbb{X}_n^*) > q^2_{\infty,1-\alpha}(\mathbb{X}_n^*)\right)\right| \\
&= \sup_{Q\in\mathcal{P}(P;\epsilon_n)} \left| \Pr_{\mathbb{X}_n\sim Q,\mathbf{W}} \left(D(\mathbb{X}_n) > q_{\infty,1-\alpha}(\mathbb{X}_n)\right) - \Pr_{\mathbb{X}_n^*\sim P,\mathbf{W}} \left(D(\mathbb{X}_n^*) > q_{\infty,1-\alpha}(\mathbb{X}_n^*)\right)\right| \\
&= \sup_{Q\in\mathcal{P}(P;\epsilon_n)} \left| \Pr_{\mathbb{X}_n\sim Q,\mathbf{W}} \left(\Delta_\theta(\mathbb{X}_n) > q_{\infty,1-\alpha}(\mathbb{X}_n)\right) - \Pr_{\mathbb{X}_n^*\sim P,\mathbf{W}} \left(\Delta_\theta(\mathbb{X}_n^*) > q_{\infty,1-\alpha}(\mathbb{X}_n^*)\right)\right| \\
&\to 0 \,,
\end{aligned}$$

where the second line holds by taking the square-root on both sides of the inequalities, and the last equality holds since $D(\mathbb{X}_n) = \Delta_\theta(\mathbb{X}_n)$ when $\theta = 0$. □

5.A.4 P-Targeted Kernel Stein Discrepancy

This section states preliminary results which will be needed to prove Theorem 5.4. The main tool we use is the following generalized definition of KSD, which was originally proposed in Shi and Mackey (2024, Definition 4).

Definition 5.3 (P-KSD)**.** *Given $P \in \mathcal{P}(\mathbb{R}^d)$ with a Lebesgue density $p \in \mathcal{C}^1$ and a reproducing kernel $k \in \mathcal{C}_b^{(1,1)}$, the* P-targeted (Langevin) kernel Stein discrepancy *(P-KSD) between two probability measures $Q, R \in \mathcal{P}(\mathbb{R}^d)$ is defined as*

$$\mathbb{S}_P(Q, R) \;:=\; \sup_{f\in\mathcal{B}} \left|\mathbb{E}_{\mathbf{X}\sim Q}[(\mathcal{A}_P f)(\mathbf{X})] - \mathbb{E}_{\mathbf{Y}\sim R}[(\mathcal{A}_P f)(\mathbf{Y})]\right| , \tag{5.48}$$

where $\mathcal{B} := \left\{h = (h_1, \ldots, h_d) : \; h_j \in \mathcal{H}_k \text{ and } \sum_{j=1}^d \|h_j\|^2_{\mathcal{H}_k} \leq 1\right\}$ is the unit ball in the d-times Cartesian product $\mathcal{H}_k^d$ of the RKHS $\mathcal{H}_k$ associated with k.

When any input probability measure, say Q, is an empirical measure based on a sample $\mathbb{X}_n = \{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^d$, we will abuse the notation by writing $\mathbb{S}_P(Q, R) = \mathbb{S}_P(\mathbb{X}_n, R)$ to emphasize the dependence on $\mathbb{X}_n$.

It can be shown that P-KSD is a *Maximum Mean Discrepancy* (MMD, Fortet and Mourier, 1953; Müller, 1997) with the Stein reproducing kernel u_P (Barp et al., 2022, Theorem 2.8). When either of the two arguments coincides with P, the P-KSD reduces to the standard KSD, as we will show in the next section. The main benefit of working with P-KSD rather than KSD is that P-KSD satisfies both symmetry and a triangle inequality (Shi and Mackey, 2024, Lemma 1), i.e., $\mathbb{S}_P(Q, R) = \mathbb{S}_P(R, Q)$ and $\mathbb{S}_P(Q, R) \leq \mathbb{S}_P(Q, R') + \mathbb{S}_P(R', R)$, for any $Q, R, R' \in \mathcal{P}(\mathbb{R}^d)$ for which they are defined. These properties allow us to show several lemmas, which will be crucial in proving the validity of our robust-KSD tests in later sections.

- Lemma 5.15 shows that the standard KSD can be written as a P-KSD.
- Lemma 5.16 shows that $\mathbb{S}_P(Q, R)$ admits a closed form involving expectations over Q and R.
- Lemma 5.17 shows that the P-KSD-projection of a probability measure onto a (standard) KSD-ball centered at P admits a closed-form expression.

- Lemma 5.18 provides a deviation bound for P-KSD when the Stein kernel is bounded. It is a counterpart of a similar bound for MMD proved in Gretton et al. (2012, Corollary 9) and will be used in Section 5.D to derive a robust-KSD test that is well-calibrated for finite samples. Although our proof follows a similar idea as in Gretton et al. (2012, Corollary 9), it is not a direct consequence of that result, because the MMD bound is for an MMD computed between *two* empirical measures, whereas our result concerns P-KSD between one empirical measure and one population measure.

The MMD has already been studied extensively in the context of both Bayesian and frequentist robust parameter estimation; see, e.g., Briol et al. (2019); Chérief-Abdellatif and Alquier (2020, 2022); Dellaporta et al. (2022); Dellaporta and Damoulas (2023); Alquier and Gerber (2024). A key assumption to ensure robustness in all of these papers is that the kernel is assumed to be bounded, and this is exactly what we are able to achieve with the KSD and P-KSD through Lemma 5.2.

5.A.4.1 Properties of P-KSD

Lemma 5.15 (KSD as P-KSD)**.** *Assume $k \in \mathcal{C}_b^{(1,1)}$ and $\mathbb{E}_{\mathbf{X}\sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2] < \infty$. For any $Q \in \mathcal{P}(\mathbb{R}^d)$ such that $D(Q,P) < \infty$, we have $\mathbb{S}_P(Q,P) = \mathbb{S}_P(P,Q) = D(Q,P)$.*

Proof of Lemma 5.15. This follows directly from the symmetry of P-KSD (Shi and Mackey, 2024, Lemma 1) and that the assumed conditions imply $\mathbb{E}_{\mathbf{X}\sim P}[(\mathcal{A}_P f)(\mathbf{X})] = 0$ for all $f \in \mathcal{B}$ (Gorham and Mackey, 2017, Proposition 1), so one of the two expectations in (5.48) vanishes when either of the two arguments coincides with P. □

Lemma 5.16 (P-KSD closed-form)**.** *Let $Q, R \in \mathcal{P}(\mathbb{R}^d)$ and $k \in \mathcal{C}_b^{(1,1)}$. If $\mathbb{E}_{\mathbf{X}\sim Q}[u_P(\mathbf{X},\mathbf{X})] < \infty$ and $\mathbb{E}_{\mathbf{Y}\sim R}[u_P(\mathbf{Y},\mathbf{Y})] < \infty$, then*

$$\mathbb{S}_P^2(Q,R) \;=\; \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}[u_P(\mathbf{X},\mathbf{X}')] + \mathbb{E}_{\mathbf{Y},\mathbf{Y}'\sim R}[u_P(\mathbf{Y},\mathbf{Y}')] - 2\mathbb{E}_{\mathbf{X}\sim Q,\mathbf{Y}\sim R}[u_P(\mathbf{X},\mathbf{Y})]\,.$$

Proof of Lemma 5.16. The assumed conditions ensure $\mathbb{S}_P^2(Q,R)$ is well-defined. Since P-KSD is an MMD with reproducing kernel u_P (Barp et al., 2022, Theorem 2.8), we can apply Gretton et al. (2012, Lemma 6) to yield this result. □

Lemma 5.17 (P-KSD projection)**.** *Let $R \in \mathcal{P}(\mathbb{R}^d)$ and assume $\mathbb{E}_{\mathbf{Y}\sim R}[u_P(\mathbf{Y},\mathbf{Y})] < \infty$. Further assume the conditions in Lemma 5.15 hold. For any $\theta > 0$,*

$$\inf_{Q\in\mathcal{B}^{\mathrm{KSD}}(P;\theta)} \mathbb{S}_P(R,Q) \;=\; \max\big(0,\; \mathbb{S}_P(R,P) - \theta\big) \;=\; \max\big(0,\; D(R,P) - \theta\big)\,.$$

Proof of Lemma 5.17. The P-KSD $\mathbb{S}_P(R,P)$ is well-defined under the assumed conditions, and hence so is $\mathbb{S}_P\big(\lambda R + (1-\lambda)P, P\big)$ for any $\lambda \in [0,1]$. Also since P-KSD is an MMD with (potentially unbounded) reproducing kernel u_P, the claim then follows by Sun and Zou (2023, Proposition 3) and noting that their proof extends directly to potentially unbounded kernels. □

Lemma 5.18 (P-KSD deviation bound)**.** *Assume k is a tilted kernel satisfying the conditions in Lemma 5.2 and $P \in \mathcal{P}(\mathbb{R}^d)$ has a density $p \in \mathcal{C}^1$. For any positive integer n, any $Q \in \mathcal{P}(\mathbb{R}^d)$ and any $\delta > 0$,*

$$\Pr_{\mathbb{X}_n \sim Q}\left(\mathbb{S}_P(\mathbb{X}_n, Q) > \sqrt{\frac{\tau_\infty}{n}} + \delta\right) \;\leq\; \exp\left(-\frac{n\delta^2}{2\tau_\infty}\right) ,$$

In particular, for any $\alpha > 0$,

$$\Pr_{\mathbb{X}_n \sim Q}\left(\mathbb{S}_P(\mathbb{X}_n, Q) > \sqrt{\frac{\tau_\infty}{n}} + \sqrt{\frac{-2\tau_\infty \log \alpha}{n}}\right) \;\leq\; \alpha \,.$$

Proof of Lemma 5.18. Let $\mathbb{X}_n = \{\mathbf{x}_i\}_{i=1}^n$ and $\mathbb{X}'_n = \{\mathbf{x}'\}_{i=1}^n$ be two samples that differ only by the first element, i.e., $\mathbf{x}_i, \mathbf{x}'_i \in \mathbb{R}^d$ and $\mathbf{x}_i = \mathbf{x}'_i$ for $i \neq 1$. Write $\mathcal{B} \coloneqq \{f \in \mathcal{H}_k^d : \|f\|_{\mathcal{H}_k^d} \leq 1\}$, and denote by Q_n, Q'_n the empirical measures formed by $\mathbb{X}_n$ and $\mathbb{X}'_n$, respectively. The difference between the P-KSD computed using $\mathbb{X}_n$ and $\mathbb{X}'_n$ can be bounded as

$$\begin{aligned}
&\left|\mathbb{S}_P(\mathbb{X}_n, Q) - \mathbb{S}_P(\mathbb{X}'_n, Q)\right| \\
&= \left|\sup_{f\in\mathcal{B}}\left(\frac{1}{n}\sum_{i=1}^n (\mathcal{A}_P f)(\mathbf{x}_i) - \mathbb{E}_{\mathbf{X}\sim Q}[\mathcal{A}_P f(\mathbf{X})]\right) - \sup_{f\in\mathcal{B}}\left(\frac{1}{n}\sum_{i=1}^n (\mathcal{A}_P f)(\mathbf{x}'_i) - \mathbb{E}_{\mathbf{X}\sim Q}[\mathcal{A}_P f(\mathbf{X})]\right)\right| \\
&\leq \sup_{f\in\mathcal{B}}\left|\frac{1}{n}\sum_{i=1}^n (\mathcal{A}_P f)(\mathbf{x}_i) - \frac{1}{n}\sum_{i=1}^n (\mathcal{A}_P f)(\mathbf{x}'_i)\right| \\
&= \frac{1}{n}\sup_{f\in\mathcal{B}}\left|(\mathcal{A}_P f)(\mathbf{x}_1) - (\mathcal{A}_P f)(\mathbf{x}'_1)\right| \\
&\leq \frac{2}{n}\sup_{\mathbf{x}\in\mathbb{R}^d}\sup_{f\in\mathcal{B}}\left|(\mathcal{A}_p f)(\mathbf{x})\right| \\
&= \frac{2}{n}\sup_{\mathbf{x}\in\mathbb{R}^d}\left(u_P(\mathbf{x},\mathbf{x})\right)^{\frac{1}{2}} ,
\end{aligned}$$

where the last line holds by noting that

$$\begin{aligned}
\sup_{f\in\mathcal{B}}\left|(\mathcal{A}_p f)(\mathbf{x})\right| &= \sup_{f\in\mathcal{B}}\left|\mathbb{E}_{\mathbf{X}\sim\delta_{\mathbf{x}}}[(\mathcal{A}_p f)(\mathbf{X})]\right| \\
&\overset{(a)}{=} D(\delta_{\mathbf{x}}, P) \\
&= \left(\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim\delta_{\mathbf{x}}}[u_P(\mathbf{X},\mathbf{X}')]\right)^{1/2} \\
&= u_P(\mathbf{x},\mathbf{x})^{1/2} ,
\end{aligned}$$

where (a) follows from the definition of KSD, and the last equality holds by the double-integral form of KSD (see, e.g., Barp et al., 2022, Corollary 2.10), which holds under the assumed kernel conditions. Moreover, Lemma 5.2 implies $\sup_{\mathbf{x}\in\mathbb{R}^d} u_P(\mathbf{x},\mathbf{x}) \leq \tau_\infty < \infty$, so, by symmetry, we have shown that varying a single data point $\mathbf{x}_i$ for *any* i changes $\mathbb{S}_P(\mathbb{X}_n, P)$ by at most a magnitude of $2\tau_\infty^{1/2}/n$. We can then apply the McDiarmid's inequality (McDiarmid et al., 1989) to yield that, for any $\delta > 0$,

$$\Pr_{\mathbf{X}_n\sim Q}\left(\mathbb{S}_P(\mathbb{X}_n, Q) - \mathbb{E}_{\mathbb{Y}_n\sim Q}[\mathbb{S}_P(\mathbb{Y}_n, Q)] > \delta\right) \;\leq\; \exp\left(-\frac{n\delta^2}{2\tau_\infty}\right) .$$

It remains to bound the expectation. Using Hölder's inequality,

$$\begin{aligned}
&\mathbb{E}_{\mathbb{Y}_n\sim Q}[\mathbb{S}_P(\mathbb{Y}_n, Q)] \\
&\leq \left(\mathbb{E}_{\mathbb{Y}_n\sim Q}[\mathbb{S}_P^2(\mathbb{Y}_n, Q)]\right)^{\frac{1}{2}} \\
&= \left(\mathbb{E}_{\mathbb{Y}_n\sim Q}\left[\mathbb{E}_{\mathbf{Y},\mathbf{Y}'\sim Q}[u_P(\mathbf{Y},\mathbf{Y}')] + \frac{1}{n^2}\sum_{1\leq i,j\leq n} u_P(\mathbf{Y}_i,\mathbf{Y}_j) - \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathbf{Y}\sim Q}[u_P(\mathbf{Y},\mathbf{Y}_i)]\right]\right)^{\frac{1}{2}} \\
&= \left(\frac{1}{n}\mathbb{E}_{\mathbf{Y}\sim Q}[u_P(\mathbf{Y},\mathbf{Y})] - \frac{1}{n}\mathbb{E}_{\mathbf{Y},\mathbf{Y}'\sim Q}[u_P(\mathbf{Y},\mathbf{Y}')]\right)^{\frac{1}{2}},
\end{aligned}$$

where the second line follows from the definition of $\mathbb{S}_P^2(\mathbb{X}_n, Q)$ and Lemma 5.16, and the last equality holds since $\mathbf{Y}_i, \mathbf{Y}_j \sim Q$ are i.i.d. To further bound the RHS of the last line, we invoke (5.30) to conclude that $\mathbb{E}_{\mathbf{Y},\mathbf{Y}'\sim Q}[u_P(\mathbf{Y},\mathbf{Y}')] \leq \mathbb{E}_{\mathbf{Y},\mathbf{Y}'\sim Q}[u_P(\mathbf{Y},\mathbf{Y})^{1/2}u_P(\mathbf{Y}',\mathbf{Y}')^{1/2}] = \mathbb{E}_{\mathbf{Y}\sim Q}[u_P(\mathbf{Y},\mathbf{Y})]$. Therefore,

$$\left(\frac{1}{n}\mathbb{E}_{\mathbf{Y}\sim Q}[u_P(\mathbf{Y},\mathbf{Y})] - \frac{1}{n}\mathbb{E}_{\mathbf{Y},\mathbf{Y}'\sim Q}[u_P(\mathbf{Y},\mathbf{Y}')]\right)^{\frac{1}{2}} \leq \frac{1}{n^{\frac{1}{2}}}\left(\mathbb{E}_{\mathbf{Y}\sim Q}[u_P(\mathbf{Y},\mathbf{Y})]\right)^{\frac{1}{2}} \leq \left(\frac{\tau_\infty}{n}\right)^{\frac{1}{2}}.$$

Combining these results shows the first claim:

$$\begin{aligned}
\Pr_{\mathbb{X}_n\sim Q}\left(\mathbb{S}_P(\mathbb{X}_n, Q) > \sqrt{\frac{\tau_\infty}{n}} + \delta\right) &\leq \Pr_{\mathbb{X}_n\sim Q}\left(\mathbb{S}_P(\mathbb{X}_n, Q) - \mathbb{E}_{\mathbb{Y}_n\sim Q}[\mathbb{S}_P(\mathbb{Y}_n, Q)] > \delta\right) \\
&\leq \exp\left(-\frac{n\delta^2}{2\tau_\infty}\right).
\end{aligned}$$

The second claims follows directly by setting the RHS to α and solving for δ. □

5.A.5 Validity of the Bootstrap Approach

We prove the validity of the bootstrap approach in Theorem 5.4. We first discuss the intuition, and then present two lemmas in Section 5.A.5.1. The proof of Theorem 5.4 will be presented in Section 5.A.5.2.

The bootstrap procedure used in robust-KSD is the same as the one in the standard KSD test, and it is not immediately obvious that this gives a valid decision threshold. For the standard KSD test, the V-statistic $D^2(\mathbb{X}_n)$ is *degenerate* under the point null $H_0 : Q = P$, i.e., $\mathbb{E}_{\mathbf{X}\sim Q}[u_P(\cdot, P)] = 0$ when $Q = P$ (Liu et al., 2016, Theorem 4.1), and classic bootstrapping methods for degenerate V-statistics (Arcones and Gine, 1992; Huskova and Janssen, 1993) shows that the bootstrap quantile $q_{\infty,1-\alpha}$ is a valid decision threshold. However, the same argument cannot be applied directly to our robust test, because our null H_0^{C} is a composite one that contains not only P but also other distributions $Q \neq P$, in which case $D^2(\mathbb{X}_n)$ is no longer degenerate as shown by Liu et al. (2016, Theorem 4.1). Nevertheless, our proof shows that $q_{\infty,1-\alpha}$ is still a correct decision threshold.

Intuition of the proof Let $\mathbb{X}_n = \{\mathbf{X}_i\}_{i=1}^n$ be a realisations from some $Q \in \mathcal{P}(\mathbb{R}^d)$. We will first show that the rejection probability of the robust-KSD test can be bounded as

$$\Pr_{\mathbb{X}_n\sim Q}\left(\Delta_\theta(\mathbb{X}_n) > q_{\infty,1-\alpha}(\mathbb{X}_n)\right) \leq \Pr_{\mathbb{X}_n\sim Q}\left(\mathbb{S}_P(\mathbb{X}_n, Q) > q_{\infty,1-\alpha}(\mathbb{X}_n)\right),$$

where $\mathbb{S}_P(\mathbb{X}_n, Q)$ is the P-KSD introduced in Section 5.A.4. We will then prove that the RHS of the above inequality converges to the prescribed test level α as $n \to \infty$ by showing that $q_{\infty,1-\alpha}(\mathbb{X}_n)$ is a valid bootstrap approximation for the $(1-\alpha)$-quantile of the distribution of $\mathbb{S}_P(\mathbb{X}_n, Q)$.

To see this, we first use the closed-form expression for P-KSD (Lemma 5.16) to write

$$\begin{aligned}\mathbb{S}_P^2(\mathbb{X}_n, Q) &= \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q_n}[u_P(\mathbf{X},\mathbf{X}')] - 2\mathbb{E}_{\mathbf{X}\sim Q_n,\mathbf{Y}\sim Q}[u_P(\mathbf{X},\mathbf{Y})] + \mathbb{E}_{\mathbf{Y},\mathbf{Y}'\sim Q}[u_P(\mathbf{Y},\mathbf{Y}')] \\ &= \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q_n}[u_P(\mathbf{X},\mathbf{X}';Q)] = \frac{1}{n^2}\sum_{1\le i,j\le n} u_P(\mathbf{X}_i,\mathbf{X}_j;Q)\,, \end{aligned} \tag{5.49}$$

where we have defined

$$u_P(\mathbf{x},\mathbf{x}';Q) := u_P(\mathbf{x},\mathbf{x}') - \mathbb{E}_{\mathbf{Y}\sim Q}[u_P(\mathbf{x},\mathbf{Y})] - \mathbb{E}_{\mathbf{Y}'\sim Q}[u_P(\mathbf{Y}',\mathbf{x}')] + \mathbb{E}_{\mathbf{Y},\mathbf{Y}'\sim Q}[u_P(\mathbf{Y},\mathbf{Y}')]\,.$$

In other words, $\mathbb{S}_P^2(\mathbb{X}_n, Q)$ is a V-statistic with symmetric function $u_P(\cdot,\cdot;Q)$.

Our key observation is that $\mathbb{S}_P^2(\mathbb{X}_n, Q)$ is the second-order term in the *Hoeffding's decomposition* (Arcones and Gine, 1992, Eq. 3.2) of the V-statistic estimate $D^2(\mathbb{X}_n, P)$ for the standard KSD defined in (5.2). This suggests that (5.3) is an appropriate bootstrap sample. Crucially, this argument does *not* require $Q = P$, unlike in the proof of the standard KSD test. We formalize this argument in the next section.

A naive alternative approach to bootstrap $\mathbb{S}_P^2(\mathbb{X}_n, Q)$ is to note that it is degenerate for any Q and use standard bootstrapping techniques for degenerate V-statistics (e.g., Arcones and Gine, 1992, Theorem 3.5). However, this approach is not applicable here since, to compute the bootstrap sample, it requires the function $u_P(\mathbf{x},\mathbf{x}';Q)$ to be computable at any $\mathbf{x},\mathbf{x}'$, which is not the case for $u_P(\mathbf{x},\mathbf{x}';Q)$ since it involves intractable expectations over Q (recall that u_P has mean zero under P, but not necessarily under Q).

5.A.5.1 Bootstrapping the P-KSD Estimate

Our proof for Theorem 5.4 relies on Arcones and Gine (1992, Lemma 3.4), which states that the asymptotic distribution of the second-order term in the Hoeffding's decomposition of a V-statistic can be approximated by a bootstrap distribution. We restate this result for the case of $\mathbb{S}_P^2(\mathbb{X}_n, Q)$ in Lemma 5.19.

Lemma 5.19 (Arcones and Gine (1992), Lemma 3.4)**.** *Let $\mathbb{X}_\infty = \{\mathbf{X}_i\}_{i=1}^\infty$ be a random sample where $\mathbf{X}_i \sim Q$ are independent, and for any n define $\mathbb{X}_n := \{\mathbf{X}_i\}_{i=1}^n$. Let Q_n be the empirical measure based on $\mathbb{X}_n$ and define $\mathbb{X}_n^* = \{\mathbf{X}_i^*\}_{i=1}^n$, where $\mathbf{X}_i^* \sim Q_n$ are independent conditionally on $\mathbb{X}_n$. Assume $\mathbb{E}_{\mathbf{X}\sim Q}[u_P(\mathbf{X},\mathbf{X})^2] < \infty$. Then, as $n \to \infty$,*

$$\sup_{t\in\mathbb{R}} \left| \Pr_{\mathbf{X}_n^*\sim Q_n}\left(n\cdot\mathbb{S}_P^2(\mathbb{X}_n^*,\,\mathbb{X}_n) \le t \mid \mathbb{X}_\infty\right) - \Pr_{\mathbb{X}_n\sim Q}\left(n\cdot\mathbb{S}_P^2(\mathbb{X}_n,\,Q) \le t\right)\right| \to 0\,.$$

Using this result, we can prove that a valid bootstrap approximation for the distribution of $\mathbb{S}_P(\mathbb{X}_n, Q)$ can be obtained by using the bootstrap sample $D_{\mathbf{W}}(\mathbb{X}_n)$ defined in (5.3). This is summarized in the

next lemma.

Lemma 5.20. *Assume* $\mathbb{E}_{\mathbf{X}\sim Q}[u_P(\mathbf{X},\mathbf{X})^2] < \infty$ *and let* $\mathbf{W} := (W_1,\dots,W_n)$ *be a random vector distributed as Multinomial*$(n; 1/n,\dots,1/n)$. *Under the notation in Lemma 5.19, the following holds as* $n \to \infty$,

$$\sup_{t\in\mathbb{R}} \left| \Pr_{\mathbf{W}}\left(\sqrt{n}\cdot D_{\mathbf{W}}(\mathbb{X}_n) \;\le\; t \mid \mathbb{X}_\infty\right) - \Pr_{\mathbb{X}_n\sim Q}\left(\sqrt{n}\cdot \mathbb{S}_P(\mathbb{X}_n, Q) \;\le\; t\right)\right| \;\to\; 0\,. \tag{5.50}$$

Proof of Lemma 5.20. As argued in (5.49), the squared P-KSD $\mathbb{S}_P^2(\mathbb{X}_n, Q)$ can be written as a V-statistic with symmetric function $u_P(\cdot,\cdot;Q)$. Moreover, direct computation shows that $\mathbb{E}_{\mathbf{X}\sim Q}[u_P(\cdot,\mathbf{X};Q)] \equiv 0$, so the symmetric function $u_P(\cdot,\cdot;Q)$ is *Q-degenerate of order 1* (Arcones and Gine, 1992, pp. 5). On the other hand, it is well-known that the weighted bootstrap form can be equivalently written as an Efron's resampled bootstrap statistic (Janssen, 1994; Dehling and Mikosch, 1994; Janssen, 1997), i.e.,

$$\begin{aligned}
D_{\mathbf{W}}^2(\mathbb{X}_n) &= \frac{1}{n^2}\sum_{1\le i,j\le n}(W_i-1)(W_j-1)u_P(\mathbf{X}_i,\mathbf{X}_j)\\
&= \frac{1}{n^2}\sum_{1\le i,j\le n} W_iW_ju_P(\mathbf{X}_i,\mathbf{X}_j) - W_iu_P(\mathbf{X}_i,\mathbf{X}_j) - W_ju_P(\mathbf{X}_i,\mathbf{X}_j) + u_P(\mathbf{X}_i,\mathbf{X}_j)\\
&\overset{d}{=} \frac{1}{n^2}\sum_{1\le i,j\le n} u_P(\mathbf{X}_i^*,\mathbf{X}_j^*) - u_P(\mathbf{X}_i^*,\mathbf{X}_j) - u_P(\mathbf{X}_i,\mathbf{X}_j^*) + u_P(\mathbf{X}_i,\mathbf{X}_j)\\
&= \frac{1}{n^2}\sum_{1\le i,j\le n} u_P(\mathbf{X}_i^*,\mathbf{X}_j^*;Q_n) \;=\; \mathbb{S}_P^2(\mathbb{X}_n^*,\mathbb{X}_n)\,,
\end{aligned}$$

where $\mathbf{X}_i^*$ and Q_n are defined in Lemma 5.19, the notation $\overset{d}{=}$ denotes equality in distribution, the second last line follows by the definition of $u_P(\cdot,\cdot;Q_n)$, and the last line. Under the assumed moment condition, we can apply Lemma 5.19 together with the above derivation to conclude that a version of (5.50) with the *squared* statistics holds, i.e.,

$$\sup_{t\in\mathbb{R}} \left| \Pr_{\mathbf{W}}\left(n\cdot D_{\mathbf{W}}^2(\mathbb{X}_n) \;\le\; t \mid \mathbb{X}_\infty\right) - \Pr_{\mathbb{X}_n\sim Q}\left(n\cdot \mathbb{S}_P^2(\mathbb{X}_n, Q) \;\le\; t\right)\right| \;\to\; 0\,.$$

The claim then follows by noting that the mapping $u \mapsto \sqrt{u}$ is everywhere continuous on $[0,\infty)$ and that weak convergence is preserved by continuous function by the Continuous Mapping Theorem (van der Vaart, 2000, Theorem 2.3). □

5.A.5.2 Proof of Theorem 5.4

We first show that $\mathbb{E}_{\mathbf{X}\sim Q}[u_P(\mathbf{X},\mathbf{X})^2] < \infty$, so in particular $\mathbb{E}_{\mathbf{X}\sim Q}[u_P(\mathbf{X},\mathbf{X})] < \infty$ and we can apply Lemma 5.17. Defining $\mathbf{s}_{p,w}(x) = w(x)\mathbf{s}_p(x)$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbf{X}\sim Q}[u_P(\mathbf{X},\mathbf{X})^2]\\
&\le\; \mathbb{E}_{\mathbf{X}\sim Q}\Big[\big(\|\mathbf{s}_{p,w}(\mathbf{x})\|_2^2 h(0) + 2\|\mathbf{s}_{p,w}(\mathbf{x})\|_2\|\nabla w(\mathbf{x})\|_2 h(0) + \|\nabla w(\mathbf{x})\|_2^2 h(0) + w(\mathbf{x})^2|\nabla^\top\nabla h(0)|\big)^2\Big]\\
&\le\; 4\Big(\mathbb{E}_{\mathbf{X}\sim Q}\big[\|\mathbf{s}_{p,w}(\mathbf{x})\|_2^4\big]h^2(0) + 2\mathbb{E}_{\mathbf{X}\sim Q}\big[\|\mathbf{s}_{p,w}(\mathbf{x})\|_2^2\|\nabla w(\mathbf{x})\|_2^2\big]h^2(0) + \mathbb{E}_{\mathbf{X}\sim Q}\big[\|\nabla w(\mathbf{x})\|_2^4\big]h^2(0)\\
&\quad + \mathbb{E}_{\mathbf{X}\sim Q}\big[w(\mathbf{x})^4\big]|\nabla^\top\nabla h(0)|^2\Big)\,,
\end{aligned}$$

where the second line follows from (5.29), and the last line holds since $(a+b+c+d)^2 \leq 4(a^2+b^2+c^2+d^2)$ for any $a, b, c, d \in \mathbb{R}$. The RHS of the last inequality is finite under the assumed conditions on k.

We are now ready to prove the first result of our theorem. For any $Q \in \mathcal{B}^{\mathrm{KSD}}(P;\theta)$, we have

$$\begin{aligned}
&\Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\left(\Delta_\theta(\mathbb{X}_n) \ > \ q_{\infty,1-\alpha}(\mathbb{X}_n)\right) \\
&= \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\left(\max(0, D(\mathbb{X}_n) - \theta) \ > \ q_{\infty,1-\alpha}(\mathbb{X}_n)\right) \\
&= \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\left(\inf_{Q' \in \mathcal{B}^{\mathrm{KSD}}(P;\theta)} \mathbb{S}_P(\mathbb{X}_n, Q') \ > \ q_{\infty,1-\alpha}(\mathbb{X}_n)\right) \\
&\leq \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\left(\mathbb{S}_P(\mathbb{X}_n, Q) \ > \ q_{\infty,1-\alpha}(\mathbb{X}_n)\right), \qquad (5.51)
\end{aligned}$$

where the second line holds by Lemma 5.17 and noting that $D(\mathbb{X}_n)$ is equivalent to the KSD between the empirical measure based on $\mathbb{X}_n$ and P, and the last line holds since $Q \in \mathcal{B}^{\mathrm{KSD}}(P;\theta)$. To show the first claim of our theorem, it suffices to show that the RHS of (5.51) converges to α as $n \to \infty$. Defining the term

$$\delta_n \ := \ \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\left(\mathbb{S}_P(\mathbb{X}_n, Q) \ > \ q_{\infty,1-\alpha}(\mathbb{X}_n)\right) - \Pr_{\mathbf{W}}\left(D_{\mathbf{W}}(\mathbb{X}_n) \ > \ q_{\infty,1-\alpha}(\mathbb{X}_n) \mid \mathbb{X}_\infty\right),$$

we can write the RHS of (5.51) as

$$\begin{aligned}
\Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\left(\mathbb{S}_P(\mathbb{X}_n, Q) \ > \ q_{\infty,1-\alpha}(\mathbb{X}_n)\right) &= \Pr_{\mathbf{W}}\left(D_{\mathbf{W}}(\mathbb{X}_n) \ > \ q_{\infty,1-\alpha}(\mathbb{X}_n) \mid \mathbb{X}_\infty\right) + \delta_n \\
&= \alpha + \delta_n,
\end{aligned}$$

where the last step holds since $q_{\infty,1-\alpha}(\mathbb{X}_n)$ is the $(1-\alpha)$-quantile of the conditional distribution of $D_{\mathbf{W}}(\mathbb{X}_n)$ given $\mathbb{X}_\infty$. Moreover, since we have shown that $\mathbb{E}_{\mathbf{X} \sim Q}[u_P(\mathbf{X}, \mathbf{X})^2] < \infty$, we can apply Lemma 5.20 to conclude that $\delta_n \to 0$ as $n \to \infty$. Combining with (5.51), we have proved that $\Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\left(\Delta_\theta(\mathbb{X}_n) \ > \ q_{\infty,1-\alpha}(\mathbb{X}_n)\right) \to \alpha$, which shows the first claim of our theorem.

Now assume $Q \notin \mathcal{B}^{\mathrm{KSD}}(P;\theta)$, so $\theta - D(Q,P) < 0$. We have

$$\begin{aligned}
&\Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\left(\Delta_\theta(\mathbb{X}_n) > q_{\infty,1-\alpha}(\mathbb{X}_n)\right) \\
&= \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\left(D(\mathbb{X}_n) - \theta > q_{\infty,1-\alpha}(\mathbb{X}_n)\right) \qquad (5.52) \\
&= \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\left(\sqrt{n}\big(D^2(\mathbb{X}_n) - D^2(Q,P)\big) > \sqrt{n}(q_{\infty,1-\alpha}(\mathbb{X}_n) + \theta)^2 - \sqrt{n}D^2(Q,P)\right), \qquad (5.53)
\end{aligned}$$

where (5.52) holds since it can be checked that $\Delta_\theta(\mathbb{X}_n) > \gamma$ if and only if $D(\mathbb{X}_n) - \theta > \gamma$ for any $\gamma \geq 0$. The argument in Liu et al. (2016, Theorem 4.1) shows that that $\sqrt{n}(D^2(\mathbb{X}_n) - D^2(Q,P))$ converges weakly to a Gaussian limit assuming $\mathbb{E}_{\mathbf{X} \sim Q}[u_P(\mathbf{X}, \mathbf{X}')^2] < \infty$, which holds since (5.30) and Jensen's inequality imply

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}, \mathbf{X}' \sim Q}[u_P(\mathbf{X}, \mathbf{X}')^2] &\leq \mathbb{E}_{\mathbf{X}, \mathbf{X}' \sim Q}[u_P(\mathbf{X}, \mathbf{X}) u_P(\mathbf{X}', \mathbf{X}')] \\
&= \left(\mathbb{E}_{\mathbf{X} \sim Q}[u_P(\mathbf{X}, \mathbf{X})]\right)^2 \\
&\leq \mathbb{E}_{\mathbf{X} \sim Q}[u_P(\mathbf{X}, \mathbf{X})^2],
\end{aligned}$$

which is finite as shown before. On the other hand, the weak convergence of $\sqrt{n}(D^2(\mathbb{X}_n) - D^2(Q,P))$

also implies $q_{\infty,1-\alpha}(\mathbb{X}_n) \to 0$ as $n \to \infty$, so when $D(Q,P) > \theta$,

$$\lim_{n\to\infty} (q_{\infty,1-\alpha}(\mathbb{X}_n) + \theta)^2 - D^2(Q,P) \;=\; \theta^2 - D^2(Q,P) \;<\; 0\,.$$

Combing these arguments shows that in the probability in (5.53), the LHS converges to a non-degenerate distribution, while the RHS tends to $-\infty$. This implies (5.53) converges to 1, proving the second claim.

5.A.6 Connections Between KSD Balls and Contamination Models

This section provides proofs for the results in Section 5.4.2.

- Section 5.A.6.1 states and proves Lemma 5.21, an intermediary result which provides a decomposition of the KSD under Huber's contamination models.
- Section 5.A.6.2 proves Proposition 5.5 using Lemma 5.21.
- Section 5.A.6.3 proves Proposition 5.6.
- Section 5.A.6.4 provides an example of using Proposition 5.6 to bound the KSD between t- and Gaussian distributions.

5.A.6.1 Statement and Proof of Lemma 5.21

Lemma 5.21. *Let $\epsilon \in [0,1]$ and $Q = (1-\epsilon)P + \epsilon R$, for any $R \in \mathcal{P}(\mathbb{R}^d)$ such that $\mathbb{E}_{\mathbf{Y}\sim R}[u_P(\mathbf{Y},\mathbf{Y})^{1/2}] < \infty$. Assume $k \in \mathcal{C}_b^{(1,1)}$ and $\mathbb{E}_{\mathbf{X}\sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2] < \infty$. Then $D(Q,P) = \epsilon D(R,P)$.*

Proof of Lemma 5.21. The assumption $\mathbb{E}_{\mathbf{Y}\sim R}[u_P(\mathbf{Y},\mathbf{Y})^{1/2}] < \infty$ implies that $D(R,P)$ is well-defined by Gorham and Mackey (2017, Proposition 2). Moreover, when $k \in \mathcal{C}_b^{(1,1)}$ and $\mathbb{E}_{\mathbf{X}\sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2] < \infty$, Gorham and Mackey (2015, Proposition 1) shows that $\mathbb{E}_{\mathbf{X}\sim P}[u_P(\mathbf{X},\cdot)] = 0$. Using this and the linearity of expectation,

$$\begin{aligned} D^2(Q,P) &= (1-\epsilon)^2 \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim P}[u_P(\mathbf{X},\mathbf{X}')] + 2(1-\epsilon)\epsilon \mathbb{E}_{\mathbf{X}\sim P,\mathbf{Y}\sim R}[u_P(\mathbf{X},\mathbf{Y})] \\ &\quad + \epsilon^2 \mathbb{E}_{\mathbf{Y},\mathbf{Y}'\sim R}[u_P(\mathbf{Y},\mathbf{Y}')] \\ &= (1-\epsilon)^2 D^2(P,P) + \epsilon^2 D^2(R,P) \\ &= \epsilon^2 D^2(R,P)\,. \end{aligned}$$

Taking square-root of both sides gives the desired result. □

5.A.6.2 Proof of Proposition 5.5

The proof follows a similar approach as Chérief-Abdellatif and Alquier (2022, Lemma 3.3). Under the assumed kernel conditions, Lemma 5.2 shows that $\sup_{\mathbf{x},\mathbf{x}'\in\mathbb{R}^d} u_P(\mathbf{x},\mathbf{x}') \le \tau_\infty < \infty$, which implies $D^2(Q,P) = \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}[u_P(\mathbf{X},\mathbf{X}')] \le \tau_\infty < \infty$ for any probability measure $Q \in \mathcal{P}(\mathbb{R}^d)$. In particular,

$D(Q, P)$ is well-defined for any $Q \in \mathcal{P}(\mathbb{R}^d)$ and the conditions in Lemma 5.21 are met. For any $Q = (1-\epsilon)P + \epsilon R$ with $Q \in \mathcal{P}(P; \epsilon_0)$, applying Lemma 5.21 shows that

$$D(Q, P) \;=\; \epsilon D(R, P) \;\leq\; \epsilon_0 D(R, P) \;\leq\; \epsilon_0 \tau_\infty^{\frac{1}{2}} \,. \tag{5.54}$$

To show that this bound is tight, we can lower bound the KSD as

$$\begin{aligned}
\sup_{Q \in \mathcal{P}(P;\epsilon_0)} D(Q, P) \;&\geq\; \sup_{\mathbf{z} \in \mathbb{R}^d, \epsilon \in [0, \epsilon_0]} D\big((1-\epsilon)P + \epsilon \delta_{\mathbf{z}}, P\big) \\
&=\; \sup_{\mathbf{z} \in \mathbb{R}^d, \epsilon \in [0, \epsilon_0]} \epsilon D(\delta_{\mathbf{z}}, P) \\
&=\; \epsilon_0 \sup_{\mathbf{z} \in \mathbb{R}^d} u_P(\mathbf{z}, \mathbf{z})^{\frac{1}{2}} \\
&=\; \epsilon_0 \tau_\infty^{\frac{1}{2}} \,,
\end{aligned}$$

where the first inequality holds since $R = \delta_{\mathbf{z}}$ is only one possible type of perturbation, the first equality follows from Lemma 5.21, the second equality holds because the supremum over ϵ is reached at ϵ_0, and the last step holds as $\sup_{\mathbf{z} \in \mathbb{R}^d} u_P(\mathbf{z}, \mathbf{z}) = \tau_\infty$ by definition. Combining with the upper bound (5.54), we have shown that $\sup_{Q \in \mathcal{P}(P;\epsilon_0)} D(Q, P) = \epsilon_0 \tau_\infty^{1/2}$.

5.A.6.3 Proof of Proposition 5.6

Under the assumed conditions on k, Lemma 5.2 shows that the Stein kernel is bounded by $\tau_\infty = \sup_{\mathbf{x} \in \mathbb{R}^d} u_P(\mathbf{x}, \mathbf{x}) < \infty$, so in particular $D(Q, P)$ is well-defined for all $Q \in \mathcal{P}(\mathbb{R}^d)$. Under the assumed integrability condition, we have $\mathbb{E}_{\mathbf{X} \sim P}[u_P(\mathbf{X}, \cdot)] = 0$ as argued in the paragraph before (5.11). We can therefore rewrite the squared KSD as

$$\begin{aligned}
D^2(Q, P) \;&=\; \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} u_P(\mathbf{x}, \mathbf{x}') q(\mathbf{x}) q(\mathbf{x}') \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{x}' \\
&=\; \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} u_P(\mathbf{x}, \mathbf{x}') (q(\mathbf{x}) - p(\mathbf{x}))(q(\mathbf{x}') - p(\mathbf{x}')) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{x}' \,.
\end{aligned}$$

Using the assumed bound $|q(\mathbf{x}) - p(\mathbf{x})| \leq \delta(\mathbf{x})$ and Lemma 5.2, we can bound the RHS of the above line by

$$\begin{aligned}
\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} u_P(\mathbf{x}, \mathbf{x}') \delta(\mathbf{x}) \delta(\mathbf{x}') \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{x}' \;\leq\; \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |u_P(\mathbf{x}, \mathbf{x}')| \delta(\mathbf{x}) \delta(\mathbf{x}') \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{x}' \;&\leq\; \tau_\infty \bigg(\int_{\mathbb{R}^d} \delta(\mathbf{x}) \, \mathrm{d}\mathbf{x} \bigg)^2 \\
&=\; \tau_\infty \delta_0^2 \,.
\end{aligned}$$

Taking the square-root of both sides implies $D(Q, P) \leq \tau_\infty^{1/2} \delta_0$.

5.A.6.4 KSD Balls and Fat Tails

The next result serves as an example for how Proposition 5.6 can be applied to design a robust KSD test when the model is Gaussian but data are drawn from t-distributions with equal mean and variance.

Proposition 5.22. *Let $P = \mathcal{N}(0,1)$ and let $Q_\nu = t_\nu\sqrt{(\nu-2)/\nu}$, where t_ν is the t-distribution with degree-of-freedom (dof) $\nu > 2$. Denote by p and q_ν their probability density functions, and by F_∞ and F_ν their cumulative distribution functions, respectively. Then $p(x)$ and $q_\nu(x)$ have exactly two intersections $a_1 < a_2$ on $(0,\infty)$. If furthermore k satisfies the conditions in Lemma 5.2, then* $\mathrm{KSD}(Q_\nu, P) \leq 4\tau_\infty^{1/2}(F_\nu(a_1) - F_\infty(a_1) + F_\infty(a_2) - F_\nu(a_2))$.

This result suggests that, to ensure quantitative robustness against the scaled t-distribution with ν degrees-of-freedom, we can choose the uncertainty radius in the robust-KSD test to be $\theta = 4\tau_\infty^{1/2}(F_\nu(a_1) - F_\infty(a_1) + F_\infty(a_2) - F_\nu(a_2))$. The intersection points a_1, a_2 do not have a closed form; instead, we approximate them by numerical solvers, which is trivial for this one-dimensional problem, and the numerical error is negligible as evidenced empirically in Figure 5.4.

Proof of Proposition 5.22. For $\nu > 2$, the change-of-variable formula shows that the scaled t-distribution Q_ν has density function $q(x) = Z_\nu q_\nu^*(x)$, where $Z_\nu := \Gamma(\frac{\nu+1}{2})/\big(\sqrt{\pi(\nu-2)}\Gamma(\frac{\nu}{2})\big)$, $q_\nu^*(x) := (1 + \frac{x^2}{\nu-2})^{-(\nu+1)/2}$, and $\Gamma(\cdot)$ is the Gamma function (see, e.g., Forbes et al., 2011, Chapter 8.1). To show $q_\nu(x)$ intersects with $p(x)$ at exactly two points on $(0,\infty)$, we first show that $q_\nu(x)$ is *strong super-Gaussian* (Palmer et al., 2010), i.e., $x \mapsto \log q_\nu(\sqrt{x})$ is convex on $[0,\infty)$. This is immediate by writing

$$\log q_\nu(\sqrt{x}) \;=\; \log Z_\nu - \frac{\nu+1}{2}\log\left(1 + \frac{x}{\nu-2}\right) ,$$

and noting that $u \mapsto \log(1+u)$ is concave on $[0,\infty)$. Palmer et al. (2010, Theorem 2) shows that any symmetric and strongly super-Gaussian density belongs to the class of *density cross inequalities* (DC_+), which are symmetric densities that cross a Gaussian density of equal variance exactly four times on $\mathbb{R}$, and take higher values than that normal density at $x = 0$ and as $x \to \infty$. Since in this case both Q_ν and P have unit variance, we conclude that q_ν and p intersect at exactly four points on $\mathbb{R}$. By symmetry of these densities about $x = 0$, this shows that q_ν and p must intersect at exactly *two* points on $(0,\infty)$.

Call these two intersections a_1, a_2 and assume without loss of generality $0 < a_1 < a_2$. It then follows that $q_\nu(x) \geq p(x)$ on $[0,a_1) \cup [a_2,\infty)$ and $q_\nu(x) \leq p(x)$ on $[a_1,a_2)$. We therefore have

$$\begin{aligned}
\delta_0 \;&:=\; \int_{\mathbb{R}} |p(x) - q_\nu(x)|\,\mathrm{d}x \\
&=\; 2\int_0^{a_1}(q_\nu(x) - p(x))\,\mathrm{d}x + 2\int_{a_1}^{a_2}(p(x) - q_\nu(x))\,\mathrm{d}x + 2\int_{a_2}^{\infty}(q_\nu(x) - p(x))\,\mathrm{d}x \\
&=\; 2(F_\nu(a_1) - F_\infty(a_1)) + 2\big(F_\infty(a_2) - F_\infty(a_1) - (F_\nu(a_2) - F_\nu(a_1))\big) \\
&\quad + 2\big((1 - F_\nu(a_2)) - (1 - F_\infty(a_2))\big) \\
&=\; 2(F_\nu(a_1) - F_\infty(a_1) + F_\infty(a_2) - F_\infty(a_1) - F_\nu(a_2) + F_\nu(a_1) + F_\infty(a_2) - F_\nu(a_2)) \\
&=\; 4(F_\nu(a_1) - F_\infty(a_1) + F_\infty(a_2) - F_\nu(a_2)) ,
\end{aligned}$$

where in the second line we have again used the symmetry of normal and t densities. Applying Proposition 5.6 completes the proof. □

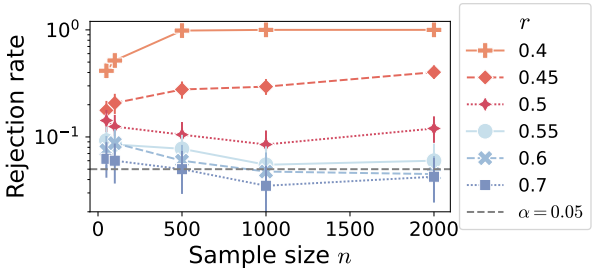


Figure 5.8: Rejection probability (in log scale) against sample size under the univariate contaminated Gaussian model with outlier $z = 10$. The contamination ratio scales as $\epsilon_n = n^{-r}$ for different values of r.

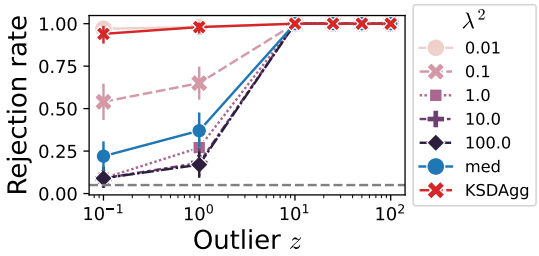


Figure 5.9: Rejection probability for an IMQ kernel with bandwidth λ. "med" is the median heuristic. "KSDAgg" is the test of Schrab et al. (2022). The dashed line shows $\alpha = 0.05$. All tests all reject H_0 for large z.

5.B Complementary Experiments

This section includes additional experimental results. Section 5.B.1 discusses the rate requirement in Theorem 5.3 and demonstrates that the KSD test is no longer qualitatively robust when this condition is not met. Section 5.B.2 reviews the robust MMD tests included in Section 5.5.1 and presents implementation details. Section 5.B.4 includes an ablation study on the choice of kernel bandwidths in the robust-KSD test. Section 5.B.6 studies its scalability with dimension. Section 5.B.7 compares robust-KSD using two different bootstrap methods: the weighted bootstrap used throughout this work, and a wild bootstrap that is more prevalent in the kernel testing literature.

5.B.1 Decline Rate of Contamination Ratio

We show empirically that the rate requirement in the qualitative robustness result Theorem 5.3 is not an artifact of the proof, i.e., the tilted-KSD test is no longer qualitatively robust to Huber's contamination models with contamination ratio ϵ_n if $\epsilon_n = n^{-r}$ for any $r \leq 1/2$, where n is the sample size. We run the standard KSD test with a tilted kernel under the contaminated Gaussian model as described in Section 5.5.1 with dimension $d = 1$ and outlier $z = 10$. The contamination ratio is chosen to be $\epsilon_n = n^{-r}$ for different choices of r, and the probability of rejection as n grows is shown in Figure 5.8. The results are averaged over 400 repetitions instead of 100 repetitions as in Section 5.5.1 to reduce numerical uncertainty. When $r > 0.5$, the rejection probability converges to the prescribed test level α, which is the limit of the rejection probability *without* contamination. This aligns with Theorem 5.3. However, this rejection probability no longer converges to α when $r \leq 0.5$, thus showing that the tilted-KSD test is no longer qualitatively robust.

5.B.2 Implementation Details for the Robust MMD Tests

This section provides implementation details for the two robust MMD-based tests included in Section 5.5.1. Both tests use the Maximum Mean Discrepancy (MMD; Gretton et al., 2012, Definition 2)

between the data-generating distribution Q and the model P as a measure of their disparity. Given a reproducing kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, the squared MMD between Q and P are defined as

$$D^2_{\mathrm{MMD}}(Q, P) \;=\; \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}[k(\mathbf{X}, \mathbf{X}')] + \mathbb{E}_{\mathbf{Y},\mathbf{Y}'\sim P}[k(\mathbf{Y}, \mathbf{Y}')] - 2\mathbb{E}_{\mathbf{X}\sim Q,\mathbf{Y}\sim P}[k(\mathbf{X}, \mathbf{Y})] \,.$$

Given independent samples $\mathbb{X}_n = \{\mathbf{X}_i\}_{i=1}^n$ from Q and $\mathbb{Y}_m = \{\mathbf{Y}_j\}_{j=1}^m$ from P, the MMD can be estimated directly by $D^2_{\mathrm{MMD}}(Q_n, P_m)$, where Q_n and P_m are the empirical measures formed by $\mathbb{X}_n$ and $\mathbb{Y}_m$, respectively (Gretton et al., 2012). The computational cost of such estimate scales with sample size n as $\mathcal{O}((n+m)^2 + mC_{\mathrm{sim}})$, where C_{sim} is the cost of simulating one datum from P, compared with $\mathcal{O}(n^2 + nC_{\mathrm{score}})$ for KSD, where C_{score} denotes the cost of a single score evaluation. In our experiment in Section 5.5.1, P is a Gaussian model, so both simulation and score evaluation are trivial and $C_{\mathrm{sim}}, C_{\mathrm{score}}$ are both small. Hence, we set $m = n$ so that the costs of MMD and of KSD are comparable. Moreover, we will use the squared-exponential kernel $k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|_2^2/(2\gamma^2))$ and set γ^2 using the median heuristic. Squared-exponential kernels are popular for MMDs and are also used in the original papers that proposed the two robust MMD tests (Sun and Zou, 2023; Schrab and Kim, 2024).

5.B.2.1 The MMD-Dev Test

The robust MMD test of Sun and Zou (2023, Eq. (38)), which we refer to as *MMD-Dev*, targets the hypotheses

$$H_0^{\mathrm{MMD}} : \; Q \in \mathcal{B}^{\mathrm{MMD}}(P; \theta_{\mathrm{MMD}}) \,, \qquad H_1^{\mathrm{MMD}} : \; Q \notin \mathcal{B}^{\mathrm{MMD}}(P; \theta_{\mathrm{MMD}}) \,, \tag{5.55}$$

where $\theta_{\mathrm{MMD}} \geq 0$ and $\mathcal{B}^{\mathrm{MMD}}(P; \theta_{\mathrm{MMD}}) = \{Q \in \mathcal{P} : D_{\mathrm{MMD}}(Q, P) \leq \theta_{\mathrm{MMD}}\}$ is the MMD ball centered at P and with radius θ_{MMD}. Given a test level $\alpha \in (0, 1)$, MMD-Dev rejects H_0^{MMD} if $D_{\mathrm{MMD}}(Q_n, P_n) > \theta_{\mathrm{MMD}} + \gamma_n$, where $\gamma_n = \sqrt{2K/n}(1 + \sqrt{-\log \alpha})$. Sun and Zou (2023, Theorem 4) shows that this test controls the Type-I error for any finite n, and is asymptotically optimal against certain alternatives. The decision threshold γ_n is derived using a McDiarmid-type deviation inequality for MMDs due to Gretton et al. (2012, Theorem 8).

In our experiments, we chose $\theta_{\mathrm{MMD}} = \epsilon_0\sqrt{2}$. This ensures that $\mathcal{B}^{\mathrm{MMD}}(P; \theta_{\mathrm{MMD}})$ contains Huber's contamination models with contamination ratios up to ϵ_0, which we show in the next result. Notably, although we can compare the MMD-Dev test with our robust-KSD test on Huber's contamination models, they are in general *not* directly comparable, since they target different null hypotheses.

Lemma 5.23 (MMD balls and Huber's model)**.** *Let k be a reproducing kernel with $0 < k(\mathbf{x}, \mathbf{x}') \leq K < \infty$, for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$. For any $\epsilon_0 \in [0, 1]$, if $\theta_{\mathrm{MMD}} = \epsilon_0\sqrt{2K}$, then $\mathcal{P}(P; \epsilon_0) \subset \mathcal{B}^{\mathrm{MMD}}(P; \theta_{\mathrm{MMD}})$, where $\mathcal{P}(P; \epsilon_0)$ is the Huber's contamination model defined in (5.6).*

Proof of Lemma 5.23. For any $Q \in \mathcal{P}(\mathbb{R}^d)$, we define its kernel mean embedding (Berlinet and Thomas-Agnan, 2004, Chapter 4) as $\mu_Q(\cdot) := \mathbb{E}_{\mathbf{X}\sim Q}[k(\mathbf{X}, \cdot)]$, which is well-defined since k is bounded. By Sriperumbudur et al. (2010b, Theorem 1), the MMD between any $Q, P \in \mathcal{P}(\mathbb{R}^d)$ can be equivalently written as $D_{\mathrm{MMD}}(Q, P) = \|\mu_Q - \mu_P\|_{\mathcal{H}_k}$, where $\mathcal{H}_k$ is the RKHS associated with k. Pick $\epsilon_0 \in [0, 1]$ and

$R \in \mathcal{P}$. For any $\epsilon \in [0, \epsilon_0]$,

$$D_{\mathrm{MMD}}((1-\epsilon)P + \epsilon R, P) = \|\mu_{(1-\epsilon)P+\epsilon R} - \mu_P\|_{\mathcal{H}_k} = \|(1-\epsilon)\mu_P + \epsilon\mu_R - \mu_P\|_{\mathcal{H}_k} = \epsilon\|\mu_R - \mu_P\|_{\mathcal{H}_k} = \epsilon D_{\mathrm{MMD}}(R, P) ,$$

where the second equality holds due to the linearity of the expectation operator. Moreover,

$$\begin{aligned} D^2_{\mathrm{MMD}}(R, P) &= \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim R}[k(\mathbf{X}, \mathbf{X}')] + \mathbb{E}_{\mathbf{Y},\mathbf{Y}'\sim P}[k(\mathbf{Y}, \mathbf{Y}')] - 2\mathbb{E}_{\mathbf{X}\sim R,\mathbf{Y}\sim P}[k(\mathbf{X}, \mathbf{Y})] \\ &\leq \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim R}[k(\mathbf{X}, \mathbf{X}')] + \mathbb{E}_{\mathbf{Y},\mathbf{Y}'\sim P}[k(\mathbf{Y}, \mathbf{Y}')] \\ &\leq 2K , \end{aligned}$$

where the last line holds since $\sup_{\mathbf{x},\mathbf{x}'\in\mathbb{R}^d} k(\mathbf{x}, \mathbf{x}') \leq K$ by assumption. This shows that $D_{\mathrm{MMD}}((1-\epsilon)P + \epsilon R, P) \leq \epsilon\sqrt{2K} \leq \epsilon_0\sqrt{2K}$, so the claim holds. □

5.B.2.2 The dcMMD Test

The dcMMD test (Schrab and Kim, 2024) targets different hypotheses. Given $\epsilon_0 \in [0, 1]$, the null assumes that the observed sample $\mathbb{X}_n$ is generated by firstly drawing i.i.d. random variables from some probability measure, then corrupting them by replacing at most a proportion of ϵ_0 of the sample by arbitrary values. The hypotheses can be formalized as follows:

$$\begin{aligned} H_0^{\mathrm{dcMMD}} &: \text{ At least } (1-\epsilon_0) \times n \text{ random variables in } \mathbb{X}_n \text{ are i.i.d. from } P. \\ H_1^{\mathrm{dcMMD}} &: \text{ Otherwise .} \end{aligned}$$

The dcMMD test rejects H_0^{dcMMD} if $D_{\mathrm{MMD}}(Q_n, P_n) > q_\alpha^{\mathrm{MMD}} + 2\epsilon_0\sqrt{2K}$, where q_α^{MMD} is the empirical quantile of B permutation samples $\{T^b_{\mathrm{MMD}}\}_{b=1}^B$, and each T^b_{MMD} is computed by *(i)* randomly permuting $\mathbb{X}_n \cup \mathbb{Y}_n$, *(ii)* partitioning the permuted set into two subsets $\mathbb{X}_n^b$ and $\mathbb{Y}_n^b$ of size n, and *(iii)* computing $T^b_{\mathrm{MMD}} \coloneqq D_{\mathrm{MMD}}(Q_n^b, P_n^b)$, where Q_n^b and P_n^b are the empirical measures based on $\mathbb{X}_n^b$ and $\mathbb{Y}_n^b$, respectively. We use $B = 500$, which is the default setup in Schrab and Kim (2024).

The dcMMD test gives stronger calibration guarantee than Huber's contamination model because it controls contamination proportion no larger than ϵ_0 for *any realization* of the sample $\mathbb{X}_n$, rather than in expectation as required by Huber's model. However, it is not necessarily stronger than KSD-balls or MMD-balls, since it does not account for the case where *all* samples are under mild perturbation.

5.B.3 Approximating the Supremum of the Stein Kernel

As described in Section 5.4.2, in our experiments we approximated the intractable supremum $\tau_\infty = \sup_{\mathbf{x}\in\mathbb{R}^d} u_P(\mathbf{x}, \mathbf{x})$ by $\max_{i=1,\ldots,n} u_P(\mathbf{x}_i, \mathbf{x}_i)$. We now study how this approximation affects the performance of the robust-KSD test. We run the same contaminated Gaussian experiment in Section 5.5.1 with $d = 1$ and outlier location $z = 10$. For this model, $\mathbf{x} \mapsto u_P(\mathbf{x}, \mathbf{x})$ is a simple, univariate function, so we can compute the ground-truth τ_∞ accurately by numerical optimization. Clearly, the ground

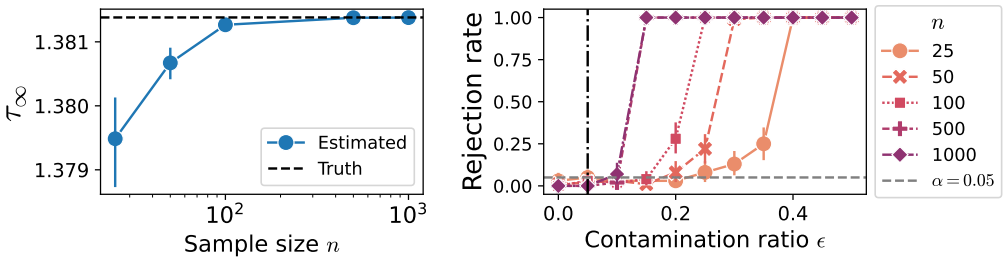


Figure 5.10: *Left.* Estimated τ_∞ using the trick described in Section 5.4.2 compared with the ground-truth; the estimate becomes more accurate with larger samples. *Right.* Rejection probability of robust-KSD under an outlier-contaminated Gaussian model; the Type-I error remains calibrated even for small sample sizes.

truth τ_∞ is equivalent to taking $n = \infty$. This is evidenced by the left plot in Figure 5.10, which shows that the finite-sample approximation $\max_{i=1,\dots,n} u_P(\mathbf{x}_i, \mathbf{x}_i)$ becomes more accurate as n increases. The right plot of Figure 5.10 shows the rejection probability of robust-KSD with θ set to control at most $\epsilon_0 = 0.05$ proportion of contamination. Remarkably, even when this approximation under-estimates τ_∞ (which is the case for small n), robust-KSD still remains well-calibrated. This is not surprising: the supremum τ_∞ represents the maximal contribution of a single datum $\mathbf{x}$ taking arbitrary values in the *entire* sample space $\mathbb{R}^d$, regardless of whether it is present in the *observed* dataset. However, in practice, it suffices to control the contribution of any single datum in the observed dataset.

5.B.4 Ablation Study for Kernel Bandwidth in the Standard KSD Test

We evaluate how the choice of the kernel bandwidth affects the performance of the *standard* KSD test. We run the same contaminated Gaussian experiment as in the previous subsection. The standard KSD test uses an IMQ kernel with bandwidth $\lambda^2 \in \Lambda \cup \{\lambda^2_{\text{med}}\}$, where $\Lambda = \{0.01, 0.1, 1, 10, 100\}$ and λ^2_{med} is the bandwidth chosen by median heuristic. We also include KSDAgg, which also uses an IMQ kernel for a fair comparison. As shown in Figure 5.9, all standard IMQ-KSD tests reject the point null with high probability for large outlier values, *regardless* of the bandwidth value. This suggests that no fixed bandwidth can ensure robustness, which is consistent with Theorem 5.1, and is because the Stein kernel is unbounded regardless of the choice of λ, thus the non-robustness issue persists. The KSDAgg test is even more sensitive to contamination than the IMQ-KSD tests. As noted in Section 5.5.1, this is because KSDAgg is designed to optimally combine multiple bandwidths to boost the test power against all alternatives, including contaminated models.

5.B.5 Gaussian Mean-Shift Experiment

We run a Gaussian mean-shift experiment to demonstrate the performance of our robust test against model deviations other than contamination. We use a Gaussian model $\mathcal{N}(0, I_d)$ in dimension $d = 50$, and draw data from $Q_{\mu_0} = \mathcal{N}(\mu_0 e_1, I_d)$, where $e_1 = (1, 0, \dots, 0)^\top \in \mathbb{R}^d$ and $\mu_0 \in \mathbb{R}$ is a mean-shift. The uncertainty radius θ is chosen to be the Tilted-KSD value corresponding to $\mu_0 = 0, 0.2$ and 0.6, respectively. The purpose of this experiment is to demonstrate that the robust-KSD test is

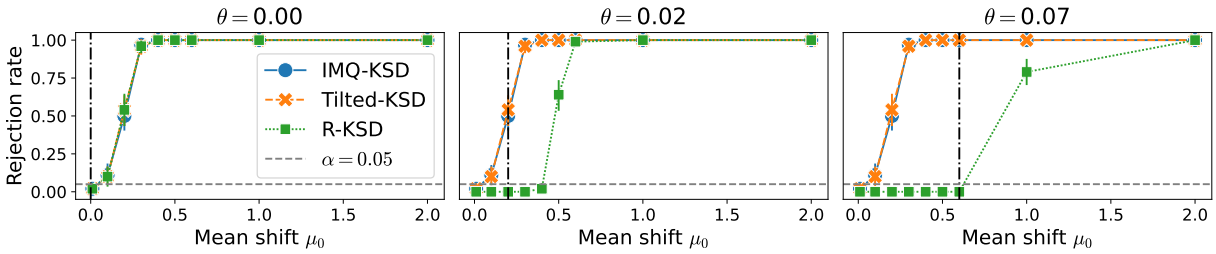


Figure 5.11: Probability of rejection against the mean-shift μ_0 under a Gaussian model in $d = 50$ dimensions. *Black dash-dot line.* Uncertainty radius θ, which is set to be the KSD value $D(Q_{\mu_0}, P)$ corresponding to different values of μ_0.

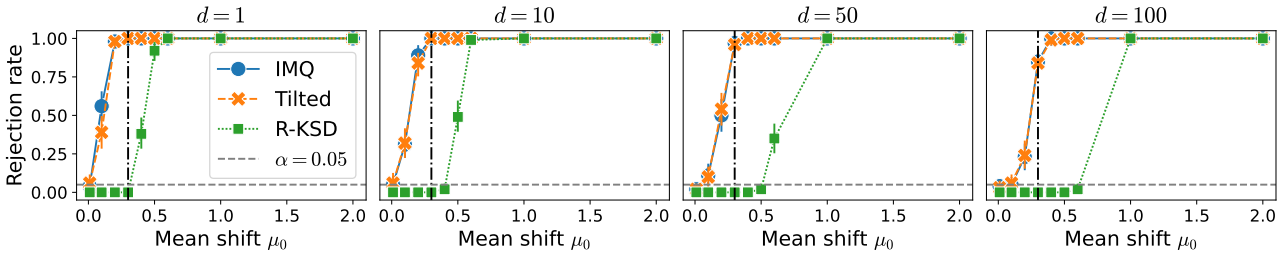


Figure 5.12: Probability of rejection against the mean shift under a Gaussian model in various dimensions. *Grey dotted line.* Test level $\alpha = 0.05$. *Black dash-dot line.* Uncertainty radius θ, which is set to be the KSD value corresponding to $\mu = 0.3$.

well-calibrated when $D(Q_{\mu_0}, P) \leq \theta$ and consistent when $D(Q_{\mu_0}, P) > \theta$. As shown in Figure 5.11, the standard tests reject with probability higher than robust-KSD. This is again expected since the standard tests are not robust to contamination. For mean-shift values not greater than the black vertical line, $Q_{\mu_0} \in \mathcal{B}^{\mathrm{KSD}}(P; \theta)$, so we are under the null hypothesis and robust-KSD rejects no more frequently than the level, showing that it is well-calibrated. For larger μ_0, robust-KSD rejects with probability approaching one, thus showing its power. Moreover, when $\mu_0 = 0$ so that $\theta = 0$, robust-KSD becomes identical to the standard Tilted-KSD test, showing that robust-KSD is indeed a generalization of the standard test.

5.B.6 Scalability with Dimension

We run the Gaussian mean-shift experiment in Section 5.B.5 in different dimensions; other experimental setups remain unchanged. The uncertainty radius for R-KSD is chosen to be the (non-squared) KSD value corresponding to mean-shift $\mu = 0.3$. Results are reported in Figure 5.12. As dimension increases, both the two standard tests and R-KSD have declining power in correctly rejecting for large values of μ. This shows that both the standard and the robust KSD tests suffer from the *curse-of-dimensionality*, a known issue for kernel-based tests (Huang et al., 2023; Reddi et al., 2015; Ramdas et al., 2015). Unsurprisingly, this issue is more prominent for the R-KSD test. This is because the KSD-ball $\mathcal{B}^{\mathrm{KSD}}(P; \theta)$ could potentially include more distributions in higher dimensions.

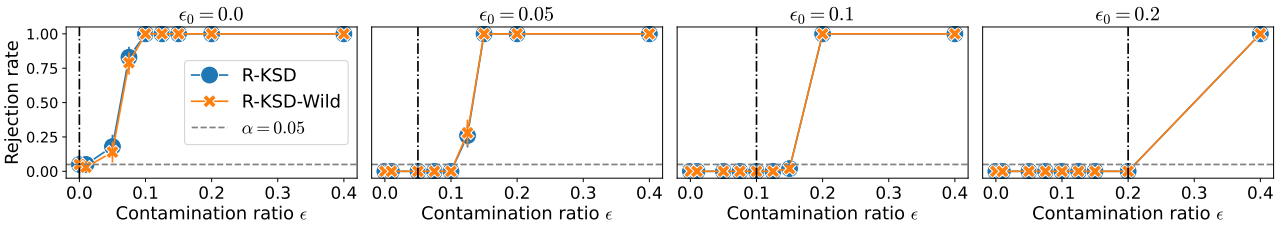


Figure 5.13: Rejection probability under an outlier-contaminated Gaussian model using the weighted bootstrap and the wild bootstrap. *Black dash-dot vertical line.* Uncertainty radius $\theta = \epsilon_0 \tau_\infty^{1/2}$, with τ_∞ estimated following Section 5.4.2.

5.B.7 Weighted Bootstrap and Wild Bootstrap

We compare the robust-KSD test using the weighted bootstrap described in Section 5.2.2 (equivalent to the Efron's bootstrap) against the wild bootstrap due to Leucht and Neumann (2013). For the standard KSD test, weighted bootstrap was used in Liu et al. (2016), while the wild bootstrap is more popular in the literature (Chwialkowski et al., 2016; Schrab et al., 2022; Liu et al., 2023). Compared with weighted bootstrap, the wild approach is more flexible as it can work for dependent samples (Chwialkowski et al., 2014), and theoretical guarantees on the power of the standard test were only proved with the wild bootstrap (Schrab et al., 2022). In this work, we have used the weighted approach since we can then leverage existing theoretical results to show the validity of the test; see Section 5.A.5.1. This is for simplicity rather than necessity.

We numerically compare these two bootstrap methods using the contaminated Gaussian model in Section 5.5.1 and show that the use of weighted bootstrap does not negatively impact the test. Figure 5.13 shows the rejection probability of the robust-KSD test using the weighted bootstrap (R-KSD) and wild bootstrap (R-KSD-Wild). The uncertainty radius is chosen to be $\theta = \epsilon_0 \tau_\infty^{1/2}$ for different values of ϵ_0, following Section 5.5.1. In particular, $\epsilon_0 = 0$ corresponds to setting $\theta = 0$, in which case robust-KSD reduces to the standard KSD. As evident from the plot, these two bootstrap methods produce almost identical results regardless of the value of θ.

5.C Extension to U-statistics

In this work, we have focused primarily on using the V-statistic (5.2) for estimating the KSD. An alternative estimator commonly used in the literature (Liu et al., 2016; Jitkrittum et al., 2017; Schrab et al., 2022; Liu et al., 2023) is the following U-statistic $\frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} u_P(\mathbf{X}_i, \mathbf{X}_j)$. In this section, we discuss briefly how to extend our key results to this U-statistic estimate.

Extending the qualitative (non-)robustness results The main ingredients for proving Theorem 5.1 are *(i)* a deviation bound for the V-statistic $D^2(\mathbb{X}_n)$ as shown in Lemma 5.7, and *(ii)* a concentration bound for the bootstrapped quantile $q^2_{\infty,1-\alpha}(\mathbb{X}_n)$ as shown in Lemma 5.8. A U-statistic counterpart of both results can be shown by following the same proof technique and replacing $D^2(\mathbb{X}_n)$ with the U-statistic. For example, with the U-statistic, a decomposition similar to (5.10) can be derived

by summing over only the non-diagonal terms $i \neq j$ and replacing normalizing factors of the form $1/n^2$ to $1/(n(n-1))$. Similarly, the proof of Theorem 5.3 can be extended to U-statistics by following the same steps in Section 5.A.3.

Extending the robust-KSD test To adapt the robust-KSD test to the U-statistic, we first note that rejecting the null if $\Delta_\theta(\mathbb{X}_n) > q_{\infty,1-\alpha}$ is equivalent to rejecting the null if $D^2(\mathbb{X}_n) > (q_{\infty,1-\alpha} + \theta)^2$, as argued in (5.52) and the paragraph thereafter. A natural approach is therefore to replace the V-statistic $D^2(\mathbb{X}_n)$ by the U-statistic, and the bootstrap quantile $q_{\infty,1-\alpha}$ by the square-root of the bootstrap quantile formed with the U-statistic. However, since the bootstrap samples (5.3) based on U-statistics can take *negative* values, the bootstrap quantile can also take negative values, rendering its square-root $q_{\infty,1-\alpha}$ undefined. One solution is to always not reject the null when this happens. This will make little impact when $D^2(Q, P)$ is large so that the bootstrap samples very likely take positive values, but will make the test conservative when $D^2(Q, P) \approx 0$ so that the bootstrap samples can take negative values with high probability.

5.D A Robust KSD Test based on Deviation Bounds

The robust-KSD test introduced in Section 5.4 is only well-calibrated when $n \to \infty$. In this appendix, we now derive a robust test that is well-calibrated with finite samples. The test rejects the null H_0^{C} in (5.8) if $\Delta_\theta(\mathbb{X}_n) > \gamma_n$, where $\gamma_n := \sqrt{\tau_\infty/n} + \sqrt{-2\tau_\infty(\log\alpha)/n}$, with τ_∞ defined in Lemma 5.2 and $\Delta_\theta(\mathbb{X}_n)$ defined in (5.9). We call this test *robust-KSD-Dev* (R-KSD-Dev).

The decision threshold γ_n is based on a deviation bound of the McDiarmid's type (McDiarmid et al., 1989), and the test can be viewed as a counterpart of the robust MMD test of Sun and Zou (2023). The next result, proved in at the end of this section, shows its finite-sample validity under H_0^{C} as well as its consistency under H_1^{C}.

Theorem 5.24. *Let $\mathbb{X}_\infty = \{\mathbf{X}_i\}_{i=1}^\infty$ be a sequence of independent random variables following Q. Suppose k is a tilted kernel satisfying the conditions in Lemma 5.2, and let $\theta \geq 0$. Then*

1. *Under H_0^{C}, we have $\sup_{Q \in \mathcal{B}^{\mathrm{KSD}}(P;\theta)} \Pr_{\mathbb{X}_n \sim Q}\left(\Delta_\theta(\mathbb{X}_n) > \gamma_n\right) \leq \alpha$.*
2. *Under H_1^{C}, if $\mathbb{E}_{\mathbf{X}_1,\mathbf{X}_2 \sim Q}[u_P(\mathbf{X}_1, \mathbf{X}_2)^2] < \infty$, then $\Pr_{\mathbb{X}_n \sim Q}\left(\Delta_\theta(\mathbb{X}_n) > \gamma_n\right) \to 1$, as $n \to \infty$.*

In Figure 5.14, we compare R-KSD-Dev with the bootstrap-based robust-KSD test, the robust MMD test and the standard tests under the outlier-contaminated Gaussian model in Section 5.5.1. We set the uncertainty radius to be $\theta = \epsilon_0 \tau_\infty^{1/2}$ for various values of ϵ_0, where τ_∞ is estimated following Section 5.4.2. As expected, R-KSD-Dev controls Type-I error against Huber's contamination with maximal proportion ϵ_0. However, when $\epsilon > \epsilon_0$, it has lower power compared with robust-KSD and the MMD-based tests. This is because R-KSD-Dev uses a deviation bound to construct the decision threshold, which is a uniform bound over all data-generating distributions and is therefore conservative.

Since the robust-KSD-Dev test has finite-sample guarantee on the Type-I error, it might be preferable to the bootstrap-based robust-KSD test when the sample size is small. On the other hand, the conservative

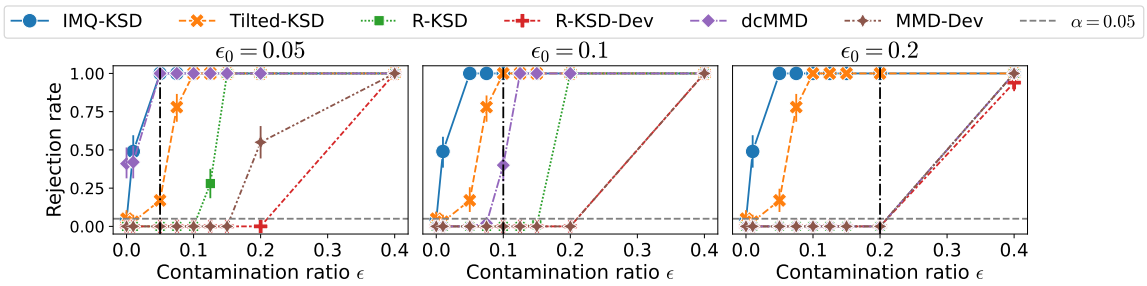


Figure 5.14: Rejection probability under an outlier-contaminated Gaussian model. *Black dash-dot vertical line.* Uncertainty radius $\theta = \epsilon_0 \tau_\infty^{1/2}$, with τ_∞ estimated following Section 5.4.2.

nature of the robust-KSD-Dev test suggests it might only be effective for identifying the most aberrant behaviors.

Proof of Theorem 5.24. Proceeding as in the proof of Theorem 5.4, for any $Q \in \mathcal{B}^{\mathrm{KSD}}(P;\theta)$, we have the inequality

$$\Pr_{\mathbb{X}_n \sim Q}\left(\Delta_\theta(\mathbb{X}_n) \;>\; \gamma_n\right) \;\leq\; \Pr_{\mathbb{X}_n \sim Q}\left(\mathbb{S}_P(\mathbb{X}_n, Q) \;>\; \gamma_n\right),$$

By the deviation bound in Lemma 5.18, the RHS is bounded by α, thus proving the first claim.

We now fix $Q \notin \mathcal{B}^{\mathrm{KSD}}(P;\theta)$ so that we are under H_1^{C}. The same argument in Section 5.A.5.2 shows that $\sqrt{n}\big(D(\mathbb{X}_n) - D(Q,P)\big)$ converges weakly to a Gaussian distribution. On the other hand,

$$\begin{aligned}\Pr_{\mathbb{X}_n \sim Q}\left(\Delta_\theta(\mathbb{X}_n) > \gamma_n\right) \;&\geq\; \Pr_{\mathbb{X}_n \sim Q}\left(D(\mathbb{X}_n) - \theta > \gamma_n\right) \\ &=\; \Pr_{\mathbb{X}_n \sim Q}\left(\sqrt{n}\big(D(\mathbb{X}_n) - D(Q,P)\big) > \sqrt{n}(\gamma_n + \theta - D(Q,P))\right).\end{aligned}$$

Since $D(Q,P) > \theta$ and $\gamma_n \to 0$ as $n \to \infty$, the term $\sqrt{n}\big(\gamma_n + \theta - D(Q,P)\big) \to -\infty$. Therefore, the probability in the last line converges to one, thus proving the second claim. □

Chapter 6

Parameter Estimation: On the Efficiency of Minimum KSD Estimation

Exponential family models are a class of parametric models widely known for their flexibility. However, the likelihood function of these models is typically only available up to a normalizing constant, as the normalization term is often intractable to compute. This limitation makes conventional parameter estimation methods based on the model likelihood infeasible. Alternative methods instead seek to estimate the parameter by fitting the (Hyvärinen) score function of the model, which does not depend on the normalizing constant. Two popular estimators based on this principle are *score matching (SM) estimators* and *minimum kernel Stein discrepancy (KSD) estimators*. Despite their shared underlying principle, little work has been done to compare their relative statistical efficiency. This chapter aims to bridge this gap. By focusing on well-specified models and kernels that admit a Mercer decomposition, we establish conditions under which the KSD estimator is less efficient than or comparably efficient as SM. Our results show that the KSD estimator can be inefficient when both the model and the kernel are sufficiently regular, hightlighting the importance of kernel selection in KSD estimation.

6.1 Introduction

Maximum likelihood estimation (MLE) has been the de-facto choice for fitting parametric models due to its optimality in term of statistical efficiency (see, e.g., van der Vaart 2000, Chapter 5). However, a crucial presumption underlying MLE is the ability to evaluate the likelihood in closed-form for given data and parameter values. This prerequisite can be overly restrictive when dealing with complex models, where the normalizing constant is often intractable.

An example of such models is exponential family models. These models are widely used in practice due to their flexibility in representing complex relationships (Besag, 1974; Yang et al., 2015; Canu and Smola, 2006). However, the normalizing constant of these models depends on the parameter to be optimized and is typically unavailable in closed-form except special cases. As a result, the conventional parameter estimation approach using the maximum likelihood principle becomes impractical. One

potential solution is to resort to estimations of the normalizing constant. However, since MLE often requires computing the normalizing constant at multiple parameter values (this is the case when, for example, gradient-based optimization is used to compute the estimator), this approach quickly becomes computationally prohibitive.

An alternative approach is to instead fit the *(Hyvärinen) score* of the model, which bypasses the need to compute the normalizing constant. Let P be a probability measure on $\mathbb{R}^d$ with Lebesgue density function p that is positively valued. The score function of P is $\mathbf{s}_p(\mathbf{x}) \coloneqq \nabla_{\mathbf{x}} \log p(\mathbf{x})$—the gradient of the log-density of the model with respect to the *data* $\mathbf{x}$. Since $\mathbf{s}_p(\mathbf{x})$ does not depend on the normalizing factor of p, if a parameter estimator is designed to depend on the model only through its score function, then it can be computed efficiently even when the normalizing constant of p is unknown. Two methods that leverage this idea are *score matching* (SM, Hyvärinen, 2005) and *minimum Kernel Stein Discrepancy (KSD) estimation* (Barp et al., 2019; Oates, 2022). Both are minimum divergence estimators (Basu et al., 2011; William, 1981) obtained by minimizing some statistical divergence based on the score function of the model.

Although both SM and KSD are suitable candidates for fitting unnormalized models, it remains unclear how their performance compares to each other. One way to quantify their relative performance is through their relative *statistical efficiency*. The efficiency of an estimator is typically quantified by its asymptotic variance, with low variance indicating a more efficient estimator. Existing research has focused only on the statistical efficiency of SM estimators (Koehler et al., 2022; Qin and Risteski, 2023; Pabbaraju et al., 2023), while the efficiency of minimum KSD estimators remains understudied. This chapter aims to fill this gap.

Outline In Section 6.2, we review the SM and minimum KSD estimators. In Section 6.4, we present our main results on the asymptotic efficiency of minimum KSD estimation, supported by concrete examples and empirical evidence. We conclude with a discussion in Section 6.5. Throughout this chapter, we will use the term "minimum KSD estimator" and the shorthand "KSD estimator" interchangeably.

Notation We introduce new notation for this chapter. Let d and m be positive integers denoting the dimensions of the sample and parameter spaces, respectively. Denote by $\otimes$ the vector outer-product, i.e., for any vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, the quantity $\mathbf{x} \otimes \mathbf{y}$ is the $d \times d$ matrix whose (i,j)-th entry is $x_i y_j$. Let $\mathbb{S}^{m-1}$ denote the unit sphere in $\mathbb{R}^m$. Given a matrix $A \in \mathbb{R}^{m \times m}$, we define $\lambda_{\max}(A) = \max_{\mathbf{u} \in \mathbb{S}^{m-1}} \|A\mathbf{u}\|_2$ and $\lambda_{\min}(A) = \min_{\mathbf{u} \in \mathbb{S}^{m-1}} \|A\mathbf{u}\|_2$ to be the maximum and minimum eigenvalues of A, respectively. Given a probability measure $Q \in \mathcal{P}(\mathbb{R}^d)$ and Q-measurable functions $f, g : \mathbb{R}^d \to \mathbb{R}^s$ for some integer s, we define the $L_2(Q)$ inner-product $\langle f, g \rangle_{L_2(Q)} = \int_{\mathbb{R}^d} f(\mathbf{x})^\top g(\mathbf{x}) Q(\mathrm{d}\mathbf{x})$, the $L_2(Q)$ norm $\|f\|_{L_2(Q)} = \langle f, f \rangle_{L_2(Q)}^{1/2}$, and the space $L_2(Q) = \{f : \mathbb{R}^d \to \mathbb{R}^s \mid \|f\|_{L_2(Q)} < \infty\}$. In particular, the dependence on s is implicit. Given $Q \in \mathcal{P}(\mathbb{R}^d)$ and a positive definite kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with associated reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$, we define the integral operator $\mathcal{T}_k : L_2(Q) \to L_2(Q)$ as

$$(\mathcal{T}_k f)(\cdot) \coloneqq \int_{\mathbb{R}^d} k(\cdot, \mathbf{y}) f(\mathbf{y}) Q(d\mathbf{y}) , \qquad f \in L_2(Q) . \tag{6.1}$$

If f is vector- or matrix-valued, then $\mathcal{T}_k f$ should be understood as applying $\mathcal{T}_k$ to each component of f.

6.2 Background

In this section, we review score-based parameter estimation. We first recall the principle of minimum divergence estimation (see also Section 2.3.3 for a brief overview). We then review the score matching and minimum KSD estimators, two estimators that are suitable for unnormalized models.

6.2.1 Minimum Divergence Estimators

Minimum divergence estimators (MDE, Pardo, 2005; Basu et al., 2011; William, 1981) are estimators of model parameters obtained by minimizing some notion of loss between the observed data and model parameters. Let $\Theta \subset \mathbb{R}^m$ be a parameter space with dimension m, and let $\mathcal{M}(\Theta) = \{p_\theta : \theta \in \Theta\}$ be a statistical model consisting of probability densities (with respect to the Lebesgue measure) parameterized by θ. Given i.i.d. random variable $\{\mathbf{X}_i\}_{i=1}^n$ drawn from some (unknown) data-generating distribution Q and given a function $\mathcal{L} : \mathbb{R}^n \times \Theta \to [0, \infty)$, the MDE is any $\theta_n^{\mathcal{L}}$ that satisfies

$$\theta_n^{\mathcal{L}} \in \operatorname{argmin}_{\theta \in \Theta} \mathcal{L}(\{\mathbf{X}_i\}_{i=1}^n, \theta) .$$

The function $\mathcal{L}$ is called the *loss function* and is often chosen as an empirical estimate of some statistical divergence $D : \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d) \to [0, \infty)$ that satisfies the *separation* and *convergence control* properties, the definitions for which can be found in Section 2.2.7.

One well-known example of MDE is the *maximum likelihood estimator* (MLE), where the loss function is the negative log-likelihood $\mathcal{L}(\{\mathbf{X}_i\}_{i=1}^n, \theta) = -\sum_{i=1}^n \log p_\theta(\mathbf{X}_i)$. This loss can be viewed as a finite-sample estimate for the Kullback-Leibler (KL) divergence up to a additive term that does not depend on θ. Other choices of statistical divergence include the Maximum Mean Discrepancy (MMD, Dziugaite et al., 2015), ϕ-divergences (Pardo, 2005), and Wasserstein metric (Frogner et al., 2015). When the divergence employed is a *distance*, namely when it is symmetric and satisfies a triangle inequality, then minimum divergence estimation is also called minimum *distance* estimation. Examples of distances include MMD and Wasserstein metric, but excludes KL divergence and ϕ-divergences.

6.2.2 Consistency and Asymptotic Normality of MDE

Under regularity conditions on $\mathcal{L}$ and p_θ, MDE enjoys desirable statistical properties such as consistency and asymptotic normality. Assume the model is *well-specified*, namely there exists $\theta_* \in \Theta$ that recovers the data-generating distribution, i.e., $Q = P_{\theta_*}$. It can be shown that the MDE $\theta_n^{\mathcal{L}}$ converges to the ground-truth parameter θ_* both in probability and almost surely as $n \to \infty$; see Newey and McFadden (1994, Theorem 2.1) and the discussions thereafter.

Moreover, Central Limit Theorems (CLTs) for $\theta_n^{\mathcal{L}}$ have been established (see Newey and McFadden 1994, Theorem 3.1 and van der Vaart 2000, Theorem 5.23). Under regularity conditions, the following convergence holds weakly as $n \to \infty$,

$$\sqrt{n}(\theta_n^{\mathcal{L}} - \theta_*) \to \mathcal{N}(0, \Gamma_{\mathcal{L}}) , \tag{6.2}$$

where $\Gamma_{\mathcal{L}} = g_{\mathcal{L}}^{-1}\Sigma_{\mathcal{L}}g_{\mathcal{L}}^{-1}$ for some $m \times m$ matrices $g_{\mathcal{L}}$ and $\Sigma_{\mathcal{L}}$ depending on $\mathcal{L}$, assuming $g_{\mathcal{L}}$ is invertible. The matrix $g_{\mathcal{L}}$ is called the *information matrix* (Amari, 2016), and the *inverse* asymptotic covariance $\Gamma_{\mathcal{L}}^{-1}$, when it exists, is called the *Godambe information matrix* (Godambe, 1991). In the remainder of this chapter, we will denote the asymptotic covariance matrices of an MDE by replacing the $\mathcal{L}$ in the subscript with the name of the divergence. For example, the asymptotic covariance for MLE will be denoted as Γ_{MLE}, and similarly for g_{MLE} and Σ_{MLE}.

The *statistical efficiency* of an estimator $\theta_n^{\mathcal{L}}$ can be quantified by the size of its asymptotic variance measured by, e.g., the operator norm $\|\Gamma_{\mathcal{L}}\|_{\mathrm{op}}$. The smaller the operator norm of the variance, the more efficient the estimator.

CLT for MLE Under regularity condition, a CLT of the form (6.2) holds for the MLE (van der Vaart, 2000, Theorem 5.39), where g_{MLE} and Σ_{MLE} coincide and equal to the *Fisher information matrix*, defined as $\mathbb{E}_{\mathbf{X}\sim Q}[\nabla_\theta \log p_\theta(\mathbf{X}) \otimes \nabla_\theta \log p_\theta(\mathbf{X})]$. The covariance is hence the inverse Fisher information matrix $\Gamma_{\mathrm{MLE}} = \left(\mathbb{E}_{\mathbf{X}\sim Q}[\nabla_\theta \log p_\theta(\mathbf{X}) \otimes \nabla_\theta \log p_\theta](\mathbf{X})\right)^{-1}$. Moreover, Cramér-Rao theorem (Cramér 1999; Rao 1945; Cover and Thomas 2006, Theorem 11.10.1) states that the inverse Fisher information matrix is also the lower bound of the covariance of any unbiased estimator, and therefore MLE is asymptotically optimal in terms of statistical efficiency (although MLE itself is not unbiased in general).

6.2.3 MDE for Unnormalized Models

We are interested in loss functions $\mathcal{L}$ that are computable even when the normalizing constant of p_θ is intractable. A wide range of models have an intractable normalizing constant; examples include *exponential family models* (Barndorff-Nielsen 1978, Part III; Bishop 2006, Section 2.4), *energy-based models* (Cho et al., 2013), and large graphical models (Jordan and Weiss, 2002). As a result, KL divergence is no longer feasible for MDE, as the normalizing constant depends on θ, and thus evaluation of $\mathcal{L}(\{\mathbf{X}_i\}_{i=1}^n, \theta)$ requires estimation of the normalizing constant, which can be prohibitively expensive. MMD and Wasserstein metric are also not feasible. This is because computing these metrics requires samples from the model p_θ, which often demands running extremely computationally expensive MCMC samplers.

To overcome these challenges, one approach leverages score-based statistical divergences, which are designed to depend on the model density only through the *score function* $s_{p_\theta}(\mathbf{x}) \coloneqq \nabla_{\mathbf{x}} \log p_\theta(\mathbf{x})$. Direct computation shows that evaluating $s_{p_\theta}(\mathbf{x})$ at a given $\mathbf{x}$ and θ does not require computing the normalizing constant, making minimum divergence estimation based on such divergences feasible even with unnormalized models. Two most popular score-based divergences are the *score matching* (SM) divergence (also called Fisher divergence; Hyvärinen, 2005, 2007) and *kernel Stein discrepancy* (Barp et al., 2019), which we review in the following subsections.

6.2.4 Minimum KSD Estimators

We review kernel Stein discrepancy (KSD) and minimum KSD estimation. Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a reproducing kernel, and let $P \in \mathcal{P}(\mathbb{R}^d)$ be a probability measure with Lebesgue density $p(\mathbf{x}) > 0$.

The squared *kernel Stein discrepancy* (KSD; Liu et al., 2016; Chwialkowski et al., 2016; Gorham and Mackey, 2017) between two probability measures $Q, P \in \mathcal{P}(\mathbb{R}^d)$, denoted as $D^2(Q, P)$, is defined in (2.8) and restated below

$$\begin{aligned} D^2(Q,P) &:= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} u_P(\mathbf{x},\mathbf{y})Q(\mathrm{d}\mathbf{x})Q(\mathrm{d}\mathbf{y}) \\ u_P(\mathbf{x},\mathbf{x}') &:= \mathbf{s}_p(\mathbf{x})^\top \mathbf{s}_p(\mathbf{x}')k(\mathbf{x},\mathbf{x}') + \mathbf{s}_p(\mathbf{x})^\top \nabla_2 k(\mathbf{x},\mathbf{x}') + \nabla_1 k(\mathbf{x},\mathbf{x}')^\top \mathbf{s}_p(\mathbf{x}') + \nabla_1^\top \nabla_2 k(\mathbf{x},\mathbf{x}') \,. \end{aligned}$$

Barp et al. (2022, Theorem 3.2) showed that $D(Q,P) = 0 \iff Q = P$ so long as $\mathbb{E}_{\mathbf{X}\sim Q}[\|\mathbf{s}_p(\mathbf{X})\|_2]$ and $\mathbb{E}_{\mathbf{X}\sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2]$ are finite and that k satisfies regularity conditions. A sufficient (but not necessary) condition for k is that it is *integrally strictly positive definite* (ISPD, Fukumizu et al., 2009; Sriperumbudur et al., 2010b), meaning that for any finite non-zero signed Borel measure ρ,

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \rho(\mathrm{d}\mathbf{x})k(\mathbf{x},\mathbf{y})\rho(\mathrm{d}\mathbf{y}) \;>\; 0 \,.$$

See Barp et al. (2019, Proposition 1) for a proof. Most kernels commonly used in practice are ISPD; these include squared-exponential (SE) kernels, inverse multi-quadric (IMQ) kernels, and Matérn kernels. On the other hand, one example of non-ISPD kernels is polynomial kernels.

When Q also admits a continuously differentiable density q, the KSD has the following equivalent form in terms of the integral operator (6.1)

$$D^2(Q,P) \;=\; \langle \mathbf{s}_p - \mathbf{s}_q, \mathcal{T}_k(\mathbf{s}_p - \mathbf{s}_q)\rangle_{L_2(Q)} \,. \tag{6.3}$$

A proof is given in Korba et al. (2020, Remark 2).

KSD depends on P only through its score function $\mathbf{s}_p$, which does not depend on the normalizing constant of p. Therefore, the squared KSD $D^2(Q,P)$ can be estimated efficiently even when $P = P_\theta$ has an unnormalized density. Given independent random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim Q$, one estimate for $D^2(Q, P_\theta)$ is the following U-statistic estimator

$$\mathcal{L}_{\mathrm{KSD}}(\{\mathbf{X}_i\}_{i=1}^n, \theta) \;=\; \frac{1}{n(n-1)} \sum_{1\le i\ne j\le n} u_P(\mathbf{X}_i, \mathbf{X}_j) \,.$$

Using $\mathcal{L}_{\mathrm{KSD}}$ as the loss function, Barp et al. (2019) proposed the minimum KSD estimator, defined as any θ_n^{KSD} satisfying

$$\theta_n^{\mathrm{KSD}} \;\in\; \arg\min_{\theta\in\Theta} \mathcal{L}_{\mathrm{KSD}}(\{\mathbf{X}_i\}_{i=1}^n, \theta) \,.$$

Barp et al. (2019) in fact presents a more general form of the KSD estimator, called *Diffusion KSD* (DKSD). Loosely speaking, DKSD generalizes KSD by including a preconditioning term in the construction of u_P. We will focus on the KSD estimator in this chapter, as it is more widely used in the literature (Kanagawa et al., 2022; Altamirano et al., 2023).

Closed-form solution for natural exponential families When the model $\mathcal{M}(\Theta)$ is a *natural exponential family*, the KSD estimator admits a closed-form solution (Barp et al., 2019). Natural

exponential family models are exponential family models with the *canonical* parametrization (Barndorff-Nielsen, 1978, pp. 113), defined as those with probability densities of the form $p_\theta(\mathbf{x}) \propto \exp(\theta^\top T(\mathbf{x}) + b(\mathbf{x}))$, where $T : \mathbb{R}^d \to \mathbb{R}^m$ is a sufficient statistic, and $b : \mathbb{R}^d \to \mathbb{R}$. Natural exponential family models have found broad applications in areas including modelling spatially interacting systems (Besag, 1974), random graph estimation (Yang et al., 2015), classification and density estimation (Canu and Smola, 2006). For such models, the minimum KSD estimator takes the form $\theta_n^{\mathrm{KSD}} = -\frac{1}{2}\Lambda_n^{-1}\eta_n$, where

$$\Lambda_n \coloneqq \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(\mathbf{X}_i, \mathbf{X}_j)(JT(\mathbf{X}_i))^\top JT(\mathbf{X}_j) \,, \tag{6.4a}$$

$$\eta_n \coloneqq \frac{2}{n^2}\sum_{i=1}^n\sum_{j=1}^n \Big[k(\mathbf{X}_i, \mathbf{X}_j)(JT(\mathbf{X}_i))^\top \nabla b(\mathbf{X}_j) + (JT(\mathbf{X}_i))^\top \nabla_2 k(\mathbf{X}_i, \mathbf{X}_j) \tag{6.4b}$$

$$+ (JT(\mathbf{X}_j))^\top \nabla_1 k(\mathbf{X}_i, \mathbf{X}_j)\Big] \,, \tag{6.4c}$$

assuming Λ_n is invertible, where $JT : \mathbb{R}^d \to \mathbb{R}^{d\times m}$ is the Jacobian of the sufficient statistic T defined as $[JT]_{ij} = \partial_{\mathbf{x}_i} T_j(\mathbf{x})$, for $i = 1, \ldots, d$ and $j = 1, \ldots, m$.

Asymptotic normality We present the asymptotic covariance of θ_n^{KSD} for natural exponential family models, which is the primary focus of this chapter, although the CLT extends to more general model families (Barp et al., 2019). Assume $\mathcal{M}(\Theta)$ is a well-specified natural exponential family model. Barp et al. (2019, Theorem 4) showed that, under regularity conditions, the θ_n^{KSD} satisfies the asymptotic normality (6.2) with covariance $\Gamma_{\mathrm{KSD}} \coloneqq g_{\mathrm{KSD}}^{-1}\Sigma_{\mathrm{KSD}}g_{\mathrm{KSD}}^{-1}$, where

$$g_{\mathrm{KSD}} = \langle JT, \mathcal{T}_k JT\rangle_{L_2(Q)} = \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} (JT(\mathbf{x}))^\top JT(\mathbf{y})k(\mathbf{x},\mathbf{y})Q(\mathrm{d}\mathbf{x})Q(\mathrm{d}\mathbf{y}) \,, \tag{6.5}$$

$$\Sigma_{\mathrm{KSD}} = \mathbb{E}_{\mathbf{X}\sim Q}[G_k(\mathbf{X}) \otimes G_k(\mathbf{X})] \,, \tag{6.6}$$

and $G_k(\mathbf{x}) \coloneqq \int_{\mathbb{R}^d} \nabla_\theta u_{p_\theta}(\mathbf{x}, \mathbf{y})Q(\mathrm{d}\mathbf{y})\big|_{\theta=\theta_*}$. A formal statement for this CLT is given in Section 6.A.

6.2.5 Minimum Score-Matching Estimators

Another score-based divergence that can be computed with unnormalized models is the *score matching* (SM) divergence (Hyvärinen, 2005), also known as the *Fisher divergence*. Given two distributions Q and P with continuously differentiable probability densities q, p, respectively, the SM divergence is

$$D_{\mathrm{SM}}(Q, P) \coloneqq \langle \mathbf{s}_p - \mathbf{s}_q, \mathbf{s}_p - \mathbf{s}_q\rangle_{L_2(Q)} = \mathbb{E}_{\mathbf{X}\sim Q}[\|\mathbf{s}_p(\mathbf{X}) - \mathbf{s}_q(\mathbf{X})\|_2^2] \,.$$

Comparing with (6.3), the SM divergence has an analogous form to the KSD except that the integral operator $\mathcal{T}_k$ is replaced by the identity operator. In Section 6.4, we will show that the asymptotic covariance matrices of the SM and KSD estimators also differ by $\mathcal{T}_k$, which is a key observation underpining our analyses.

The SM divergence can be computed efficiently without the normalizing constant of p. Indeed, using integration by parts, Hyvärinen (2005) shows that the SM divergence can be written equivalently as

$D_{\mathrm{SM}}(q,p) = \mathbb{E}_{\mathbf{X}\sim Q}[\|\mathbf{s}_p(\mathbf{X})\|_2^2 + \Delta \log p(\mathbf{X})] + C_Q$, where C_Q is a constant depending on Q but not on P, the operator Δ denotes the Laplacian operator defined as $\Delta f = \sum_{i=1}^d \partial_i^2 f(\mathbf{x})$ for $f : \mathbb{R}^d \to \mathbb{R}$, and $\partial_i f(\mathbf{x})$ denotes the derivative of f with respect to the i-th coordinate of $\mathbf{x}$. Therefore, when $P = P_\theta$ takes a parametric form, C_Q does not depend on θ, and minimizing $D_{\mathrm{SM}}(Q, P_\theta)$ over θ is equivalent to minimizing $\mathbb{E}_{\mathbf{X}\sim Q}[\|\mathbf{s}_{p_\theta}(\mathbf{X})\|_2^2 + \Delta \log p_\theta(\mathbf{X})]$, which does not depend on the normalizing constant of p_θ as neither $\mathbf{s}_{p_\theta}(\mathbf{x})$ nor $\Delta \log p_\theta(\mathbf{x})$ does. The SM estimator is then defined as any θ_n^{SM} satisfying

$$\theta_n^{\mathrm{SM}} \in \arg\min_{\theta\in\Theta} \mathcal{L}_{\mathrm{SM}}(\{\mathbf{X}_i\}_{i=1}^n, \theta) , \quad \text{where } \mathcal{L}_{\mathrm{SM}}(\{\mathbf{X}_i\}_{i=1}^n, P_\theta) := \frac{1}{n}\sum_{i=1}^n \left[\|\mathbf{s}_{p_\theta}(\mathbf{X}_i)\|_2^2 + \Delta \log p_\theta(\mathbf{X}_i)\right] . \tag{6.7}$$

Closed-form solution for natural exponential families When P_θ belongs to a natural exponential family model, SM also admits a close-form solution $\theta_n^{\mathrm{SM}} = -\Xi_n^{-1}\omega_n$, where

$$\Xi_n := \frac{1}{n}\sum_{i=1}^n (JT(\mathbf{X}_i))^\top JT(\mathbf{X}_i) , \qquad \omega_n := \frac{1}{n}\sum_{i=1}^n \left[(JT(\mathbf{X}_i))^\top \nabla b(\mathbf{X}_i) + \sum_{j=1}^m e_j \Delta T_j(\mathbf{X}_i)\right] ,$$

assuming Ξ_n is invertible, and e_j is the j-th standard basis vector in $\mathbb{R}^d$ taking value 1 in its j-th entry and 0 otherwise.

Asymptotic normality An asymptotic normality holds for the SM estimator (Forbes and Lauritzen, 2015; Barp et al., 2019). We again present the result for natural exponential families, although it holds more generally. Under regularity conditions, the SM estimator $\hat{\theta}_n^{\mathrm{SM}}$ satisfies the CLT (6.2) with covariance matrix $\Gamma_{\mathrm{SM}} = g_{\mathrm{SM}}^{-1}(\theta_*)\Sigma_{\mathrm{SM}}g_{\mathrm{SM}}^{-1}(\theta_*)$, where

$$g_{\mathrm{SM}} = \langle JT, JT\rangle_{L_2(P_{\theta_*})} = \mathbb{E}_{\mathbf{X}\sim P_{\theta_*}}[(JT(\mathbf{X}))^\top JT(\mathbf{X})] \tag{6.8}$$

$$\Sigma_{\mathrm{SM}} = \mathbb{E}_{\mathbf{X}\sim P_{\theta_*}}[G(\mathbf{X}) \otimes G(\mathbf{X})] , \tag{6.9}$$

and $G(\mathbf{x}) := \nabla_\theta\left(\|\mathbf{s}_{p_\theta}(\mathbf{x})\|_2^2 + \Delta \log p_\theta(\mathbf{x})\right)\big|_{\theta=\theta_*}$. This CLT result is stated formally in Section 6.A for completeness.

KSD versus SM estimation Both KSD and SM are computable with unnormalized models. SM requires evaluation of the Laplacian $\Delta \log p_\theta$, so it assumes the model density is twice continuously differentiable. In comparison, KSD only requires first order differentiability, making it applicable to models with less smooth densities, e.g., Bessel distributions (Barp et al., 2019). Another difference between these methods is their computational cost: a single evaluation of the loss function takes $\mathcal{O}(n^2 d)$ for KSD and $\mathcal{O}(nd^2)$ for SM (Barp et al., 2019). Consequently, KSD is less computationally expensive in small-data, high-dimensional regimes, while SM is cheaper for large-size, low-dimensional data.

6.3 Related Work

Most work in the literature focuses on the relative statistical efficiency of the SM estimator and the MLE. This relative efficiency is defined by the ratio $\|\Gamma_{\mathrm{SM}}\|_{\mathrm{op}}/\|\Gamma_{\mathrm{MLE}}\|_{\mathrm{op}}$, and Koehler et al. (2022)

showed that $\|\Gamma_{\mathrm{SM}}\|_{\mathrm{op}}/\|\Gamma_{\mathrm{MLE}}\|_{\mathrm{op}}$ can be characterized by the regularity of the geometry of the model density. Specifically, they showed that the efficiency of SM is comparable to that of MLE for models with a small Poincaré constant (e.g., Gaussian models), whereas the SM estimator can be much less efficient if the Poincaré constant is large (e.g., for mixtures of Gaussians with well-separated modes). These results are extended by Qin and Risteski (2023) to a generalized SM loss that incorporates the injection of Gaussian noises into the observed data, an approach commonly adopted to facilitate parameter estimation for generative models (Song and Ermon, 2019; Song et al., 2020). Moreover, Pabbaraju et al. (2023) constructed concrete examples of models for which SM is comparably efficient as MLE but much more computationally efficient to compute. Despite these studies on the SM estimator, the minimum KSD estimator has attracted little attention from the literature.

6.4 Efficiency Bounds via a Kernel Decomposition

In this section, we derive bounds on the statistical efficiency of minimum KSD estimators, assuming the kernel admits a Mercer decomposition and the model is well-specified. Specifically, we will assume a Mercer decomposition defined as follows (see, e.g., Cucker and Zhou, 2007, Theorem 4.10).

Definition 6.1 (Mercer decomposition). *A scalar-valued kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is said to have a Mercer decomposition with respect to a probability measure Q if there exists a sequence of pairs $\{(\lambda_l, \phi_l)\}_{l=1}^{\infty}$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ are non-negative scalars, and ϕ_l are orthonormal functions, i.e., $\langle \phi_l, \phi_r \rangle_{L_2(Q)} = 1$ if $l = r$ and 0 otherwise, such that*

$$k(\mathbf{x}, \mathbf{y}) \;=\; \sum_{l=1}^{\infty} \lambda_l \phi_l(\mathbf{x}) \phi_l(\mathbf{y}) \;, \tag{6.10}$$

where the convergence holds absolutely for each $\mathbf{x}, \mathbf{y} \in \mathcal{R}^d$ and uniformly on $\mathcal{X} \times \mathcal{X}$ for all compact $\mathcal{X} \subset \mathbb{R}^d$.

If the Mercer decomposition (6.10) exists, then $\{(\lambda_l, \phi_l)\}_{l=1}^{\infty}$ are eigenvalue-eigenfunction pairs for the kernel integral operator $\mathcal{T}_k$ defined in (6.1), i.e., $\mathcal{T}_k \phi_l = \lambda_l \phi_l$ for all l. For kernels satisfying mild assumptions, the existence of a Mercer decomposition is guaranteed by Mercer's theorem (Cucker and Zhou, 2007, Theorem 4.10; Sun, 2005, Theorem 2). These assumptions are met by many commonly used kernels, including the SE, IMQ, and Matérn kernels. Explicit expressions for the pairs $\{(\lambda_l, \phi_l)\}_{l=1}^{\infty}$ are non-trivial to compute, but they have been derived for specific examples, e.g., when Q is a Gaussian distribution and k is the SE kernel (Zhu et al., 1998, Section 4). This is given in the next example.

Example 6.1 (SE kernels). *Let k be the SE kernel $k(x, y) = \exp(-(x-y)^2/(2\gamma^2))$ with bandwidth $\gamma > 0$. Then k admits a Mercer decomposition with respect to the standard normal distribution $Q = \mathcal{N}(0, 1)$ with eigenvalues and eigenfunctions defined as (see, e.g., Zhu et al. 1998, Section 4)*

$$\lambda_{l,\gamma} \;=\; \sqrt{\frac{2a}{\alpha_\gamma}} \beta_\gamma^{l-1} \;, \qquad \phi_{l,\gamma}(x) \;=\; \exp\big(-(c_\gamma - a)x^2\big) H_{l-1}\big(\sqrt{2c_\gamma}\, x\big) \;, \qquad l = 1, 2, \ldots \;,$$

where $H_l(x) = (-1)^l \exp\big(x^2\big) \frac{\mathrm{d}^l}{\mathrm{d}x^l} \exp\big(-x^2\big)$ is the l-th (physicists') Hermite polynomial *(Gradshteyn et al., 1980, Section 8.95), $a = 1/4, \beta_\gamma = b_\gamma/\alpha_\gamma, c_\gamma \;=\; (a^2 + 2ab_\gamma)^{1/2}, \alpha_\gamma = a + b_\gamma + c$ and $b_\gamma = 1/(2\gamma^2)$.*

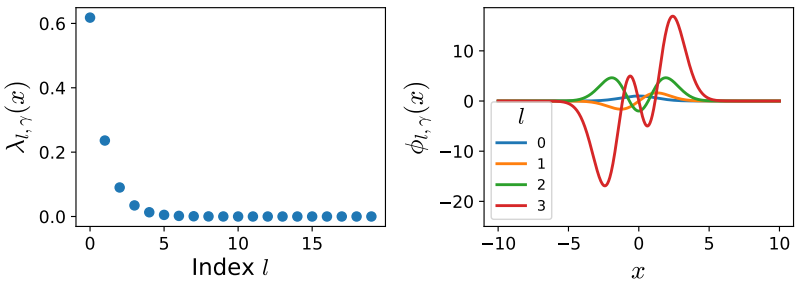


Figure 6.1: Eigendecomposition of the SE kernel with bandwidth $\gamma = 1$ against a standard Gaussian measure. *Left.* First 20 eigenvalues $\lambda_{l,\gamma}$. *Right.* First 4 eigenfunctions $\phi_{l,\gamma}(x)$.

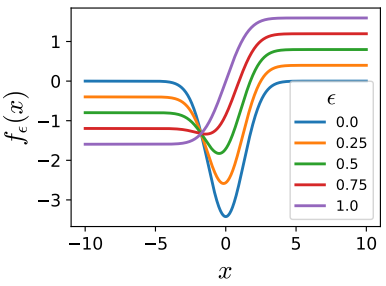


Figure 6.2: The second coordinate $f_\epsilon(x)$ of the sufficient statistic in the augmented Gaussian model (Example 6.2), for different values of ϵ.

In particular, $\beta_\gamma < 1$, so $\lambda_{l,\gamma} \to 0$ as $l \to \infty$. The first few eigenvalues and eigenfunctions are plotted in Figure 6.1.

The decay rate of the eigenvalues is closely related to the smoothness of the kernel—the eigenvalues λ_l decay rapidly for smooth kernels (such as squared-exponential kernels) and more slowly for less smooth kernels (such as Matérn kernels); see, e.g., Rasmussen and Williams (2006, Chapter 4). As we will show in Section 6.4.1, the minimum KSD estimator can be inefficient when the eigenvalues decay rapidly.

Throughout, we also assume the model P_θ belongs to a natural exponential family (Section 6.2.4) that satisfies the following regularity conditions.

Assumption 6.1. *Suppose $p_\theta(\mathbf{x}) \propto \exp\left(\theta^\top T(\mathbf{x}) + b(\mathbf{x})\right)$, where the Jacobian JT of T has linearly independent rows, and assume*

M.1 *$\|JT\|_{\text{op}}, \|\nabla_{\mathbf{x}} b\|_2 \in L_2(Q)$, and $\|\Delta T_j\|_2 \in L_2(Q)$, for all $j = 1, \ldots, m$,*

M.2 *$[JT]_{ij} = \partial_i T_j \in \text{Span}\{\phi_l\}_{l\geq 1}$, for all $i = 1, \ldots, d$ and $j = 1, \ldots, m$, where $\partial_i T_j$ denotes the partial derivative of T with respect to the i-th entry of $\mathbf{x}$. Specifically, $\text{Span}\{\phi_l\}_{l\geq 1} = \left\{f : \exists (\alpha_l)_{l\geq 1} \subset \mathbb{R} \text{ such that } \left\|f - \sum_{l=1}^{L} \alpha_l \phi_l\right\|_{L_2(Q)} \to 0 \text{ as } L \to \infty\right\}$; see Kreyszig (1991, Chapter 2.3).*

We briefly discuss the role of each assumption. Assumption M.1 impose regularity on the model, and similar conditions also appear in the asymptotic normality results for KSD and SM estimators proved in Barp et al. (2019). In Section 6.A, we state these CLT results for completeness and show that they hold under Assumption M.1 and some additional assumptions. Assumption M.1 holds when the functions $\|JT\|_{\text{op}}, \|\nabla_{\mathbf{x}} b\|_2$ and $\|\Delta T_j\|_2$ grow at most polynomially and Q has sub-Gaussian tails. When p_θ is a Gaussian distribution with mean parameterized by θ, they hold whenever Q has finite first moment.

Assumption M.2 ensures we can decompose JT using the eigenfunctions from Definition 6.1 as

$$[JT(\mathbf{x})]_{ij} \;=\; \partial_i T_j(\mathbf{x}) = \sum_{l=1}^{\infty} \langle \phi_l, \partial_i T_j \rangle_{L_2(Q)} \phi_l(\mathbf{x}) \;, \tag{6.11}$$

which holds absolutely and uniformly on compact subsets of $\mathbb{R}^d$. When $\{\phi_l\}_{l\geq 1}$ forms an orthogonal basis to $L_2(Q)$, then Assumption M.2 holds so long as $[JT]_{ij} \in L_2(Q)$ for all i, j. When instead $\{\phi_l\}_{l\geq 1}$ only spans a proper subspace of $L_2(Q)$, which would happen if for example the number of basis functions

ϕ_l is finite, then Assumption M.2 can be viewed as a "regularity" condition that requires each entry $[JT]_{ij}$ to lie in the linear span of these basis functions.

Crucially, equation (6.11) will then allow us to compare the asymptotic covariances of SM and of KSD: assuming the Mercer decomposition (6.10), direct computation gives

$$\mathcal{T}_k[JT]_{ij}(\mathbf{x}) \;=\; \sum_{l=1}^{\infty} \lambda_l \langle \phi_l, \partial_i T_j \rangle_{L_2(Q)} \phi_l(\mathbf{x}) \;, \tag{6.12}$$

for each i and j. Comparing with (6.11), the kernel integral operator $\mathcal{T}_k$ effectively weighs the projection of $\partial_i T_j$ along each eigenfunction ϕ_l by λ_l. In particular, choosing $\mathcal{T}_k \equiv \mathrm{id}$ in (6.12), where id is the identity operator, gives (6.11). This observation will underpin our analyses in this section.

Before presenting the main results, we define the following notation concerning the projections of JT onto spaces defined by the eigenfunctions $\{\phi_l\}_{l\geq 1}$. For each $l = 1, 2, \ldots$, we define the $d \times m$ matrix A_l where $[A_l]_{ij} \coloneqq \langle [JT]_{ij}, \phi_l \rangle_{L_2(Q)} = \langle \partial_i T_j, \phi_l \rangle_{L_2(Q)}$, for $i = 1, \ldots, d$ and $j = 1, \ldots, m$. Intuitively, each $[A_l]_{ij}$ is the projection of $[JT]_{ij}$ onto the direction ϕ_l. For a positive integer L, we denote by Π_L the linear projection operator mapping a scalar-valued function $f \in \mathrm{Span}\{\phi_l\}_{l\geq 1}$ onto the subspace $\mathrm{Span}\{\phi_l\}_{l=1}^{L}$, i.e., the span of the first L eigenfunctions of k, and by $\Pi_L^{\perp}$ the projection operator onto the subspace $\mathrm{Span}\{\phi_l\}_{l>L}$ spanned by the remaining eigenfunctions. For matrix-valued functions F, we overload the notation by writing $\Pi_L F$ to mean Π_L applied to each entry of F. It follows that $JT = \Pi_L JT + \Pi_L^{\perp} JT$ if $[JT]_{ij} \in \mathrm{Span}\{\phi_l\}_{l=1}^{\infty}$ for all i, j.

Moreover, for all $\mathbf{w} \in \mathbb{R}^m$, Lemma 6.5 in Section 6.B.1 shows that the information tensor of the SM estimator can be decomposed as $\mathbf{w}^\top g_{\mathrm{SM}} \mathbf{w} = \mathbf{w}^\top g_{\mathrm{SM},L} \mathbf{w} + \mathbf{w}^\top g^{\perp}_{\mathrm{SM},L} \mathbf{w}$, where

$$g_{\mathrm{SM},L} \;=\; \langle \Pi_L JT, \Pi_L JT \rangle_{L_2(Q)} \;=\; \sum_{l=1}^{L} \|A_l \mathbf{w}\|_2^2 \;, \tag{6.13a}$$

$$g^{\perp}_{\mathrm{SM},L} \;=\; \langle \Pi_L^{\perp} JT, \Pi_L^{\perp} JT \rangle_{L_2(Q)} \;=\; \sum_{l=L+1}^{\infty} \|A_l \mathbf{w}\|_2^2 \;, \tag{6.13b}$$

and similarly for the KSD estimator: $\mathbf{w}^\top g_{\mathrm{KSD}} \mathbf{w} = \mathbf{w}^\top g_{\mathrm{KSD},L} \mathbf{w} + \mathbf{w}^\top g^{\perp}_{\mathrm{KSD},L} \mathbf{w}$, where

$$g_{\mathrm{KSD},L} \;\coloneqq\; \langle \Pi_L JT, \mathcal{T}_k \Pi_L JT \rangle_{L_2(Q)} \;=\; \sum_{l=1}^{L} \lambda_l \|A_l \mathbf{w}\|_2^2 \;, \tag{6.14a}$$

$$g^{\perp}_{\mathrm{KSD},L} \;\coloneqq\; \langle \Pi_L^{\perp} JT, \mathcal{T}_k \Pi_L^{\perp} JT \rangle_{L_2(Q)} \;=\; \sum_{l=L+1}^{\infty} \lambda_l \|A_l \mathbf{w}\|_2^2 \;. \tag{6.14b}$$

Finally, given a positive integer L, we define the following vector-valued functions

$$G_{k,L}(\mathbf{x}) \;\coloneqq\; \sum_{l=1}^{L} \lambda_l A_l^\top (\mathcal{A}^0_{p_\theta} \phi_l)(\mathbf{x}) \;, \qquad G^{\perp}_{k,L}(\mathbf{x}) \;\coloneqq\; \sum_{l=L+1}^{\infty} \lambda_l A_l^\top (\mathcal{A}^0_{p_\theta} \phi_l)(\mathbf{x}) \;, \tag{6.15}$$

$$G_{L}(\mathbf{x}) \;\coloneqq\; \sum_{l=1}^{L} A_l^\top (\mathcal{A}^0_{p_\theta} \phi_l)(\mathbf{x}) \;, \qquad G^{\perp}_{L}(\mathbf{x}) \;\coloneqq\; \sum_{l=L+1}^{\infty} A_l^\top (\mathcal{A}^0_{p_\theta} \phi_l)(\mathbf{x}) \tag{6.16}$$

where $\mathcal{A}^0_{p_\theta}$ denotes a modified Langevin Stein operator mapping scalar-valued functions to vector-valued ones and is defined as $(\mathcal{A}^0_{p_\theta} f)(\mathbf{x}) = \mathbf{s}_{p_\theta}(\mathbf{x}) f(\mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x})$, where $f : \mathbb{R}^d \to \mathbb{R}$. This is in contrast to the standard Langevin Stein operator (2.6), which maps vector fields to scalar-valued functions. Intuitively, $G_{k,L}$ and $G^{\perp}_{k,L}$ are obtained by decomposing G_k (defined below (6.6)) onto $\mathrm{Span}\{\phi_l\}_{l=1}^{L}$ and $\mathrm{Span}\{\phi_l\}_{l=L}^{\infty}$, respectively, and a similar interpretation applies to $G_L, G^{\perp}_L$. Indeed, it can be shown that $G_k(\mathbf{x}) = G_{k,L}(\mathbf{x}) + G^{\perp}_{k,L}(\mathbf{x})$ and $G(\mathbf{x}) = G_L(\mathbf{x}) + G^{\perp}_L(\mathbf{x})$; see Lemma 6.6 in Section 6.B.1.

6.4.1 KSD with Overly Smoothing Kernels is Inefficient

We now investigate when minimum KSD is less efficient than SM, and present a concrete example with an augmented Gaussian model to verify our result. Specifically, our result reveals that this can happen when the "signals" in the function JT is smoothed out by the kernel.

This is formalized in the following result. We consider a fixed data-generating distribution Q, and a natural exponential family model $\mathcal{M}(\Theta)$. We assume $\mathcal{M}(\Theta)$ is well-specified, so $Q = P_{\theta_*}$ for some $\theta_* \in \mathcal{M}(\Theta)$. The next result gives conditions under which we can derive a lower bound on $\|\Gamma_{\mathrm{KSD}}\|_{\mathrm{op}}$. The proof is deferred to Section 6.B.2.

Theorem 6.1. *Suppose Assumption 6.1 holds. Assume k has a Mercer decomposition with respect to $Q = P_{\theta_*}$ with eigenvalue-eigenfunction pairs $\{(\lambda_l, \phi_l)\}_{l=1}^{\infty}$, and assume $g_{\mathrm{KSD}} = \langle JT, \mathcal{T}_k JT\rangle_{L_2(P_{\theta_*})}$ is invertible. Let $\mathbf{w} \in \mathrm{argmin}_{\mathbf{u}\in\mathbb{S}^{m-1}} \mathbf{u}^\top g_{\mathrm{SM},1}\mathbf{u}$, where $g_{\mathrm{SM},1}$ is defined in (6.13) with $L = 1$. Assume the first two eigenvalues in the Mercer decomposition satisfy*

$$\frac{\lambda_2}{\lambda_1} \;\le\; \min\left\{ \frac{\|A_1^\top A_1 \mathbf{w}\|_2}{\sum_{l=2}^{\infty} \|A_l^\top A_l \mathbf{w}\|_2}\,, \quad \frac{\left\|(A_1\mathbf{w})^\top (\mathcal{A}^0_{p_{\theta_*}}\phi_1)\right\|_{L_2(Q)}}{2\sum_{l=2}^{\infty}\left\|(A_l\mathbf{w})^\top (\mathcal{A}^0_{p_{\theta_*}}\phi_l)\right\|_{L_2(Q)}} \right\}. \tag{6.17}$$

Moreover, assume there exists $C > 0$ such that

$$\mathbb{E}_{\mathbf{X}\sim P_{\theta_*}}\left[(\mathcal{A}^0_{p_{\theta_*}}\phi_1)(\mathbf{X}) \otimes (\mathcal{A}^0_{p_{\theta_*}}\phi_1)(\mathbf{X})\right] \;\succcurlyeq\; C I_m\,. \tag{6.18}$$

Then $\|\Gamma_{\mathrm{KSD}}\|_{\mathrm{op}} \ge 16^{-1} C \lambda_{\min}(g_{\mathrm{SM},1})^{-1}$, where $\lambda_{\min}(g_{\mathrm{SM},1})^{-1}$ is the smallest eigenvalue of $g_{\mathrm{SM},1}$.

This result implies that the asymptotic variance of the KSD estimator can explode to infinity when $\lambda_{\min}(g_{\mathrm{SM},1})$ is small. By (6.13) with $L = 1$, we can write $g_{\mathrm{SM},1} = \langle \Pi_1 JT, \Pi_1 JT\rangle_{L_2(Q)}$, where $\Pi_1 JT$ is the projection of JT onto the span of ϕ_1. Therefore, $\lambda_{\min}(g_{\mathrm{SM},1})$ is small when $\Pi_L JT$ is almost singular, which can happen when each entry of JT is almost orthogonal to ϕ_1 so that $\Pi_L JT \approx 0$. Theorem 6.1 then highlights the potential inefficiency of KSD: for a model satisfying the stated assumptions, if JT is almost orthogonal to ϕ_1, and the eigenvalues of the kernel k decays so fast that the projection of JT along the other eigenfunctions is also "smoothed-out", then $\|\Gamma_{\mathrm{KSD}}\|_{\mathrm{op}}$ will take a large value.

We now provide interpretations on the assumptions in Theorem 6.1.

- Assumption 6.1 and the invertibility of g_{KSD} are needed to establish the asymptotic normality of the KSD and SM estimators (Barp et al., 2019). Assumption 6.1 is discussed in detail in the paragraphs following it. As for the invertibility of g_{KSD}, a sufficient condition is that JT has linearly independent rows and k is ISPD (Barp et al., 2019, Lemma 3 of Appendix D.3.1).

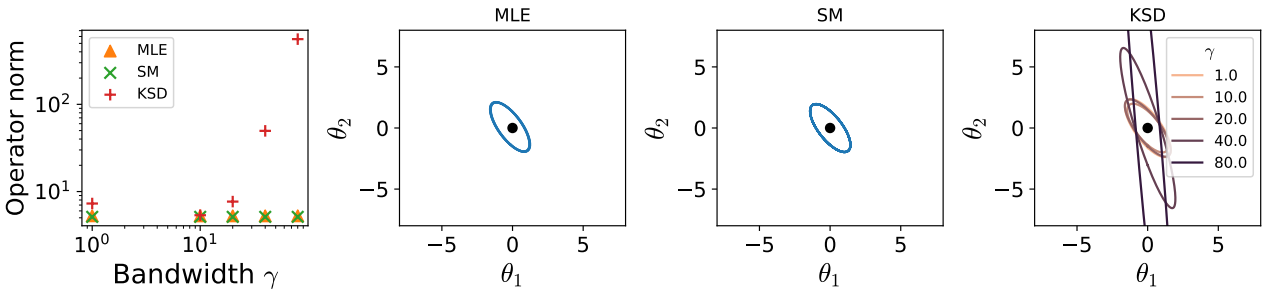


Figure 6.3: *Left.* Operator norm of the asymptotic covariance matrix in the augmented Gaussian-location model. *Middle left, middle right and right.* Level sets for the asymptotic normal distribution predicted by CLT. The ground-truth parameter $\theta_* = (0,0)^\top$ is shown by the black dot. The KSD estimator becomes inefficient as γ grows.

- Equation (6.17) is a spectral gap condition. Intuitively, it says that the eigen-spectrum of the kernel decays so fast that the projection of JT becomes negligible along all directions but ϕ_1. When P_{θ_*} is a standard Gaussian distribution and k is the squared-exponential kernel with bandwidth γ (cf. Example 6.1), this condition holds for large enough γ; see the proof of Corollary 6.1.1.
- The condition that (6.18) holds for some $C > 0$ is equivalent to the positive-definiteness of the matrix $\mathbb{E}_{\mathbf{X}\sim P_{\theta_*}}\left[(\mathcal{A}^0_{p_{\theta_*}}\phi_1)(\mathbf{X}) \otimes (\mathcal{A}^0_{p_{\theta_*}}\phi_1)(\mathbf{X})\right]$.
- In (6.18), only the first eigenfunction ϕ_l appears, because, under the assumed conditions, it dominates the remaining eigenfunctions, and the function JT can be approximated by its projection onto the subspace spanned by ϕ_1.

Overall, verifying these assumptions in practice can be difficult, thus limiting the utility of this result. Nevertheless, we present the following example and show that the assumptions in Theorem 6.1 hold.

Example 6.2 (Augmented Gaussian-location model)**.** *Let the data-generating distribution be the standard univariate Gaussian $Q = \mathcal{N}(0, I_d)$. We are interested in estimating its mean using the model $\mathcal{M}(\Theta) = \{p_\theta \propto \exp(-(x-\theta_1)^2/2 + \theta_2 f_\epsilon(x)) : \theta = (\theta_1, \theta_2) \in \mathbb{R}^d\}$, where f_ϵ is any function with derivative $f'_\epsilon = \epsilon\phi_{1,\gamma_\infty} + (1-\epsilon)\phi_{2,\gamma_\infty}$, with $\epsilon \in [0,1]$ and $\gamma_\infty > 0$ to be specified. The parameter space is $\Theta = \mathbb{R}^2$. This is a well-specified model with $Q = P_{\theta_*}$ where $\theta_* = (0,0)^\top$. For this model, $T(x) = (x, f_\epsilon(x))$ with Jacobian $JT(x) = (1, f'_\epsilon(x))$. Figure 6.2 shows the function f_ϵ for different values of ϵ.*

Intuitively, this model adds an extra sufficient statistic f_ϵ to a Gaussian-location model, the purpose of which is to demonstrate the detrimental effect of kernel smoothing. Specifically, choosing the direction $\mathbf{u} = (0,1)^\top$, we have due to orthonormality that

$$A_l\mathbf{u} \;=\; \langle f'_\epsilon, \phi_{l,\gamma_\infty}\rangle_{L_2(Q)} \;=\; \begin{cases} \epsilon, & l = 1\,, \\ 1-\epsilon, & l = 2\,, \\ 0\,, & l \geq 3\,. \end{cases} \tag{6.19}$$

We can then use Theorem 6.1 to derive a lower bound on the asymptotic covariance of the KSD estimator for this model. The proof proceeds by verifying and is deferred to Section 6.B.3.

Corollary 6.1.1. *Let P_θ be the augmented Gaussian model defined in (6.2) and let k be the squared-exponential kernel with bandwidth $\gamma > 0$. Then there exists $\gamma_0 > 0$ such that for all $\gamma \geq \gamma_0$, there exists a constant $C > 0$ such that (6.17) and (6.18) hold. Moreover, $\|\Gamma_{\mathrm{KSD}}\|_{\mathrm{op}} \geq 16^{-1} C \epsilon^{-2}$.*

In particular, the asymptotic covariance of the minimum KSD estimator $\|\Gamma_{\mathrm{KSD}}\|_{\mathrm{op}}$ can take large values when choosing ϵ to be small. This is verified numerically in Figure 6.3. We compute the minimum KSD estimator and compare its asymptotic variance with that of the SM estimator and MLE. We choose $\epsilon = 0.1$ and $\gamma_\infty = 80$ as the ground-truth model. The bandwidth γ of the kernel is varied from 1 to γ_∞. We plot the operator norm of the asymptotic covariance matrices of the MLE, SM and KSD estimators, and the level sets of their asymptotic normal densities from their CLT results. The level set of an estimator is defined as $\{\theta \in \Theta : (\theta - \theta_*)^\top \Gamma_{\mathcal{L}}^{-1} (\theta - \theta_*) = C\}$, where $\Gamma_{\mathcal{L}}$ is the covariance matrix of the asymptotic Gaussian distribution (6.2) of the estimator, and $C > 0$ is a fixed constant set to be identical for all estimators for a fair comparison. Loosely speaking, the size of these level sets indicate the size of the asymptotic covariance. In all experiments, the asymptotic covariance matrix $\Gamma_{\mathcal{L}}$ of these estimators are estimated by Monte Carlo estimation using 10,000 samples.

We can see that the asymptotic covariance of the minimum KSD estimator explodes in some direction due to presence of the augmented sufficient statistic f_ϵ. The reason is precisely that, for KSD, the projection of the function $\mathbf{x} \mapsto JT(\mathbf{x})$ onto $\mathrm{Span}\{\phi_1\}$ vanishes as γ grows, while in the space orthogonal to $\mathrm{Span}\{\phi_1\}$, the projection of JT also reduces to zero due to kernel smoothing. As a result, the KSD estimator becomes considerably less efficient than SM estimator and MLE for large γ.

6.4.2 KSD with Less Smoothing kernels is Comparably Efficient as SM

Theorem 6.1 implies that when the projections of JT along certain directions is smoothed out by the kernel, then minimum KSD estimation can be extremely inefficient. A natural question is then how does minimum KSD estimation perform when the kernel retains most of the signals of JT? We summarize this in the next result, which states that in this case, the efficiency of KSD estimation is not so much worse than that of the SM. Its proof is given in Section 6.B.4, and we present a concrete example with Gaussian models to verify this result.

Theorem 6.2. *Suppose Assumption 6.1 holds. Assume k has a Mercer decomposition with respect to P_{θ_*} with eigenvalue-eigenfunction pairs $\{(\lambda_l, \phi_l)\}_{l=1}^\infty$, and assume $g_{\mathrm{KSD}} = \langle JT, \mathcal{T}_k JT \rangle_{L_2(P_{\theta_*})}$ is invertible. Moreover, assume that* (i) *there exists a positive integer L such that $g_{\mathrm{SM},L}^\perp \preccurlyeq g_{\mathrm{SM},L}$,* (ii) *$2G_1^\perp \preccurlyeq G_1$, and* (iii) *$2G_{k,1}^\perp \preccurlyeq G_{k,1}$. Then*

$$\|\Gamma_{\mathrm{KSD}}\|_{\mathrm{op}} \;\leq\; 36 \frac{\lambda_1^2}{\lambda_L^2} \left(\frac{\lambda_{\max}(g_{\mathrm{SM}})}{\lambda_{\min}(g_{\mathrm{SM}})} \right)^2 \|\Gamma_{\mathrm{SM}}\|_{\mathrm{op}} \,.$$

This result suggests that, if the eigen-spectrum of k decays slowly so that $\lambda_1/\lambda_L \approx 1$, then the minimum KSD estimator has the same asymptotic efficiency as that of the SM estimator up to a factor of $36(\lambda_{\max}(g_{\mathrm{SM}})/\lambda_{\min}(g_{\mathrm{SM}}))^2$. The ratio $\lambda_{\max}(g_{\mathrm{SM}})/\lambda_{\min}(g_{\mathrm{SM}})$ is the condition number of the information matrix g_{SM} of the SM estimator, and can be large when g_{SM} is almost singular.

We now discuss the assumptions in this theorem.

- Condition (i) can be viewed as a smoothness assumption on the model density, specifically the sufficient statistic T. It requires most of the "signals" in JT lies in $\mathrm{Span}\{\phi_l\}_{l=1}^L$, the span of the first L eigenfunctions. Such L always exists whenever $\|g_{\mathrm{SM}}\|_{\mathrm{op}} < \infty$. Indeed, for any $\mathbf{w} \in \mathbb{S}^{m-1}$ and any positive integer L,

$$\mathbf{w}^\top g_{\mathrm{SM},L}^{\perp} \mathbf{w} \;=\; \sum_{l>L} \|A_l \mathbf{w}\|_2^2 \;=\; \mathbf{w}^\top \Big(\sum_{l>L} A_l^\top A_l \Big) \mathbf{w} \;\le\; \Big\| \sum_{l>L} A_l^\top A_l \Big\|_{\mathrm{op}} ,$$

  where the first equality holds due to a simplification for the information matrix proved in Lemma 6.5 in Section 6.B.2, and the last step holds by taking supremum over $\mathbf{w} \in \mathbb{S}^{m-1}$. Since the sum in the operator norm is the tail of the infinite series $\sum_{l=1}^\infty A_l^\top A_l$, the RHS converges to zero as $L \to \infty$ provided that $\|g_{\mathrm{SM}}\|_{\mathrm{op}} = \|\sum_{l=1}^\infty A_l^\top A_l\|_{\mathrm{op}} < \infty$. We can then choose L sufficiently large so that the RHS is smaller than $\mathbf{w}^\top g_{\mathrm{SM}} \mathbf{w}$ for all $\mathbf{w} \in \mathbb{S}^{m-1}$. More specifically, for any two positive integers $L' \le L$, we have $0 \le \mathbf{w}^\top g_{\mathrm{SM},L'} \mathbf{w} = \sum_{l \le L'} \|A_l \mathbf{w}\|_2^2 \le \sum_{l \le L} \|A_l \mathbf{w}\|_2^2 = \mathbf{w}^\top g_{\mathrm{SM},L} \mathbf{w}$ for all $\mathbf{w}$, so $0 \preccurlyeq g_{\mathrm{SM},L'} \preccurlyeq g_{\mathrm{SM},L}$. Moreover, since $0 \prec g_{\mathrm{SM}}$ by assumption, we must have $0 \prec g_{\mathrm{SM},L'}$ for some L', and in particular $\lambda_{\min}(g_{\mathrm{SM},L}) > 0$. Therefore, we can choose $L = L(L')$ large enough so that $\|\sum_{l>L} A_l^\top A_l\|_{\mathrm{op}} \le \lambda_{\min}(g_{\mathrm{SM},L'}) \le \lambda_{\min}(g_{\mathrm{SM},L})$, which implies that $g_{\mathrm{SM},L}^{\perp} \preccurlyeq g_{\mathrm{SM},L}$.

- Condition (ii) is a smoothness assumption on JT. To see this, for $\mathbf{w} \in \mathbb{S}^{m-1}$ we can bound

$$(\mathbf{w}^\top G_1^{\perp} \mathbf{w})^{1/2} \;=\; \Big\| \sum_{l>1} (A_l \mathbf{w})^\top \mathcal{A}_{p_\theta}^0 \phi_l \Big\|_{L_2(P_{\theta_*})} \;\le\; \sum_{l>1} \|A_l\|_{\mathrm{op}} \|\mathcal{A}_{p_\theta}^0 \phi_l\|_{L_2(P_{\theta_*})} ,$$

  assuming the RHS is convergent, where the last inequality follows from a triangular inequality and a Cauchy-Schwarz inequality. On the other hand, $(\mathbf{w}^\top G_1 \mathbf{w})^{1/2} = \|(A_1 \mathbf{w})^\top \mathcal{A}_{p_\theta}^0 \phi_1\|_{L_2}$. Hence, for assumption (ii) to hold, it suffices to assume $\|A_l\|_{\mathrm{op}}$ is small enough for all $l > 1$ so that

$$\sqrt{2} \sum_{l>1} \|A_l\|_{\mathrm{op}} \|\mathcal{A}_{p_\theta}^0 \phi_l\|_{L_2(P_{\theta_*})} \;\le\; \inf_{\|u\|_2=1} \|(A_1 \mathbf{w})^\top \mathcal{A}_{p_\theta}^0 \phi_1\|_{L_2(P_{\theta_*})} . \tag{6.20}$$

  One way to interpret this condition is that it requires each entry of JT has sufficient signals along the direction ϕ_1, so that A_1 is large in operator norm.

- Condition (iii) has a similar interpretation as condition (ii), except for an extra weighting factor due to the kernel. A similar argument as before leads to the following sufficient condition for assumption (iii):

$$\sqrt{2} \sum_{l>1} \lambda_l \|A_l\|_{\mathrm{op}} \|\mathcal{A}_{p_\theta}^0 \phi_l\|_{L_2(P_{\theta_*})} \;\le\; \inf_{\|u\|_2=1} \lambda_1 \|(A_1 \mathbf{w})^\top \mathcal{A}_{p_\theta}^0 \phi_1\|_{L_2(P_{\theta_*})} .$$

  Since $\lambda_1 \ge \lambda_l$ for all l by assumption, it is evident that (6.20) is sufficient for both assumption (iii) and (ii). More generally, the above inequality holds when the spectral values of k decay sufficiently fast, so that λ_1 and λ_2 have a spectral gap of a similar form as (6.17).

Similar to the assumptions in Theorem 6.1, the conditions assumed in Theorem 6.2 are technical and verifying them in practice can be challenging. Nevertheless, we provide a Gaussian-location example and show that it fulfills the assumptions in Theorem 6.2.

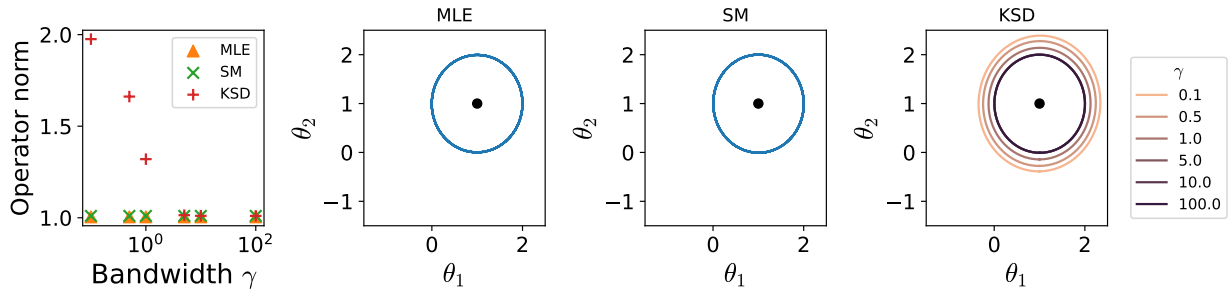


Figure 6.4: *Left.* Operator norm of the asymptotic covariance matrix in the Gaussian-location model. *Middle left, middle right and right.* Level sets for the asymptotic normal distribution predicted by CLT. The ground-truth parameter $\theta_* = (1, 1)^\top$ is shown by the black dot. The KSD estimator is as efficient as the MLE and SM for large values of γ.

Example 6.3 (Gaussian-location model)**.** *Let the data-generating process be a d-dimensional Gaussian distribution $Q = \mathcal{N}(\theta_*, I_d)$ with mean θ_* and identity covariance matrix. We are interested in estimating θ_* using the well-specified model $\mathcal{M}(\Theta) = \{p_\theta(\mathbf{x}) \propto \exp(-\|\mathbf{x} - \theta\|_2^2/2) : \theta \in \mathbb{R}^d\}$, where the parameter space is $\Theta = \mathbb{R}^d$ and coincides with the sample space. In this case, $T(\mathbf{x}) = \mathbf{x}$ and $JT(\mathbf{x}) = I_d$.*

Under this model, the assumptions in Theorem 6.2 are fulfilled with the constant kernel $k_{\text{const}}(x, y) = 1$. This is shown in the next result, proved in Section 6.B.5.

Corollary 6.2.1. *Let $\mathcal{M}(\Theta)$ be a d-dimensional Gaussian-location model defined in Example 6.3, and let $k_{\text{const}}(\mathbf{x}, \mathbf{y}) = 1$ be the constant kernel taking value 1. Then the assumptions* (i)-(iii) *in Theorem 6.2 hold. In particular, assumption* (i) *holds with $L = 1$. Moreover, in this case, $\|\Gamma_{\text{KSD}}\|_{\text{op}} \leq 36\|\Gamma_{\text{SM}}\|_{\text{op}}$.*

Unlike SE kernels, constant kernels are *not* ISPD and thus not commonly used in practice. However, this example serves to illustrate that non-ISPD kernels can still offer statistical benefits by guaranteeing the efficiency of the KSD estimator.

We provide empirical evidence for Corollary 6.2.1. In our experiment, we choose $d = 2$ and set the ground-truth parameter to be $\theta_* = (1, 1)^\top$. We choose an SE kernel $k(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|_2^2/(2\gamma^2))$ instead of a constant kernel, and investigate the statistical efficiency of the minimum KSD estimator for different values of the bandwidth γ. As γ increases, the SE kernel converges pointwise to the constant kernel $k_{\text{const}}(\mathbf{x}, \mathbf{y}) = 1$, so by Corollary 6.2.1, we expect the minimum KSD estimator to become comparably efficient as the SM estimator for large γ. The numerical results are shown in Figure 6.4, where γ is varied from 0.1 to 100. All asymptotic covariance matrices are again estimated by Monte Carlo estimation using $10,000$ samples. For large values of γ, the statistical efficiency of the KSD, measured in terms of $\|\Gamma_{\text{KSD}}\|_{\text{op}}$, is comparable to that of the SM and the MLE, which is consistent with Corollary 6.2.1.

6.5 Conclusion and Discussion

In this chapter, we have studied the statistical efficiency of the minimum KSD estimator. Assuming a sufficiently regular, well-specified model and a kernel decomposition, we established conditions under

which the minimum KSD estimation is less than or comparably efficient as the SM estimation. Our results highlight that the minimum KSD estimator can be inefficient with overly smoothing kernels, which are characterized by rapidly decaying eigenvalues in a Mercer decomposition. Conversely, kernels with a slowly decaying eigen-spectrum can offer comparable efficiency as the SM estimator. Our findings highlight the challenges in designing efficient kernels for KSD in practice, and suggest that SM is preferable over KSD when the model is regular and (at least "almost") well-specified.

Our analyses raise new research questions. First, in this chapter we have restricted our attention to well-specified models. A natural question is whether the same conclusion extends to misspecified models? This is particularly interesting given that a main advantage of KSD is the flexibility of choosing the reproducing kernel to enforce robustness against different types of model misspecification, such as outlier contamination and misspecified tails (Barp et al., 2019; Liu and Briol, 2024). Second, constructing checkable criteria for when KSD estimation should be preferred over SM estimation can have practical benefits. As noted in Section 6.4, the conditions outlined in Theorem 6.1 and Theorem 6.2 can be challenging to verify in practice. Deriving a set of simple, practically verifiable conditions can provide more insights into the applicability of KSD in real-life scenarios.

6.A Asymptotic Theory for Minimum Score-based Estimators

We state the central limit theorems for the minimum KSD and SM estimators, adapted from Barp et al. (2019, Theorem 4 and 6).

6.A.1 Asymptotic Normality for Minimum KSD Estimators

We first present the central limit theorem for the minimum KSD estimator in Theorem 6.3. A special case where P_θ belongs to natural exponential family models, which are the focus of this chapter, will be given in Corollary 6.3.1. We say that a kernel k has *bounded derivatives up to order 2* if its m-times partial derivatives (defined before Theorem 2.4) exist and are bounded for $m = 1, 2$.

Theorem 6.3 (CLT for KSD)**.** *Let k be a bounded kernel (not necessarily ISPD) with bounded derivatives up to order 2. Suppose that θ_n^{KSD} exists for sufficiently large n and $\theta_n^{\mathrm{KSD}} \to \theta_*$ in probability, and that there exists a neighborhood $\mathcal{N} \subset \Theta$ containing θ_* such that the function $\theta \mapsto \mathcal{L}_{\mathrm{KSD}}(\{\mathbf{X}_i\}_{i=1}^n, \theta)$ is twice continuously differentiable for all $\theta \in \mathcal{N}$ and Q-almost-all $\mathbf{x} \in \mathbb{R}^d$. Moreover, assume there exist functions $f \in L_2(Q)$ and $g \in L_1(Q)$ such that the following holds for all $\theta \in \mathcal{N}$ and Q-almost-all $\mathbf{x} \in \mathbb{R}^d$,*

1. $\|\mathbf{s}_{p_\theta}(\mathbf{x})\|_2 + \|\partial_{\theta_j} \mathbf{s}_{p_\theta}(\mathbf{x})\|_2 \leq f(\mathbf{x})$, *for all* $j = 1, \ldots, m$.
2. $\|\partial_{\theta_i}\partial_{\theta_j} \mathbf{s}_{p_\theta}(\mathbf{x})\|_2 + \|\partial_{\theta_i}\partial_{\theta_j}\partial_{\theta_l} \mathbf{s}_{p_\theta}(\mathbf{x})\|_2 \leq g(\mathbf{x})$.

Suppose also that the information matrix g_{KSD} defined in (6.5) is invertible. Then

$$\sqrt{n}\big(\theta_n^{\mathrm{KSD}} - \theta_*\big) \xrightarrow{\mathrm{d}} \mathcal{N}\big(0, g_{\mathrm{KSD}}^{-1} \Sigma_{\mathrm{KSD}} g_{\mathrm{KSD}}^{-1}\big) \,, \tag{6.21}$$

where Σ_{KSD} *is defined in* (6.6).

Corollary 6.3.1. *Let* k *be a bounded kernel with bounded derivatives up to order 2. Assume* $p_\theta(\mathbf{x}) \propto \exp(\theta^\top T(\mathbf{x}) + b(\mathbf{x}))$*, where the Jacobian* JT *of* T *has linearly independent rows, and* $\|JT\|_{\mathrm{op}}, \|\nabla_{\mathbf{x}} b\|_2 \in L_2(Q)$*. Also assume that the KSD information matrix* $g_{\mathrm{KSD}} = \langle JT, \mathcal{T}_k JT\rangle_{L_2(Q)}$ *defined in* (6.5) *is invertible. Then* (6.21) *holds.*

Proof of Corollary 6.3.1. Let $\mathbf{X}_i \sim Q$ be independent random variables, for $i = 1, \ldots, n$. Then the KSD estimator takes the form $\theta_n^{\mathrm{KSD}} = -\frac{1}{2}\Lambda_n^{-1}\eta_n$, where Λ_n and η_n are defined in (6.4). In particular, Λ_n is a finite-sample estimate of g_{KSD}. Since $\|JT\|_{\mathrm{op}} < \infty$, Hoeffding (1961, Theorem 1) implies that the components of Λ_n converges to the components of g_{KSD} almost surely. Also since the matrix inverse is a continuous map, by the continuous mapping theorem Λ_n^{-1} exists for sufficiently large n and converges almost surely to g_{KSD}^{-1}. This shows that θ_n^{KSD} exists for large enough n and converges almost surely to θ_*.

Moreover, we can apply the same argument in the proof of Theorem 1 in Barp et al. (2019, Appendix D.3.1) with identity diffusion term ($m(\mathbf{x}) = I_d$, in their notation) to conclude that $\theta_n^{\mathrm{KSD}} \to \theta_*$ in probability. Moreover, since with natural exponential family models we have $\mathbf{s}_{p_\theta}(\mathbf{x}) = JT(\mathbf{x})\theta + \nabla_{\mathbf{x}}(\mathbf{x})$, $\partial_{\theta_j}\mathbf{s}_{p_\theta}(\mathbf{x}) = \nabla_{\mathbf{x}} T_j(\mathbf{x})$ and $\partial_{\theta_i}\partial_{\theta_j}\mathbf{s}_{p_\theta}(\mathbf{x}) = 0$, the conditions on $\mathbf{s}_{p_\theta}$ assumed in Theorem 6.3 hold. Applying Theorem 6.3 shows the claim. □

6.A.2 Asymptotic Normality for SM Estimators

We now present the CLT for the SM estimator, as well as a special case for natural exponential family models as a corollary.

Theorem 6.4 (CLT for SM)**.** *Suppose that* θ_n^{KSD} *exists for sufficiently large* n *and* $\theta_n^{\mathrm{SM}} \to \theta_*$ *in probability, and that the function* $\theta \mapsto \mathbf{s}_{p_\theta}(\mathbf{x})$ *is twice continuously differentiable in a closed neighbourhood* $\mathcal{N} \subset \Theta$ *containing* θ_* *and* Q*-almost-all* $\mathbf{x} \in \mathbb{R}^d$*. Moreover, assume there exist functions* f, g *such that* $f^2 \in L_2(Q), g \in L_1(Q)$ *and, for all* $\theta \in \mathcal{N}$ *and* Q*-almost-all* $\mathbf{x} \in \mathbb{R}^d$*,*

1. $\|\mathbf{s}_{p_\theta}(\mathbf{x})\|_2 + \|\partial_{\theta_i}\mathbf{s}_{p_\theta}(\mathbf{x})\|_2 + \|\partial_{\theta_j}\partial_{\mathbf{x}_i}\mathbf{s}_{p_\theta}(\mathbf{x})\|_2 \le f(\mathbf{x})$.
2. $\|\partial_{\theta_j}\mathbf{s}_{p_\theta}(\mathbf{x})\|_2^2 + \big(\|\mathbf{s}_{p_\theta}(\mathbf{x})\|_2 + 1\big)\|\partial_{\theta_j}\partial_{\theta_l}\mathbf{s}_{p_\theta}(\mathbf{x})\|_2 + \|\partial_{\theta_j}\partial_{\theta_l}\partial_{\mathbf{x}_i}\mathbf{s}_{p_\theta}(\mathbf{x})\|_2 \le g(\mathbf{x})$.

Suppose also that the information matrix g_{SM} *defined in* (6.8) *is invertible. Then*

$$\sqrt{n}\big(\theta_n^{\mathrm{SM}} - \theta_*\big) \xrightarrow{\mathrm{d}} \mathcal{N}\big(0, g_{\mathrm{SM}}^{-1}\Sigma_{\mathrm{SM}} g_{\mathrm{SM}}^{-1}\big)\,, \tag{6.22}$$

where Σ_{SM} *is defined in* (6.9).

Corollary 6.4.1. *Assume* $p_\theta(\mathbf{x}) \propto \exp(\theta^\top T(\mathbf{x}) + b(\mathbf{x}))$*, where the Jacobian* JT *of* T *has linearly independent rows, and* $\|JT\|_{\mathrm{op}}, \|\nabla_{\mathbf{x}} b\|_2, \|\Delta T_j\|_2 \in L_2(Q)$*, for all* $j = 1, \ldots, m$*. Then* (6.22) *holds.*

Proof of Corollary 6.4.1. Under the assumed conditions, Theorem 13 in Barp et al. (2019, Appendix D.3.3) applied with identity diffusion term ($m(\mathbf{x}) = I_d$) shows that $\theta_n^{\mathrm{SM}} \to \theta_*$ in probability. Using a similar argument as in the proof of Corollary 6.3.1, the conditions on $\mathbf{s}_{p_\theta}$ assumed in Theorem 6.4 hold. Applying Theorem 6.4 then shows the claim. □

6.B Proofs of Theoretical Results

This section contains the proof of the main results in this chapter. We start by showing several preliminary results in Section 6.B.1. These will be used to prove Theorem 6.1 and Theorem 6.2, the proofs of which are given in Section 6.B.2 and Section 6.B.4, respectively.

6.B.1 Preliminary Results

We first present and prove several preliminary results. These results provide a decomposition of the terms in the asymptotic asymptotic covariance matrices of the KSD and SM estimators; see (6.6) and (6.9).

- Lemma 6.5 provides a decomposition for the information matrices of KSD and SM, which are defined in (6.5) and (6.8), respectively.
- Lemma 6.6 provides an alternative expression for the terms $G_k(\mathbf{x})$ and $G(\mathbf{x})$ defined in (6.15) and (6.16), respectively.
- Using this result, Lemma 6.7 gives a decomposition of the middle terms Σ_{KSD} and Σ_{SM} in the asymptotic covariance of the minimum KSD and SM estimators, respectively.

Lemma 6.5. *Suppose Assumption 6.1 holds. Assume k has a Mercer decomposition with respect to P_{θ_*} and with eigenvalue-eigenfunction pairs $\{(\lambda_l, \phi_l)\}_{l=1}^{\infty}$. Then, for all $\mathbf{w} \in \mathbb{R}^m$,*

$$\mathbf{w}^\top g_{\mathrm{KSD}} \mathbf{w} = \sum_{l=1}^{\infty} \lambda_l \|A_l \mathbf{w}\|_2^2 , \qquad \mathbf{w}^\top g_{\mathrm{SM}} \mathbf{w} = \sum_{l=1}^{\infty} \|A_l \mathbf{w}\|_2^2 ,$$

where $[A_l]_{ij} = \langle [JT]_{ij}, \phi_l \rangle_{P_{\theta_*}}$. Moreover, for any integer $L \geq 1$,

$$\mathbf{w}^\top g_{\mathrm{KSD},L} \mathbf{w} = \sum_{l=1}^{L} \lambda_l \|A_l \mathbf{w}\|_2^2 , \qquad \mathbf{w}^\top g^{\perp}_{\mathrm{KSD},L} \mathbf{w} = \sum_{l=L+1}^{\infty} \lambda_l \|A_l \mathbf{w}\|_2^2$$

$$\mathbf{w}^\top g_{\mathrm{SM},L} \mathbf{w} = \sum_{l=1}^{L} \|A_l \mathbf{w}\|_2^2 , \qquad \mathbf{w}^\top g^{\perp}_{\mathrm{SM},L} \mathbf{w} = \sum_{l=L+1}^{\infty} \|A_l \mathbf{w}\|_2^2 .$$

Proof of Lemma 6.5. By M.2 of Assumption 6.1, for all $i = 1, \ldots, d$ and $j = 1, \ldots, m$, we have

$$[JT]_{ij} = \sum_{l=1}^{\infty} \langle [JT]_{ij}, \phi_l \rangle_{L_2(P_{\theta_*})} \phi_l = \sum_{l=1}^{\infty} [A_l]_{ij} \phi_l , \tag{6.23}$$

where the last line holds since $[A_l]_{ij} = \langle [JT]_{ij}, \phi_l \rangle_{P_{\theta_*}}$ by definition. Using this identity, the information matrix of the minimum score matching estimator can be simplified as

$$g_{\mathrm{SM}} \;=\; \langle JT,\, JT \rangle_{L_2(P_{\theta_*})} \;=\; \left\langle \sum_{l=1}^{\infty} A_l \phi_l,\, \sum_{r=1}^{\infty} A_r \phi_r \right\rangle_{L_2(Q)} \;=\; \sum_{l=1}^{\infty} A_l^\top A_l^2 \;,$$

where in the last step we have used the orthonormality of the eigenfunctions $\langle \phi_l, \phi_r \rangle_{L_2(Q)} = \delta_{lr}$. This shows that $\mathbf{w}^\top g_{\mathrm{SM}} \mathbf{w} = \sum_{l=1}^{\infty} \|A_l \mathbf{w}\|_2^2$.

We then show the expression for Σ_{KSD}. The decomposition (6.23) gives

$$\begin{aligned}
\sum_{l=1}^{\infty} \lambda_l A_l^\top A_l \;&=\; \left\langle \sum_{l=1}^{\infty} A_l \phi_l,\, \sum_{r=1}^{\infty} \lambda_r A_r \phi_r \right\rangle_{L_2(Q)} \\
&=\; \left\langle JT,\, \sum_{r=1}^{\infty} \lambda_r A_r \phi_r \right\rangle_{L_2(Q)} \\
&=\; \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} (JT(\mathbf{x}))^\top JT(\mathbf{y}) \sum_{l=1}^{\infty} \lambda_l(\mathbf{x}) \phi(\mathbf{y}) \phi(\mathbf{y}) P_{\theta_*}(\mathrm{d}\mathbf{x}) P_{\theta_*}(\mathrm{d}\mathbf{y}) \qquad (6.24) \\
&=:\; S \;,
\end{aligned}$$

where the first equality holds again due to the orthonormality of $\{\phi_l\}_{l=1}^{\infty}$, the second step follows from (6.23), and in the last line we have substituted the definition of A_l and used Fubini-Toneli Theorem to exchange the sum and the integrals. If the Mercer decomposition (6.10) holds uniformly on $\mathbb{R}^d \times \mathbb{R}^d$, then (6.24) would be identical to $g_{\mathrm{KSD}} = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} (JT(\mathbf{x}))^\top JT(\mathbf{y}) k(\mathbf{x}, \mathbf{y}) P_{\theta_*}(\mathrm{d}\mathbf{x}) P_{\theta_*}(\mathrm{d}\mathbf{y})$ by directly substituting the decomposition of $k(\mathbf{x}, \mathbf{y})$. However, (6.10) only holds on compact subsets. We instead show $S = g_{\mathrm{KSD}}$ by decomposing the double integral into compact and non-compact domains.

To do so, we define the shorthand notation

$$\Delta(\mathbf{x}, \mathbf{y}) \;=\; k(\mathbf{x}, \mathbf{y}) - \sum_{l=1}^{\infty} \lambda_l(\mathbf{x}) \phi(\mathbf{y}) \phi(\mathbf{y}) \;, \quad \text{and} \quad \Delta_T(\mathbf{x}, \mathbf{y}) \;=\; (JT(\mathbf{x}))^\top JT(\mathbf{y}) \Delta(\mathbf{x}, \mathbf{y}) \;.$$

Moreover, for each positive integer L, we define the ball $B_L \coloneqq \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 \le L\}$. By the partition $\mathbb{R}^d = B_L \cup \mathbb{R}^d \backslash B_l$ and a triangle inequality,

$$\begin{aligned}
\|g_{\mathrm{KSD}} - S\|_{\mathrm{op}} \;&=\; \left\| \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \Delta_T(\mathbf{x}, \mathbf{y}) P_{\theta_*}(\mathrm{d}\mathbf{x}) P_{\theta_*}(\mathrm{d}\mathbf{y}) \right\|_{\mathrm{op}} \\
&\le\; \left\| \int_{B_L} \int_{B_L} \Delta_T(\mathbf{x}, \mathbf{y}) P_{\theta_*}(\mathrm{d}\mathbf{x}) P_{\theta_*}(\mathrm{d}\mathbf{y}) \right\|_{\mathrm{op}} + 2 \left\| \int_{B_L} \int_{\mathbb{R}^d \backslash B_L} \Delta_T(\mathbf{x}, \mathbf{y}) P_{\theta_*}(\mathrm{d}\mathbf{x}) P_{\theta_*}(\mathrm{d}\mathbf{y}) \right\|_{\mathrm{op}} \\
&\quad + \left\| \int_{\mathbb{R}^d \backslash B_L} \int_{\mathbb{R}^d \backslash B_L} \Delta_T(\mathbf{x}, \mathbf{y}) P_{\theta_*}(\mathrm{d}\mathbf{x}) P_{\theta_*}(\mathrm{d}\mathbf{y}) \right\|_{\mathrm{op}} . \qquad (6.25)
\end{aligned}$$

We claim that the terms on the RHS converge to zero as $L \to \infty$. The first term vanishes as the Mercer decomposition implies that $\Delta(\mathbf{x}, \mathbf{y}) = 0$ uniformly on the compact subset B_L. For the second term, since $\int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \|\Delta_T(\mathbf{x}, \mathbf{y})\|_{\mathrm{op}} P_{\theta_*}(\mathrm{d}\mathbf{x}) P_{\theta_*}(\mathrm{d}\mathbf{y}) < \infty$ under the assumed conditions, applying Hölder's

inequality and the dominated convergence theorem gives

$$\begin{aligned}
&\left\| \int_{B_L} \int_{\mathbb{R}^d \setminus B_L} \Delta_T(\mathbf{x}, \mathbf{y}) P_{\theta_*}(\mathrm{d}\mathbf{x}) P_{\theta_*}(\mathrm{d}\mathbf{y}) \right\|_{\mathrm{op}} \\
&\leq \left\| \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \mathbb{1}\{\mathbf{x} \in B_L\} \mathbb{1}\{\mathbf{y} \in \mathbb{R}^d \setminus B_L\} \Delta_T(\mathbf{x}, \mathbf{y}) P_{\theta_*}(\mathrm{d}\mathbf{x}) P_{\theta_*}(\mathrm{d}\mathbf{y}) \right\|_{\mathrm{op}} \\
&\leq \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \mathbb{1}\{\mathbf{x} \in B_L\} \mathbb{1}\{\mathbf{y} \in \mathbb{R}^d \setminus B_L\} \|\Delta_T(\mathbf{x}, \mathbf{y})\|_{\mathrm{op}} P_{\theta_*}(\mathrm{d}\mathbf{x}) P_{\theta_*}(\mathrm{d}\mathbf{y}) \;\to\; 0 \,,
\end{aligned}$$

as $L \to \infty$. A similar argument shows that the third term in (6.25) also converges to zero. This shows that $S = g_{\mathrm{KSD}}$, and therefore $\mathbf{w}^\top g_{\mathrm{KSD}} \mathbf{w} = \sum_{l=1}^\infty \lambda_l \|A_l \mathbf{w}\|_2^2$. Finally, given any integer $L \geq 1$, the expressions for $g_{\mathrm{KSD},L}, g^\perp_{\mathrm{KSD},L}$ and $g_{\mathrm{SM},L}, g^\perp_{\mathrm{SM},L}$ can be shown by noting that

$$\Pi_L [JT]_{ij} \;=\; \sum_{l=1}^L \langle [JT]_{ij}, \phi_l \rangle_{L_2(P_{\theta_*})} \phi_l \;=\; \sum_{l=1}^L [A_l]_{ij} \phi_l \,,$$

and

$$\Pi^\perp_L [JT]_{ij} \;=\; \sum_{l=L+1}^\infty \langle [JT]_{ij}, \phi_l \rangle_{L_2(P_{\theta_*})} \phi_l \;=\; \sum_{l=L+1}^\infty [A_l]_{ij} \phi_l \,,$$

and by following the previous argument with JT replaced by $\Pi_L JT$ or $\Pi^\perp_L JT$. □

Lemma 6.6. *Suppose Assumption 6.1 holds. Assume k has a Mercer decomposition with respect to P_{θ_*} and with eigenvalue-eigenfunction pairs $\{(\lambda_l, \phi_l)\}_{l=1}^\infty$. Then*

$$G_k(\mathbf{x}) \;=\; \sum_{l=1}^\infty \lambda_l A_l^\top (\mathcal{A}^0_{p_\theta} \phi_l)(\mathbf{x}) \,, \qquad G(\mathbf{x}) \;=\; \sum_{l=1}^\infty A_i^\top (\mathcal{A}_{p_\theta} \phi_l)(\mathbf{x}) \,.$$

Moreover, for any integer $L \geq 1$, we have $G_k(\mathbf{x}) = G_{k,L}(\mathbf{x}) + G^\perp_{k,L}(\mathbf{x})$ and $G(\mathbf{x}) = G_L(\mathbf{x}) + G^\perp_L(\mathbf{x})$, where $G_{k,L}(\mathbf{x}), G^\perp_{k,L}(\mathbf{x})$ are defined in (6.15), and $G_L(\mathbf{x}), G^\perp_L(\mathbf{x})$ are defined in (6.16).

Proof of Lemma 6.6. We first derive the expression for $G_k(\mathbf{x})$. Differentiating the Stein kernel with respect to θ and noting that k does not depend on θ, we have

$$\begin{aligned}
\nabla_\theta u_{p_\theta}(\mathbf{x}, \mathbf{y}) \;=\; & JT(\mathbf{y})^\top k(\mathbf{x}, \mathbf{y}) \mathbf{s}_{p_\theta}(\mathbf{x}) + (JT(\mathbf{x}))^\top k(\mathbf{x}, \mathbf{y}) \mathbf{s}_{p_\theta}(\mathbf{y}) && \text{(6.26a)} \\
& + JT(\mathbf{y})^\top \nabla_1 k(\mathbf{x}, \mathbf{y}) + (JT(\mathbf{x}))^\top \nabla_2 k(\mathbf{x}, \mathbf{y}) \,. && \text{(6.26b)}
\end{aligned}$$

Moreover, for the natural exponential model assumed in Assumption 6.1, the score function takes the form $\mathbf{s}_{p_{\theta_*}}(\mathbf{x}) = JT(\mathbf{x})\theta_* + \nabla_{\mathbf{x}} b(\mathbf{x})$. Hence, by a triangle inequality

$$\begin{aligned}
\mathbb{E}_{\mathbf{X} \sim P_{\theta_*}} [\|\mathbf{s}_{p_{\theta_*}}(\mathbf{X})\|_2] \;&\leq\; \mathbb{E}_{\mathbf{X} \sim P_{\theta_*}} [\|JT(\mathbf{X})\theta\|_2] + \mathbb{E}_{\mathbf{X} \sim P_{\theta_*}} [\|\nabla_{\mathbf{x}} b(\mathbf{x})\|_2] \\
&\leq\; \mathbb{E}_{\mathbf{X} \sim P_{\theta_*}} [\|JT(\mathbf{X})\|_{\mathrm{op}} \|_2] \|\theta\|_2 + \mathbb{E}_{\mathbf{X} \sim P_{\theta_*}} [\|\nabla_{\mathbf{x}} b(\mathbf{x})\|_2] \;<\; \infty \,,
\end{aligned}$$

where the first step follows from a triangle inequality, the second step holds since $\|JT(\mathbf{x})\theta\|_2 \leq \|JT(\mathbf{x})\|_{\mathrm{op}} \|\theta\|_2$ for all θ, and the last inequality holds as both terms belong to $L_2(P_{\theta_*})$ by Assump-

tion M.1. Therefore, when $Q = P_{\theta_*}$, we can apply Stein identity (Gorham and Mackey, 2017, Proposition 1) to each component and conclude that the sum of the second and the third terms in (6.26) integrates to zero for all $\mathbf{x}$, i.e.,

$$\int_{\mathbb{R}^d} (JT(\mathbf{x}))^\top k(\mathbf{x},\mathbf{y})\mathbf{s}_{p_\theta}(\mathbf{y}) + JT(\mathbf{y})^\top \nabla_1 k(\mathbf{x},\mathbf{y}) Q(\mathrm{d}\mathbf{y}) \;=\; 0\,.$$

This identity implies

$$\begin{aligned} G_k(\mathbf{x}) \;=\; \int_{\mathbb{R}^d} \nabla_\theta u_{p_\theta}(\mathbf{x},\mathbf{y}) Q(\mathrm{d}\mathbf{y}) \;&=\; \int_{\mathbb{R}^d} JT(\mathbf{y})^\top k(\mathbf{x},\mathbf{y})\mathbf{s}_{p_\theta}(\mathbf{x}) + JT(\mathbf{y})^\top \nabla_1 k(\mathbf{x},\mathbf{y}) Q(\mathrm{d}\mathbf{y}) \\ &=\; \int_{\mathbb{R}^d} JT(\mathbf{y})^\top \big(\mathcal{A}^0_{p_\theta} k(\cdot,\mathbf{y})\big)(\mathbf{x}) Q(\mathrm{d}\mathbf{y})\,, \end{aligned} \tag{6.27}$$

where $\mathcal{A}^0_{p_\theta} k(\cdot,\mathbf{y})$ denotes the modified Langevin Stein operator applied to the function $\mathbf{x} \mapsto k(\mathbf{x},\mathbf{y})$. Since $\int_{\mathbb{R}^d} \|JT(\mathbf{y})^\top \big(\mathcal{A}^0_{p_\theta} k(\cdot,\mathbf{y})\big)(\mathbf{x})\|_2 Q(\mathrm{d}\mathbf{y}) < \infty$ by assumption, we can substitute the Mercer decomposition of k (Definition 6.1) and use a similar argument as in the proof of Lemma 6.5 (specifically, the part concerning the expression of Σ_{KSD}) to further simplify (6.27) as

$$\begin{aligned} G_k(\mathbf{x}) \;&=\; \int_{\mathbb{R}^d} JT(\mathbf{y})^\top \left(\mathcal{A}^0_{p_\theta} \sum_{l=1}^\infty \lambda_l \phi_l(\cdot)\phi_l(\mathbf{y})\right)(\mathbf{x}) Q(\mathrm{d}\mathbf{y}) \\ &=\; \sum_{l=1}^\infty \lambda_l \left(\int_{\mathbb{R}^d} \phi_l(\mathbf{y}) JT(\mathbf{y}) Q(\mathrm{d}\mathbf{y})\right)^\top (\mathcal{A}^0_{p_\theta}\phi_l)(\mathbf{x}) \\ &=\; \sum_{l=1}^\infty \lambda_l A_l^\top (\mathcal{A}^0_{p_\theta}\phi_l)(\mathbf{x})\,, \end{aligned}$$

where in the second step we have exchanged the sum and the integral using Fubini-Tonelli theorem (Bartle, 1995, Section 10.9, 10.10), and the last step holds by the definition of A_l.

We now derive the expression for $G(\mathbf{x})$. The assumption $[JT]_{ij} = \partial_i T_j \in \mathrm{Span}\{\phi_l\}_{l\geq 1}$ implies that the decomposition $JT = \sum_{l\geq 1} A_l \phi_l$ holds. Using this decomposition,

$$\begin{aligned} G(\mathbf{x}) \;=\; \nabla_\theta\big(\|\mathbf{s}_{p_\theta}(\mathbf{x})\|_2^2 + \Delta \log p_\theta(\mathbf{x})\big) \;&=\; (JT(\mathbf{x}))^\top \mathbf{s}_{p_\theta}(\mathbf{x}) + \langle \nabla, JT(\mathbf{x})\rangle \\ &=\; \sum_{l=1}^\infty A_l^\top \left(\phi(\mathbf{x})\mathbf{s}_{p_\theta}(\mathbf{x}) + \nabla\phi(\mathbf{x})\right) \;=\; \sum_{l=1}^\infty A_i^\top (\mathcal{A}_{p_\theta}\phi_l)(\mathbf{x})\,, \end{aligned}$$

Finally, the decomposition of G_k and G into $G_{k,L}$, $G^\perp_{k,L}$, G_L and $G^\perp_L$ follows by truncating the sum in the above expressions at L. □

Lemma 6.7. *Suppose Assumption 6.1 holds. Assume k has a Mercer decomposition with respect to P_{θ_*} and with eigenvalue-eigenfunction pairs $\{(\lambda_l,\phi_l)\}_{l=1}^\infty$. Let $Q = P_{\theta_*}$. Then, for all $\mathbf{w} \in \mathbb{R}^m$,*

$$\mathbf{w}^\top \Sigma_{\mathrm{KSD}} \mathbf{w} \;=\; \left\| \sum_{l=1}^\infty \lambda_l (A_l\mathbf{w})^\top (\mathcal{A}^0_{p_\theta}\phi_i) \right\|^2_{L_2(Q)}, \qquad \mathbf{w}^\top \Sigma_{\mathrm{SM}} \mathbf{w} \;=\; \left\| \sum_{l=1}^\infty (A_l\mathbf{w})^\top (\mathcal{A}^0_{p_\theta}\phi_i) \right\|^2_{L_2(Q)}.$$

Proof of Lemma 6.7. Pick $\mathbf{w} \in \mathbb{R}^d$. Using the form of G_k derived in Lemma 6.6,

$$\mathbf{w}^\top \Sigma_{\mathrm{KSD}} \mathbf{w} \;=\; \|G_k^\top \mathbf{w}\|_{L_2(Q)}^2 \;=\; \Big\| \sum_{l=1}^{\infty} \lambda_l (A_l \mathbf{w})^\top (\mathcal{A}^0_{p_\theta} \phi_l) \Big\|^2_{L_2(Q)} ,$$

as claimed. Similarly, we use the form of G in Lemma 6.6 to obtain

$$\mathbf{w}^\top \Sigma_{\mathrm{SM}} \mathbf{w} \;=\; \|G^\top \mathbf{w}\|_{L_2(Q)}^2 \;=\; \Big\| \sum_{l=1}^{\infty} (A_l \mathbf{w})^\top (\mathcal{A}^0_{p_\theta} \phi_l) \Big\|^2_{L_2(Q)} .$$

□

6.B.2 Proof for Theorem 6.1

For a vector $\mathbf{w} \in \operatorname{argmin}_{\mathbf{u} \in \mathbb{S}^{m-1}} \mathbf{u}^\top g_{\mathrm{SM},1} \mathbf{u}$, we define $\mathbf{u}_\mathbf{w} \coloneqq g_{\mathrm{KSD}} \mathbf{w} / \|g_{\mathrm{KSD}} \mathbf{w}\|_2$. We first lower-bound $\|\Gamma_{\mathrm{KSD}}\|_{\mathrm{op}}$ as

$$\|\Gamma_{\mathrm{KSD}}\|_{\mathrm{op}} \;=\; \sup_{u \in \mathbb{S}^{m-1}} \mathbf{u}^\top \Gamma_{\mathrm{KSD}} \mathbf{u} \;\geq\; \mathbf{u}_\mathbf{w}^\top \Gamma_{\mathrm{KSD}} \mathbf{u}_\mathbf{w} \;=\; \mathbf{u}_\mathbf{w}^\top g_{\mathrm{KSD}}^{-1} \Sigma_{\mathrm{KSD}} g_{\mathrm{KSD}}^{-1} \mathbf{u}_\mathbf{w} \;=\; \mathbf{w}^\top \Sigma_{\mathrm{KSD}} \mathbf{w} \|g_{\mathrm{KSD}} \mathbf{w}\|_2^{-2} . \tag{6.28}$$

The proof proceeds by lower bounding $\mathbf{w}^\top \Sigma_{\mathrm{KSD}} \mathbf{w}$ and upper bounding $\|g_{\mathrm{KSD}} \mathbf{w}\|_2$. To lower bound $\mathbf{w}^\top \Sigma_{\mathrm{KSD}} \mathbf{w}$, we apply the decomposition proved in Lemma 6.7 and use a triangular inequality to yield

$$\begin{aligned}
\big(\mathbf{w}^\top \Sigma_{\mathrm{KSD}} \mathbf{w}\big)^{1/2} \;&=\; \Big\| \sum_{l=1}^{\infty} \lambda_l (A_l \mathbf{w})^\top (\mathcal{A}^0_{p_{\theta_*}} \phi_l) \Big\|_{L_2(Q)} \\
&\geq\; \|\lambda_1 (A_1 \mathbf{w})^\top (\mathcal{A}^0_{p_{\theta_*}} \phi_1)\|_{L_2(Q)} - \Big\| \sum_{l=2}^{\infty} \lambda_l (A_l \mathbf{w})^\top (\mathcal{A}^0_{p_{\theta_*}} \phi_l) \Big\|_{L_2(Q)} \\
&\geq\; \frac{1}{2} \big\|\lambda_1 (A_1 \mathbf{w})^\top (\mathcal{A}^0_{p_{\theta_*}} \phi_1)\big\|_{L_2(Q)} ,
\end{aligned} \tag{6.29}$$

where $A_l = \int_{\mathbb{R}^d} JT(\mathbf{x}) \phi_l(\mathbf{x}) Q(\mathrm{d}\mathbf{x})$ is defined in the paragraph before (6.13), and in the last line we have used the following inequalities

$$\begin{aligned}
\Big\| \sum_{l=2}^{\infty} \lambda_l (A_l \mathbf{w})^\top (\mathcal{A}^0_{p_{\theta_*}} \phi_l) \Big\|_{L_2(Q)} \;\leq\; \sum_{l=2}^{\infty} \lambda_l \|(A_l \mathbf{w})^\top (\mathcal{A}^0_{p_{\theta_*}} \phi_l)\|_{L_2(Q)} \;&\leq\; \lambda_2 \sum_{l=2}^{\infty} \|(A_l \mathbf{w})^\top (\mathcal{A}^0_{p_{\theta_*}} \phi_l)\|_{L_2(Q)} \\
&\leq\; \frac{1}{2} \|\lambda_1 (A_1 \mathbf{w})^\top (\mathcal{A}^0_{p_{\theta_*}} \phi_1)\|_{L_2} ,
\end{aligned}$$

where the first inequality follows by a triangle inequality, the second inequality holds since the eigenvalues are non-increasing and non-negative by assumption, i.e., $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$, and the last line holds by the assumed spectral gap (6.17). On the other hand, applying Lemma 6.5 yields

$$g_{\mathrm{KSD}} \;=\; \sum_{l=1}^{\infty} \lambda_l A_l^\top A_l^2 \;=\; \lambda_l A_1^\top A_1^2 + \sum_{l=2}^{\infty} \lambda_l A_l^\top A_l^2 \;=\; g_{\mathrm{KSD},1} + g_{\mathrm{KSD},1}^{\perp} ,$$

where the last step holds by the definition of $g_{\mathrm{KSD},1}$ and $g^{\perp}_{\mathrm{KSD},1}$ in (6.14). Using a triangle inequality then gives the following bound

$$\|g_{\mathrm{KSD}}\mathbf{w}\|_2 \;=\; \|g_{\mathrm{KSD},1}\mathbf{w} + g^{\perp}_{\mathrm{KSD},1}\mathbf{w}\|_2 \;\le\; \|g_{\mathrm{KSD},1}\mathbf{w}\|_2 + \|g^{\perp}_{\mathrm{KSD},1}\mathbf{w}\|_2 \;\le\; 2\lambda_1\|A_1^\top A_1\mathbf{w}\|_2\,, \tag{6.30}$$

where the last inequality holds since $\|g_{\mathrm{KSD},1}\mathbf{w}\|_2 = \lambda_1\|A_1^\top A_1\mathbf{w}\|_2$ and

$$\|g^{\perp}_{\mathrm{KSD},1}\mathbf{w}\|_2 \;=\; \Big\|\sum_{l=2}^{\infty}\lambda_l A_l^\top A_l\mathbf{w}\Big\|_2 \overset{(a)}{\le} \sum_{l=2}^{\infty}\|\lambda_l A_l^\top A_l\mathbf{w}\|_2 \overset{(b)}{\le} \lambda_2\sum_{l=2}^{\infty}\|A_l^\top A_l\mathbf{w}\|_2 \overset{(c)}{\le} \lambda_1\|A_1^\top A_1\mathbf{w}\|_2\,,$$

where (a) follows from a triangle inequality, (b) holds again since the eigenvalues are non-increasing and non-negative, and (c) holds due to the spectral gap (6.17). Moreover, using the orthonormality of the eigenfunctions, we can decompose $g_{\mathrm{SM},1}$ using a similar argument in the proof of Lemma 6.5 to conclude that $g_{\mathrm{SM},1} = \langle \Pi_1 JT, \Pi_1 JT\rangle_{L_2(Q)} = A_1^\top A_1$. Also since $\mathbf{w} \in \operatorname{argmin}_{\mathbf{u}\in\mathbb{S}^{m-1}}\mathbf{u}^\top g_{\mathrm{SM},1}\mathbf{u}$ by construction, we have

$$\|A_1^\top A_1\mathbf{w}\|_2 \;=\; \|g_{\mathrm{SM},1}\mathbf{w}\|_2 \;=\; \lambda_{\min}(g_{\mathrm{SM},1}) \;=\; \mathbf{w}^\top g_{\mathrm{SM},1}\mathbf{w} \;=\; \mathbf{w}^\top A_1^\top A_1\mathbf{w} \;=\; \|A_1\mathbf{w}\|_2^2\,.$$

Substituting this into (6.30) gives

$$\|g_{\mathrm{KSD}}\|_{L_2(Q)} \;\le\; 2\lambda_1\|A_1^\top A_1\mathbf{w}\|_2 \;=\; 2\lambda_1\|A_1\mathbf{w}\|_2^2\,, \tag{6.31}$$

where the last equality holds since $\mathbf{w} \in \operatorname{argmin}_{\mathbf{u}\in\mathbb{S}^{m-1}}\mathbf{u}^\top g_{\mathrm{SM},1}\mathbf{u}$ by construction. Combining the inequalities (6.28), (6.29) and (6.31), we have

$$\begin{aligned}
\|\Gamma_{\mathrm{KSD}}\|_{\mathrm{op}} \;\ge\; \mathbf{w}^\top\Sigma_{\mathrm{KSD}}\mathbf{w}\|g_{\mathrm{KSD}}\mathbf{w}\|_2^{-2} \;&\ge\; \frac{1}{4}\Big\|\lambda_1(A_1\mathbf{w})^\top(\mathcal{A}^0_{p_{\theta_*}}\phi_1)\Big\|_{L_2}^2\big(2\lambda_1\|A_1\mathbf{w}\|_2^2\big)^{-2} \\
&=\; \frac{1}{16}\Big\|(A_1\mathbf{w})^\top(\mathcal{A}^0_{p_{\theta_*}}\phi_1)\Big\|_{L_2}^2\|A_1\mathbf{w}\|_2^{-4}\,, \\
&=\; \frac{1}{16}\Big\|\big(A_1\mathbf{w}\|A_1\mathbf{w}\|_2^{-1}\big)^\top(\mathcal{A}^0_{p_{\theta_*}}\phi_1)\Big\|_{L_2}^2\|A_1\mathbf{w}\|_2^{-2} \\
&\ge\; \frac{1}{16}\Big(\inf_{\|\mathbf{u}\|_2=1}\|\mathbf{u}^\top(\mathcal{A}^0_{p_{\theta_*}}\phi_1)\|_{L_2}^2\Big)\|A_1\mathbf{w}\|_2^{-2} \\
&\ge\; \frac{1}{16}C\|A_1\mathbf{w}\|_2^{-2}\,,
\end{aligned}$$

where the second-last line follows by taking infimum over $u \in \mathbb{S}^{m-1}$, and the last inequality holds as the assumption $\mathbb{E}_Q\big[(\mathcal{A}^0_{p_{\theta_*}}\phi_1)(\mathcal{A}^0_{p_{\theta_*}}\phi_1)^\top\big] \succcurlyeq CI_m$ implies that $\inf_{\|\mathbf{u}\|_2=1}\|\mathbf{u}^\top\mathcal{A}_{p^\nu_{\theta_*}}\phi_1\|_{L_2(Q)} \ge C$ for all ν.

6.B.3 Proof of Corollary 6.1.1

We verify the assumptions in Theorem 6.1 and apply this theorem to conclude the claimed result. For this model, P_{θ_*} with $\theta_* = (0,0)$ reduces to the standard Gaussian distribution. Therefore, for k being the SE kernel, a Mercer decomposition holds with $\{(\lambda_l,\phi_l)\}_{l=1}^{\infty}$ defined in Example 6.1. To verify Assumption 6.1, since $JT = (1, f'_\epsilon(x))$ and $b(x) = -x$ for this model, direct computation shows that the integrability conditions in Assumption M.1 hold. Moreover, Assumption M.2 holds since JT are

linear combinations of $\phi_{1,\gamma}, \phi_{2,\gamma}$ by construction, thus showing Assumption 6.1.

The condition (6.18) holds as direct computation shows that $\mathcal{A}^{\nu}_{p_{\theta_*}}\phi_1$ is not a zero function, so there must exist $C > 0$ such that $\mathbb{E}_{\mathbf{X}\sim Q}[(\mathcal{A}_{p_{\theta_*}}\phi_1)(\mathbf{X}) \otimes (\mathcal{A}_{p_{\theta_*}}\phi_1)(\mathbf{X})] \succcurlyeq CI_m$. To show (6.17), the RHS of (6.17) for this model becomes

$$\frac{\lambda_2}{\lambda_1} \leq \min\left\{ \frac{\|A_1^\top A_1 \mathbf{w}\|_2}{\|A_2^\top A_2 \mathbf{w}\|_2}, \quad \frac{\|(A_1\mathbf{w})^\top(\mathcal{A}_{p_{\theta_*}}\phi_1)\|_{L_2(Q)}}{2\|(A_2\mathbf{w})^\top(\mathcal{A}_{p_{\theta_*}}\phi_2)\|_{L_2(Q)}} \right\},$$

where $\mathbf{w} \in \operatorname{argmin}_{u\in\mathbb{S}^{m-1}} \mathbf{u}^\top g_{\mathrm{SM},1}\mathbf{u}$. Since both A_1 and $\mathcal{A}_{p_{\theta_*}}\phi_1$ are non-zero, the RHS of the above inequality is positively valued. Moreover, as $\gamma \to \infty$, the first eigenvalue $\lambda_{1,\gamma} = \sqrt{2a/\alpha_\gamma}$ is bounded while $\lambda_2 = \beta_\gamma\sqrt{2a/\alpha_\gamma} \to 0$. Therefore, (6.17) holds for sufficiently large γ. We hence conclude the assumptions in Theorem 6.1 hold for sufficiently large γ.

Finally, using (6.19) we can also compute that $\lambda_{\min}(g_{\mathrm{SM},1}) \leq \mathbf{u}^\top g_{\mathrm{SM},1}\mathbf{u} = \|A_1\mathbf{u}\|_2^2 = \epsilon^2$ and $\mathbf{u}^\top g^{\perp}_{\mathrm{SM},1}\mathbf{u} = \sum_{l>1}\|A_l\mathbf{u}\|_2^2 = (1-\epsilon)^2$. Therefore, for γ large enough, we can apply Theorem 6.1 to conclude that $\|\Gamma_{\mathrm{KSD}}\|_{\mathrm{op}} \geq 16^{-1}C\lambda_{\min}(g_{\mathrm{SM},1})^{-1} \geq 16^{-1}C\epsilon^{-2}$ for some $C > 0$.

6.B.4 Proof for Theorem 6.2

We first show two lemmas, which will be used to prove Theorem 6.2. The first result, Lemma 6.8, bounds the middle terms Σ_{KSD} and Σ_{SM} in the asymptotic covariance matrices of the minimum KSD and SM estimators. The second result, Lemma 6.9 gives an equivalent expression for the operator norm of a matrix of the form ABA, where $A, B \in \mathbb{R}^{m\times m}$. The proof of Theorem 6.2 is provided at the end of this section.

Lemma 6.8. *Assume* (i) $2G^{\perp}_{k,1} \preccurlyeq G_{k,1}$, *and* (ii) $2G^{\perp}_1 \preccurlyeq G_1$, *then* $\Sigma_{\mathrm{KSD}} \preccurlyeq 9\lambda_1^2\Sigma_{\mathrm{SM}}$.

Proof of Lemma 6.8. For all $\mathbf{w} \in \mathbb{S}^{m-1}$, we can use triangular inequality to bound

$$\begin{aligned}
\left(\mathbf{w}^\top\Sigma_{\mathrm{KSD}}\mathbf{w}\right)^{1/2} &= \Big\|\sum_{i\geq 1}\lambda_i(A_1\mathbf{w})^\top(\mathcal{A}^0_{p_\theta}\phi_i)\Big\|_{L_2} \leq \lambda_1\|(A_1\mathbf{w})^\top(\mathcal{A}^0_{p_\theta}\phi_i)\|_{L_2} + \Big\|\sum_{i>1}\lambda_i(A_1\mathbf{w})^\top(\mathcal{A}^0_{p_\theta}\phi_i)\Big\|_{L_2} \\
&= \mathbf{w}^\top G_{k,1}\mathbf{w} + \mathbf{w}^\top G^{\perp}_{k,1}\mathbf{w} \\
&\leq \frac{3}{2}\mathbf{w}^\top G_{k,1}\mathbf{w} = \frac{3}{2}\lambda_1\|(A_1\mathbf{w})^\top(\mathcal{A}^0_{p_\theta}\phi_i)\|_{L_2},
\end{aligned}$$

where the first equality follows from Lemma 6.7, and the last inequality holds due to assumption *(i)*. Using a similar argument,

$$\begin{aligned}
\left(\mathbf{w}^\top\Sigma_{\mathrm{SM}}\mathbf{w}\right)^{1/2} &= \Big\|\sum_{i\geq 1}(A_1\mathbf{w})^\top(\mathcal{A}^0_{p_\theta}\phi_i)\Big\|_{L_2} \geq \|(A_1\mathbf{w})^\top(\mathcal{A}^0_{p_\theta}\phi_i)\|_{L_2} - \Big\|\sum_{i>1}(A_1\mathbf{w})^\top(\mathcal{A}^0_{p_\theta}\phi_i)\Big\|_{L_2} \\
&= \mathbf{w}^\top G_1\mathbf{w} - \mathbf{w}^\top G^{\perp}_1\mathbf{w} \\
&\geq \frac{1}{2}\mathbf{w}^\top G_1\mathbf{w} \\
&= \frac{1}{2}\|(A_1\mathbf{w})^\top(\mathcal{A}^0_{p_\theta}\phi_i)\|_{L_2} \geq \frac{1}{3\lambda_1}\left(\mathbf{w}^\top\Sigma_{\mathrm{KSD}}\mathbf{w}\right)^{1/2}.
\end{aligned}$$

This shows that $\mathbf{w}^\top \Sigma_{\mathrm{KSD}} \mathbf{w} \le 9\lambda_1^2 \mathbf{w}^\top \Sigma_{\mathrm{SM}} \mathbf{w}$, so the claim follows. □

Lemma 6.9. *Given symmetric matrices $A, B \in \mathbb{R}^{m\times m}$, if A is invertible, then* $\sup_{\|\mathbf{w}\|_2=1} \mathbf{w}^\top ABA\mathbf{w} = \sup_{\|\mathbf{u}\|_2=1} \mathbf{u}^\top B\mathbf{u} \|A^{-1}\mathbf{u}\|_2^{-2}$.

Proof of Lemma 6.9. Given $\mathbf{w}$ with $\|\mathbf{w}\|_2 = 1$, define $\mathbf{u} \coloneqq A\mathbf{w}/\|A\mathbf{w}\|_2 \in \mathbb{S}^{m-1}$. Then

$$\begin{aligned} \mathbf{u}^\top B\mathbf{u} \|A^{-1}\mathbf{u}\|_2^{-2} &= \big(\mathbf{w}^\top ABA\mathbf{w} \|A\mathbf{w}\|_2^{-2}\big) \big\|\mathbf{w}/\|A\mathbf{w}\|_2\big\|_2^{-2} = \big(\mathbf{w}^\top ABA\mathbf{w} \|A\mathbf{w}\|_2^{-2}\big) \|A\mathbf{w}\|_2^2 \|\mathbf{w}\|_2^{-2} \\ &= \mathbf{w}^\top ABA\mathbf{w} \,, \end{aligned}$$

where the last line holds since $\|\mathbf{w}\|_2 = 1$ by construction. This shows that $\sup_{\|\mathbf{w}\|_2=1} \mathbf{w}^\top ABA\mathbf{w} \le \sup_{\|\mathbf{u}\|_2=1} u^\top Bu \|A^{-1}u\|_2^{-2}$. To show equality, for any $\mathbf{u} \in \mathbb{S}^{m-1}$ we define $\mathbf{w} \coloneqq A^{-1}\mathbf{u}/\|A^{-1}\mathbf{u}\|_2 \in \mathbb{S}^{m-1}$. Then a similar argument shows that $\mathbf{w}^\top ABA\mathbf{w} = \mathbf{u}^\top B\mathbf{u} \|A^{-1}\mathbf{u}\|_2^{-2}$, and hence taking supremum over $\mathbf{u}$ gives $\sup_{\|\mathbf{w}\|_2=1} \mathbf{w}^\top ABA\mathbf{w} \ge \sup_{\|\mathbf{u}\|_2=1} u^\top Bu \|A^{-1}u\|_2^{-2}$. This completes the proof. □

Proof of Theorem 6.2. We upper bound $\|\Gamma_{\mathrm{KSD}}\|_{\mathrm{op}}$ as

$$\begin{aligned} \sup_{\|\mathbf{w}\|_2=1} \mathbf{w}^\top \Gamma_{\mathrm{KSD}} \mathbf{w} &= \sup_{\|\mathbf{w}\|_2=1} \mathbf{w}^\top g_{\mathrm{KSD}}^{-1} \Sigma_{\mathrm{KSD}} g_{\mathrm{KSD}}^{-1} \mathbf{w} \\ &\overset{(a)}{=} \sup_{\|\mathbf{u}\|_2=1} \mathbf{u}^\top \Sigma_{\mathrm{KSD}} \mathbf{u} \|g_{\mathrm{KSD}}\mathbf{u}\|_2^{-2} \\ &\overset{(b)}{\le} 9\lambda_1^2 \sup_{\|\mathbf{u}\|_2=1} \mathbf{u}^\top \Sigma_{\mathrm{SM}} \mathbf{u} \|g_{\mathrm{KSD}}\mathbf{u}\|_2^{-2} \\ &= 9\lambda_1^2 \sup_{\|\mathbf{u}\|_2=1} \big(g_{\mathrm{SM}}\mathbf{u}\big)^\top g_{\mathrm{SM}}^{-1} \Sigma_{\mathrm{SM}} g_{\mathrm{SM}}^{-1} \big(g_{\mathrm{SM}}\mathbf{u}\big) \|g_{\mathrm{KSD}}\mathbf{u}\|_2^{-2} \\ &= 9\lambda_1^2 \sup_{\|\mathbf{v}\|_2=1} \mathbf{v}^\top g_{\mathrm{SM}}^{-1} \Sigma_{\mathrm{SM}} g_{\mathrm{SM}}^{-1} \mathbf{v} \|g_{\mathrm{KSD}} g_{\mathrm{SM}}^{-1} \mathbf{v}\|_2^{-2} \\ &\overset{(c)}{=} 9\lambda_1^2 \sup_{\|\mathbf{v}\|_2=1} \mathbf{v}^\top \Gamma_{\mathrm{SM}} \mathbf{v} \|g_{\mathrm{KSD}} g_{\mathrm{SM}}^{-1} \mathbf{v}\|_2^{-2} \\ &\le 9\lambda_1^2 \bigg(\sup_{\|\mathbf{v}\|_2=1} \mathbf{v}^\top \Gamma_{\mathrm{SM}} \mathbf{v} \bigg) \bigg(\sup_{\|\mathbf{v}\|_2=1} \|g_{\mathrm{KSD}} g_{\mathrm{SM}}^{-1} \mathbf{v}\|_2^{-2} \bigg) \,, \end{aligned} \tag{6.32}$$

where in *(a)* and *(c)* we have applied Lemma 6.9, and *(b)* follows from Lemma 6.8, which holds under assumptions *(i)* and *(ii)*. Now, $\sup_{\|\mathbf{v}\|_2=1} \mathbf{v}^\top \Gamma_{\mathrm{SM}} \mathbf{v} = \|\Gamma_{\mathrm{op}}\|_2$. To simplify the last term in (6.32), we again apply Lemma 6.9 to yield

$$\begin{aligned} \sup_{\|\mathbf{v}\|_2=1} \|g_{\mathrm{KSD}} g_{\mathrm{SM}}^{-1} \mathbf{v}\|_2^{-2} = \sup_{\|\mathbf{u}\|_2=1} \|g_{\mathrm{KSD}}\mathbf{u}\|_2^{-2} \|g_{\mathrm{SM}}\mathbf{u}\|_2^2 &\le \Big(\inf_{\|\mathbf{u}\|_2=1} \|g_{\mathrm{KSD}}\mathbf{u}\|_2^2 \Big)^{-1} \sup_{\|\mathbf{u}\|_2=1} \|g_{\mathrm{SM}}\mathbf{u}\|_2^2 \\ &= \big(\lambda_{\min}(g_{\mathrm{KSD}})\big)^{-2} \lambda_{\max}(g_{\mathrm{SM}})^2 \,. \end{aligned} \tag{6.33}$$

It then remains to bound $\lambda_{\min}(g_{\mathrm{KSD}})$. Under the stated assumption, there exists a positive integer L

such that, for any $\mathbf{w} \in \mathbb{S}^{m-1}$,

$$\begin{aligned}\mathbf{w}^\top g_{\mathrm{SM}} \mathbf{w} \;=\; \sum_{i\geq 1} \|A_i \mathbf{w}\|_2^2 \;=\; \sum_{i\leq L} \|A_i \mathbf{w}\|_2^2 + \sum_{i>L} \|A_i \mathbf{w}\|_2^2 \;&=\; \mathbf{w}^\top g_{\mathrm{SM},L} \mathbf{w} + \mathbf{w}^\top g^{\perp}_{\mathrm{SM},L} \mathbf{w} \\ &\leq\; 2\mathbf{w}^\top g_{\mathrm{SM},L} \mathbf{w} \\ &\leq\; 2\lambda_L^{-1} \sum_{i\leq L} \lambda_i \|A_i \mathbf{w}\|_2^2 \\ &\leq\; 2\lambda_L^{-1} \sum_{i\geq 1} \lambda_i \|A_i \mathbf{w}\|_2^2 \;=\; 2\lambda_L^{-1} \mathbf{w}^\top g_{\mathrm{KSD}} \mathbf{w}\,,\end{aligned}$$

where the second last line holds since $\lambda_1 \geq \lambda_2 \geq \ldots$. Therefore, $\lambda_{\min}(g_{\mathrm{SM}}) \leq 2\lambda_L^{-1} \lambda_{\min}(g_{\mathrm{KSD}})$. Combining this with (6.32) and (6.33), we conclude that

$$\sup_{\|\mathbf{w}\|_2=1} \mathbf{w}^\top \Gamma_{\mathrm{KSD}} \mathbf{w} \;\leq\; 36 \frac{\lambda_1^2}{\lambda_L^2} \left(\frac{\lambda_{\max}(g_{\mathrm{SM}})}{\lambda_{\min}(g_{\mathrm{SM}})} \right)^2 \|\Gamma_{\mathrm{SM}}\|_{\mathrm{op}}\,.$$

□

6.B.5 Proof of Corollary 6.2.1

Define $\phi_1 \equiv 1$, and let $\{\phi_l\}_{l=2}^{\infty}$ be any sequence of orthonormal functions that are orthogonal to ϕ_1 under P_{θ_*}. Furthermore, define $\lambda_1 = 1$ and $\lambda_l = 0$ for $l \geq 2$. It is then clear that the eigenvalue-eigenfunction pairs $\{(\lambda_l, \phi_l)\}_{l=1}^{\infty}$ form a Mercer decomposition for $k(\mathbf{x}, \mathbf{y}) = 1$. Moreover, the sufficient statistic of the Gaussian-location model is $T(\mathbf{x}) = \mathbf{x}$, its Jacobian is $JT(\mathbf{x}) = I_d = \phi_1(\mathbf{x}) I_d$, and $\nabla_{\mathbf{x}} b(\mathbf{x}) = -\mathbf{x}$. Direct computation then shows that Assumption M.1 holds. Assumption M.2 also holds since $[JT]_{ij} = \phi_1$, which clearly lies in $\mathrm{Span}\{\phi_l\}_{l=1}^{\infty}$.

Furthermore, since $JT = \phi_1 I_d$, we have $A_l = \langle JT, \phi_l \rangle_{P_{\theta_*}} = \delta_{1l} I_d$, and

$$g_{\mathrm{SM},1} \;=\; I_d\,, \quad g^{\perp}_{\mathrm{SM},1} \;=\; 0\,.$$

This implies that assumption *(i)* in Theorem 6.2 holds with $L = 1$. To verify *(ii)*, for any $\mathbf{w} \in \mathbb{R}^d$, we have $\mathbf{w}^\top G_1 \mathbf{w} = \|(A_1 \mathbf{w}) \mathcal{A}_{p_{\theta_*} \phi_1}\|_{L_2(P_{\theta_*})} = \|\mathbf{w}^\top \mathcal{A}_{p_{\theta_*} \phi_1}\|_{L_2(P_{\theta_*})} > 0$ since $\mathcal{A}_{p_{\theta_*}} \phi_1$ is non-zero by direct computation. On the other hand, $G_1^{\perp} = \|\sum_{l>1} (A_l \mathbf{w})^\top \mathcal{A}^0_{p_\theta} \phi_l\|_{L_2(P_{\theta_*})} = 0$ as $A_l = 0$ for all $l \geq 2$, so *(ii)* holds. A similar argument shows that *(iii)* also holds.

Finally, we can apply Theorem 6.2 to conclude

$$\|\Gamma_{\mathrm{KSD}}\|_{\mathrm{op}} \;\leq\; 36 \frac{\lambda_1^2}{\lambda_L^2} \left(\frac{\lambda_{\max}(g_{\mathrm{SM}})}{\lambda_{\min}(g_{\mathrm{SM}})} \right)^2 \|\Gamma_{\mathrm{SM}}\|_{\mathrm{op}} \;=\; 36 \|\Gamma_{\mathrm{SM}}\|_{\mathrm{op}}\,,$$

where the last step holds since $g_{\mathrm{SM}} = I_d$ so that $\lambda_{\max}(g_{\mathrm{SM}}) = \lambda_{\min}(g_{\mathrm{SM}}) = 1$, and since we can choose $L = 1$, so that $\lambda_1 / \lambda_L = 1$.

Chapter 7

Summary and Future Directions

The increasing complexity of statistical models has driven the need for new tools in computational statistics, particularly for unnormalized models. This thesis focuses on the recent development of a class of novel methods for making statistical inference with unnormalized models. The central tool employed is kernel Stein discrepancy (KSD), a statistical divergence that combines kernel methods with Stein's method to enable tractable computation even when the normalizing constant of the model likelihood is unknown. This thesis has made several contributions to the development of KSD-based algorithms, by identifying the limitations of existing methods and proposing novel solutions.

- In Chapter 3, we demonstrated that Stein variational gradient descent (SVGD), a method for constructing finite-sample approximations of a given target distribution, can suffer from the curse of dimensionality. Motivated by existing work on projection-based dimensionality reduction methods, we proposed a new algorithm, Grassmann Stein variational gradient descent (GSVGD), that combines SVGD with linear projections to address this issue. This contribution bridges a gap in the literature by providing a flexible method for approximating high-dimensional distributions, particularly those exhibiting sparse structures.

- In Chapter 4, we focused on goodness-of-fit (GOF) testing. We showed that existing KSD-based GOF tests can fail when the target distribution has well-separated modes, an issue known as the *blindness to well-separated modes*. As a solution, we propose a class of new methods, called perturbed KSD (pKSD), that perturbs the observed sample before applying the KSD test in order to boost the test power against multi-modal alternatives. The perturbation is designed to be Markov processes that leave the target distribution invariant but perturbs alternatives, and we provide numerical evidence for its superior performance compared to existing methods. This contribution was the first to address the blindness of KSD in the context of GOF testing.

- In Chapter 5, we continue the topic of GOF testing, but focus instead on the robustness of KSD tests. We analyze the performance of the KSD test under different forms of model misspecification, including outlier contamination and misspecified tails. Both scenarios appear frequently in practice. We show that the KSD test can be sensitive to outliers even if the proportion of outliers in the observed data is small. As a result, standard KSD tests fail to control Type-I error in

this case, thus limiting their applicability in practice where data are subject to contamination. Motivated by this observation, we proposed a robust KSD test that guarantees robustness at little additional computational cost. Our novel test is also the first GOF test that is both robust and applicable to unnormalized models.

- In Chapter 6, we shift our focus to parameter estimation for unnormalized models. We study the statistical efficiency of existing KSD-based estimators, and show that they can be inferior to its competitors when the reproducing kernel in KSD has a rapidly decaying eigen-spectrum. Our analysis highlights the importance of kernel choice in minimum KSD estimation and reveals its limitation compared with alternative kernel-free approaches, such as score matching.

We hope that these contributions will inspire further research in this field and help address the challenges posed by modern statistical modelling given the increasing demand for large-scale, complex models. Research directions specific to each proposed method are discussed in the "Conclusion and Discussion" section of the corresponding chapters. Here, we highlight some general themes.

Privacy and fairness Understanding the privacy- and fairness-preserving properties of KSD-based algorithms is crucial for their responsible use in real-world applications. However, literature on the privacy or fairness of KSD-based methods is still limited. On the other hand, privacy has been shown to be closely related to *robustness* (Dwork and Lei, 2009), and the robustness of KSD has received more attention in the literature; see, for example, Barp et al. (2019); Matsubara et al. (2022) and Chapter 5. We therefore hope that these findings, including our contribution in Chapter 5, can lay the foundations for future research on the privacy and fairness aspects of KSD-based methods.

Interpretability A common criticism of KSD is its lack of interpretability. Although the KSD provides a valid notion of dissimilarity between probability distributions, it does not directly offer insights into what aspects of the distributions differ, such as mismatches in specific moments. Unlike Maximum Mean Discrepancy or many other integral probability metrics, which have a clearer interpretation through differences in expectations of certain test functions, the interpretation of KSD is much less clear due to the extra complexity added by the Stein operator. Jitkrittum et al. (2017) discussed in their experiment section how model diagnosis can be achieved by investigating the *witness function* of KSD, namely the function that attains the maximum of the KSD objective (2.8). Further development in this area could be helpful in providing a more intuitive understanding of this discrepancy.

Applications in deep learning The applications of KSD are mostly restricted to models with a low- or moderate-dimensional sample space, especially in the parameter estimation or goodness-of-fit fields. Some studies have explored its potential for high-dimensional data in the context of deep learning. For example, Grathwohl et al. (2020) used extensions of KSD to train deep generative models, while Schrab et al. (2022) explored testing the goodness-of-fit of large normalizing flows. However, KSD often underperforms when compared to other divergences like the score-matching loss (Song and Ermon, 2019; Song et al., 2020) or maximum mean discrepancy (Frogner et al., 2015). Future research could focus on identifying the causes of these limitations and developing extensions of KSD that can achieve state-of-the-art performance. This is particularly relevant in light of the recent boom in the applications of large language models and deep generative models in real-world scenarios.

Bibliography

Absil, P.-A., Mahony, R., and Sepulchre, R. (2008). *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ.

Adamczak, R. (2015). A note on the Hanson-Wright inequality for random vectors with dependencies. *Electronic Communications in Probability*, 20(none):1–13.

Afzali, E. and Muthukumarana, S. (2023). Gradient-free kernel conditional Stein discrepancy goodness-of-fit testing. *Machine Learning with Applications*, 12:100463.

Ahn, S., Chen, Y., and Welling, M. (2013). Distributed and adaptive darting Monte Carlo through regenerations. In *International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 108–116, Scottsdale, Arizona, USA. PMLR.

Alquier, P. and Gerber, M. (2024). Universal robust regression via maximum mean discrepancy. *Biometrika*, 111(1):71–92.

Altamirano, M., Briol, F.-X., and Knoblauch, J. (2023). Robust and scalable Bayesian online changepoint detection. In *International Conference on Machine Learning*, pages 642–663. PMLR.

Altamirano, M., Briol, F.-X., and Knoblauch, J. (2024). Robust and conjugate Gaussian process regression. In *International Conference on Machine Learning*, pages 1155–1185.

Aluffi-Pentini, F., Parisi, V., and Zirilli, F. (1985). Global optimization and stochastic differential equations. *Journal of optimization theory and applications*, 47(1):1–16.

Amari, S.-i. (2016). *Information Geometry and Its Applications*, volume 194. Springer.

Amin, A. N., Weinstein, E. N., and Marks, D. S. (2023). A kernelized Stein discrepancy for biological sequences. In *International Conference of Machine Learning*, pages 718–767.

Anastasiou, A., Barp, A., Briol, F.-X., Ebner, B., Gaunt, R. E., Ghaderinezhad, F., Gorham, J., Gretton, A., Ley, C., Liu, Q., Mackey, L., Oates, C. J., Reinert, G., and Swan, Y. (2023). Stein's method meets computational statistics: A review of some recent developments. *Statistical Science*, 38(1):120–139.

Arcones, M. A. and Gine, E. (1992). On the bootstrap of U and V statistics. *The Annals of Statistics*, 20(2):655–674.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404.

Asuncion, A. and Newman, D. (2007). UCI machine learning repository.

Ba, J., Erdogdu, M. A., Ghassemi, M., Sun, S., Suzuki, T., Wu, D., and Zhang, T. (2022). Understanding the variance collapse of SVGD in high dimensions. In *International Conference on Learning Representations*.

Ba, J., Erdogdu, M. A., Ghassemi, M., Suzuki, T., Sun, S., Wu, D., and Zhang, T. (2019). Towards characterizing the high-dimensional bias of kernel-based particle inference algorithms. *AABI 2019, Symposium on Advances in Approximate Bayesian Inference, 2019*.

Balasubramanian, K., Banerjee, S., and Ghosal, P. (2024). Improved finite-particle convergence rates for stein variational gradient descent. *arXiv preprint arXiv:2409.08469*.

Barbour, A. D. (1988). Stein's method and Poisson process convergence. *Journal of Applied Probability*, 25(A):175–184.

Barbour, A. D. (1990). Stein's method for diffusion approximations. *Probability theory and related fields*, 84(3):297–322.

Barker, A. A. (1965). Monte Carlo calculations of the radial distribution functions for a proton-electron plasma. *Australian Journal of Physics*, 18:119.

Barndorff-Nielsen, O. (1978). *Information and Exponential Families: In Statistical Theory*. John Wiley & Sons.

Barp, A., Briol, F.-X., Duncan, A., Girolami, M., and Mackey, L. (2019). Minimum Stein discrepancy estimators. *Advances in Neural Information Processing Systems*, 32:12964–12976.

Barp, A., Simon-Gabriel, C.-J., Girolami, M., and Mackey, L. (2022). Targeted separation and convergence with kernel discrepancies. *arXiv preprint arXiv:2209.12835*.

Bartle, R. G. (1995). *The Elements of Integration and Lebesgue Measure*. John Wiley & Sons, New York.

Basu, A. and Sarkar, S. (1994). The trade-off between robustness and efficiency and the effect of model smoothing in minimum disparity inference. *Journal of Statistical Computation and Simulation*, 50(3-4):173–185.

Basu, A., Shioya, H., and Park, C. (2011). *Statistical Inference: The Minimum Distance Approach*. CRC press.

Baum, J., Kanagawa, H., and Gretton, A. (2023). A kernel Stein test of goodness of fit for sequential models. In *International Conference on Machine Learning*, pages 1936–1953. PMLR.

Baxendale, P. et al. (1976). Measures and Markov processes on function spaces. *Mémoires de la Société Mathématique de France*, 46(131-141):3.

Belopolskaya, Y. I. and Dalecky, Y. L. (2012). *Stochastic Equations and Differential Geometry*, volume 30. Springer Science & Business Media.

Bendokat, T., Zimmermann, R., and Absil, P.-A. (2020). A Grassmann manifold handbook: Basic geometry and computational aspects. *arXiv preprint arXiv:2011.13699*.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science+Business Media, New York.

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019). On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225.

Billingsley, P. (1999). *Convergence of Probability Measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, 2 edition.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer New York, NY, 1 edition.

Blackard, K. L., Rappaport, T. S., and Bostian, C. W. (1993). Measurements and models of radio frequency impulsive noise for indoor wireless communications. *IEEE Journal on Selected Areas in Communications*, 11(7):991–1001.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Bonnini, S., Corain, L., Marozzi, M., and Salmaso, L. (2014). *Nonparametric Hypothesis Testing: Rank and Permutation Methods with Applications in R*. John Wiley & Sons.

Boumal, N. (2020). *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press.

Bourne, D. E. (1977). *Vector Analysis and Cartesian Tensors*. Chapman and Hall/CRC, 2 edition.

Briol, F.-X., Barp, A., Duncan, A. B., and Girolami, M. (2019). Statistical inference for generative models with maximum mean discrepancy. *arXiv preprint arXiv:1906.05944*.

Cambanis, S., Huang, S., and Simons, G. (1981). On the theory of elliptically contoured distributions. *Journal of Multivariate Analysis*, 11(3):368–385.

Canu, S. and Smola, A. (2006). Kernel methods and the exponential family. *Neurocomputing*, 69(7-9):714–720.

Carmeli, C., De Vito, E., and Toigo, A. (2006). Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem. *Analysis and Applications*, 4(04):377–408.

Carmeli, C., De Vito, E., Toigo, A., and Umanitá, V. (2010). Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61.

Casella, G. and Berger, R. (2002). *Statistical Inference*. CRC Press, 2 edition.

Chen, L. H., Goldstein, L., and Shao, Q.-M. (2010). *Normal Approximation by Stein's Method*. Springer Science & Business Media.

Chen, P. and Ghattas, O. (2020). Projected Stein variational gradient descent. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1947–1958. Curran Associates, Inc.

Chen, W. Y., Barp, A., Briol, F.-X., Gorham, J., Girolami, M., Mackey, L., and Oates, C. (2019). Stein point Markov chain Monte Carlo. In Chaudhuri, K. and Salakhutdinov, R., editors, *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1011–1021. PMLR.

Chen, W. Y., Mackey, L., Gorham, J., Briol, F.-X., and Oates, C. (2018). Stein points. In Dy, J. and Krause, A., editors, *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 844–853. PMLR.

Chen, Z., Ding, L., Huang, J., Chu, Z., Dai, Q., and Wang, H. (2023). Unsupervised anomaly detection & diagnosis: A Stein variational gradient descent approach. In *ACM International Conference on Information and Knowledge Management*, Cikm '23, pages 3783–3787, New York, NY, USA. Association for Computing Machinery.

Chérief-Abdellatif, B.-E. and Alquier, P. (2020). MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. In *Proceedings of the 2nd Symposium on Advances in Approximate Bayesian Inference (AABI)*, pages 1–21.

Chérief-Abdellatif, B.-E. and Alquier, P. (2022). Finite sample properties of parametric MMD estimation: Robustness to misspecification and dependence. *Bernoulli. Official Journal of the Bernoulli Society for Mathematical Statistics and Probability*, 28(1):181–213.

Cho, K. H., Raiko, T., and Ilin, A. (2013). Gaussian-Bernoulli deep Boltzmann machine. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Chung, F. and Lu, L. (2002). Connected components in random graphs with given expected degree sequences. *Annals of combinatorics*, 6(2):125–145.

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016). A kernel test of goodness of fit. In Balcan, M. F. and Weinberger, K. Q., editors, *International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2606–2615, New York, New York, USA. PMLR.

Chwialkowski, K. P., Sejdinovic, D., and Gretton, A. (2014). A wild bootstrap for degenerate kernel tests. *Advances in neural information processing systems*, 27.

Cichocki, A. and Amari, S.-i. (2010). Families of alpha- beta- and gamma- divergences: Flexible and robust measures of similarities. *Entropy. An International and Interdisciplinary Journal of Entropy and Information Studies*, 12(6):1532–1568.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley-Interscience, 2 edition.

Cramér, H. (1999). *Mathematical Methods of Statistics*, volume 26. Princeton university press, reprint of the 1946 original edition.

Csiszár, I., Shields, P. C., et al. (2004). Information theory and statistics: A tutorial. *Foundations and Trends® in Communications and Information Theory*, 1(4):417–528.

Cucker, F. and Zhou, D. X. (2007). *Learning Theory: An Approximation Theory Viewpoint*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge.

Cui, T., Law, K. J., and Marzouk, Y. M. (2016). Dimension-independent likelihood-informed MCMC. *Journal of Computational Physics*, 304:109–137.

Dabak, A. G. and Johnson, D. (1994). Geometrically based robust detection. *Proceedings of the Conference on Information Sciences and Systems, Johns Hopkins University, Baltimore*, pages 73–77.

D'Agostino, Ralph B (2017). *Goodness-of-Fit-Techniques*. Routledge, 1 edition.

Dang, K.-D., Quiroz, M., Kohn, R., Minh-Ngoc, T., and Villani, M. (2019). Hamiltonian Monte Carlo with energy conserving subsampling. *Journal of machine learning research*, 20.

de Figueiredo, L. P., Grana, D., Roisenberg, M., and Rodrigues, B. B. (2019). Multimodal Markov chain Monte Carlo method for nonlinear petrophysical seismic inversion. *Geophysics*, 84(5):M1–M13.

De Visscher, L., De Baets, B., and Baetens, J. M. (2024). A critical review of common pitfalls and guidelines to effectively infer parameters of agent-based models using Approximate Bayesian Computation. *Environmental Modelling & Software*, 172:105905.

Dehling, H. and Mikosch, T. (1994). Random quadratic forms and the bootstrap for U-statistics. *Journal of Multivariate Analysis*, 51(2):392–413.

Del Barrio, E., Cuesta-Albertos, J. A., Matrán, C., and Rodríguez-Rodríguez, J. M. (1999). Tests of goodness of fit based on the L2-Wasserstein distance. *Annals of Statistics*, pages 1230–1239.

Dellaporta, C. and Damoulas, T. (2023). Robust Bayesian inference for Berkson and classical measurement error models. *arXiv:2306.01468*.

Dellaporta, C., Knoblauch, J., Damoulas, T., and Briol, F.-X. (2022). Robust Bayesian inference for simulator-based models via the MMD posterior bootstrap. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pages 943–970.

Detommaso, G., Cui, T., Marzouk, Y., Spantini, A., and Scheichl, R. (2018). A Stein variational Newton method. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Duncan, A., Nüsken, N., and Szpruch, L. (2023). On the geometry of Stein variational gradient descent. *Journal of Machine Learning Research*, 24(56):1–39.

Duran-Martin, G., Altamirano, M., Shestopaloff, A. Y., Knoblauch, J., Jones, M., Briol, F.-X., and Murphy, K. (2024). Outlier-robust Kalman filtering through generalised Bayes. In *International Conference on Machine Learning*, pages 12138–12171.

Durrett, R. (2019). *Probability: Theory and Examples*, volume 49. Cambridge university press.

Dwork, C. and Lei, J. (2009). Differential privacy and robust statistics. In *ACM Symposium on Theory of Computing*, pages 371–380.

Dziugaite, G. K., Roy, D. M., and Ghahramani, Z. (2015). Training generative neural networks via maximum mean discrepancy optimization. In *Conference on Uncertainty in Artificial Intelligence*, pages 258–267, Arlington, Virginia, USA. AUAI Press.

Efron, B. (2022). *Exponential Families in Theory and Practice*. Institute of Mathematical Statistics Textbooks. Cambridge University Press, Cambridge.

Fauß, M. and Zoubir, A. M. (2016). Old bands, new tracks—Revisiting the band model for robust hypothesis testing. *IEEE Transactions on signal Processing*, 64(22):5875–5886.

Fauß, M., Zoubir, A. M., and Poor, H. V. (2021). Minimax robust detection: Classic results and recent advances. *IEEE Transactions on Signal Processing*, 69:2252–2283.

Feng, Y., Wang, D., and Liu, Q. (2017). Learning to draw samples with amortized Stein variational gradient descent. *arXiv e-prints*, page arXiv:1707.06626.

Fernandez, T., Rivera, N., Xu, W., and Gretton, A. (2020). Kernelized Stein discrepancy tests of goodness-of-fit for time-to-event data. In *International Conference on Machine Learning*, pages 3112–3122. PMLR.

Fischer, A. and Igel, C. (2014). Training restricted Boltzmann machines: An introduction. *Pattern Recognition*, 47(1):25–39.

Fisher, M. A. and Oates, C. J. (2023). Gradient-free kernel Stein discrepancy. In *Advances in Neural Information Processing Systems*, volume 36, pages 23855–23885. Curran Associates, Inc.

Forbes, C., Evans, M., Hastings, N., and Peacock, B. (2011). *Statistical Distributions*. John Wiley & Sons, 4 edition.

Forbes, P. G. and Lauritzen, S. (2015). Linear estimating equations for exponential families with application to Gaussian linear concentration models. *Linear Algebra and its Applications*, 473:261–283.

Fortet, R. and Mourier, E. (1953). Convergence de la répartition empirique vers la répartition théorique. *Annales scientifiques de l'École Normale Supérieure*, 3e série, 70(3):267–285.

Frénay, B. and Verleysen, M. (2013). Classification in the presence of label noise: A survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869.

Frogner, C., Zhang, C., Mobahi, H., Araya, M., and Poggio, T. A. (2015). Learning with a Wasserstein loss. *Advances in neural information processing systems*, 28.

Fukumizu, K., Gretton, A., Lanckriet, G., Schölkopf, B., and Sriperumbudur, B. K. (2009). Kernel choice and classifiability for RKHS embeddings of probability distributions. *Advances in neural information processing systems*, 22.

Fukumizu, K., Song, L., and Gretton, A. (2013). Kernel Bayes' rule: Bayesian inference with positive definite kernels. *The Journal of Machine Learning Research*, 14(1):3753–3783.

Gao, R., Xie, L., Xie, Y., and Xu, H. (2018). Robust hypothesis testing using Wasserstein uncertainty sets. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. (2020). Bayesian workflow. *arXiv preprint arXiv:2011.01808*.

Geman, S. and Hwang, C.-R. (1986). Diffusions for global optimization. *SIAM Journal on Control and Optimization*, 24(5):1031–1043.

Gerber, P. R. and Polyanskiy, Y. (2024). Likelihood-free hypothesis testing. *IEEE Transactions on Information Theory*.

Gilks, W., Richardson, S., and Spiegelhalter, D. (1995). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis.

Godambe, V. P. (1991). *Estimating Functions*. Oxford University Press.

Golub, G. H. and van Loan, C. F. (2013). *Matrix Computations*. JHU Press, 4 edition.

Gong, W., Li, Y., and Hernández-Lobato, J. M. (2021a). Sliced kernelized Stein discrepancy. In *International Conference on Learning Representations*.

Gong, W., Zhang, K., Li, Y., and Hernández-Lobato, J. M. (2021b). Active slices for sliced Stein discrepancy. *arXiv preprint arXiv:2102.03159*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Gorham, J., Duncan, A. B., Vollmer, S. J., and Mackey, L. (2019). Measuring sample quality with diffusions. *The Annals of Applied Probability*, 29(5):2884–2928.

Gorham, J. and Mackey, L. (2015). Measuring sample quality with Stein's method. *Advances in Neural Information Processing Systems*, 28.

Gorham, J. and Mackey, L. (2017). Measuring sample quality with kernels. In *International Conference on Machine Learning*, pages 1292–1301. PMLR.

Gotze, F. (1991). On the rate of convergence in the multivariate CLT. *The Annals of Probability*, pages 724–739.

Gradshteyn, I., Ryzhik, I., and Jeffrey, A. (1980). *Table of Integrals, Series, and Products*. Academic Press.

Grams, W. F. and Serfling, RJ. (1973). Convergence rates for U-statistics and related statistics. *The Annals of Statistics*, pages 153–160.

Grathwohl, W., Wang, K.-C., Jacobsen, J.-H., Duvenaud, D., and Zemel, R. (2020). Learning the Stein discrepancy for training and evaluating energy-based models without sampling. In *International Conference on Machine Learning*, pages 3732–3747. PMLR.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Green, P. J. and Hastie, D. I. (2009). Reversible jump MCMC. *Genetics*, 155(3):1391–1403.

Green, P. J., Łatuszyński, K., Pereyra, M., and Robert, C. P. (2015). Bayesian computation: A summary of the current state, and samples backwards and forwards. *Statistics and Computing*, 25:835–862.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2006). A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems*, volume 19.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. (2007). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, volume 20.

Gül, G. and Zoubir, A. M. (2016). Robust hypothesis testing with α-divergence. *IEEE Transactions on Signal Processing*, 64(18):4737–4750.

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. (2017). Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pages 1352–1361. PMLR.

Hafner, R. (1993). Construction of minimax-tests for bounded families of probability-densities. *Metrika. International Journal for Theoretical and Applied Statistics*, 40(1):1–23.

Hagrass, O., Sriperumbudur, B., and Balasubramanian, K. (2024). Minimax optimal goodness-of-fit testing with kernel stein discrepancy. *arXiv preprint arXiv:2404.08278*.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.

Han, J., Ding, F., Liu, X., Torresani, L., Peng, J., and Liu, Q. (2020). Stein variational inference for discrete distributions. In *International Conference on Artificial Intelligence and Statistics*, pages 4563–4572. PMLR.

Han, J. and Liu, Q. (2018). Stein variational gradient descent without gradient. In *International Conference on Machine Learning*, pages 1900–1908. PMLR.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Hird, M., Livingstone, S., and Zanella, G. (2020). A fresh take on 'Barker dynamics' for MCMC. In *International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pages 169–184. Springer.

Hodgkinson, L., Salomone, R., and Roosta, F. (2020). The reproducing Stein kernel approach for post-hoc corrected sampling. *arXiv preprint arXiv:2001.09266*.

Hoeffding, W. (1961). The strong law of large numbers for U-statistics. Technical report, North Carolina State University. Dept. of Statistics.

Hoffman, M. D., Gelman, A., et al. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Hofmann, T., Schölkopf, B., and Smola, A. J. (2008). Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220.

Hogg, R. V., Tanis, E. A., and Zimmerman, D. L. (1977). *Probability and Statistical Inference*, volume 993. Macmillan New York.

Huang, K. H., Liu, X., Duncan, A., and Gandy, A. (2023). A high-dimensional convergence theorem for U-statistics with applications to kernel-based testing. In Neu, G. and Rosasco, L., editors, *Conference on Learning Theory*, volume 195, pages 3827–3918. PMLR.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Huber, P. J. (1965). A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, pages 1753–1758.

Huber, P. J. and Ronchetti, E. M. (2011). *Robust Statistics*. John Wiley & Sons.

Huggins, J. and Mackey, L. (2018). Random feature Stein discrepancies. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Huskova, M. and Janssen, P. (1993). Consistency of the generalized bootstrap for degenerate U-statistics. *The Annals of Statistics*, 21(4):1811–1823.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709.

Hyvärinen, A. (2007). Some extensions of score matching. *Computational Statistics & Data Analysis*, 51(5):2499–2512.

Ihler, A., Fisher, J., Moses, R., and Willsky, A. (2005). Nonparametric belief propagation for self-localization of sensor networks. *IEEE Journal on Selected Areas in Communications*, 23(4):809–819.

Jager, L. and Wellner, J. A. (2007). Goodness-of-fit tests via phi-divergences. *The Annals of Statistics*, 35(5):2018–2053.

Janssen, P. (1994). Weighted bootstrapping of U-statistics. *Journal of Statistical Planning and Inference*, 38(1):31–41.

Janssen, P. (1997). Bootstrapping U-statistics. *South African Statistical Journal*, 31(2):185–216.

Jitkrittum, W., Xu, W., Szabó, Z., Fukumizu, K., and Gretton, A. (2017). A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems*, volume 30, pages 261–270, Red Hook, NY, USA. Curran Associates Inc.

Jordan, M. I. and Weiss, Y. (2002). Probabilistic inference in graphical models. *Handbook of neural networks and brain theory*.

Kanagawa, H., Gretton, A., and Mackey, L. (2022). Controlling moments with kernel Stein discrepancies. *arXiv preprint arXiv:2211.05408*.

Kanagawa, H., Jitkrittum, W., Mackey, L., Fukumizu, K., and Gretton, A. (2023). A kernel Stein test for comparing latent variable models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):986–1011.

Kassam, S. (1981). Robust hypothesis testing for bounded classes of probability densities (corresp.). *IEEE Transactions on Information Theory*, 27(2):242–247.

Key, O., Gretton, A., Briol, F.-X., and Fernandez, T. (2021). Composite goodness-of-fit tests with kernels. *arXiv preprint arXiv:2111.10275*.

Knoblauch, J., Jewson, J., and Damoulas, T. (2022). An optimization-centric view on Bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23(132):1–109.

Koehler, F., Heckett, A., and Risteski, A. (2022). Statistical efficiency of score matching: The view from isoperimetry. *arXiv preprint arXiv:2210.00726*.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge didistribuzione. *Giorn Dell'inst Ital Degli Att*, 4:89–91.

Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., and Rohde, G. (2019). Generalized sliced Wasserstein distances. In *Advances in Neural Information Processing Systems*, volume 32, pages 261–272.

Korba, A., Aubin-Frankowski, P.-C., Majewski, S., and Ablin, P. (2021). Kernel Stein discrepancy descent. In *International Conference on Machine Learning*, pages 5719–5730. PMLR.

Korba, A., Salim, A., Arbel, M., Luise, G., and Gretton, A. (2020). A non-asymptotic analysis for Stein variational gradient descent. In *Advances in Neural Information Processing Systems*, volume 33, pages 4672–4682.

Kreyszig, E. (1991). *Introductory Functional Analysis with Applications*, volume 17. John Wiley & Sons.

Lambert, D. (1982). Qualitative robustness of tests. *Journal of the American Statistical Association*, 77(378):352–357.

Lan, S., Streets, J., and Shahbaba, B. (2014). Wormhole Hamiltonian Monte Carlo. In *AAAI Conference on Artificial Intelligence*, volume 28.

LeCam, L. (1973). Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53.

Lehmann, E. L. and Romano, J. P. (2022). *Testing Statistical Hypotheses*. Springer Cham, 4 edition.

Leucht, A. and Neumann, M. H. (2013). Dependent wild bootstrap for degenerate U- and V-statistics. *Journal of Multivariate Analysis*, 117:257–280.

Levin, D. A. and Peres, Y. (2017). *Markov Chains and Mixing Times*, volume 107. American Mathematical Society.

Levy, B. C. (2008). Robust hypothesis testing with a relative entropy tolerance. *IEEE Transactions on Information Theory*, 55(1):413–421.

Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S., and Corander, J. (2017). Fundamentals and recent developments in approximate Bayesian computation. *Systematic Biology*, 66(1):e66–e82.

Liu, A., Liang, Y., and den Broeck, G. V. (2020). Off-policy deep reinforcement learning with analogous disentangled exploration. *In International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

Liu, B., Rocco, P., Zanchettin, A. M., Zhao, F., Jiang, G., and Mei, X. (2024a). A real-time hierarchical control method for safe human–robot coexistence. *Robotics and Computer-Integrated Manufacturing*, 86:102666.

Liu, C. and Zhu, J. (2018). Riemannian Stein variational gradient descent for Bayesian inference. In *AAAI Conference on Artificial Intelligence*, pages 3627–3634, New Orleans, Louisiana USA. AAAI press.

Liu, M., Grana, D., and Mukerji, T. (2024b). Geostatistical inversion for subsurface characterization using Stein variational gradient descent with autoencoder neural network: An application to geologic carbon sequestration. *Journal of Geophysical Research: Solid Earth*, 129(7):e2024JB029073.

Liu, Q. (2017). Stein variational gradient descent as gradient flow. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Liu, Q., Lee, J., and Jordan, M. (2016). A kernelized Stein discrepancy for goodness-of-fit tests. In Balcan, M. F. and Weinberger, K. Q., editors, *International Conference on Machine Learning*, volume 48, pages 276–284, New York, New York, USA. PMLR.

Liu, Q. and Wang, D. (2016). Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in Neural Information Processing Systems*, volume 29.

Liu, X. and Briol, F.-X. (2024). On the robustness of kernel goodness-of-fit tests. *arXiv preprint arXiv:2408.05854*.

Liu, X., Duncan, A. B., and Gandy, A. (2023). Using perturbation to improve goodness-of-fit tests based on kernelized Stein discrepancy. In *International Conference on Machine Learning*, pages 21527–21547. PMLR.

Liu, X., Zhu, H., Ton, J.-F., Wynne, G., and Duncan, A. (2022). Grassmann Stein variational gradient descent. In *International Conference on Artificial Intelligence and Statistics*, volume 151, pages 2002–2021. PMLR.

Liu, Y., Ramachandran, P., Liu, Q., and Peng, J. (2017). Stein variational policy gradient. In *Uncertainty in Artificial Intelligence (UAI)*.

Livingstone, S. and Zanella, G. (2021). The Barker proposal: Combining robustness and efficiency in gradient-based MCMC. *Journal of the Royal Statistical Society: Series B*.

Lloyd, J. R. and Ghahramani, Z. (2015). Statistical model criticism using kernel two sample tests. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Lu, J., Lu, Y., and Nolen, J. (2019). Scaling limit of the Stein variational gradient descent: The mean field regime. *SIAM Journal on Mathematical Analysis*, 51(2):648–671.

Mahalanobis, P. (1936). On the generalized distance in statistics. *Proceedings of the National Academy of Sciences (India)*, 2:49–55.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and computing*, 22(6):1167–1180.

Marzouk, Y., Moselhy, T., Parno, M., and Spantini, A. (2016). An introduction to sampling via measure transport. *arXiv preprint arXiv:1602.05023*.

Massey Jr, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78.

Matsubara, T., Knoblauch, J., Briol, F.-X., and Oates, C. J. (2022). Robust generalised Bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):997–1022.

McDiarmid, C. et al. (1989). On the method of bounded differences. *Surveys in Combinatorics*, 141(1):148–188.

Melchior, J., Wang, N., and Wiskott, L. (2017). Gaussian-binary restricted Boltzmann machines for modeling natural image statistics. *PLoS One*, 12(2):e0171015.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer Science & Business Media.

Mohamed, S. and Lakshminarayanan, B. (2016). Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*.

Moore, E. H. (1935). *General Analysis*, volume 1 of *Memoirs of the American Philosophical Society*. American Philosophical Society, Philadelphia.

Morelli, M., Hauth, J., Guardone, A., Huan, X., and Zhou, B. Y. (2023). A rotorcraft in-flight ice detection framework using computational aeroacoustics and Bayesian neural networks. *Structural and Multidisciplinary Optimization*, 66(9):197.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443.

Murphy, S. A. and Van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95(450):449–465.

Murray, I., Ghahramani, Z., and MacKay, D. (2012). MCMC for doubly-intractable distributions. *arXiv preprint arXiv:1206.6848*.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245.

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, NY, 2. ed. edition.

Noughabi, H. A. and Balakrishnan, N. (2016). Tests of goodness of fit based on Phi-divergence. *Journal of Applied Statistics*, 43(3):412–429.

Oates, C. J. (2022). Minimum kernel discrepancy estimators. In *International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, pages 133–161. Springer.

Oates, C. J., Girolami, M., and Chopin, N. (2017). Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(3):695–718.

Ouimet, F. (2020). Explicit formulas for the joint third and fourth central moments of the multinomial distribution. *arXiv preprint arXiv:2006.09059*.

Pabbaraju, C., Rohatgi, D., Sevekari, A., Lee, H., Moitra, A., and Risteski, A. (2023). Provable benefits of score matching. *arXiv preprint arXiv:2306.01993*.

Palmer, J. A., Kreutz-Delgado, K., and Makeig, S. (2010). Strong Sub- and Super-Gaussianity. In Vigneron, V., Zarzoso, V., Moreau, E., Gribonval, R., and Vincent, E., editors, *Latent Variable Analysis and Signal Separation*, pages 303–310, Berlin, Heidelberg. Springer Berlin Heidelberg.

Pardo, L. (2005). *Statistical Inference Based on Divergence Measures*, volume 170. Chapman and Hall/CRC.

Paulsen, V. I. and Raghupathi, M. (2016). *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge University Press.

Peskun, P. H. (1973). Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60(3):607–612.

Pompe, E., Holmes, C., and Łatuszyński, K. (2020). A framework for adaptive MCMC targeting multimodal distributions. *The Annals of Statistics*, 48(5):2930–2952.

Postman, M., Huchra, J. P., and Geller, M. J. (1986). Probes of large-scale structure in the Corona Borealis region. *Astronomical Journal (ISSN 0004-6256), vol. 92, Dec. 1986, p. 1238-1247.*, 92:1238–1247.

Prokhorov, Y. V. (1956). Convergence of random processes and limit theorems in probability theory. *Theory of Probability & Its Applications*, 1(2):157–214.

Pu, Q., Chen, Y., Zhou, M., Ng, J. K.-Y., and Cai, R. (2024). Bayesian meta-learning: Toward fast adaptation in neural network positioning techniques. *IEEE Internet of Things Journal*, 11(8):14924–14937.

Pu, Y., Gan, Z., Henao, R., Li, C., Han, S., and Carin, L. (2017). VAE learning via Stein variational gradient descent. In *Advances in Neural Information Processing Systems*, volume 30.

Qin, Y. and Priebe, C. E. (2017). Robust hypothesis testing via L_q-likelihood. *Statistica Sinica*, pages 1793–1813.

Qin, Y. and Risteski, A. (2023). Fit like you sample: Sample-efficient generalized score matching from fast mixing Markov chains. *arXiv preprint arXiv:2306.09332*.

Ramdas, A., Reddi, S. J., Póczos, B., Singh, A., and Wasserman, L. (2015). On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. In *Bulletin of the Calcutta Mathematical Society*, volume 37 of *3*, pages 81–91.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

Reddi, S., Ramdas, A., Póczos, B., Singh, A., and Wasserman, L. (2015). On the high dimensional power of a linear-time two sample test under mean-shift alternatives. In *Artificial Intelligence and Statistics*, pages 772–780. PMLR.

Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538. PMLR.

Riabiz, M., Chen, W. Y., Cockayne, J., Swietach, P., Niederer, S. A., Mackey, L., and Oates, C. J. (2022). Optimal thinning of MCMC output. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4):1059–1081.

Rieder, H. (1982). Qualitative robustness of rank tests. *The Annals of Statistics*, 10(1):205–211.

Rizzoni, G. and Min, PS. (1991). Detection of sensor failures in automotive engines. *IEEE transactions on vehicular technology*, 40(2):487–500.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer New York, New York, NY, 2nd ed. 2004. edition.

Robert, C. P. et al. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, volume 2. Springer.

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 85(411):617–624.

Ross, N. (2011). Fundamentals of Stein's method. *Probability Surveys*, 8(none):210–293.

Saitoh, S., Sawano, Y., et al. (2016). *Theory of Reproducing Kernels and Applications*. Springer.

Salim, A., Sun, L., and Richtarik, P. (2022). A convergence theory for SVGD in the population limit under Talagrand's inequality T1. In *International Conference on Machine Learning*, pages 19139–19152. PMLR.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT press.

Schrab, A., Guedj, B., and Gretton, A. (2022). KSD aggregated goodness-of-fit test. In *Advances in Neural Information Processing Systems*, volume 35, pages 32624–32638.

Schrab, A. and Kim, I. (2024). Robust kernel hypothesis testing under data corruption. *arXiv preprint arXiv:2405.19912*.

Schrab, A., Kim, I., Albert, M., Laurent, B., Guedj, B., and Gretton, A. (2023). MMD aggregated two-sample test. *Journal of Machine Learning Research*, 24(194):1–81.

Scovel, C., Hush, D., Steinwart, I., and Theiler, J. (2010). Radial kernels and their reproducing kernel Hilbert spaces. *Journal of Complexity*, 26(6):641–660.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013). Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291.

Serfling, R. J. (2009). *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons.

Shao, X. (2010). The dependent wild bootstrap. *Journal of the American Statistical Association*, 105(489):218–235.

Sharma, A. B., Golubchik, L., and Govindan, R. (2010). Sensor faults: Detection methods and prevalence in real-world datasets. *ACM Transactions on Sensor Networks (TOSN)*, 6(3):1–39.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge.

Shi, J., Liu, C., and Mackey, L. (2021). Sampling with mirrored Stein operators. *arXiv preprint arXiv:2106.12506*.

Shi, J. and Mackey, L. (2024). A finite-particle convergence rate for Stein variational gradient descent. In *Advances in Neural Information Processing Systems*, volume 36.

Song, K.-S. (2002). Goodness-of-fit tests based on Kullback-Leibler discrimination information. *IEEE Transactions on Information Theory*, 48(5):1103–1117.

Song, Y. and Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32.

Song, Y. and Kingma, D. P. (2021). How to train your energy-based models. *arXiv preprint arXiv:2101.03288*.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2020). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

South, L. F., Riabiz, M., Teymur, O., and Oates, C. J. (2022). Postprocessing of MCMC. *Annual Review of Statistics and Its Application*, 9(1):529–555.

Sriperumbudur, B., Fukumizu, K., Gretton, A., Hyv, A., Kumar, R., et al. (2017). Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18(57):1–59.

Sriperumbudur, B., Fukumizu, K., and Lanckriet, G. (2010a). On the relation between universality, characteristic kernels and RKHS embedding of measures. In *International Conference on Artificial Intelligence and Statistics*, volume 9, pages 773–780, Chia Laguna Resort, Sardinia, Italy. PMLR.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. (2011). Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(7).

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. (2010b). Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561.

Staneva, V. and Younes, L. (2017). Learning shape trends: Parameter estimation in diffusions on shape manifolds. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 38–46.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, volume 6, pages 583–603. University of California Press.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer Science & Business Media.

Stewart, J. (1976). Positive definite functions and generalizations, an historical survey. *Rocky Mountain Journal of Mathematics*, 6(3):409–434.

Sun, H. (2005). Mercer theorem for RKHS on noncompact sets. *Journal of Complexity*, 21(3):337–349.

Sun, L., Karagulyan, A., and Richtarik, P. (2023a). Convergence of Stein variational gradient descent under a weaker smoothness condition. In *International Conference on Artificial Intelligence and Statistics*, pages 3693–3717. PMLR.

Sun, Z., Barp, A., and Briol, F.-X. (2023b). Vector-valued control variates. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *International Conference on Machine Learning*, volume 202, pages 32819–32846. PMLR.

Sun, Z. and Zou, S. (2023). Kernel robust hypothesis testing. *IEEE Transactions on Information Theory*.

Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272.

Tak, H., Meng, X.-L., and van Dyk, D. A. (2018). A repelling–attracting Metropolis algorithm for multimodality. *Journal of Computational and Graphical Statistics*, 27(3):479–490.

van der Vaart, A. W. (2000). *Asymptotic Statistics*, volume 3. Cambridge University Press.

Van Erven, T. and Harremos, P. (2014). Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

Vasicek, O. (1976). A test for normality based on sample entropy. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 38(1):54–59.

Villani, C. (2008). *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media.

Vincent, P. (2011). A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674.

Wainwright, M. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wang, D. and Liu, Q. (2019). Nonlinear Stein variational gradient descent for learning diversified mixture models. In Chaudhuri, K. and Salakhutdinov, R., editors, *International Conference on Machine Learning*, volume 97, pages 6576–6585. PMLR.

Wang, D., Zeng, Z., and Liu, Q. (2018). Stein variational message passing for continuous graphical models. In Dy, J. G. and Krause, A., editors, *International Conference on Machine Learning*, volume 80, pages 5206–5214, Stockholmsmässan, Stockholm, Sweden. PMLR.

Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning*, pages 681–688. Citeseer.

Wenliang, L. K. and Kanagawa, H. (2020). Blindness of score-based methods to isolated components and mixing proportions. *arXiv preprint arXiv:2008.10087*.

William, C. P. (1981). Minimum distance estimation: A bibliography. *Communications in Statistics-Theory and Methods*, 10(12):1205–1224.

Wynne, G., Kasprzak, M., and Duncan, A. B. (2022). A spectral representation of kernel Stein discrepancy with application to goodness-of-fit tests for measures on infinite dimensional Hilbert spaces. *arXiv preprint arXiv:2206.04552*.

Xie, Y., Chauris, H., and Desassis, N. (2024). Stochastic full waveform inversion with deep generative prior for uncertainty quantification. *arXiv preprint arXiv:2406.04859*.

Xu, W. and Matsuda, T. (2020). A Stein goodness-of-fit test for directional distributions. In *International Conference on Artificial Intelligence and Statistics*, pages 320–330. PMLR.

Xu, W. and Matsuda, T. (2021). Interpretable Stein goodness-of-fit tests on Riemannian manifold. In *International Conference on Machine Learning*, pages 11502–11513. PMLR.

Xu, W. and Reinert, G. (2021). A Stein goodness-of-test for exponential random graph models. In *International Conference on Artificial Intelligence and Statistics*, pages 415–423. PMLR.

Yan, L. and Zou, X. (2021). Gradient-free Stein variational gradient descent with kernel approximation. *Applied Mathematics Letters*, 121:107465.

Yang, E., Ravikumar, P., Allen, G. I., and Liu, Z. (2015). Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research*, 16(1):3813–3847.

Yang, J., Liu, Q., Rao, V., and Neville, J. (2018). Goodness-of-fit testing for discrete distributions via Stein discrepancy. In *International Conference on Machine Learning*, pages 5561–5570. PMLR.

Yang, J., Rao, V., and Neville, J. (2019). A Stein–Papangelou goodness-of-fit test for point processes. In *International Conference on Artificial Intelligence and Statistics*, pages 226–235. PMLR.

Yang, P. and Chen, B. (2018). Robust Kullback-Leibler divergence and universal hypothesis testing for continuous distributions. *IEEE Transactions on Information Theory*, 65(4):2360–2373.

Yoon, J., Kim, T., Dia, O., Kim, S., Bengio, Y., and Ahn, S. (2018). Bayesian model-agnostic meta-learning. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Zhang, L., Lin, J., Liu, B., Zhang, Z., Yan, X., and Wei, M. (2019). A review on deep learning applications in prognostics and health management. *IEEE access : practical innovations, open solutions*, 7:162415–162438.

Zhang, M., Hayes, P., Bird, T., Habib, R., and Barber, D. (2020). Spread divergence. In *International Conference on Machine Learning*, volume 119, pages 11106–11116. PMLR.

Zhang, M., Key, O., Hayes, P., Barber, D., Paige, B., and Briol, F.-X. (2022). Towards healing the blindness of score matching. *arXiv preprint arXiv:2209.07396*.

Zhang, N., Ding, S., Zhang, J., and Xue, Y. (2018). An overview on Restricted Boltzmann Machines. *Neurocomputing*, 275:1186–1199.

Zhao, Y., Jiang, H., Cai, Y., and Deng, Y. (2024). A generalized data assimilation architecture of digital twin for complex process industrial systems. *Measurement Science and Technology*, 35(6):066003.

Zhu, H., Williams, C. K., Rohwer, R., and Morciniec, M. (1998). Gaussian regression and optimal finite dimensional linear models. *Neural Networks and Machine Learning*.

Zhuo, J., Liu, C., Shi, J., Zhu, J., Chen, N., and Zhang, B. (2018). Message passing Stein variational gradient descent. In *International Conference on Machine Learning*, pages 6018–6027. PMLR.