Hierarchical Gaussian Processes with Wasserstein-2 Kernels

Sebastian G. Popescu
Imperial College London
s.popescu16@ic.ac.uk

David J. Sharp
Imperial College London
david.sharp@ic.ac.uk

James H. Cole
University College London
james.cole@ucl.ac.uk

Ben Glocker
Imperial College London
b.glocker@ic.ac.uk

Abstract

We investigate the usefulness of Wasserstein-2 kernels in the context of hierarchical Gaussian Processes. Stemming from an observation that stacking Gaussian Processes severely diminishes the model's ability to detect outliers, which when combined with non-zero mean functions, further extrapolates low variance to regions with low training data density, we posit that directly taking into account the variance in the computation of Wasserstein-2 kernels is of key importance towards maintaining outlier status as we progress through the hierarchy. We propose two new models operating in Wasserstein space which can be seen as equivalents to Deep Kernel Learning and Deep GPs. Through extensive experiments, we show improved performance on large scale datasets and improved out-of-distribution detection on both toy and real data.

1 Introduction

Deep Gaussian Processes (DGPs) (Damianou and Lawrence, 2013) are a multi-layered generalization of Gaussian Processes (GPs) that inherit the advantages of GPs, namely calibrated predictive uncertainty and data-efficient learning. This makes them attractive in domains where data is sparse, such as in medical imaging or in safety critical applications such as self driving cars. The sequential embedding of the input through stacked layers of GPs solves the issue of having to hand tune kernels for specific tasks and implicitly embeds non-stationarity in the final output. Even though DGPs can be used in conjunction with the inducing point framework introduced in Hensman et al. (2013), this does not entail tractable inference as it is the case with shallow GPs. Recent implementations using stochastic approximate inference techniques have succeeded in using DGPs in medium and large datasets (Bui et al., 2016; Salimbeni and Deisenroth, 2017; Havasi et al., 2018; Yu et al., 2019). In this work we make use of the framework introduced in Salimbeni and Deisenroth (2017).

Recent work (Ustyuzhaninov et al., 2019) has questioned the validity of uncertainties present in the hidden layers of DGPs, showing that approximate inference schemes using variational Gaussian distributions result in all but the last GP collapsing to deterministic transformations in the case of noiseless data. Such pathological behaviour should be avoided as it undermines the utility of layered GPs. In this paper we further investigate the status of hidden layer uncertainties in DGP, showing failure cases and we propose a solution by reinterpreting already existing models in Wasserstein-2 space.

1.1 Contributions

This paper makes the following contributions:

- We provide empirical evidence that as the number of inducing point increases, the distributional variance in hidden layers collapses in the case of zero-mean DGP;
- We show that using PCA mean functions solves this problem, but this procedure induces low variance perpendicular to the eigenvectors;
• We derive theoretical requirements for zero-mean DGP to propagate high variance through the hierarchy;
• We propose equivalents to Deep Kernel Learning (DKL) (Wilson et al., 2016) and DGPs that operate on Wasserstein-2 space and we show them to be a better fit for out-of-distribution detection, alongside better testing time results for large-scale datasets.

2 BACKGROUND

In this section, we lay the theoretical foundations for the models that are introduced subsequently.

2.1 Optimal Transport

The Wasserstein space on \( \mathbb{R} \) can be defined as the set \( W_2(\mathbb{R}) \) of probability measures on \( \mathbb{R} \) with a finite moment of order two. We denote by \( \Pi(\mu, \nu) \) the set of all probability measures \( \Pi \) over the product set \( \mathbb{R} \times \mathbb{R} \) with marginals \( \mu \) and \( \nu \), which are probability distributions in \( W_2(\mathbb{R}) \). The transportation cost between two measures \( \mu \) and \( \nu \) is defined as:

\[
T_2(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int |x - y|^2 d\pi(x, y)
\]

This transportation cost allows us to endow the set \( W_2(\mathbb{R}) \) with a metric by defining the quadratic Wasserstein distance between \( \mu \) and \( \nu \) as:

\[
W_2(\mu, \nu) = T_2(\mu, \nu)^{1/2}
\]

2.2 Wasserstein-2 distance kernels on Gaussian measure space

Thi Thien Trang et al. (2019) have provided a proof of the following theorem which we reiterate in the supplementary material.

**Theorem 1** Let \( k_W : W_2(\mathbb{R}) \times W_2(\mathbb{R}) \to \mathbb{R} \) be the Wasserstein-2 RBF kernel defined as following:

\[
k_{W_2}(\mu, \nu) = \sigma^2 \exp \left( -\frac{W_2^2(\mu, \nu)}{l^2} \right),
\]

then \( k_{W_2}(\mu, \nu) \) is a positive definite kernel.

Multiplication of positive definite kernels results again in a positive definite kernel, hence we arrive at the automatic relevance determination kernel based on Wasserstein-2 distances:

\[
k_{W_2}([\mu_d]_{d=1}^D, [\nu_d]_{d=1}^D) = \sigma^2 \exp \left( \sum_{d=1}^D -\frac{W_2^2(\mu_d, \nu_d)}{l^2_d} \right)
\]

**Wasserstein Distance between Gaussian distributions** The Wasserstein-2 distance between two multivariate Gaussian distributions \( \mathcal{N}(m_1, \Sigma_1) \) and \( \mathcal{N}(m_2, \Sigma_2) \) has been shown by Dowson and Landau (1982) to have the following form:

\[
\| m_1 - m_2 \|_2^2 + \text{Tr} \left[ \Sigma_1 + \Sigma_2 - 2 \left( \Sigma_1^{1/2} \Sigma_2 \Sigma_1^{1/2} \right)^{1/2} \right]
\]

2.3 Sparse Gaussian Processes

We denote the output vector \( Y \), where each entry \( Y_i \) is a noisy observation of the function \( F(x_i) \) for all input points \( X = (x_i)_{i=1}^n \). We place a \( GP(m, k) \) prior on the stochastic function \( F \). We introduce inducing points \( Z = (z_i)_{i=1}^m \) with inducing point function values \( U = (u_i)_{i=1}^m \). Under standard Gaussian identities we have

\[
p(Y|F) = \mathcal{N}(Y|F, \beta),
\]

\[
p(F|U; X, Z) = \mathcal{N}(F|K_{nm}^E, K_{nn}^{E^{-1}} u, K_{nn}^E - K_{nm}^E K_{mn}^{E^{-1}} K_{nn}^E; X, Z)
\]

\[
p(U; Z) = \mathcal{N}(U|0, K_{nn}^E),
\]
We define \( K^E \) as the standard radial basis function with automatic relevance determination seen in GP literature (Rasmussen, 2004):

\[
k^E_{i,j}([x_{i,d}]_{d=1}^D, [x_{j,d}]_{d=1}^D) = \sigma^2 \exp \left( -\frac{(x_{i,d} - x_{j,d})^2}{l_d^2} \right) \quad (9)
\]

By the definition of a sparse GP, the joint density is

\[
p(Y, F, U) = p(F|U)p(U) \prod_{i=1}^n p(Y_i|F_i) \quad (10)
\]

We follow the variational inference framework introduced in Hensman et al. (2013) and maximize the lower bound on the marginal likelihood

\[
L = E_{q(F, U)}[\log \frac{p(Y, F, U)}{q(F, U)}] \quad (11)
\]

where the variational posterior is chosen as \( q(F, U) = p(F|U; X, Z)q(U) \), where \( q(U) = \mathcal{N}(U|m, S) \). Here, \( m \) and \( S \) are free variational parameters. Due to the Gaussian nature of both terms we can marginalize \( U \) to arrive at

\[
q(F) = \int p(F|U)q(U) = \mathcal{N}(F|\hat{U}, \hat{\Sigma}) \quad (12)
\]

where

\[
\hat{U} = K_{n,m}K^{-1}_{mm}m \\
\hat{\Sigma} = K_{nn} - K_{nm}K^{-1}_{mm}[K_{mm} - S]K^{-1}_{mm}K_{mn} \quad (13, 14)
\]

Then the evidence lower bound (ELBO) can be rewritten as

\[
L = E_{q(F)}(\log p(Y|F)) - E_q(U) \left( \log \frac{q(U)}{p(U)} \right) \quad (15)
\]

2.4 Disentangling Uncertainties in Gaussian Processes

As noted by Hensman et al. (2017), the variational posterior over function values in equation 12 can be divided into two components:

\[
f(\cdot) = h(\cdot) + g(\cdot) \quad (16)
\]

\[
h(\cdot) = \mathcal{N}(h(0|K_{nn} - K_{nm}K^{-1}_{mm}K_{mn})) \\
g(\cdot) = \mathcal{N}(g|K_{nm}K^{-1}_{mm}m, K_{nm}K^{-1}_{mm}SK^{-1}_{mm}K_{mn}) \quad (17, 18)
\]

The \( g(\cdot) \) component represents the within data distribution variance as given by the current number of inducing points after taking into account aleatoric uncertainty, which is given by the noise term in the likelihood. In this paper it will be denoted as within-data uncertainty as it is given by the basis functions that control the shape of the mean, thereby it could be interpreted as the variance stemming from the parametric side of the GP. Malinin and Gales (2018) have introduced the concept of “distributional uncertainty” in the field of Bayesian deep learning. Succinctly, it is defined as the uncertainty pertaining to the mismatch between the data manifold at training and testing time. This concept is similar to the role played by \( h(\cdot) \) which captures the shift from within to outside the data manifold as it is given by the current number of inducing points.

For a given multivariate normal \( \mathcal{N}(M, V) \) with \( V \) being a diagonal covariance matrix, the differential entropy is defined as \( 0.5 \times \sum_{d=1}^D \log V_d \), with \( V_d \) being the variance associated to the \( d \)-th dimension. This has the effect to measure the overall uncertainty in logit space, with higher values representing more uncertainty.

3 RELIABLE PROPAGATION OF DISTRIBUTIONAL VARIANCE IN ZERO-MEAN DGP

Ustyuzhaninov et al. (2019) have proved for noiseless data that the various stochastic layers in the hierarchy of a DGP collapse to deterministic transformations. In this section, we analyze what are the necessary theoretical requirements for a zero-mean DGP to maintain comparatively higher variance for an outlier data point throughout the hierarchy.
We consider the posterior DGP as a composition of functions \( q_L = q_L \circ \ldots \circ q_1 \) with a certain \( q_l \) being given by:

\[
q(F_l) = \int p(F_l|U_l)q(U_l) = N(F|\tilde{U}_l, \Sigma_l)
\]

where \( \tilde{U}_l \) and \( \Sigma_l \) are as in equations 13 and 14.

We consider the case of a two layer DGP, with the forthcoming analysis being easily extended to more layers.

Following the layout of a framework introduced in Girard et al. (2002) for obtaining Gaussian approximations of GPs with uncertain inputs, we have the following adapted case:

\[
q(F_2) = \int q(F_2|F_1)q(F_1(x))dF_1
\]

The above integral is analytically intractable. By applying the Law of Iterated Expectations and the Law of Conditional Variance we arrive at:

\[
m(q(F_2)) = E_{F_1}[E_{q(F_2)}q(F_2|F_1)] \tag{21}
\]

\[
v(q(F_2)) = E_{F_1}[V_{q(F_2)}q(F_2|F_1) + V_{q(F_2)}[E_{F_1}q(F_2|F_1)]] \tag{22}
\]

which after additional derivations detailed in the supplementary material arrives at:

\[
m(F_2) = \tilde{U}_2(\tilde{U}_1(x)) \tag{23}
\]

\[
v(F_2) = \tilde{\Sigma}_2(\tilde{U}_1(x)) + \Sigma_1(x) \ast \left[ \frac{1}{2} \ast \left( \frac{\partial^2 \tilde{\Sigma}_2(F_1)}{\partial F_1} \right)^2 + \left( \frac{\partial \tilde{U}_2(F_1)}{\partial F_1} \right)^2 \right] \tag{24}
\]

In Ustyuzhaninov et al. (2019), the authors have empirically proven that \( \lim_{M \to \infty} \frac{\partial^2 \tilde{\Sigma}_2(F_1)}{\partial F_1} = 0 \), hence that term is dropped in subsequent analysis. Intuitively, the term follows a sinusoidal path with the amplitude converging to 0 as the number of inducing points is increased.

We consider a data point \( x_{out} \) in input space, which is sufficiently far away from the data training manifold such that we have \( \Sigma_1(x_{out}) = \sigma^2 \) and \( \tilde{U}_1(x_{out}) = 0 \). Moreover, we also take \( x_{in} \) to be a high density point with regards to the data training manifold, with \( \Sigma_1(x_{in}) = V_{in} \) and \( \tilde{U}_1(x_{out}) = M_{in} \). We mention that \( M_{in} \) can be thought as a point close to 0.0 and \( V_{in} \leq \sigma^2 \), by assumption that the variances of the inducing points have been reduced compared to their prior. We assume that the inducing point locations for the second layer \( Z_2 \) are equidistantly placed in the interval \([-2\sigma, 2\sigma]\) and the variance of the inducing point values are taken to be equal to simplify the problem.

For \( x_{out} \) and \( x_{in} \) using equation 24 we obtain:

\[
v(F_2(x_{out})) = \tilde{\Sigma}_2(0) + \sigma^2 \ast \left( \frac{\partial \tilde{U}_2(F_1)}{\partial F_1} \right)^2 \bigg|_{F_1=0} \tag{25}
\]

\[
v(F_2(x_{in})) = \tilde{\Sigma}_2(M_{in}) + V_{in} \ast \left( \frac{\partial \tilde{U}_2(F_1)}{\partial F_1} \right)^2 \bigg|_{F_1=M_{in}} \tag{26}
\]

The desired behaviour for the second layer is \( v(F_2(x_{in})) \leq v(F_2(x_{out})) \), more explicitly:

\[
\tilde{\Sigma}_2(M_{in}) + V_{in} \ast \left( \frac{\partial \tilde{U}_2(x)}{\partial x} \right)^2 \bigg|_{F_1=M_{in}} \leq \tilde{\Sigma}_2(0) + \sigma^2 \ast \left( \frac{\partial \tilde{U}_2(x)}{\partial x} \right)^2 \bigg|_{F_1=0} \tag{27}
\]

Since we have assumed that the inducing point locations are equidistantly placed in a grid with centre zero, which is a sufficiently similar situation to what one might observe in practice, then \( \tilde{\Sigma}_2(0) \approx \tilde{\Sigma}_2(M_{in}) \) as we have also assumed that the posterior variances of the inducing point values are equal.
Hence the inequality \( v(F_2(x_{in})) \leq v(F_2(x_{out})) \) holds if and only if:

\[
\left( \frac{\partial^2 \tilde{U}_2(F_1)}{\partial F_1^2} \right)_{F_1=M_{in}}^2 \leq \frac{\sigma^2}{V_{in}}
\]  

(28)

Intuitively, this inequality can be easily satisfied if \( \left( \frac{\partial^2 \tilde{U}_2(F_1)}{\partial F_1^2} \right)_{F_1=0} \) reaches its peak at \( F_1 = 0 \). However, this requirement is not guaranteed to be satisfied in DGP in practice. In the supplementary material we show how to impose this constraint on a two layer DGP with the aid of variational inducing point first order derivatives.

4 HIERARCHICAL GAUSSIAN PROCESSES WITH WASSERSTEIN-2 KERNELS

In this section we present the two models introduced in this paper, which can be seen as the equivalents of DKL and DGP in Wasserstein-2 space.

4.1 Distributional Gaussian Processes with Wasserstein-2 Sequential Embeddings

A Deep Gaussian Process (Damianou and Lawrence, 2013) is defined as a stack of shallow GPs acting as the prior:

\[
p(y) = p(Y|F_L) \prod_{l=1}^{L} p(F_l|F_{l-1}, U_l; Z_{l-1})p(U_l)
\]  

(29)

where for brevity of notation we denote \( Z_0 = X \). The sparse GPs between hidden layers are treated as being noiseless. As the prior is analytically intractable to integrate, Salimbeni and Deisenroth (2017) have suggested to sample from each hidden layer of the DGP in order to obtain unbiased stochastic gradients.

Here, we diverge from this approach by treating all layers except the last one as a set of deterministic operations that provide the support for the Distributional GP making predictions.

We obtain the following joint density prior:

\[
p(Y|F_L) p(F_L|U_L; Z_{L-1}, F_{L-1}) \prod_{l=1}^{L} p(U_l)
\]  

(30)

where \( p(F_L|U_L; Z_{L-1}, F_{L-1}) \) is given by equation 7 with Wasserstein-2 kernels instead.

We would like to highlight that for \( 1 \leq l \leq L - 1 \), \( F_l \) act as features for the next kernel as opposed to random variables that need to be integrated out. We have included pseudocode for these steps in the supplementary material.

We introduce a factorized posterior between layers and dimensions of the following form:

\[
q(F_L, \{U_l\}_{l=1}^{L}) = p(F_L|U_L; Z_{L-1}) \prod_{l=1}^{L} q(U_l)
\]  

(31)

where \( q(U_l) \) is taken to be a multivariate Gaussian with mean \( m_l \) and variance \( S_l \). Using Jensen’s inequality we arrive at the evidence lower bound:

\[
L = \mathbb{E}_{q(F_L, \{U_l\}_{l=1}^{L})} \frac{p(Y, F_L, \{U_l\}_{l=1}^{L})}{q(F_L, \{U_l\}_{l=1}^{L})}
\]  

(32)

which after basic mathematical operations we arrive at the ELBO:

\[
L = \mathbb{E}_{q(F_L, \{U_l\}_{l=1}^{L})} p(Y|F_L) - \sum_{l=1}^{L} KL(q(U_l)|p(U_l))
\]  

(33)
4.2 Distributional Deep Gaussian Processes

One potential drawback with operating in the Wasserstein-2 space of Gaussian measures in the "deep kernel" formulation of Distributional Gaussian Processes with Wasserstein-2 Sequential Embeddings (DistGP-WE) is that the predictive variances might rather act as a feature more than their probabilistic interpretation, which is the propagation of uncertainty. We consider the case where the uncertainty component is explicitly imposed in the architecture. To this end, in this subsection we introduce hybrid kernels on both Wasserstein-2 and Euclidean space within the general framework of DGPs.

We introduce the hybrid kernel:

$$k_{hybrid}(\mu_i, \mu_j) = k^E(x_i, x_j) \exp \left( \sum_{d=1}^{D} \frac{W_2^2(\mu_i,d,\mu_j,d)}{l_d^2} \right)$$

where \(x_i, x_j\) are samples from Gaussian measures \(\mu_i, \mu_j\).

By combining the kernels stemming from Euclidean and Wasserstein-2 space, we explicitly impose on the predictive variances their original role which is uncertainty propagation, while still preserving the added flexibility of operating in the Wasserstein-2 space of Gaussian measures.

Under this new type of kernel we have the following joint density prior for the Distributional Deep Gaussian Processes (DDGP):

$$p(Y, \{F_l, U_l\}_{l=1}^L) = p(Y | F_L) \prod_{l=2}^L p(F_l | F_{l-1}, U_l; Z_{l-1}) p(U_l) \ p(F_l | U_l; X) p(U_l)$$

We mention that as opposed to the joint density prior of DistGP-WE, which has just \(F_L\) as a latent variable, thereby being the probabilistic equivalent of DKL (Wilson et al., 2016), DDGPs are DGPs as every layer is a latent variable that needs to be integrated out. Moreover, for \(2 \leq l \leq L\), \(Z_l\) is given by a multivariate Gaussian with a diagonal covariance matrix.

We arrive at the following lower evidence bound for DDGPs:

$$L = E_{q(F_L,U_L)} p(Y | F_L) - \sum_{l=1}^L KL(q(U_l) | p(U_l))$$

4.3 Propagation of outlier status in Distributional Deep Gaussian Processes

In this subsection, we reiterate the analysis done in section 3 but for DDGPs. We assume the same notation as in the aforementioned section. We assume that the inducing point mean locations for the second layer \(Z_2\) are equidistantly placed in the interval \([-2\sigma, 2\sigma]\) and we take the variance of the locations to be \(Z_{var} \leq \sigma^2\).

The desired behaviour for the second layer is \(v(F_2(x_{in})) \leq v(F_2(x_{out}))\), more explicitly:

$$\Sigma_2(M_{in}) + V_{in} \left( \frac{\partial \tilde{U}_2(F_1)}{\partial F_1} \right)^2 \bigg|_{F_1=M_{in}} \leq \Sigma_2(0) + \sigma^2 \left( \frac{\partial \tilde{U}_1(F_1)}{\partial F_1} \right)^2 \bigg|_{F_1=0}$$

In stark contrast to zero mean DGP, where under our current set of assumptions we get \(\Sigma_2(M_{in}) \approx \Sigma_2(0)\), this is not the case for DDGPs. Expanding on the variance terms \(\Sigma_2(M_{in})\) and \(\Sigma_2(0)\) looking at their distributional uncertainty:

$$\sigma^2 - K_{M_{in},M_{in}} K_{M_{in},M_{in}}^{-1} K_{M_{in}} \leq \sigma^2 - K_{M_{in},M_{in}}^{-1} K_{M_{in}}$$

To simplify the analysis we can consider just one inducing point, lying halfway between 0 and \(M_{in}\) at \(Z_m\) and with \(Z_{var} \ll \sigma^2\). After some additional operations detailed in the supplementary material we arrive at:

$$W_2^2(N(0, \sigma^2), N(Z_m, Z_{var})) \geq W_2^2(N(M_{in}, Var_{in}), N(Z_m, Z_{var}))$$
Figure 1: Distributional Uncertainty of DGP trained on a toy dataset with data between -5.0 and 5.0 with a varying number of inducing points. X axis denotes input space.

Figure 2: Decomposition of uncertainty into within-data (parametric) and distributional (non-parametric). Violin plots represent the variational distributions of the inducing points, with the x-axis denoting the location, and the y-axis the mean value of the inducing point. The x axis is taken to be the space of the previous layer.

The above inequality holds, as $\sigma^2 \gg Z_{var}$ and $Var_{in} \approx Z_{var}$. Consequently, in opposition to DGPs which require a specific pattern of first order derivatives surrounding 0, DDGPs rely less on this constraint as the above inequality is already providing a significant difference in distributional variance between outlier and inlier points. We provide the intuition behind why inducing point locations’ variance terms tend to converge to the within-data distribution variance (18) of inlier points in the supplementary material.

5 COLLAPSE OF DISTRIBUTIONAL VARIANCE IN DGP

Duvenaud et al. (2014) highlighted a certain pathology when stacking GPs, respectively the increasingly more non-injective mappings as the network increases in number of layers. In this subsection we highlight another pathology of DGPs with zero-mean functions, namely, the progressive collapse of distributional uncertainty for larger number of inducing points. This erroneous behaviour is also present in relatively small networks with just two hidden layers (Figure 1).

To understand what is causing this pathology, we take a simple case study of a DGP with two hidden layers trained on a toy regression dataset (Figure 2). Taking a clear outlier in input space, say the data point situated at -7.5, it gets correctly identified as an outlier in the mapping from input space to hidden layer space as given by its distributional variance. However, its outlier property dissipates in the next layer after sampling, as it gets mapped to regions where points inside the data manifold get confidently mapped between -2.5 and 2.5 in hidden layer space. Consequently, what was initially correctly identified as an outlier will now have its final distributional uncertainty close to zero. Adding further layers, will only compound this pathology.

6 PCA MEAN FUNCTIONS AND DISTRIBUTIONAL UNCERTAINTY

In this subsection we investigate the effect that PCA mean function has on distributional uncertainty. For this, we use the “banana” dataset and train our proposed models with 2 hidden layers and 10 inducing points. The DGP distributional uncertainty is guided by the direction of the PCA projection of the data, thereby widely extrapolating low uncertainty regions to spatial locations where there is no data. For DistGP-WE we can observe that the PCA projection does not
influence distributional variance in the first hidden layers and only slightly in the second. On the other hand, DDGPs with PCA mean functions reliably notices the dataset shift in hidden layers and in output space (Figure 3).

7 EXPERIMENTS & RESULTS

We evaluate our models on a range of regression and classification benchmark tasks from the UCI machine learning dataset repository. All of the experiments were run with the same initializations, with the goal of comparison with DKL and DGP for 2, respectively 5 layers. We used 100 inducing points at each layer, with 5 hidden units. Further details are given in the supplementary material.

Figure 3: Layer-wise moments of specified models trained on the “banana” dataset.

Figure 4: Left: Results on UCI regression datasets. Right: Results on UCI classification datasets, alongside MNIST and Fashion-MNIST. All subplots: 10 different runs of each model with different initialization seeds are taken into the composition of each boxplot. Higher values indicate better model fit.
Intuitively, the variance terms optimized for the inducing point locations will converge towards relatively small values, so as to obtain reliable inducing point location samples for the Euclidean part. With this in mind, the Wasserstein-2 framework solves this problem as the Wasserstein-2 RBF kernel explicitly takes into account the variance. Inituitively, even if the mean component of the Gaussian measure space is dominated by the PCA projection, which could coincide with the projection values of the training set, if the variance component is higher then the Wasserstein-2 kernel will collapse to zero if the data is noiseless (Ustyuzhaninov et al., 2019).

DGP are not able to detect the shift in the data manifold, whereas our models exhibit a steady increase in distributional differential entropy as deformations are increased. In contrast, a hard to classify digit would instead have low distributional differential entropy and high predictive entropy.

DCGP are not able to detect the shift in the data manifold, whereas our models exhibit a steady increase in distributional differential entropy (Figure 5). Additional results on Out-of-Distribution detection are provided in the supplementary material.

7.1 Outlier detection on image data

In this subsection we adapt DDGP and DGP-WE to the architecture introduced in Deep Convolutional GP (DCGP) (Blomqvist et al., 2018), obtaining the Distributional Deep Convolutional GP (DDCGP), respectively the Convolutionally Wasserstein-2 Embedded Distributional GP (Conv-WE-DistGP). The goal is to investigate their outlier detection capabilities on real data. MorphoMNIST (Castro et al., 2019) enables the systematic deformation of MNIST digits using morphological operations. We use swelling deformations with a strength of 3 and increasing radius from 3 to 14. As MorphoMNIST digits lie outside of the data manifold, we expect to see an increase in both predictive entropy and distributional differential entropy as deformations are increased. In contrast, a hard to classify digit would instead have low distributional differential entropy and high predictive entropy.

DGP rely just on the variance term from the inducing points in the next layer. This leads to the deactivation of \( h(\cdot) \) in all layers, except the first one. Hence, the DGP relies just on the variance term from the \( g(\cdot) \) (17) component from the first hidden layer onwards. This in turn collapses to zero if the data is noiseless (Ustyuzhaninov et al., 2019).

The PCA based projections stemming from the training data as introduced in Salimbeni and Deisenroth (2017) help to solve these issues. However, these also come with a high cost, as the PCA projections will wildly extrapolate the regions of low distributional variance in directions perpendicular to the eigenvectors. The proposed DDGP/DistGP-WE framework solves this problem as the Wasserstein-2 RBF kernel explicitly takes into account the variance. Intuitively, even if the mean component of the Gaussian measure space is dominated by the PCA projection, which could coincide with the projection values of the training set, if the variance component is higher then the Wasserstein-2 kernel will highlight this dataset shift with high distributional variance. DDGP explicitly propagates uncertainty, as we are sampling from Gaussian measures stemming from the hidden layers to compute the Euclidean part of the hybrid kernel. Intuitively, the variance terms optimized for the inducing point locations will converge towards relatively small values, so as to obtain reliable inducing point location samples for the Euclidean part. With this in mind, the Wasserstein-2 distances between a previously identified outlier will be higher compared to an inlier point. In the case of DistGP-WE,
the uncertainty gets implicitly propagated through the deterministic hierarchy up to the latent space that the final Distributional GP is operating on. An empirical example for this phenomenon is provided in the supplementary material.

We have shown under a set of assumptions that distributional variance can be correctly propagated under the framework introduced in Salimbeni and Deisenroth (2017). However, the requirement that the first derivative predictive mean function attains maximum values at the point where previously identified outliers get mapped is not guaranteed in practice. However, this requirement can be hard coded into a DGP by explicitly imposing high values for the first order derivative of the GP at the desired location. Toy example can be found in the supplementary material, leaving this as an avenue for further research.

Lastly, a drawback of the method resides in the lack of an improvement in test time log likelihood on small to medium-scale datasets using either DDGP or DGP-WE. However, on large-scale datasets the aforementioned models outperformed their Euclidean counterparts. One possible explanation is that for complex datasets, the explicit inclusion of the variance term in computing the kernel can increase the representational capacity of network.

**Conclusion** We have empirically shown that hierarchical GP models with Wasserstein-2 kernels are better suited for outlier detection on both toy and real datasets, while also showing improvements at testing time in large-scale regression and classification datasets.

**References**


1 Supplementary Material

1.1 EXPERIMENTS DETAILS

For all datasets we randomly selected 20% as the testing set, with the remainder being used for training. All implementations use the RBF kernel with automatic relevance determination for the DGP/DKL models, whereas for the DDGP/DGP-WE models the standard RBF kernel on Euclidean space is used just for the first layer, with the remaining layers using the Hybrid/Wasserstein-2 kernel. In terms of model architecture, we use 5 hidden units for each hidden level, with 100 inducing inputs. All models are optimized until convergence with a mini-batch of size 32. Results are provided for 2 and 5 layers.

For all the UCI experiments (including MNIST and Fashion-MNIST) we have used the following approach:

Euclidean Kernels All Euclidean space kernels used the standard ARD RBF, using a lengthscale parameter per input dimension, initialized to 1.351. We initialize the variance of the kernel with the same value.

Wasserstein Kernels All Wasserstein space kernels used the Wasserstein-2 distance ARD RBF introduced in the paper, using a lengthscale parameter per input dimension, initialized to the 1.351. We initialize the variance of the kernel with the same value.

Euclidean inducing points For all experiments we used 100 inducing points for the first layer, which is mapping from input space to first hidden layer space. We initialize the inducing point locations to the k-means of the training data. For the Gaussian variational approximations of the inducing point values we initialized the mean by uniformly sampling between -2.0 and 2.0. The Cholesky decomposition of the variance was initialized to be the diagonal matrix multiplied by $1e^{-5}$.

Wasserstein inducing points For all experiments we used 100 inducing points in the hidden layers of the multi-layered GP architectures. Since these inducing point are operating in Wasserstein space, the inducing point locations are actually distributions in this case. We initialize the mean of the inducing points locations by uniformly sampling between -2.0 and 2.0. The initial variance of the inducing point locations is taken to be the diagonal matrix multiplied by $1e^{-1}$. This low initialization of the variance term was chosen so that the DDGP/DGP-WE will implicitly learn to produce low variance terms at all hidden layers for data points inside the data manifold. A detailed discussion on this topic is provided in a subsequent section. For variational approximations of the inducing point values, we take the same initialization as for their Euclidean equivalents, with the sole mention that for the last layer of DDGP/DGP-WE, the Cholesky decomposition of the variance term of the inducing point values is taken to be a diagonal matrix of ones.

Optimization on UCI datasets All parameters were optimized using the Adam optimizer with a learning rate of 0.001. We used a batch size of 32 and trained for 10,000 iterations or until convergence.

Likelihood For regression tasks the likelihood variance was initialized to 1.0

1.2 ADDITIONAL DERIVATIONS

1.2.1 Proof of positive definitiveness of generalized radial basis kernels with Wasserstein distances

Thi Thien Trang et al. (2019) provides a detailed proof that a generalized radial basis function kernel with Wasserstein distances is positive definite. We recall here the main results.

Schoenberg’s Theorem. Let $F: \mathbb{R}^+ \Rightarrow \mathbb{R}^+$ be strictly monotonically decreasing, and $K$ a negative definite kernel. Then $(x,y) \Rightarrow F(K(x,y))$ is a positive definite kernel.

Proposition. The function $W_2^{2H}$ is a negative definite kernel if and only if $0 < H \leq 1$.

A complete proof of this proposition can be found in Kolouri et al. (2016), where they prove this result for absolutely continuous distributions in $W_2(\Omega)$, where $\Omega \subset \mathbb{R}$ is a compact subset.

The following theorem follows from Schoenberg’s Theorem and the above proposition.

Theorem 2 Let $k_W: W_2(\mathbb{R}) \times W_2(\mathbb{R}) \rightarrow \mathbb{R}$ be the Wasserstein-2 RBF kernel defined as following:

$$k_W(\mu, \nu) = \sigma^2 \exp \left( -\frac{W_2^{2H}(\mu, \nu)}{l^2} \right),$$

(1)
then $k_{W^2}(\mu, \nu)$ is a positive definite kernel for $0 < H \leq 1$.
Throughout this paper we take $H = 1.0$. For each dimension of our hidden layer in the DGP we use a different $k_{1V}$ kernel. Due to the fact that multiplication of positive definite kernels results again in a positive definite kernel, we arrive at the Automatic Relevance Determination kernel based on Wasserstein-2 distances:

$$k_{W^2}([\mu_d]_{d=1}^D, [\nu_d]_{d=1}^D) = \sigma^2 \exp \sum_{d=0}^D \frac{-W^2_H (\mu_d, \nu_d)}{\sigma^2}$$  \hspace{1cm} (2)

### 1.2.2 Additional derivation for section 3

We consider the case of a two layer DGP, with the forthcoming analysis being easily extended to more layers.
Following the layout of a framework introduced in Girard et al. (2002) for obtaining Gaussian approximations of GPs with uncertain inputs, we have the following adapted case:

$$q(F_2)(x) = \int q(F_2|F_1; \bar{U}_2, \bar{\Sigma}_2)q(F_1; \bar{U}_1(x), \bar{\Sigma}_1(x))dF_1$$  \hspace{1cm} (3)

The above integral is analytically intractable, with our current aim is to obtain an analytical Gaussian approximation. By applying the Law of Iterated Expectations and the Law of Conditional Variance we arrive at:

$$m(F_2) = \mathbb{E}_{F_1} \mathbb{E}_{q(F_2)} q(f_2|F_1; \bar{U}_2, \bar{\Sigma}_2) = \mathbb{E}_{F_1} \bar{U}_2(F_1)$$  \hspace{1cm} (4)

$$v(F_2) = \mathbb{E}_{F_1} [\text{var}_{q(F_2)} q(F_2|F_1; \bar{U}_2, \bar{\Sigma}_2)] + \text{var}_{q(F_2)}[\mathbb{E}_{F_1} q(F_2|F_1; \bar{U}_2, \bar{\Sigma}_2)]$$  

$$= \mathbb{E}_{F_1} \bar{\Sigma}_2(F_1) + \text{var}_{F_1} [\mathbb{E}_{q(F_2)} \bar{U}_2(F_1)]$$  \hspace{1cm} (5)

The Gaussian approximation framework continues by using a first order Taylor expansion around $\bar{U}_1(x)$ giving raise to:

$$\bar{U}_2(F_1) = \bar{U}_2(\bar{U}_1(x)) + \frac{\partial \bar{U}_2(F_1)}{\partial F_1}(F_1 - \bar{U}_1(x))$$  \hspace{1cm} (6)

which by taking the expectation w.r.t. to $\mathbb{E}_{F_1}$ results in

$$\mathbb{E}_{F_1} \bar{U}_2(F_1) = \bar{U}_2(\bar{U}_1(x))$$  \hspace{1cm} (7)

For the variance term we require the second order Taylor expansion $\bar{\Sigma}_1(x)$ giving raise to:

$$\bar{\Sigma}_2(F_1) = \bar{\Sigma}_2(\bar{U}_1) + \frac{\partial^2 \bar{\Sigma}_2(F_1)}{\partial F_1^2}(F_1 - \bar{U}_1(x)) + \frac{1}{2} \frac{\partial^2 \bar{\Sigma}_2(F_1)}{\partial F_1^2}(F_1 - \bar{U}_1(x))^2$$  \hspace{1cm} (8)

After some manipulations we arrive at:

$$v(F_2) = \bar{\Sigma}_2(\bar{U}_1(x)) + \bar{\Sigma}_1(x) * \left[ \frac{1}{2} \frac{\partial^2 \bar{U}_2(F_1)}{\partial F_1^2} + \left( \frac{\partial \bar{U}_2(F_1)}{\partial F_1} \right)^2 \right]$$  \hspace{1cm} (9)

### 1.2.3 Additional derivation for subsection 4.3

The desired behaviour for the second layer is $v(F_2(x_{in})) \leq v(F_2(x_{out}))$, more explicitly:

$$\bar{\Sigma}_2(M_{in}) + V_{in} \left( \frac{\partial \bar{U}_2(F_1)}{\partial F_1} \right)^2 \bigg|_{F_1 = M_{in}} \leq \bar{\Sigma}_2(0) + \sigma^2 \left( \frac{\partial \bar{U}_1(F_1)}{\partial F_1} \right)^2 \bigg|_{F_1 = 0}$$  \hspace{1cm} (10)
In stark contrast to zero mean DGP, where under our current set of assumptions we get $\tilde{\Sigma}^2(m_{in}) \approx \tilde{\Sigma}^2(0)$, this is not the case for DDGPs. Expanding on the variance terms $\Sigma^2(m_{in})$ and $\Sigma^2(0)$ looking at their distributional uncertainty:

$$\sigma^2 - K_{m_{in}m}K_{m_{in}m}^{-1}K_{m_{in}0} \leq \sigma^2 - K_{0m}K_{m_{in}m}^{-1}K_{m_{in}0} \quad (11)$$

To simplify the analysis we can consider just one inducing point, lying halfway between 0 and $m_{in}$ at $Z_m$ and with $Z_{var} << \sigma^2$. Using the hybrid kernel formulation introduced in equation 2 and re-arranging the terms in the above equation considering the matrices are scalars we arrive at:

$$\sigma^2 \exp \frac{-W^2_2(N(0, \sigma^2), N(Z_m, Z_{var}))}{l^2} * \exp \frac{-(0 - Z_m)^2}{l^2} \geq \sigma^2 \exp \frac{-W^2_2(N(m_{in}, var_{in}), N(Z_m, Z_{var}))}{l^2} * \exp \frac{(m_{in} - Z_m)^2}{l^2} \quad (12)$$

which after eliminating redundant terms and taking the logarithm we obtain:

$$W^2_2(N(0, \sigma^2), N(Z_m, Z_{var})) + (0 - Z_m)^2 \geq W^2_2(N(m_{in}, var_{in}), N(Z_m, Z_{var})) + (m_{in} - Z_m)^2 \quad (13)$$

### 1.3 DEEP GAUSSIAN PROCESSES WITH VARIATIONAL FIRST ORDER DERIVATIVES

This section comes as an additional material with regards to the issues highlighted in the discussion section. Further to our analysis on the theoretical requirements necessary to propagate high uncertainty for outlier point we subsequently introduce a method by which we artificially impose the requirement.

Differentiation is a linear operator, hence the first order derivative of a GP is still a GP. Previous work on using derivatives has mostly focused on observed derivatives Solak et al. (2003); Eriksson et al. (2018). We refer the reader to McHutchon (2013) for an in-depth technical report on differentiating GPs.

In this subsection we assume the same notation as in the background section. We have the following prior over $F_l$ and $F'_l$:

$$p(F_l, F'_l) = \mathcal{N} \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K_{ff} & K_{f'f'} \\ K_{f'f} & K_{f'f'} \end{pmatrix} \right]$$

where by $F'_l$ is the first derivative of the $l$-th layer in a DGP.

We denote by $Z'_l$ the inducing point locations for the first order derivatives of the $l$-th layer and by $m_i$ and $S_i$ the variational mean and variance of the inducing point values of the first order derivatives.

This DGP with Variational First Order Derivatives has the following prior:

$$p(y) = p(Y|F_l) \prod_{l=1}^{L} p(F_l|F_{l-1}, U_l, U'_l; Z_{l-1}, Z'_{l-1}) p(U_l) p(U'_l) \quad (14)$$

where $p(F_l|F_{l-1}, U_l, U'_l; Z_{l-1}, Z'_{l-1})$ has the following moments:

$$m(F_l) = K_{nm}^{mixed}(F_{l-1})K_{m0}^{mixed-L}U_{L}^{mixed}$$

$$\nu(F_l) = K_{nm}(F_{l-1}) - K_{nm}^{mixed}(F_{l-1})K_{m0}^{mixed-L}K_{mn}^{mixed}(F_{l-1})$$

where we have:
where we have the following individual elements in the matrices $K_{nm}^{\text{mixed}}$ and $K_{mm}^{\text{mixed}}$:

\[
K_{i'j} = \frac{\sigma^2}{l_d^2} (x_{i,d} - x_{j,d}) \exp \sum_{d=1}^{D} -\frac{(x_{i,d} - x_{j,d})^2}{l_d^2}
\]

\[
K_{i'j'} = \frac{\sigma^2}{l_d^4} (l_d^2 - (x_{i,d} - x_{j,d})^2) \exp \sum_{d=1}^{D} -\frac{(x_{i,d} - x_{j,d})^2}{l_d^2}
\]

which correspond to the first and second derivative of the standard squared exponential kernel taken with respect to the $d$-th dimension of the data.

Rest of the derivation up to computing the final loss functions is similar to the rationale developed in Salimbeni and Deisenroth (2017). The lower evidence bound for DGPs with Variational First Order Derivatives (DGP-VFOD) is as follow:

\[
L = \mathbb{E}_{q(F_l, U_l^{\text{mixed}})} \left[ \mathbb{P}(Y | F_L) - \sum_{l=1}^{L} KL(q(U_l^{\text{mixed}}) || p(U_l^{\text{mixed}})) \right]
\]

### 1.3.1 Explicitly imposing the first derivative requirement to propagate uncertainty

Having defined this new framework for imposing specific first order derivatives in the hierarchy of a DGP, we test the hypothesis that if $\left( \frac{\partial U_2(F_1)}{\partial F_1} \right)^2$ achieves its maximum at $F_1 = 0$ in the second hidden layer, then high variance stemming from the first layer will be correctly propagated. To this end, in the following experiment we use a varying number of inducing points for the observations, whose locations are constrained to be between the intervals [-5.0, -0.5] and [0.5, 5.0]. Furthermore, to artificially impose the aforementioned theoretical requirement, we use an additional 10 inducing points for derivatives. Their location is uniformly sampled and constrained to be between [-0.125, 0.125] and their variational mean values are taken to be the maximum of the optimized mean values of the inducing points pertaining to observations plus 2.0. Consequently, we ensure that $\left( \frac{\partial U_2(F_1)}{\partial F_1} \right)^2$ attains its maximum around 0.

From Figure 1 we can observe that in regions with no data, such as the intervals in input space [-15.0, -7.5] and [7.5, 15.0], for zero mean DGPs there is a shift between inlier and outlier distributional variance values, however the distributional uncertainty is less stable compared to zero mean DGP-VFOD, where we can observe that the space highlighted with higher values is fully compact.

**Remark** This experiments can be naturally extended to more than two layer DGPs. However, the artificial imputation of first order derivative values for layers higher in the hierarchy becomes more difficult as we would not have the simple case of doing it around 0, but surrounding the value where outliers from the second layer get mapped on average in the third layer. We leave this as an avenue for further research into correctly propagating high distributional variance in zero-mean DGPs.

### 1.4 IMPLICIT PROPAGATION OF UNCERTAINTY IN DISTGP-WE

DistGP-WE implicitly propagate uncertainty in the hidden layers, even though the variance terms are not explicitly used in their natural Bayesian formulation and are used rather as features for the Wasserstein-2 RBF kernel in the upcoming layer. This is achieved through the following uncertainty transmission mechanism.
Outlier points in input space are correctly highlighted with mean zero and variance equal to the variance of the kernel. Inside the data manifold points will have a comparatively lower variance term in the first hidden layer. During optimization, the variance term of the distributional inducing points from the second hidden layer will be forced to have low values to match the distribution of the within data manifold points, hence the Wasserstein-2 RBF kernel will correctly highlight outlier points also in the second hidden layer. We specifically encourage this implicit propagation of outlier status by initializing the distributional inducing points’ variance terms close to zero. Moving on from the second hidden layer to the third one, the same mechanism implicitly propagates high uncertainty in regions of input space which are outside the data manifold.

Figure 2 illustrates the implicit transmission of uncertainty mechanism. We can notice that green Gaussian distributions (corresponding to outlier points in input space) have higher variance compared to the dark red Gaussian distributions (corresponding to randomly sampled testing points). We can also notice that the orange Gaussian distributions (corresponding to the distributional inducing point locations) tend to have lower variance, hence better matching in distribution the testing points as opposed to the outlier points.

Pathological uncertainty propagation behaviour caused by wrong initialization  We would like to mention that initialization plays an important step in the correct propagation of uncertainty. We can take a wrong initialization scheme, such as starting with high variance for the distributional inducing points. What will happen in practice is that inside the data manifold points are going to have high within-data uncertainty so as to match the variance of the distributional inducing points. Generally, outside of the manifold points will have low within-data uncertainty but high distributional uncertainty. Thereby, moving from the second layer to the third layer, the Wasserstein-2 RBF kernel will not notice a dataset shift in the variance terms. Hence, outlier status will be lost from the third layer onwards.
Remark DDGs will not suffer from the same adverse effects as DGP-WE in case of wrong initialization. This is due to the fact that we are also sampling from the variational distributions associated to the distributional inducing points in order to compute the Euclidean part of the hybrid kernel. Hence, the variance of the distributional inducing points will converge towards relatively low values so as to obtain stable inducing point locations.

1.5 MEAN FUNCTIONS IN HIERARCHICAL GP WITH WASSERSTEIN-2 KERNELS

In this section we investigate the role that different mean functions have on the priors at each layer of the hierarchy and how these priors constrain the posterior functions.

1.5.1 Zero-Mean Function DistGP-WE

The first layer is a SGP given by $p(F_1) = \mathcal{N}(0, K_{F_0}F_0)$, hence the prior has the capability to capture correlations between data points. However, upon constructing the second layer, which is now based on distributional SGPs, we only take the diagonal of the $K_{F_0}F_0$ covariance matrix. Therefore, correlations between different data points are ignored as we arrive at the prior on $F_2$:

$$p(F_2) = \mathcal{N} \left[ \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_2^2 & \cdots & \sigma_2^2 \\ \vdots & \ddots & \vdots \\ \sigma_2^2 & \cdots & \sigma_2^2 \end{pmatrix} \right]$$

Consequently, this type of construction will always default to a prior given by a stochastic process with the amplitude given by the kernel variance of the second layers.
1.5.2 PCA mean function in DistGP-WE

One approach suggested in Salimbeni and Deisenroth (2017) to address the issue of non-injective DGPs was to add a linear mean function to the hidden layers of a DGP based on the first components of the SVD decomposition of the training data. In this paper we use the same approach but to a different end, namely for introducing correlations between data points in the hidden layers of the DistGP-WE.

We introduce the following notation:

$$λ_l(n) = K_{F_l,n,Z_l}K_{Z_l,Z_l}(U_l - PCA(Z_l))$$

$$U_l = 0$$

Hence, the first layer prior $p(F_1)$ has the following form:

$$\mathcal{N} \left( \begin{pmatrix} \lambda_1(1) + PCA(F_{0,1}) \\ \vdots \\ \lambda_1(n) + PCA(F_{0,n}) \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \cdots & \kappa_{1,N} \\ \vdots & \ddots & \vdots \\ \kappa_{N,N} & \cdots & \sigma_1^2 \end{pmatrix} \right)$$

Compared to the zero-mean function version introduced in the previous paragraph, we now have different mean functions for data points due to the PCA projections. This has the effect of introducing correlations between data points in all further hidden layers, such that the prior for the second layer $p(F_2)$ is defined as:
\[
\mathcal{N}
\left[
\begin{pmatrix}
\lambda_2(1) + \text{PCA}(F_{1,1}) \\
\vdots \\
\lambda_2(n) + \text{PCA}(F_{1,n})
\end{pmatrix},
\begin{pmatrix}
\sigma_2^2 & \cdots & K_{n,n} \\
\vdots & \ddots & \vdots \\
K_{n,1} & \cdots & \sigma_2^2
\end{pmatrix}
\right]
\]

where

\[
K_{i,j} = k_{W_2}(\mathcal{N}(\lambda_1(i) + \text{PCA}(F_{0,i}), \sigma_1^2), \mathcal{N}(\lambda_1(j) + \text{PCA}(F_{0,j}), \sigma_1^2))
\]

One can now notice that the correlations between data points are controlled by

\[
\| [\lambda_1(i) + \text{PCA}(F_{0,i})] - [\lambda_1(j) + \text{PCA}(F_{0,j})] \|^2_2
\]

Figure 3 illustrates that the layer-wise priors are structured, which in turn has the effect of smoothing the posterior functions at each layer. Moreover, the introduction of the PCA mean function for DistGP-WE manages to transform the prior from a white noise process to a structured one.

**Remark** For DDGP, the introduction of PCA mean functions does not influence the prior samples to the same extent, as its hybrid kernel operating in Euclidean and Wasserstein-2 space implies sampling at each layer, which introduces spatial correlations.

### 1.6 ADDITIONAL RESULTS ON OUT-OF-DISTRIBUTION DETECTION

**Deep Convolutional Gaussian Processes architecture** We use 250 inducing points at each layer. Every hidden layer is taken to have five channels. We use the Adam optimizer starting with a learning rate of 0.01 and decreasing by 0.1 every 20,000 iterations. We optimized all models trained on MNIST for 100,000 iterations.

#### 1.6.1 Outlier Detection on Morpho-MNIST

For better understanding the morphological changes occurring with swelling of different radii we provide some examples for the digit 4. We would like to remind the reader that Deep Convolutional GP operating in Euclidean space are not able to detect these changes via distributional uncertainty. Conversely, Distributional Deep Convolutional GP and Convolutionally Wasserstein-2 Embedded Distributional GP are able to obtain a steadily increasing distributional variance on these morphological changes.

#### 1.6.2 Outlier Detection on Rotated MNIST

To further assess the outlier detection capabilities of our methods, we employ the experiments introduced in Gal and Ghahramani (2016) by successively rotating digits from MNIST. We expect to see an increase in both entropy and distributional differential entropy as digits are rotated. For our experiment we rotate digits 6 and 8. When the digit is rotated by around 180 degrees the entropy and distributional differential entropy should revert back closer to initial levels, as it will resembles digits 9, respectively 8. From Figure 5 we can notice that in terms of distributional differential entropy the same patterns are exhibited by all models. However, DeepConvGP seem to indicate a shift between the original images and the images rotated by 180 degrees in terms of predictive entropy. This shift is not observed for the other models.

![Figure 4: Examples of increasing swelling applied to MNIST images.](image-url)
Figure 5: Entropy and Distributional Differential Entropy as digits 6 and 8 are rotated.

Figure 6: Histograms of distributional differential entropy computed on testing set of MNIST and NotMNIST, respectively SVHN. Higher values indicated an increased uncertainty.

1.6.3 Outlier detection on Different Digit Fonts

In these experiments we assess the capacity of our model to detect dataset shift by training it on MNIST and looking at the uncertainty measures computed on the testing set of MNIST and the entire NotMNIST and SVHN datasets. The hypothesis is that we ought to see higher differential entropy for distributional uncertainty for the digits stemming from a wide array of fonts present in NotMNIST as none of the fonts are handwritten. The same pattern is expected to occur also on SVHN, as these are stree house numbers.

From Figure 6 we can notice that DistDeepConvGP and Conv-WE-DistGP manage to better separate MNIST from NotMNIST compared to DeepConvGP. In terms of detecting digits from SVHN, DistDeepConvGP and Conv-WE-DistGP exhibit non-overlapping distributional differential entropy histograms, whereas DeepConvGP has a slight overlap with MNIST digits.
References


