Editorial: Special Issue on the Optimization Models and Algorithms for Data Science

Panos Parpas, Danny Ralph and Wolfram Wiesemann

In the modern information age, data and decisions are more strongly linked than ever before. The recent emergence of ‘data science’ as an interdisciplinary field that aims to distil insights from data and that informs data-driven decision-making bears witness to the importance of rigorous, evidence-based decision-making that crosses the boundaries of statistics, computer science, machine learning and optimization. This special issue focuses on some of the recent data science-related advances in optimization theory. The contributions to this special issue include optimization models and algorithms that deal with large data sets, as well as applications from business analytics and machine learning that have emerged in the last decades as data accumulates from multiple sources.

In the first paper (*Max-Norm Optimization for Robust Matrix Recovery*), Ethan X. Fang, Han Liu, Kim-Chuan Toh and Wen-Xin Zhou develop a new estimator for the matrix completion problem under arbitrary sampling schemes. The matrix completion problem aims to reconstruct an unknown matrix given information from a small number of noise-contaminated entries. Current methods assume that the indices of the observed entries follow a uniform distribution. This assumption is restrictive in many applications, including the well-known Netflix problem. To address this issue, the authors propose an estimator that incorporates both a max-norm and nuclear-norm regularization and solve the resulting model using ADMM (Alternating Direction Method of Multipliers). Through numerical experiments with real datasets, the authors show improvements over other estimators without a significant increase in computational effort.

In the second paper (*Primal and Dual Predicted Decrease Approximation Methods*), Amir Beck, Edouard Pauwels and Shoham Sabach develop the notion of ‘predicted decrease approximation’ to provide a unified convergence analysis for various existing algorithms for constrained convex optimization. To this end, the authors consider the problem of minimizing smooth functions over compact convex sets using linear oracles. Methods based on linear oracles perform inexpensive computations at each iteration and are widely used in data science applications. The authors unify the convergence analysis of generalized conditional gradient, proximal gradient, greedy coordinate descent for separable constraints and working set methods for linear equality constraints with bounds. The analysis is performed using the concept of ‘predicted decrease’ that ensures that an algorithmic step is ‘at least as good’ as a conditional gradient step. The authors show that the dual application of this approach leads to primal-dual convergence guarantees that hold even if the primal model is only partially strongly convex. Their numerical experiments illustrate the performance of the different algorithms in various applications from data science.

In the third paper (*Near-Optimal Stochastic Approximation for Online Principal Component Estimation*), Chris Junchi Li, Mengdi Wang, Han Liu and Tong Zhang study the challenging problem of online algorithms for estimating the principal component by processing streaming data. The authors formulate this problem as a non-convex stochastic optimization problem. They then

propose to analyze the online PCA problem as iterations of stochastic approximation. Using this viewpoint, the authors prove nearly optimal finite-sample error bounds for the online PCA problem with sub-Gaussian samples.

In the fourth paper (*Consistent Learning by Composite Proximal Thresholding*), Patrick L. Combettes, Saverio Salzo and Silvia Villa address the fundamental data science problem of extracting information from observations using a linear model. They propose an optimization model that incorporates regularization and that can take advantage of prior information such as sparsity and hard constraints. This is based on inexact computation of the proximity operators of regularizers which feeds into an implementable proximal forward-backward thresholding gradient algorithm. The authors show that solutions obtained by minimizing their model yield consistent estimators.

In the fifth paper (*Interpreting Latent Variables in Factor Models via Convex Optimization*), Armeen Taeb and Venkat Chandrasekaran address the difficult issue of attributing meaning to latent variables, ζ , which are used to describe statistical relationships between the measured variables, y ; and in particular, how to associate a low rank relationship between y and additional (measured) covariates, x , that captures as much of the behaviour of the latent variables as possible. The main results are statistically consistent estimates of the underlying model parameters and associated probabilistic bounds that show the interplay between regularization parameters and the number of observations. A two-step algorithm for this embeds convex optimization in a grid search over possible model parameters to identify the ranks of various matrix outputs. Numerical examples are used to show the selection of the most relevant factors from lists of candidate covariates.

In the sixth paper (*Wait-and-Judge Scenario Optimization*), Marco C. Campi and Simone Garatti consider convex optimization problems with probabilistically described constraints, and they study the generalization level of the associated scenario optimization problems which only contain a sample of the constraints. Rather than relating the generalization level of the scenario solution to the number of optimization variables, as has been done in the past, the paper relates the generalization level to the number of support constraints of the solution to the scenario optimization problem. This new approach leads to generalization bounds that largely outperform those carried out based on the number of optimization variables.

In the final paper (*Data-Driven Inverse Optimization with Incomplete Information*), Peyman M. Esfahani, Soroosh Shafieezadeh-Abadeh, Grani A. Hanasusanto and Daniel Kuhn study inverse optimization problems where an observer learns the preferences of an agent solving a parametric optimization problem that depends on an exogenous signal. The authors assume that the observer has imperfect information, that the agent suffers from bounded rationality or implementation errors, or that the observer is affected by measurement noise. The resulting inverse optimization problem is formulated as a distributionally robust optimization program that offers rigorous out-of-sample guarantees and that can be exactly reformulated as (or safely approximated by) tractable convex programs. Numerical tests show that the proposed approach frequently outperforms the state-of-the-art approaches in out-of-sample results.

We would like to thank the authors as well as the referees for their invaluable contributions to this issue.