Statistical Optimization in High Dimensions

Huan Xu
National University of Singapore
mpexuh@nus.edu.sg

Constantine Caramanis
University of Texas at Austin
cmcaram@ece.utexas.edu

Shie Mannor
Technion, Israel
shie@ee.technion.ac.il

Abstract

We consider optimization problems whose parameters are known only approximately, based on a noisy sample. Of particular interest is the high-dimensional regime, where the number of samples is roughly equal to the dimensionality of the problem, and the noise magnitude may greatly exceed the magnitude of the signal itself. This setup falls far outside the traditional scope of Robust and Stochastic optimization. We propose three algorithms to address this setting, combining ideas from statistics, machine learning, and robust optimization. In the important case where noise artificially increases the dimensionality of the parameters, we show that combining robust optimization and dimensionality reduction can result in high-quality solutions at greatly reduced computational cost.

1 Introduction

Optimization has become a corner stone of machine learning research and practice. Indeed, the machine learning community has benefited from theory (in particular convex duality, e.g. [7, 10]), algorithm (e.g., [11, 6]), and software [8, 13], of optimization. On the other hand, insights and algorithms from machine learning have yet to make a significant impact on optimization. This paper pursues precisely this avenue, harnessing recent advances in high-dimensional statistics.

We consider solving an optimization problem where its parameters are known only through potentially noisy observation. Many problems fall under this general setting, particularly as optimization is increasingly used to deal with large-scale problems with data-driven constraints. A large class of such problems arises from user satisfaction tasks, where an objective is maximized subject to the constraints of keeping as many users' perceived performance above a threshold, as possible. User preferences are typically observed through very noisy processes, such as user surveys or collaborative filtering, and while typically soft constraints, are often modeled as hard constraints in optimization problems. Many problems in engineering share similar qualities. Of particular relevance is the vast family of problems where the system behaviors, and hence optimization constraints, are only learned via observation through many noisy or potentially unreliable sensors. Environmental monitoring, multiple-object tracking, and related problems all fall under this general umbrella. This paper attempts to bring to the table tools from statistics and machine learning, to study precisely this problem: how can we approach an optimization problem whose constraints are highly corrupted.

Optimization with noisy or corrupted parameters traditionally falls under the purview of stochastic and robust optimization [1, 3, 5]. Consequently, techniques from both fields of optimization have seen significant impact in statistics and machine learning [12]. On the other hand, the focus of machine learning on over-fitting, and the arsenal of tools developed, have not seen commensurate influence on optimization. Indeed, robust optimization takes an uncertainty set as a primitive, essentially overlooking the issue of data altogether; stochastic optimization often assumes (partial)

knowledge of the distribution (e.g., the distribution itself, or perhaps some of its moments), and thus has not explored issues of sample complexity to the degree done in machine learning.

In this paper, we consider optimization under uncertainty, in the data-driven and *high dimensional regime* where our knowledge of the constraint parameters comes from samples, the dimensionality of the problem and hence the noise is very high, and hence the magnitude of the noise can greatly exceed the magnitude of the true parameters. Ignoring issues of overfitting and dimensionality in such a setting can present potentially catastrophic consequences for both the solution of the problem, as well as computational complexity. Reversing the typical arrows of influence, we leverage results from statistics and machine learning, to inform optimization.

2 Problem Setup

The general problem we consider is the following: we wish to solve the convex problem

$$\begin{aligned} \text{Minimize:}_{\mathbf{x}\in\mathcal{X}} \quad & f_0(\mathbf{x}) \\ \text{Subject to:} \quad & f(\mathbf{x}, \mathbf{a}_i) \leq 0, \quad i = 1, \dots m. \end{aligned}$$

where $\mathcal{X}$ is a known convex feasible set representing structural constraints, f and f_0 are convex, but where the parameters $\{\mathbf{a}_i\}$ are known only through noisy samples, hence representing data-driven constraints. That is, we observe $\{\tilde{\mathbf{a}}_i\}_{i=1}^m$, generated according to $\tilde{\mathbf{a}}_i = \mathbf{a}_i + \mathbf{n}_i$, where $\mathbf{a}_i$ are unknown parameters, and $\mathbf{n}_i$ are iid Gaussian noise $\mathcal{N}(0, \sigma^2 I)$. We are particularly interested in the high-dimensional regime where the dimensionality, p, is approximately equal to m.

We focus on the case of linear optimization, and without loss of generality, consider only uncertain constraints: $\mathbf{a}_i^\top \mathbf{x} \leq b_i$. For $\mathbf{n}_i \sim \mathcal{N}(0, \sigma^2 I)$, $\|\mathbf{n}_i\| = \Theta(\sqrt{p}\sigma)$, hence the magnitude of the corrupting noise may dwarf the magnitude of the true parameter.

Given this setting, estimating or even approximating each true constraint parameter $\mathbf{a}_i$ is hopeless. The contribution of this paper is to show that nevertheless, there is a way forward. We propose three distinct formulations that approximate this problem. We give bounds on the performance of each. Our third formulation, is geared to the setting where the true parameters $\{\mathbf{a}_i\}$ lie in a low-dimensional space, but this special structure is obscured by the added noise. In this case, our approach combines robust optimization and dimensionality reduction, and provides drastic improvements in computation time.

The first formulation, which we call the *nominal method*, takes a (surprisingly) naive approach: it simply replaces the unknown true parameter with its noisy observation. Thus, one solves

$$\textbf{Nominal Method:} \quad \begin{cases} \text{Minimize:}_{\mathbf{x}\in\mathcal{X}} & \mathbf{c}^\top \mathbf{x} \\ \text{Subject to:} & \tilde{\mathbf{a}}_i^\top \mathbf{x} \leq b_i, \quad i = 1, \dots m. \end{cases} \tag{1}$$

We show that the optimal solution, $\mathbf{x}_o^*$, to the nominal method will not violate the majority of the true constraints with a large gap and hence is already a reasonable candidate solution. Note that under this guarantee, it is still possible that $\mathbf{x}_o^*$ violates most or all constraints, with a small gap. Thus, if the decision maker is less sensitive to the gap of the constraint violation, but instead cares more about the number of constraints satisfied, the nominal method may not be appropriate.

The second formulation, which we call the *robust method*, borrows an idea borrowed from *robust optimization* [2, 4, 14] to address exactly this setup. The basic idea is since $\tilde{\mathbf{a}}_i$ is a noisy copy of the true parameter, we require the constraint to hold for all parameters "close" to $\tilde{\mathbf{a}}_i$. This leads to the following formulation for fixed $\gamma > 0$.

$$\textbf{Robust Method:} \quad \begin{cases} \text{Minimize:}_{\mathbf{x}\in\mathcal{X}} & \mathbf{c}^\top \mathbf{x} \\ \text{Subject to:} & (\tilde{\mathbf{a}}_i + \boldsymbol{\delta}_i)^\top \mathbf{x} \leq b_i, \, \forall \|\boldsymbol{\delta}_i\|_2 \leq \gamma, \quad i = 1, \dots m. \end{cases} \tag{2}$$

Note that larger γ leads to a solution that violates fewer constraints, at the cost of being more conservative. Interestingly, while the noise satisfies $\|\mathbf{n}_i\|_2 = \Theta(\sqrt{p}\sigma)$, we show that it is sufficient to pick $\gamma = \Theta(\sigma)$ to guarantee that the *majority of constraints are satisfied*. That is, by protecting against order-wise smaller protection, the robust method significantly improves the feasibility of the solution, even though the true parameters is not "close" to the observed parameter.

The third method focuses on the setting where the true parameters $\mathbf{a}_1, \dots, \mathbf{a}_m$ lie on a d-dimensional subspace where $d \ll p$. We call this the *dimensionality reduction method*. We first perform Principal

Component Analysis (PCA) [9], and let $\mathbf{w}_1^*, \ldots, \mathbf{w}_d^*$ be the d principal components of $\tilde{\mathbf{a}}_1, \ldots, \tilde{\mathbf{a}}_m$. Next we project $\tilde{\mathbf{a}}_i$ onto the span of $\mathbf{w}_1^*, \ldots, \mathbf{w}_d^*$, denoting the projection by $\hat{\mathbf{a}}_i$. Then we solve the following Robust Optimization problem.

$$\textbf{PCA Method:} \quad \begin{cases} \text{Minimize:}_{\mathbf{x}\in\mathcal{X}} & \mathbf{c}^\top \mathbf{x} \\ \text{Subject to:} & (\hat{\mathbf{a}}_i + \boldsymbol{\delta}_i)^\top \mathbf{x} \le b_i, \quad \forall \|\boldsymbol{\delta}_i\|_2 \le \gamma; \quad i = 1, \ldots, m. \end{cases} \tag{3}$$

The main advantage of this formulation is computational: by reducing the dimensionality, the computational cost is reduced compared to the robust method.

Our work diverges in an important way from the traditional setup of optimization under uncertainty (e.g., [4, 5]). The classical setup (high-dimensionality and noise magnitude aside) assume we observe parameters $\mathbf{a}_i$, but then the solution $\mathbf{x}^*$ is judged against perturbed parameters $\mathbf{a}_i + \mathbf{n}_i$, thus rendering the solution *independent* of the noise. We find this to be a poor model of reality, where noise could potentially skew the solution itself, not just degrade its performance. Indeed, in our setting, in all methods presented, the solution is *dependent* on the noise. In terms of the analysis, it is this fact that presents the main technical challenges.

3 Technical Guarantees

In this section we provide technical guarantees for the three methods mentioned. Due to space constraints, all proofs are omitted. We first show that the optimal solution to the nominal problem, $\mathbf{x}_o^*$, satisfies the following property: the number of constraints that are violated with a large gap is small.

Theorem 1. *Let $\mathbf{x}_o^*$ be an optimal solution to the nominal method, i.e., Formulation (1). Then with probability at least $1-\theta$, for any $c \in \mathbb{R}^+$, the following holds:*[1]

$$\frac{1}{m}\sum \mathbf{1}(\mathbf{a}_i^\top \mathbf{x}_o^* > b_i + c) \le \frac{\sigma\|\mathbf{x}_o^*\|_2(1+\sqrt{\tau}+\sqrt{-2\log\theta/m})}{c}.$$

While Theorem 1 bounds the magnitude of the constraint violation, it is still possible that the solution of the nominal method violates every constraint (maybe slightly). In contrast, we next show that the solution of the robust method is guaranteed to satisfy most of the constraints.

Theorem 2. *Fix $\gamma > 0$. Let $\mathbf{x}_r^*$ be an optimal solution to Formulation (2). Then we have with probability at least $1-\theta$,*

$$\frac{1}{m}\sum_{i=1}^{m} \mathbf{1}(\mathbf{a}_i^\top \mathbf{x}_r^* > b_i) \le \frac{\sigma(1+\sqrt{\tau}+\sqrt{-2\log\theta/m})}{\gamma}.$$

Besides feasibility, conservatism of the solution is an equally important property of a formulation. We next quantifies the conservatism of the robust approach. Specifically we consider a solution to the following problem assuming that $\mathbf{a}_i$ are indeed known,

$$\begin{array}{ll} \text{Minimize:} & \mathbf{c}^\top \mathbf{x} \\ \text{Subject to:} & \sup\limits_{\|\boldsymbol{\delta}_i\|_2 \le \tilde{\gamma}} (\mathbf{a}_i + \boldsymbol{\delta}_i)^\top \mathbf{x} \le b_i; \quad i = 1, \ldots, m. \end{array} \tag{4}$$

Hence Formulation (4) can be regarded as an ideal formulation with an additional conservatism $\tilde{\gamma}$. The next theorem shows that a solution to Formulation (4) satisfies the majority of constraints of the robust approach, and hence the latter is not overly conservative.

Theorem 3. *Suppose $\overline{\gamma} > \gamma$, Let $\overline{\mathbf{x}}$ be the optimal solution to Problem (4), then with probability $1-\theta$, we have*

$$\frac{1}{m}\sum_{i=1}^{m} \mathbf{1}(\sup_{\|\boldsymbol{\delta}_i\|_2 \le \gamma} (\tilde{\mathbf{a}}_i + \boldsymbol{\delta}_i)^\top \overline{\mathbf{x}} > b_i) \le 1 - \Phi((\overline{\gamma}-\gamma)/\sigma) + \sqrt{\frac{-\log\theta}{2m}}.$$

[1] Here and in the sequel, unless otherwise stated, the probability is taken over random realizations of the observations.

If the true parameters $\mathbf{a}_1, \ldots, \mathbf{a}_m$ belong to a low-dimensional subspace, one can perform PCA to approximately recover this space together with the parameters, and solve an optimization problem based on the approximated parameters $\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_m$. We now analyze the performance of this dimensionality-reduction based algorithm.

Theorem 4. *Let $\mathbf{x}_d^*$ is the optimal solution to Formulation (3), then with probability $1-\theta$, we have*

$$\frac{1}{m}\sum_{i=1}^{m}(\mathbf{a}_i^\top \mathbf{x}_d^* > b_i) \leq 5\sqrt{d}(1+\sqrt{\tau}+\sqrt{-2\log\theta/m})\frac{\sigma\nu}{\gamma^2} + \frac{d\sigma^2(1+\sqrt{\tau}+\sqrt{-2\log\theta/m})^2}{\gamma^2}.$$

Suppose $(\mathbf{a}_1, b_1), \ldots, (\mathbf{a}_m, b_m)$ are indeed iid sampling of an unknown distribution μ supported on a d-dimensional subspace, then we can bound the probability that $\mathbf{x}_d^*$ violates a new constraint, randomly generated from the same distribution. We remark that bound only depends on the intrinsic dimensionality d.

Corollary 1. *Let $\mathbf{x}_d^*$ be the optimal solution to Formulation (3), then with probability $1-2\theta$, we have*

$$\begin{aligned} Pr_{(\mathbf{a},b)\sim\mu}(\mathbf{a}^\top \mathbf{x}_d^* > b) \leq & \sqrt{\frac{4}{m}\big(d+1\big)\ln(\frac{2em}{d+1}) + \ln(\frac{\delta}{4})\big)} \\ & +5\sqrt{d}(1+\sqrt{\tau}+\sqrt{-2\log\theta/m})\frac{\sigma\nu}{\gamma^2} + \frac{d\sigma^2(1+\sqrt{\tau}+\sqrt{-2\log\theta/m})^2}{\gamma^2}. \end{aligned} \quad (5)$$

We next investigate the conservatism of the dimensionality reduction approach.

Theorem 5. *Fix $\tilde{\gamma} > \gamma$ and let $\overline{\mathbf{x}}$ be the optimal solution to Formulation (4). Then the following holds with probability $1-\theta$:*

$$\begin{aligned} & \frac{1}{m}\sum_{i=1}^{m}\mathbf{1}\left(\sup_{\|\hat{\boldsymbol{\delta}}_i\|_2\leq\gamma}(\hat{\mathbf{a}}+\hat{\boldsymbol{\delta}}_i)^\top\overline{\mathbf{x}} > b_i\right) \\ \leq \quad & 5\sqrt{d}(1+\sqrt{\tau}+\sqrt{-2\log\theta/m})\frac{\sigma\nu}{(\tilde{\gamma}-\gamma)^2} + \frac{d\sigma^2(1+\sqrt{\tau}+\sqrt{-2\log\theta/m})^2}{(\tilde{\gamma}-\gamma)^2}. \end{aligned}$$

References

[1] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

[2] A. Ben-Tal and A. Nemirovski. Robust solutions of uncertain linear programs. *Operations Research Letters*, 25(1):1–13, August 1999.

[3] D. Bertsimas, D. B. Brown, and C. Caramanis. Theory and applications of robust optimization. To appear in *SIAM Review*, 2011.

[4] D. Bertsimas and M. Sim. The price of robustness. *Operations Research*, 52(1):35–53, January 2004.

[5] J. R. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer-Verlag, New York, 1997.

[6] J-F. Cai, E. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20:1956–1982, 2008.

[7] E. J. Candès and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

[8] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21. http://cvxr.com/cvx, 2011.

[9] I. T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics, Berlin: Springer, 1986.

[10] S. Shalev-Shwartz and Y. Singer. A primal-dual perspective of online learning algorithms. *Machine Learning*, 69(2-3):115–142, 2007.

[11] S. Shalev-Shwartz and N. Srebro. SVM optimization: Inverse dependence on training set size. In *Proceedings of the 22nd international conference on Machine learning*, 2008.

[12] S. Sra, S. Nowozin, and J. Wright. *Optimization in Machine Learning*. MIT Press, 2011.

[13] J.F. Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11–12:625–653, 1999. Special issue on Interior Point Methods (CD supplement with software).

[14] H. Xu, C. Caramanis, and S. Mannor. Robust regression and Lasso. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1801–1808, 2009.