Statistical Optimization in High Dimensions

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

Constantine Caramanis
Electrical and Computer Engineering
The University of Texas at Austin
caramanis@mail.utexas.edu

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

Abstract

We consider optimization problems whose parameters are known only approximately, based on noisy samples. 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 cornerstone of machine learning research and practice. Indeed, the machine learning community has benefited from theory (in particular convex duality, e.g. Candès & Tao, 2007; Tropp, 2006; Shalev-Shwartz & Singer, 2007), algorithms (e.g., Shalev-Shwartz & Srebro, 2008; Li & Zhang, 2009; Cai et al., 2008), and software (Grant & Boyd, 2011; Sturm, 1999), for optimization. On the other hand, insights and algorithms from machine learning have yet to make commensurate impact on optimization. This paper pursues precisely this avenue, harnessing recent advances in machine learning and high-dimensional statistics.

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

We consider solving an optimization problem where its parameters are known only through potentially noisy observations. 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, they 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 or noisy.

Optimization with noisy or corrupted parameters traditionally falls under the purview of stochastic and robust optimization (Prékopa, 1995; Ben-Tal et al., 2009; Bertsimas et al., 2011; Birge & Louveaux, 1997). Consequently, techniques from both fields of optimization have seen significant impact in statistics and machine learning (Ben-Tal et al., 2009). 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

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

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: $\mathbf{a}_i^\top \mathbf{x} \le b_i$. 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

Nominal Method:

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

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 from *robust optimization* (Ben-Tal & Nemirovski, 1999; Bertsimas & Sim, 2004; Xu et al., 2009a) 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 $\gamma > 0$.

Robust Method:

$$\text{Minimize:}_{\mathbf{x}\in\mathcal{X}}\ \mathbf{c}^\top\mathbf{x} \tag{2}$$
$$\text{S. t. : } (\tilde{\mathbf{a}}_i+\boldsymbol{\delta}_i)^\top\mathbf{x}\le b_i,\ \forall\|\boldsymbol{\delta}_i\|_2\le\gamma,\ i=1,\dots m.$$

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 perturbation, the robust method significantly improves the feasibility of the solution, even though the true parameters is not "close" to the observed parameter. Interestingly, the robust constraint , is equivalent to $\tilde{\mathbf{a}}_i^\top\mathbf{x}+\gamma\|\mathbf{x}\|_2\le b_i$ which is a constraint with a regularization term. The latter has been broadly applied in various machine learning algorithms. For example, if we consider a quadratic objective function, then the resulting robust method is indeed a variant of support vector machines (Xu et al., 2009b; Shivaswamy et al., 2006).

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) (Jolliffe, 1986), and let $\mathbf{w}_1^*,\dots,\mathbf{w}_d^*$ be the d principal components of $\tilde{\mathbf{a}}_1,\dots,\tilde{\mathbf{a}}_m$. Next we project $\tilde{\mathbf{a}}_i$ onto the span of $\mathbf{w}_1^*,\dots,\mathbf{w}_d^*$, denoting the projection by $\hat{\mathbf{a}}_i$. Then we solve the following Robust

Optimization problem.

PCA Method:

$$\text{Minimize:}_{\mathbf{x}\in\mathcal{X}}\ \mathbf{c}^\top\mathbf{x} \qquad (3)$$
$$\text{S. t. : } (\hat{\mathbf{a}}_i+\boldsymbol{\delta}_i)^\top\mathbf{x}\le b_i,\ \forall\|\boldsymbol{\delta}_i\|_2\le\gamma,\ i=1,\ldots,m.$$

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., Bertsimas & Sim, 2004; Birge & Louveaux, 1997). 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 A Technical Lemma

The centerpiece of our analysis is random matrix theory, and specifically an estimation of the largest singular value of random matrices, as shown in the following proposition. We define a quantity that we frequently use in the sequel: $\tau\triangleq\max(p/m,1)$.

Proposition 1. *Let $\mathbf{n}_i$ be iid following $\mathcal{N}(0,\sigma^2I_p)$. Then for any $\delta\in[0,1]$, with probability $1-\theta$ we have*

$$\sup_{\mathbf{w}\in\mathbb{R}^p,\|\mathbf{w}\|_2\le1}\tfrac{1}{m}\textstyle\sum_{i=1}^m(\mathbf{w}^\top\mathbf{n}_i)^2 \le\sigma^2(1+\sqrt{\tau}+\sqrt{-2\log\theta/m})^2.$$

Proof. We require the following lemma, which is essentially Theorem II.13 of Davidson and Szarek (2001).

Lemma 1. *Let Γ be an $m\times p$ matrix, whose entries are IID $\mathcal{N}(0,1)$. Denote $v_1=\min(m,p)$ and $v_2=\max(m,p)$. Let $s_1(\Gamma)$ be the largest singular value of Γ, then we have*

$$Pr(s_1(\Gamma)>\sqrt{v_1}+\sqrt{v_2}+\sqrt{v_2}\epsilon)\le\exp(-v_2\epsilon^2/2).$$

Note that $\sup_{\mathbf{w}\in\mathbb{R}^p,\|\mathbf{w}\|_2\le1}\frac{1}{m}\sum_{i=1}^m(\mathbf{w}^\top\mathbf{n}_i)^2 = \sigma^2[s_1(\Gamma)]^2/m$. The proposition thus holds. □

4 The Nominal Method

In this section we 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]

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

Proof. Proposition 1 implies that with probability $1-\theta$, the following holds uniformly over all $\mathbf{x}\in\mathbb{R}^p$,

$$\sum_{i=1}^m(\mathbf{x}^\top\mathbf{n}_i)^2\le m\|\mathbf{x}\|_2^2\sigma^2(1+\sqrt{\tau}+\sqrt{-2\log\theta/m})^2.$$

Since $\mathbf{x}_o^*$ is an optimal solution to Formulation (1),

$$\begin{aligned}
0 &= \sum_{i=1}^m\max\left(\mathbf{a}_i^\top\mathbf{x}_o^*-b_i,0\right)\\
&= \sum_{i=1}^m\max\left((\tilde{\mathbf{a}}_i-\mathbf{n}_i)^\top\mathbf{x}_o^*-b_i,0\right)\\
&\le \sum_{i=1}^m\max\left(\tilde{\mathbf{a}}_i^\top\mathbf{x}_o^*-b_i,0\right)+\sum_{i=1}^m|\mathbf{x}_o^{*\top}\mathbf{n}_i|\\
&\le \sqrt{m\sum_{i=1}^m(\mathbf{x}_o^{*\top}\mathbf{n}_i)^2}\\
&\le m\sigma\|\mathbf{x}_o^*\|_2(1+\sqrt{\tau}+\sqrt{-2\log\theta/m}),
\end{aligned}$$

which implies the theorem. □

5 The Robust Method

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, in this section we 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 $1-\theta$,*

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

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

Proof. Since $\mathbf{x}_r^*$ is an optimal solution to Formulation (2), then we have for $i = 1, \ldots, m$,

$$(\tilde{\mathbf{a}}_i + \boldsymbol{\delta}_i)^\top \mathbf{x}_r^* \le b_i, \ \forall \|\boldsymbol{\delta}_i\|_2 \le \gamma,$$

which leads to $\sum_{i=1}^m \max\left(\tilde{\mathbf{a}}_i^\top \mathbf{x}_r^* - b_i + \gamma\|\mathbf{x}_r^*\|_2, 0\right) = 0$. Note that for any i

$$\begin{aligned}
&\max\left(\mathbf{a}_i^\top \mathbf{x}_r^* - b_i, 0\right) \\
=&\max\left((\tilde{\mathbf{a}}_i - \mathbf{n}_i)^\top \mathbf{x}_r^* - b_i, 0\right) \\
\le&\max\left(\tilde{\mathbf{a}}_i^\top \mathbf{x}_r^* - b_i + \gamma\|\mathbf{x}_r^*\|_2, 0\right) \\
&\quad + \max(|\mathbf{x}_r^{*\top} \mathbf{n}_i| - \gamma\|\mathbf{x}_r^*\|_2, 0) \\
=&\max(|\mathbf{x}_r^{*\top} \mathbf{n}_i| - \gamma\|\mathbf{x}_r^*\|_2, 0).
\end{aligned}$$

Furthermore, similarly to the proof of Theorem 1, we have with probability $1 - \theta$

$$\sum_{i=1}^m |\mathbf{x}_r^{*\top} \mathbf{n}_i| \le m\sigma\|\mathbf{x}_r^*\|_2(1 + \sqrt{\tau} + \sqrt{-2\log\theta/m}).$$

Combining these, we have with probability at least $1 - \theta$,

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^m \mathbf{1}(\mathbf{a}_i^\top \mathbf{x} > b_i) &\le \frac{1}{m}\sum_{i=1}^m \mathbf{1}(|\mathbf{x}_r^{*\top} \mathbf{n}_i| > \gamma\|\mathbf{x}_r^*\|_2) \\
&\le \frac{\sigma(1 + \sqrt{\tau} + \sqrt{-2\log\theta/m})}{\gamma}.
\end{aligned}$$

This establishes the theorem. □

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

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

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*

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

Proof. First notice that $\overline{\mathbf{x}}$ does not depend on the noise $\mathbf{n}_1, \ldots, \mathbf{n}_m$. If $\overline{\mathbf{x}} = \mathbf{0}$, then claim trivially holds. Hence we assume $\overline{\mathbf{x}} \neq \mathbf{0}$, and let $\mathbf{w} \triangleq \overline{\mathbf{x}}/\|\overline{\mathbf{x}}\|_2$. Fix $i \in [1 : m]$, we have

$$\tilde{\mathbf{a}}_i^\top \overline{\mathbf{x}} = \mathbf{a}_i^\top \overline{\mathbf{x}} + \mathbf{n}_i^\top \overline{\mathbf{x}} = \mathbf{a}_i^\top \overline{\mathbf{x}} + (\mathbf{w}^\top \mathbf{n}_i)\mathbf{w}^\top \overline{\mathbf{x}}.$$

Since $\mathbf{w}$ is independent to $\mathbf{n}_i$, we have $\Pr(\mathbf{w}^\top \mathbf{n}_i > (\overline{\gamma} - \gamma)) = 1 - \Phi((\overline{\gamma} - \gamma)/\sigma)$. Notice that $\mathbf{1}(\mathbf{w}^\top \mathbf{n}_i > (\overline{\gamma} - \gamma))$ is a binomial random variable. By independence of $\{\mathbf{n}_j\}_{j=1}^m$,

$$\begin{aligned}
&\Pr\left\{\frac{1}{m}\sum_{i=1}^m \mathbf{1}(\mathbf{w}^\top \mathbf{n}_i \ge \overline{\gamma} - \gamma) \ge 1 - \Phi((\overline{\gamma} - \gamma)/\sigma) + \epsilon\right\} \\
&\le \exp(-2m\epsilon^2).
\end{aligned}$$

Equivalently, with probability $1 - \theta$

$$\frac{1}{m}\sum_{i=1}^m \mathbf{1}(\mathbf{w}^\top \mathbf{n}_i \ge (\overline{\gamma} - \gamma)) \le 1 - \Phi((\overline{\gamma} - \gamma)/\sigma) + \sqrt{\frac{-\log\theta}{2m}}.$$

Note that by definition $\sup_{\|\boldsymbol{\delta}_i\|_2 \le \tilde{\gamma}} (\mathbf{a}_i + \boldsymbol{\delta}_i)^\top \mathbf{x} \le b_i$. Hence the event $\{\sup_{\|\boldsymbol{\delta}_i\|_2 \le \gamma} (\tilde{\mathbf{a}}_i + \boldsymbol{\delta}_i)^\top \overline{\mathbf{x}} > b_i\}$ implies that $\{\mathbf{w}^\top \mathbf{n}_i \ge \overline{\gamma} - \gamma\}$. Hence we have Equation (5) holds. □

6 The Dimensionality Reduction Method

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 $\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_m$. In this section we analyze the performance of this dimensionality-reduction based algorithm. To lighten notations, we define the following: $C_0 \triangleq \text{conv}\{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$, $\hat{C} \triangleq \text{conv}\{\hat{\mathbf{a}}_1, \ldots, \hat{\mathbf{a}}_m\}$, $\hat{C}_\gamma \triangleq \{\mathbf{c} + \mathbf{b} | \mathbf{c} \in \hat{C}, \|\mathbf{b}\|_2 \le \gamma\}$. $\mathbb{P}_{\hat{\Omega}}(\cdot)$ is the projection onto the subspace spanned by $\mathbf{w}_1^*, \ldots, \mathbf{w}_d^*$, and $\mathbb{P}_{\Omega_0}(\cdot)$ is the projection onto the original subspace. Thus we have $\mathbb{P}_{\Omega_0}(\mathbf{a}_i) = \mathbf{a}_i$, and similarly $\mathbb{P}_{\hat{\Omega}}(\hat{\mathbf{a}}_i) = \hat{\mathbf{a}}_i$. Finally, define

$$\nu \triangleq \left(\frac{1}{m}\sum_{i=1}^m \|\mathbf{a}_i\|_2^2\right)^{1/2}.$$

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

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

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*

$$\begin{aligned} Pr_{(\mathbf{a},b)\sim\mu}(\mathbf{a}^\top \mathbf{x}_d^* > b) \leq & \sqrt{\frac{4}{m}\big((d+1)\ln(\frac{2em}{d+1}) + \ln(\frac{\delta}{4})\big)} \\ & + 5\sqrt{d}(1+\sqrt{\tau}+\sqrt{-2\log\theta/m})\frac{\sigma\nu}{\gamma^2} \\ & + \frac{d\sigma^2(1+\sqrt{\tau}+\sqrt{-2\log\theta/m})^2}{\gamma^2}. \end{aligned} \tag{6}$$

Proof. It is known (e.g.,Anthony & Bartlett, 1999; van der Vaart & Wellner, 2000) that the VC dimension of the indicator functions of d-dimensional half-spaces is $d+1$, i.e.,

$$\begin{aligned} & VC\left(\{f_{\mathbf{v},\mathbf{b}}(\cdot)|\mathbf{v}\in\mathbb{R}^d, b\in\mathbb{R}\}\right) = d+1; \\ & \text{where } f_{\mathbf{v},\mathbf{b}}(\mathbf{x}) \equiv \mathbf{1}(\mathbf{v}^\top\mathbf{x} > b); \quad \forall \mathbf{x}\in\mathbb{R}^d. \end{aligned}$$

Therefore, by VC theory, we have the following holds with probability (of sampling) at least $1-\theta$:

$$\begin{aligned} & \sup_{\mathbf{z}\in\mathbb{R}^p, \mathbb{P}_\Omega(\mathbf{z})=\mathbf{z}} \left\{\mathbb{E}_{(\mathbf{a},b)\sim\mu}\mathbf{1}(\mathbf{a}^\top\mathbf{z} > b) - \frac{1}{m}\sum_{i=1}^m \mathbf{1}(\mathbf{a}_i^\top\mathbf{z} > b_i)\right\} \\ & \leq \sqrt{\frac{4}{m}\big((d+1)\ln(\frac{2em}{d+1}) + \ln(\frac{\theta}{4})\big)}. \end{aligned}$$

Here we used the fact that $(\mathbf{a}_i, b_i)$ are iid follows μ, and that μ is supported on the d-dimensional. This implies the corollary. □

6.1 Proof of Theorem 4

We now prove Theorem 4. We first show that under certain condition $\mathbf{a}_i$ will be close to $\hat{\mathbf{a}}_i$, which indeed implies the feasibility.

Lemma 2. *Suppose for some $c\in[0,1]$ and $\alpha > 0$, we have*

$$\sum_{i=1}^m \|\mathbb{P}_{\hat{\Omega}}(\mathbf{a}_i)\|_2^2 \geq (1-c)\sum_{i=1}^m \|\mathbf{a}_i\|_2^2; \tag{7}$$

$$\sup_{\mathbf{w}:\,\|\mathbf{w}\|_2=1} \frac{1}{m}\sum_{i=1}^m (\mathbf{w}^\top\mathbf{n}_i)^2 \leq \alpha. \tag{8}$$

Then we have

$$\frac{1}{m}\sum_{i=1}^m \mathbf{1}(\mathbf{a}_i^\top\mathbf{x}_d^* > b) \leq \frac{c\sum_{i=1}^m \|\mathbf{a}_i\|_2^2 + md\alpha}{m\gamma^2}.$$

Proof. Recall that $\hat{\mathbf{a}}_i = \mathbb{P}_{\hat{\Omega}}(\tilde{\mathbf{a}}_i) = \mathbb{P}_{\hat{\Omega}}(\mathbf{a}_i + \mathbf{n}_i)$. Denote $\check{\mathbf{a}}_i = \mathbf{a}_i - \mathbb{P}_\Omega(\mathbf{a}_i)$, then Equation (7) leads to

$$\sum_{i=1}^m \|\check{\mathbf{a}}_i\|_2^2 \leq c\sum_{i=1}^m \|\mathbf{a}_i\|_2^2.$$

On the other hand, notice that $\mathbb{P}_{\hat{\Omega}}$ is a projection onto a d-dimensional subspace. Hence Equation (8) leads to

$$\sum_{i=1}^m \|\mathbb{P}_{\hat{\Omega}}(\mathbf{n}_i)\|_2^2 \leq d \sup_{\mathbf{w}:\,\|\mathbf{w}\|_2=1}\sum_{i=1}^m (\mathbf{w}^\top\mathbf{n}_i)^2 \leq dm\alpha.$$

Thus we have

$$\begin{aligned} & \frac{1}{m}\sum_{i=1}^m \mathbf{1}(\mathbf{a}_i^\top\mathbf{x} \leq b_i) \geq \frac{1}{m}\sum_{i=1}^m \mathbf{1}(\mathbf{a}_i \in \hat{C}_\gamma) \\ & \geq \frac{1}{m}\sum_{i=1}^m \mathbf{1}(\|\mathbf{a}_i - \hat{\mathbf{a}}_i\| \leq \gamma) \\ & \overset{(a)}{=} \frac{1}{m}\sum_{i=1}^m \mathbf{1}\left(\|\mathbb{P}_{\hat{\Omega}}(\mathbf{a}_i) - \hat{\mathbf{a}}_i\|_2^2 + \|\mathbf{a}_i - \mathbb{P}_{\hat{\Omega}}(\mathbf{a}_i)\|_2^2 \leq \gamma^2\right) \\ & = \frac{1}{m}\sum_{i=1}^m \mathbf{1}\left(\|\mathbb{P}_{\hat{\Omega}}(\mathbf{n}_i)\|_2^2 + \|\check{\mathbf{a}}_i\|_2^2 \leq \gamma^2\right) \\ & \overset{(b)}{\geq} 1 - \frac{\sum_{i=1}^m \left(\|\mathbb{P}_{\hat{\Omega}}(\mathbf{n}_i)\|_2^2 + \|\check{\mathbf{a}}_i\|_2^2\right)}{m\gamma^2} \\ & \geq 1 - \frac{c\sum_{i=1}^m \|\mathbf{a}_i\|_2^2 + md\alpha}{m\gamma^2}. \end{aligned}$$

Here, (a) holds because $\mathbb{P}_{\hat{\Omega}}(\mathbf{a}_i) - \hat{\mathbf{a}}_i$ and $\mathbf{a}_i - \mathbb{P}_{\hat{\Omega}}(\mathbf{a}_i)$ are orthogonal to each other; (b) follows from Markov inequality. The lemma thus holds. □

Thus, to establish Theorem 4, we only need to find c and α that satisfies Equation (7) and (8).

Lemma 3. *Suppose Equation (7) holds, and let $\beta \triangleq \frac{1}{m}\sum_{i=1}^m \|\mathbf{a}_i\|_2^2/(d\alpha)$. Then*

$$\sum_{i=1}^m \|\mathbb{P}_{\hat{\Omega}}(\mathbf{a}_i)\|_2^2 \geq (1 - 4\sqrt{1/\beta} - 1/\beta)\sum_{i=1}^m \|\mathbf{a}_i\|_2^2,$$

Proof. We have that

$$\begin{aligned} & \frac{1}{m}\sum_{i=1}^m \|\mathbb{P}_{\Omega_0}(\mathbf{a}_i + \mathbf{n}_i)\|_2^2 \\ = & \frac{1}{m}\sum_{i=1}^m \left\{\|\mathbf{a}_i\|_2^2 + \|\mathbb{P}_{\Omega_0}(\mathbf{n}_i)\|_2^2 + 2(\mathbf{a}_i)^\top(\mathbb{P}_{\Omega_0}(\mathbf{n}_i))\right\} \\ \overset{(a)}{\geq} & \frac{1}{m}\sum_{i=1}^m \|\mathbf{a}_i\|_2^2 - 2\sqrt{[\frac{1}{m}\sum_{i=1}^m \|\mathbf{a}_i\|_2^2][\frac{1}{m}\sum_{i=1}^m \|\mathbb{P}_{\Omega_0}(\mathbf{n}_i)\|_2^2]} \\ \overset{(b)}{\geq} & \frac{1}{m}\sum_{i=1}^m \|\mathbf{a}_i\|_2^2 - 2\sqrt{d\alpha\frac{1}{m}\sum_{i=1}^m \|\mathbf{a}_i\|_2^2}. \end{aligned}$$

Here, (a) follows from the inequality $(\sum_i a_i b_i)^2 \le \sum_i a_i^2 \sum_i b_i^2$, and (b) holds due to $\frac{1}{m}\sum_{i=1}^m \|\mathbb{P}_{\Omega_0}(\mathbf{n}_i)\|_2^2 \le d \sup_{\mathbf{w}\in\mathbb{R}^m} \frac{1}{m}\sum_{i=1}^m (\mathbf{w}^\top \mathbf{n}_i)^2 \le d\alpha$. Similarly we have

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^m \|\mathbb{P}_{\hat{\Omega}}(\mathbf{a}_i+\mathbf{n}_i)\|_2^2 \\
=&\frac{1}{m}\sum_{i=1}^m \left\{\|\mathbb{P}_{\hat{\Omega}}(\mathbf{a}_i)\|_2^2 + \|\mathbb{P}_{\hat{\Omega}}(\mathbf{n}_i)\|_2^2 + 2(\mathbf{a}_i)^\top(\mathbb{P}_{\Omega_0}(\mathbf{n}_i))\right\} \\
\le&\frac{1}{m}\sum_{i=1}^m \|\mathbb{P}_{\hat{\Omega}}(\mathbf{a}_i)\|_2^2 + d\alpha + 2\sqrt{d\alpha\frac{1}{m}\sum_{i=1}^m \|\mathbb{P}_{\hat{\Omega}}(\mathbf{a}_i)\|_2^2}.
\end{aligned}$$

Since by definition, $\frac{1}{m}\sum_{i=1}^m \|\mathbb{P}_{\hat{\Omega}}(\mathbf{a}_i+\mathbf{n}_i)\|_2^2 \ge \sum_{i=1}^m \|\mathbb{P}_{\Omega_0}(\mathbf{a}_i+\mathbf{n}_i)\|_2^2$, and $\beta = \frac{1}{m}\sum_{i=1}^m \|\mathbf{a}_i\|_2^2/(d\alpha)$, we have

$$\begin{aligned}
&(1-2\sqrt{1/\beta})(\frac{1}{m}\sum_{i=1}^m \|\mathbf{a}_i\|_2^2) \\
\le&\frac{1}{m}\sum_{i=1}^m \|\mathbf{a}_i\|_2^2 - 2\sqrt{d\alpha\frac{1}{m}\sum_{i=1}^m \|\mathbf{a}_i\|_2^2} \\
\le&\frac{1}{m}\sum_{i=1}^m \|\mathbb{P}_{\hat{\Omega}}(\mathbf{a}_i)\|_2^2 + d\alpha + 2\sqrt{d\alpha\frac{1}{m}\sum_{i=1}^m \|\mathbb{P}_{\hat{\Omega}}(\mathbf{a}_i)\|_2^2} \\
\le&\frac{1}{m}\sum_{i=1}^m \|\mathbb{P}_{\hat{\Omega}}(\mathbf{a}_i)\|_2^2 + (\frac{1}{m}\sum_{i=1}^m \|\mathbf{a}_i\|_2^2)/\beta \\
&\quad + 2\sqrt{1/\beta}(\frac{1}{m}\sum_{i=1}^m \|\mathbf{a}_i\|_2^2).
\end{aligned}$$

Re-arranging the terms establishes the lemma. □

Now, observing that the assumption in Equation (7) of Lemma 2 indeed follows from Proposition 1, we collect these pieces to prove Theorem 4.

Proof of Theorem 4. By Proposition 1, we have with probability $1-\theta$,

$$\sup_{\mathbf{w}\in\mathbb{R}^p, \|\mathbf{w}\|_2\le 1} \frac{1}{m}\sum_{i=1}^m (\mathbf{w}^\top \mathbf{n}_i)^2 \le \sigma^2(1+\sqrt{\tau}+\sqrt{-2\log\theta/m})^2.$$

Thus, let $\beta = \frac{1}{m}\sum_{i=1}^m \|\mathbf{a}_i\|_2^2/(d\sigma^2(1+\sqrt{\tau}+\sqrt{-2\log\theta/m})^2)$. By Lemma 3 we have

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^m \|\mathbb{P}_{\hat{\Omega}}(\mathbf{a}_i)\|_2^2 &\ge \frac{1}{m}(1-4\sqrt{1/\beta}-1/\beta)\sum_{i=1}^m \|\mathbf{a}_i\|_2^2 \\
&\ge \frac{1}{m}(1-5\sqrt{1/\beta})\sum_{i=1}^m \|\mathbf{a}_i\|_2^2.
\end{aligned}$$

Theorem 4 holds by applying Lemma 2 with $\alpha = \sigma^2(1+\sqrt{\tau}+\sqrt{-2\log\delta/m})^2$ and $c = 5\sqrt{1/\beta}$, and noting that $\beta = \nu^2/d\alpha$. □

6.2 Conservatism

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 $1-\theta$:*

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

Proof. We have that $\hat{\mathbf{a}}_i^\top\overline{\mathbf{x}} \le \mathbf{a}_i^\top\overline{\mathbf{x}} + \sup_{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|_2\le\|\mathbf{a}_i-\hat{\mathbf{a}}_i\|_2} \boldsymbol{\delta}^\top\overline{\mathbf{x}}$, and thus by optimality of $\overline{\mathbf{x}}$,

$$\{\|\mathbf{a}_i-\hat{\mathbf{a}}_i\|_2 \le \tilde{\gamma}-\gamma\} \implies \{\sup_{\|\boldsymbol{\delta}_i\|_2\le\gamma} (\hat{\mathbf{a}}_i+\hat{\boldsymbol{\delta}}_i)^\top\overline{\mathbf{x}} \le b_i\}.$$

This leads to

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^m \mathbf{1}\left(\sup_{\|\boldsymbol{\delta}_i\|_2\le\gamma} (\hat{\mathbf{a}}_i+\hat{\boldsymbol{\delta}}_i)^\top\overline{\mathbf{x}} > b_i\right) \\
\le\ & \frac{1}{m}\sum_{i=1}^m \mathbf{1}\left(\|\mathbf{a}_i-\hat{\mathbf{a}}_i\|_2 > \tilde{\gamma}-\gamma\right) \\
=\ & \frac{1}{m}\sum_{i=1}^m \mathbf{1}\left(\|\mathbf{a}_i-\hat{\mathbf{a}}_i\|_2^2 > (\tilde{\gamma}-\gamma)^2\right).
\end{aligned}$$

By Markov inequality we have the right-hand-side is upper-bounded by

$$\begin{aligned}
&\frac{\frac{1}{m}\sum_{i=1}^m \|\mathbf{a}_i-\hat{\mathbf{a}}_i\|_2^2}{(\tilde{\gamma}-\gamma)^2} \\
\overset{(a)}{=}\ & \frac{\frac{1}{m}\sum_{i=1}^m [\|\mathbf{a}_i-\mathbb{P}_{\hat{\Omega}}(\mathbf{a}_i)\|_2^2 + \|\mathbb{P}_{\hat{\Omega}}(\mathbf{n}_i)\|_2^2]}{(\tilde{\gamma}-\gamma)^2} \\
\overset{(b)}{=}\ & \frac{\frac{1}{m}\sum_{i=1}^m [\|\mathbf{a}_i\|_2^2 - \|\mathbb{P}_{\hat{\Omega}}(\mathbf{a}_i)\|_2^2] + \frac{1}{m}\sum_{i=1}^m \|\mathbb{P}_{\hat{\Omega}}(\mathbf{n}_i)\|_2^2}{(\tilde{\gamma}-\gamma)^2}.
\end{aligned}$$

Here, (a) follows from $\hat{\mathbf{a}}_i = \mathbb{P}_{\hat{\Omega}}(\mathbf{a}_i+\mathbf{n}_i)$, and the fact that $\mathbf{a}_i - \mathbb{P}_{\hat{\Omega}}(\mathbf{a}_i)$ and $\mathbb{P}_{\hat{\Omega}}(\mathbf{n}_i)$ are orthogonal; (b) follows from the the fact that $\mathbb{P}_{\hat{\Omega}}$ is an orthogonal projection. Suppose Equation (7) and (8) from Lemma 2 holds. Then follow a similar argument as in the proof of Lemma 2, the right-hand-side is upper bounded by

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^m \mathbf{1}\left(\sup_{\|\hat{\boldsymbol{\delta}}_i\|_2\le\gamma} (\hat{\mathbf{a}}+\hat{\boldsymbol{\delta}}_i)^\top\overline{\mathbf{x}} > b_i\right) \\
\le\ & \frac{c\sum_{i=1}^m \|\mathbf{a}_i\|_2^2 + md\alpha}{m(\tilde{\gamma}-\gamma)^2}.
\end{aligned}$$

Now we can apply Proposition 1 to bound α in Equation (8), and by Lemma 3 to bound c in Equation (7). The theorem thus follows. □

6.3 Low-dimensional computation

Suppose for this section that the set $\mathcal{X}$ is the entire space, i.e., there are no additional constraints besides the data-driven ones. (This assumption is needed to facilitate the analysis but can be ignored in practice by projecting the set to which the solution belongs to.) Formulation (3) is a Second Order Cone Program (SOCP) in $\mathbb{R}^p$, as the robust method, and can often be computationally expensive. However, in this section we show that solving Formulation (3) can be reduced to a SOCP in $\mathbb{R}^d$. This in turn provides significant computational advantage over the robust formulation.

Theorem 6. *Let $\hat{\mathbf{c}} \triangleq \mathbb{P}_{\hat{\Omega}}(\mathbf{c})$, and $\check{\mathbf{c}} \triangleq \mathbf{c} - \hat{\mathbf{c}}$. Let $\alpha \triangleq \|\check{\mathbf{c}}\|_2$. Then the optimal solution of Formulation (3) is given by $\mathbf{x}_d^* = \hat{\mathbf{x}}_d^* + \check{\mathbf{x}}_d^*$, where $\check{\mathbf{x}}_d^* = -r^*\check{\mathbf{c}}/\alpha$, and $(\hat{\mathbf{x}}_d^*, r^*)$ is the optimal solution to*

$$\begin{array}{rll} \textit{Minimize:}_{\hat{\mathbf{x}}, r\in\mathbb{R}} & \hat{\mathbf{c}}^\top\hat{\mathbf{x}} - r\alpha & (9)\\ \textit{Subject to:} & \hat{\mathbf{a}}_i^\top\hat{\mathbf{x}} + \gamma\sqrt{\|\hat{\mathbf{x}}\|_2 + r^2} \le b_i;\ \forall i; \\ & r \ge 0; \\ & \mathbb{P}_{\hat{\Omega}}(\hat{\mathbf{x}}) = \hat{\mathbf{x}}. \end{array}$$

Proof. Note that we can decompose any $\mathbf{x}$ into the sum of two parts, one belongs to $\mathbb{P}_{\hat{\Omega}}$, and the other one is orthogonal to $\mathbb{P}_{\hat{\Omega}}$. Thus, formulation (3) is equivalent to

$$\begin{array}{rl} \text{Minimize:}_{\hat{\mathbf{x}},\check{\mathbf{x}}} & \hat{\mathbf{c}}^\top\hat{\mathbf{x}} + \check{\mathbf{c}}^\top\check{\mathbf{x}} \\ \text{Subject to:} & \hat{\mathbf{a}}_i^\top\hat{\mathbf{x}} + \gamma\sqrt{\|\hat{\mathbf{x}}\|_2^2 + \|\check{\mathbf{x}}\|_2^2} \le b_i,\ \forall i; \\ & \mathbb{P}_{\hat{\Omega}}(\check{\mathbf{x}}) = 0; \\ & \mathbb{P}_{\hat{\Omega}}(\hat{\mathbf{x}}) = \hat{\mathbf{x}}. \end{array}$$

Introducing slack variable r, we have the following equivalent formulation

$$\begin{array}{rll} \text{Minimize:}_{\hat{\mathbf{x}},\check{\mathbf{x}}} & \hat{\mathbf{c}}^\top\hat{\mathbf{x}} + \check{\mathbf{c}}^\top\check{\mathbf{x}} & (10)\\ \text{Subject to:} & \hat{\mathbf{a}}_i^\top\hat{\mathbf{x}} + \gamma\sqrt{\|\hat{\mathbf{x}}\|_2 + r^2} \le b_i,\ \forall i; \\ & \|\check{\mathbf{x}}\| \le r; \\ & \mathbb{P}_{\hat{\Omega}}(\check{\mathbf{x}}) = 0; \\ & \mathbb{P}_{\hat{\Omega}}(\hat{\mathbf{x}}) = \hat{\mathbf{x}}. \end{array}$$

Notice that for any r and $\hat{\mathbf{x}}$, the corresponding optimal $\check{\mathbf{x}} = -r\check{\mathbf{c}}/\alpha$. Substituting this into Formulation (10) implies the theorem. □

Notice that all terms in Formulation (9) belong to the d-dimensional subspace $\mathbb{P}_{\hat{\Omega}}$, and hence can be represented by d-dimensional vectors. Thus, solving (9) is indeed solving an SOCP in $\mathbb{R}^d$.

7 Simulation

In this section we report some simulation results to illustrate the proposed methods.[2] We randomly generate a $m \times p$ constraint matrix with rank d, using Matlab command $randn(m,d) * randn(d,p)$. The cost vector also belongs to this d-dimensional subspace, and b_i is set as 1. We then perturb each entry of the cost vector and the constraint matrix by iid Gaussian noise $\mathcal{N}(0,\sigma^2)$. We compare the performance of four methods: the nominal method, the robust method (we set $\gamma = \sigma/2$), the PCA-nominal method (Formulation 3 with $\gamma = 0$), and the PCA-robust method (Formulation 3 with $\gamma = \sigma/2$). We fix the dimensionality $p = 100$, and vary the number of constraints m from 100 to 400. Three performance criteria are compared, namely (a) magnitude of violation, (b) fraction of violated constraints, and (c) objective value. For all three criteria, a small value means a better performance. We repeat the experiment for different noise levels ($\sigma = 0.1$, 0.5 and 1). For each parameter set, 50 experiments are performed. The result is reported in Figure 1. The results show that when m is relatively large, all methods perform well. On the other hand, for smaller m, both the nominal method and the robust method is worse, and often incurs huge constraint violation. One possible explanation is that when m is relatively small, because of the noise, the nominal problem may be ill-conditioned and the "optimal" solution can deviate significantly or becomes even unbounded.

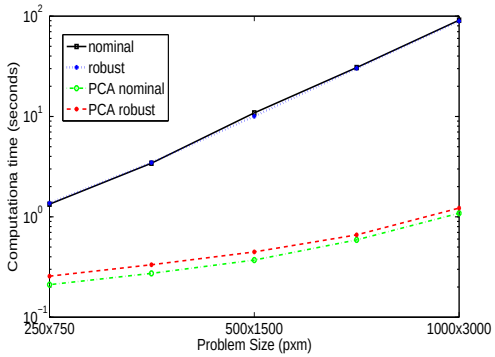
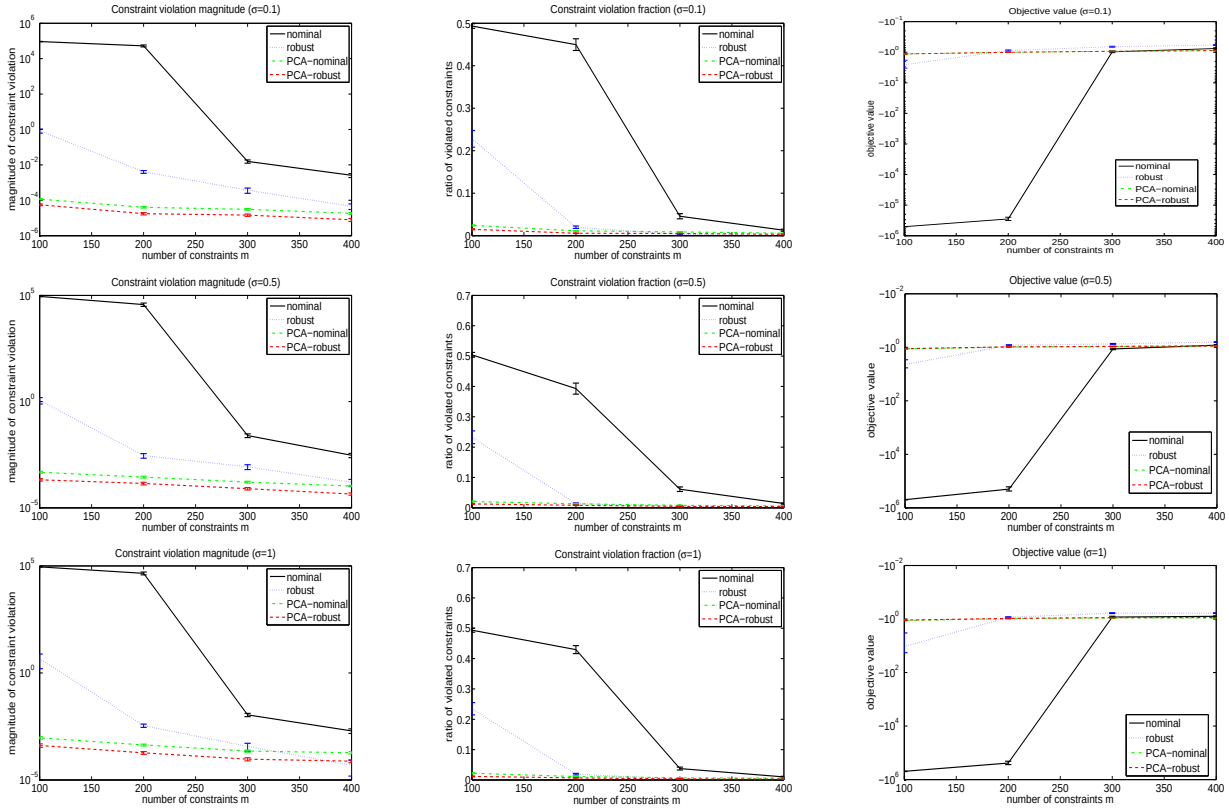


Figure 2: Computation time for different approaches.

We further compare the computational time of each method for different problem sizes, while d is fixed as 5. To avoid the solution to the nominal problem being unbounded, we let the number of constraints equals to three times the number of variables. The computation was done on a Dell desktop using Matlab, and Sedumi as the solver. For each parameter set, 10 runs were

[2]The code of the experiment is available online at http://guppy.mpe.nus.edu.sg/~mpexuh/code/sta-opt.zip.



(a) Magnitude of Violation (b) Ratio of Violated Constraints (c) Objective Value

Figure 1: Performance comparison of nominal method (black), robust method (blue), PCA-nominal (green), and PCA-robust (red). We fix the dimensionality $p = 100$, and vary the number of constraints m from 100 to 400. Three performance criteria are compared: (a) the magnitude of violation; (b) the fraction of violated constraints; and (c) the objective value. The first, second and third rows are for $\sigma = 0.1$, $\sigma = 0.5$ and $\sigma = 1$, respectively.

conducted, and the average running time is reported in Figure 2. It is clear that the nominal method and the robust method do not scale well when the problem size increase. Indeed, when the problem size doubles, the computational time to the nominal and the robust method increases about 8 times. On the other hand, the PCA based methods scales much better.

8 Discussion

In this paper we investigate linear programming under uncertainty where the parameters are observed via noisy sampling. We propose to (approximately) solve such problems using dimensionality reduction techniques, in particular PCA. We provide theoretic justifications as well as empirical evidence to support the proposed method. Our main thrust is to bring to bear tools from statistics and machine learning to inform optimization. There are some natural extensions. These include the consideration of more general structure for the samples. For instance, instead of assuming samples are generated from a low-dimensional subspace, one can consider the case where the samples are generated by a low-dimensional manifold, or a union of multiple subspaces, which would call for dimensionality reduction techniques other than PCA. Another interesting extension is to investigate general (i.e., non-linear) convex optimization problems under uncertainty.

Acknowledgements

H. Xu acknowledges the support from NUS startup grant R-265-000-384-133. The research of C. Caramanis was partially supported by NSF grants EFRI-0735905, EECS-1056028, and DTRA grant HDTRA 1-08-0029. The research of S. Mannor was partially supported by the Israel Science Foundation (contract 890015).

References

Anthony, M., & Bartlett, P. L. (1999). *Neural network learning: Theoretical foundations*. Cambridge University Press.

Ben-Tal, A., El Ghaoui, L., & Nemirovski, A. (2009). *Robust optimization*. Princeton University Press.

Ben-Tal, A., & Nemirovski, A. (1999). Robust solutions of uncertain linear programs. *Operations Research Letters*, *25*, 1–13.

Bertsimas, D., Brown, D. B., & Caramanis, C. (2011). Theory and applications of robust optimization. To appear in *SIAM Review*.

Bertsimas, D., & Sim, M. (2004). The price of robustness. *Operations Research*, *52*, 35–53.

Birge, J. R., & Louveaux, F. (1997). *Introduction to stochastic programming*. Springer-Verlag, New York.

Cai, J.-F., Candès, E., & Shen, Z. (2008). A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, *20*, 1956–1982.

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

Davidson, K., & Szarek, S. (2001). Local operator theory, random matrices and banach spaces. *Handbook on the Geometry of Banach Spaces* (pp. 317–366). Elsvier.

Grant, M., & Boyd, S. (2011). CVX: Matlab software for disciplined convex programming, version 1.21. http://cvxr.com/cvx.

Jolliffe, I. T. (1986). *Principal component analysis*. Springer Series in Statistics, Berlin: Springer.

Li, J. L. L., & Zhang, T. (2009). Sparse online learning via truncated gradient. *Journal of Machine Learning Research*, *10*, 777–801.

Prékopa, A. (1995). *Stochastic programming*. Kluwer.

Shalev-Shwartz, S., & Singer, Y. (2007). A primal-dual perspective of online learning algorithms. *Machine Learning*, *69*, 115–142.

Shalev-Shwartz, S., & Srebro, N. (2008). SVM optimization: Inverse dependence on training set size. *Proceedings of the 22nd international conference on Machine learning*.

Shivaswamy, P. K., Bhattacharyya, C., & Smola, A. J. (2006). Second order cone programming approaches for handling missing and uncertain data. *Journal of Machine Learning Research*, *7*, 1283–1314.

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

Tropp, J. A. (2006). Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, *51*, 1030–1051.

van der Vaart, A. W., & Wellner, J. A. (2000). *Weak convergence and empirical processes*. Springer-Verlag, New York.

Xu, H., Caramanis, C., & Mannor, S. (2009a). Robust regression and Lasso. *Advances in Neural Information Processing Systems 21* (pp. 1801–1808).

Xu, H., Caramanis, C., & Mannor, S. (2009b). Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, *10*, 1485–1510.