Optimization with uncertain data

John Duchi
Notes for EE364b, Spring 2015

May 9, 2015

Contents

1 Robust optimization

Robust (convex) optimization problems are a class of convex optimization problems in which we take the somewhat agnostic view that our problem data is not exact. As in essentially no situations in which one has actually collected data can we represent things exactly, this explicit modeling of data uncertainty can prove extremely useful, and we will see several examples of this in this note.

Abstractly, robust convex optimization problems are formulated with an *uncertainty set* $\mathcal{U}$, objective convex objective $f_0 : \mathbf{R}^n \to \mathbf{R}$, and functions $f_i : \mathbf{R}^n \times \mathcal{U} \to \mathbf{R}$ such that $f_i(\cdot, u)$ is convex for each $u \in \mathcal{U}$. With this abstract formulation in mind, the most general form for robust convex optimization is as follows: we wish to solve

$$\begin{aligned} &\text{minimize} \;\; f_0(x) \\ &\text{subject to} \;\; f_i(x, u) \leq 0 \text{ for all } u \in \mathcal{U}, i = 1, \ldots, m. \end{aligned} \tag{1}$$

The problem (1) is a convex optimization problem, as it is clearly equivalent to

$$\begin{aligned} &\text{minimize} \;\; f_0(x) \\ &\text{subject to} \;\; \sup_{u \in \mathcal{U}} f_i(x, u) \leq 0 \;\; i = 1, \ldots, m, \end{aligned}$$

and suprema of collections of convex functions are convex. We can always consider the constraint functions f_i individually, as they must all be satisfied for all u. Moreover, we always assume that the objective functions f_0 are fixed and not subject to uncertainty; if this is not the case, we may replace the objective with its worst case value $\sup_{u \in \mathcal{U}} f_0(x, u)$, and then write this in epigraph form by introducing the variable t: minimize t subject to $f_0(x, u) \leq t$ for all $u \in \mathcal{U}$. Additionally, we never include equality constraints, as "robust" equality constraints make little sense: how could satisfy $(a + u)^T x = b$ for all $u \in \mathcal{U}$ for only very restrictive sets $\mathcal{U}$.

Now, we must answer three questions about the above formulation: first, is it useful? Second, is it computable? And third, how should we choose the uncertainty sets $\mathcal{U}$ (perhaps to help with the first two issues)?

1.1 An example and failure of classical optimization

We begin by considering a linear program originally described by Ben-Tal, El Ghaoui, and Nemirovski [BTGN09], with variables $c \in \mathbf{R}^n$, $A \in \mathbf{R}^{m \times n}$, and $b \in \mathbf{R}^m$, which comes from a medical production example. We have a cost vector $c = [100 \;\; 199.9 \;\; -5500 \;\; -6100]^T$ corresponding to costs and profits of selling two drugs, and constraints $Ax \leq b$ on their

production, with

$$A = \begin{bmatrix} -.01 & -.02 & .5 & .6 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 90 & 100 \\ 0 & 0 & 40 & 50 \\ 100 & 199.9 & 700 & 800 \\ & & -I_4 & \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} 0 \\ 1000 \\ 2000 \\ 800 \\ 100000 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Here, $-.01$ and $-.02$ correspond to percentages of chemicals in the raw materials used for making components of the different drugs; if we modify them only slightly, things can get out of hand very quickly.

Indeed, let us assume that the prices .01 and .02 vary by at most .5% and 2%, and that the variability is distributed uniformly in $.01 \pm .00005$ and $.02 \pm .0004$. In Figure 1, we show the results of a Monte Carlo simulation in which we drew 10^4 samples uniformly varying the $-.01$ and $-.02$ entries of A as described, where we correct production (i.e. modify the nominal solution x) by reducing x_3 and x_4 to address increases in A_{11} or A_{12} so that we still satisfy the constraint $\sum_{i=1}^4 A_{1i}x_i \leq 0$. We plot the frequency of relative changes in optimal value $c^T x$ after this (tiny) random perturbation; in at least 25% of the experiments, we lose at least 15% of the profits associated with production (and often 20% or more). This suggests that a reformulation of our problem to address the uncertainty is warranted; in fact, if we instead solve

$$\begin{array}{c} \text{minimize } c^T x \\ \text{subject to } (A + \Delta)x \leq b \end{array}$$

for all matrices Δ with $|\Delta_{11}| \leq .00005$ and $|\Delta_{12}| \leq .0004$, with $\Delta_{ij} = 0$ otherwise, we find an optimal solution whose degradation is at most 6% over the nominal.

1.2 Robustness can be hard

As written in the abstract formulation (1), it is not clear whether we should expect to be able to solve robust optimization problems. In general, in spite of the convexity of the objective, it is not the case that all robust convex optimization problems can be solved. Consider the following example. Suppose we would like decide whether the following convex (quadratic even) inequality may be satisfied:

$$\|Ax + Bu\|_2 \leq 1, \quad \text{for all } u \text{ s.t. } \|u\|_\infty \leq 1. \tag{2}$$

Unfortunately, by taking $A = 0$, this amounts to checking whether $u^T B^T B u \leq 1$ for all vectors u such that $\|u\|_\infty \leq 1$; as this is convex in u over the compact set $\{u \in \mathbf{R}^n \mid \|u\|_\infty \leq 1\}$, the maximum must be attained at one of the extreme points $u \in \{-1, 1\}^n$. But then to check whether the constraint (2) is feasible, we must be able to maximize non-convex quadratic functions over the hypercube, which is NP-hard even to do approximately [Hås01].

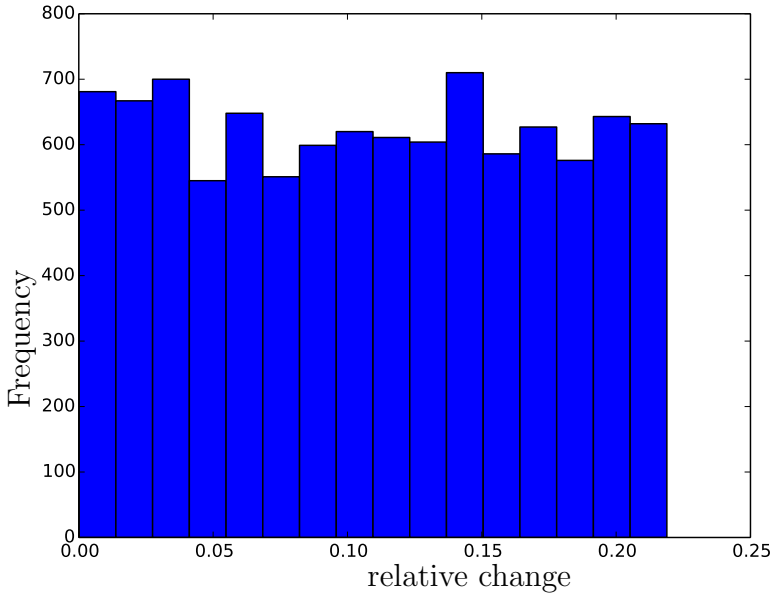

Figure 1: Frequency of fluctations to a certain level for the non-robust production planning problem with .5% fluctuations in material proportions.

1.3 What we consider

With these examples in mind, we arrive at three major questions for robust optimization:

(1) Why should we attempt to be robust?

(2) What problems can we actually solve robustly?

(3) How should we choose the uncertainty sets $\mathcal{U}$?

For the first question, one reason is clear: because data may be uncertain, and it is unreasonable to assume we have perfectly represented any real number $a \in \mathbf{R}$. (And with even tiny relative changes, as in § 1.1, we can huge swings in solution quality.) There are other reasons as well, which we discuss subsequently. We can use robustness as a proxy for dealing with randomness whose values we do not know, and in some cases robust optimization techniques allow us to tractably approximate non-convex probabilistic constraints. The second question we illustrate via examples in linear programming, second order cone programs, and semidefinite programs; however, as the constraint $\|Ax + Bu\|_2 \leq 1$ for all $u \in [-1, 1]^n$ makes clear, we must be careful. A major theme in efficient representation of robust uncertainty sets is duality and dual representations, which allows us to turn infinite (or semi-infinite) sets of constraints into a few simpler inequalities. Lastly, choosing uncertainty sets is an important question, as it effects both whether we can efficiently represent a robust problem, and choosing too large an uncertainty set $\mathcal{U}$ can yield robust solutions that are so conservative—the resulting solutions of such low quality for the objective f_0 at hand—as to lose essentially any advantage over non-robust optimization. One important

tool here, which we will see, is the use of probabilistic guarantees to choose uncertainty sets $\mathcal{U}$.

2 Robust linear programs

We first consider robust linear programs, which provide the simplest set of examples for tractable robust optimization formulations. An important question in robust optimization is in which scenarios the robust formulation remains in the same "problem class" as the original non-robust (certain) problem. That is, when is a robust linear program still a (similarly-sized) linear program, when is a robust second order cone problem still an SOCP, when is a robust SDP still an SDP, when is a robust geometric program a geometric program. As a running example, we will use extensions of a simple portfolio optimization problem to illustrate the ideas.

2.1 Robust LPs as LPs

For our first set of robust problems, we consider polyhedral uncertainty sets, which allow the simplest "tractable" representations. In particular, we consider the problem of solving

$$\begin{aligned} &\text{minimize } c^T x \\ &\text{subject to } (A+U)x \succeq b \text{ for } U \in \mathcal{U}. \end{aligned}$$

Because we can always represent the robust formulation constraint-wise, we will only consider the single robust inequality

$$(a+u)^T x \leq b \text{ for all } u \in \mathcal{U}. \tag{3}$$

The simplest set $\mathcal{U}$ is to have interval-based uncertainty, that is, we know that $u_j \in [-\delta, \delta]$ for each coordinate j (*i.e.* $\|u\|_\infty \leq \delta$). In this case, a direct calculation gives that

$$\sup_{u \in \mathcal{U}} (a+u)^T x = a^T x + \delta \|x\|_1 ,$$

so that the robust inequality (3) is equivalent to the linear inequality $a^T x + \delta \|x\|_1 \leq b$. Similarly, if $\mathcal{U}$ is an ℓ_1-ball, we have a robust formulation of the form $a^T x + \delta \|x\|_\infty \leq b$. Both of these inequalities place additional restrictions on the acceptable values for x, which has the effect of "robustifying" solutions to the linear program.

A more general version of this setting is *polyhedral uncertainty*, where for a matrix $F \in \mathbf{R}^{m \times n}$ and vector $g \in \mathbf{R}^m$, we have the robust inequality

$$(a+u)^T x \leq b \text{ for } u \in \mathcal{U} = \{u \in \mathbf{R}^n \mid Fu + g \succeq 0\} .$$

In this case, duality plays an important role in giving a tractable representation of the above semi-infinite inequality (we transform "for-all" quantifiers into existence quantifiers). Indeed,

writing the Lagrangian for the maximization problem over $u^T x$ by introducing the variable $\lambda \succeq 0$, we have

$$L(u, \lambda) = x^T u + \lambda^T (Fu + g) \quad \text{and} \quad \sup_u L(u, \lambda) = \begin{cases} +\infty & \text{if } F^T \lambda + x \neq 0 \\ \lambda^T g & \text{if } F^T \lambda + x = 0. \end{cases}$$

In particular, as all the inequality constraints are linear, strong duality obtains, and we have that

$$\sup_{u \in \mathcal{U}} u^T x = \inf\{\lambda^T g \mid F^T \lambda + x = 0, \lambda \succeq 0\}.$$

Thus the robust linear inequality with polyhedral uncertainty is equivalent to the three linear equations

$$a^T x + \lambda^T g \leq b, \quad F^T \lambda + x = 0, \quad \lambda \succeq 0.$$

So for any polyhedral uncertainty set $\mathcal{U}$, we may write the resulting robust linear program as a standard linear program of essentially the same size (modulo a few additional constraints) as the original linear program.

Example 1 *Portfolio optimization.* In the portfolio optimization problem, we seek to invest in a collection of n assets $i = 1, \ldots, n$, each of which has random return R_i with expectation $\mathbf{E}\, R_i = \mu_i \geq 1$ (our wealth is multiplied by a factor R_i if we invest in asset i), where we assume that $\mathbf{E}\, R_1 = \mu_1 \geq \mu_2 \geq \ldots \geq \mu_n$. We let x_i denote the proportion of resources invested in asset i, and maximizing our expected returns corresponds to the certainty-equivalent problem

$$\text{maximize } \mu^T x \text{ subject to } \mathbf{1}^T x = 1, \ x \succeq 0. \tag{4}$$

The solution to the problem (4) is clearly to put all the resources into the asset with highest mean, μ_1.

This ignores variability in the solutions, which of course is an important part of any optimization scheme with random data. Now, suppose we know that each asset i varies in a range $\mu_i \pm \mathsf{u}_i$, where $u_1 \geq u_2 \geq \ldots \geq u_n = 0$, where we assume the last asset is given by investing all of our money in the bank. Then the (most conservative) robust solution to the portfolio optimization problem is to maximize our worst case return subject to our uncertainties:

$$\text{maximize } \sum_{i=1}^n \inf_{u \in [-\mathsf{u}_i, \mathsf{u}_i]} (\mu_i + u) x_i \text{ subject to } \mathbf{1}^T x = 1, \ x \succeq 0.$$

Clearly, this is equivalent to the problem

$$\text{maximize } \mu^T x - \sum_{i=1}^n \mathsf{u}_i |x_i| \text{ subject to } \mathbf{1}^T x = 1, \ x \succeq 0.$$

We now fix values for the μ_i and u_i and study the robust and non-robust solution. We set $\mu_i = 1.05 + .3\frac{n-i}{n}$ and the uncertainty bounds $\mathsf{u}_i = .05 + .5\frac{n-i}{n}$, except that $u_n = 0$. Then by inspection, the solutions to the non-robust problem are $x_{\text{nom}} = e_1$, the first basis vector (we

invest fully in the riskiest but highest return asset), and the robust problem we put all of our money in the least risky asset, $x_{\rm rob} = e_n$. So, for example, we we assume that each asset has return $R_i = \mu_i + \mathsf{u}_i$ with probability $1/2$ and return $R_i = \mu_i - \mathsf{u}_i$ with probability $1/2$, in the non-robust case, there is a 50% chance we lose about 50% of our wealth, while in the robust case, we simply increase wealth by 5%.

2.2 Robust LPs as SOCPs

We can allow other types of uncertainty in the parameters a in our linear inequalities. Another standard uncertainty type is norm-based uncertainty in the vectors a, meaning there exists a matrix $P \in \mathbf{R}^{n\times m}$ such that that our uncertain inequality is of the form

$$(a + Pu)^T x \le b \;\; \text{for } u \in \mathcal{U} = \{u \in \mathbf{R}^m \mid \|u\| \le 1\},$$

where $\|\cdot\|$ is some norm. We may directly take a supremum over all such u, which gives us the dual norm constraint

$$a^T x + \left\|P^T x\right\|_* \le b,$$

which follows because $\sup_{\|u\|\le 1}(Pu)^T x = \sup_{\|u\|\le 1} u^T P^T x = \|P^T x\|_*$ by definition of the dual norm. So, for example, we may use the ellipsoidal uncertainty set given by $\mathcal{U} = \{u \in \mathbf{R}^m \mid \|u\|_2 \le 1\}$, which gives the second order cone inequality

$$a^T x + \left\|P^T x\right\|_2 \le b.$$

Example 2 *Portfolio optimization* (Example 1 continued). Let us put ourselves back in the setting of Example 1, where we assume that our returns $R_i \in [\mu_i - u_i, \mu_i + u_i]$ and satisfy $\mathbf{E}\, R_i = \mu_i$ for each i. Now, rather than guaranteeing a certain return—being extremely conservative—suppose we want to guarantee a return that holds with probability $\ge 1 - \epsilon$ for some small $\epsilon > 0$ (this is known as *value at risk*). That is, we would like to solve

$$\text{maximize } \; t \;\; \text{subject to } \; \mathbf{Prob}\left(\sum_{i=1}^n R_i x_i \ge t\right) \ge 1 - \epsilon.$$

This is non-convex, but we may approximate it with appropriate uncertainty sets.

In the robust formulation of Example 1, we chose $\epsilon = 0$ to give a guaranteed return of μ_n, the bank interest rate. Somewhat more careful control of the probability of failure gives better returns for small (non-zero) ϵ. First we note that by Hoeffding's inequality (see § 7 for derivations of these guarantees and more on the choice of uncertainty sets), we have that for R_i bounded as above, for any $t \ge 0$, and any fixed vector $x \in \mathbf{R}^n$,

$$\mathbf{Prob}\left(\sum_{i=1}^n (R_i - \mu_i) x_i \le -t\right) \le \exp\left(-\frac{t^2}{2\sum_{i=1}^n x_i^2 \mathsf{u}_i^2}\right).$$

Written slightly differently,

$$\mathbf{Prob}\left[\sum_{i=1}^n R_i x_i \le \mu^T x - t\left(\sum_{i=1}^n \mathsf{u}_i^2 x_i^2\right)^{\frac{1}{2}}\right] \le \exp\left(-\frac{t^2}{2}\right),$$

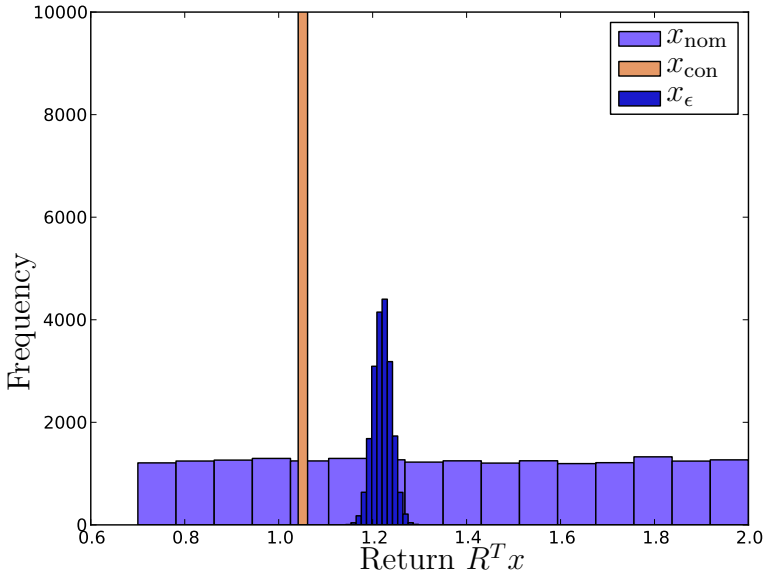

Figure 2: Histogram of random returns for model described in examples 1 and 2.

and taking $t = \sqrt{2\log\frac{1}{\epsilon}}$, we obtain that $\sum_{i=1}^n R_i x_i \geq \mu^T x - \sqrt{2\log\frac{1}{\epsilon}}\,\|\mathbf{diag}(\mathsf{u})x\|_2$ with probability at least $1-\epsilon$. This gives us the robust problem (much less conservative than that of Example 1)

$$\text{maximize } \mu^T x - \sqrt{2\log\frac{1}{\epsilon}}\,\|\mathbf{diag}(\mathsf{u})x\|_2 \quad \text{subject to } \mathbf{1}^T x = 1,\ x \succeq 0.$$

This corresponds to the uncertainty set given by the diagonally scaled ℓ_2 ball

$$\mathcal{U} = \left\{ u \in \mathbf{R}^n \mid \left\|\mathbf{diag}(u)^{-1/2}u\right\|_2 \leq \sqrt{2\log\frac{1}{\epsilon}} \right\}$$

in the robust inequality $(\mu + u)^T x \leq t$ for all $u \in \mathcal{U}$.

Now we perform a small simulation study comparing the quality of the solutions given by the nominal solution $x_{\text{nom}} = e_1$, the extremely conservative solution $x_{\text{con}} = e_n$, and the robust solution x_ϵ with value-at-risk $\epsilon = 2 \times 10^{-4}$. In Figure 2, we plot the results of a simulation study in which returns are chosen uniformly at random from $[\mu_i - \mathsf{u}_i, \mu_i + \mathsf{u}_i]$. We see that the nominal solution has better maximum returns (and average), but has some substantial failures as well, while the solution with fixed value at risk $\epsilon = 2 \cdot 10^{-4}$ has some variability, but we are guaranteed to have return at least 1.08 with probability at least $1 - \epsilon$.

2.3 LPs with conic uncertainty

The most general form of constraints we consider for linear programs are general conic uncertainty sets; these include as special cases all of the examples in this section, among

many others, including semidefinite uncertainty. Recall that a set $K \subset \mathbf{R}^m$ is a *convex cone* if for all $x \in K$, we have $tx \in K$ for all $t \geq 0$, and K is convex. We also let $K^* = \{v \in \mathbf{R}^m \mid v^T x \geq 0 \text{ for all } x \in K\}$ denote the *dual cone* to K. We define the generalized (cone) inequality $x \succeq_K y$ if $x - y \in K$ (recall [BV04, Chapters 4.6 and 5.9]), and we consider the robust inequality

$$(a+u)^T x \leq b \;\; \text{for } u \in \mathcal{U} = \{u \in \mathbf{R}^n \mid Fu + g \succeq_K 0\}, \tag{5}$$

where $F \in \mathbf{R}^{m \times n}$ and $g \in \mathbf{R}^n$ are fixed. We now show how, if we can represent K efficiently, duality can let us write the infinite collection of inequalities (5) as a small set of standard convex (cone) inequalities.

To begin, we assume that the inequality $Fu + g \succeq_K 0$ can be satisfied strictly, that is, a Slater condition holds (*i.e.* there is a $\bar{u} \in \mathbf{R}^n$ such that $F\bar{u} + g \in \mathbf{relint}\, K$). In this case, we have that strong duality obtains for the problem of maximizing $u^T x$ over $u \in \mathcal{U}$, and writing the Lagrangian for this (concave) problem, we have for $\lambda \succeq_{K^*} 0$

$$L(u, \lambda) = u^T x + \lambda^T (Fu + g) \;\; \text{and} \;\; \sup_u L(u, \lambda) = \begin{cases} +\infty & \text{if } \; x + F^T \lambda \neq 0 \\ \lambda^T g & \text{if } x + F^T \lambda = 0. \end{cases}$$

By strong duality, we have $\sup_{u \in \mathcal{U}} u^T x = \inf_{\lambda \succeq_{K^*} 0} \{g^T \lambda \mid x + F^T \lambda = 0\}$, and the latter infimum is attained at some $\lambda^\star$, so that inequality (5) is equivalent to the three inequalities

$$a^T x + \lambda^T g \leq b, \;\; \lambda \succeq_{K^*} 0, \;\; x + F^T \lambda = 0.$$

Example 3 *Linear programs with semidefinite uncertainty.* Suppose we have symmetric matrices $A_0, A_1, \ldots, A_m \in \mathbf{S}^m$, and for a single constraint $a^T x \leq b$ we have the robust counterpart

$$(a + Pu)^T x \leq b \quad \text{for all } u \text{ s.t. } A_0 + \sum_{i=1}^m u_i A_i \succeq 0$$

for some matrix $P \in \mathbf{R}^{n \times m}$. Using the preceding derivation with positive semidefinite cone $K = \{X \in \mathbf{S}^m \mid X \succeq 0\}$ (which is self-dual), then assuming there exists some $\bar{u}$ such that $A_0 + \sum_{i=1}^m A_i u_i \succ 0$, the uncertain inequality is equivalent to the existence of a positive semidefinite matrix $\Lambda \succeq 0$ such that

$$a^T x + \mathbf{Tr}(\Lambda A_0) \leq b, \;\; \text{and} \;\; P^T x + \begin{bmatrix} \mathbf{Tr}(\Lambda A_1) \\ \vdots \\ \mathbf{Tr}(\Lambda A_m) \end{bmatrix} = 0.$$

---

3 Robust cone programs

In this section, we consider second-order (like) cone programs, giving examples in which duality or direct calculations allow us to efficiently represent our uncertainty, that is, to

write the robust problem using standard convex formulations. The starting point for all of our derivations in this section is a (minor extension of) the standard second order (Lorentz) cone, where we consider the constraint

$$\|Ax + b\|_2 \leq c^T x + d, \tag{6}$$

where $A \in \mathbf{R}^{m \times n}$, $b \in \mathbf{R}^m$, $c \in \mathbf{R}^n$, and $d \in \mathbf{R}$, and we denote the rows of A by $a_i \in \mathbf{R}^n$.

3.1 SOCPs with interval uncertainty

The simplest—yet still general—robustification of the constraint (6) is to uncertainty $\Delta \in \mathbf{R}^{m \times n}$ to the matrix A where the uncertainty is uniformly bounded as $|\Delta_{ij}| \leq \delta$ for some $\delta > 0$. We can also easily add—building off of the conic representations in § 2.3—conic uncertainty to the vectors on the right hand side, which leaves us with the robust set of inequalities

$$\|(A + \Delta)x + b\|_2 \leq (c + u)^T x + d \;\text{ for all } \Delta \text{ s.t. } \|\Delta\|_\infty \leq \delta, u \in \mathcal{U}, \tag{7}$$

where $U = \{u : Fu + g \succeq_K 0\}$ for a convex cone K.

By breaking inequality (7) into two inequalities, namely, $\|(A + \Delta)x + b\|_2 \leq t$ and $t \leq (c + u)^T x + d$ for some $t \in \mathbf{R}_+$, we can address each side in turn. The right hand side of inequality (7) is handled identically to the results in § 2.3 on robust linear programs: we have for any t that $t \leq (c + u)^T x + d$ for all $u \in \mathcal{U}$ if and only if we can find solutions to the three inequalities

$$t \leq c^T x + d - \lambda^T g, \; x = F^T \lambda, \; \lambda \succeq_{K^*} 0.$$

For the left hand side $\|(A + \Delta)x + b\|_2 \leq t$, we have by direct calculation that

$$\begin{aligned}
\sup_{\Delta : |\Delta_{ij}| \leq \delta} \|(A + \Delta)x + b\|_2 &= \sup_{\Delta : |\Delta_{ij}| \leq \delta} \left(\sum_{i=1}^m [(a_i + \Delta_i)^T x + b_i]^2 \right)^{\frac{1}{2}} \\
&= \sup_{\Delta \in \mathbf{R}^{m \times n}} \left\{ \|z\|_2 \mid z_i = a_i^T x + \Delta_i^T x + b_i, \|\Delta_i\|_\infty \leq \delta \right\} \\
&= \inf \left\{ \|z\|_2 \mid z_i \geq |a_i^T x + b| + \delta \|x\|_1 \right\}.
\end{aligned}$$

Combining these results, we find that the infinite collection of inequalities (7) is equivalent to a small number of convex inequalities and linear equalities in the variables $z \in \mathbf{R}^m$, $\lambda \succeq_{K^*} 0$, and $x \in \mathbf{R}^n$:

$$\begin{aligned}
\|z\|_2 \leq c^T x - \lambda^T g + d, \;\; x = F^T \lambda, \; \lambda \succeq_{K^*} 0, \\
z_i \geq |a_i^T x + b_i| + \delta \|x\|_1 \;\; \text{for } i = 1, \ldots, m.
\end{aligned}$$

Based on the decomposition of the right and left sides and the results on robust linear inequalities, for these types of conic problems we really need only consider robustifying inequalities of the form $\|Ax + b\|_2 \leq t$ for $x \in \mathbf{R}^n$ and $t \in \mathbf{R}_+$.

3.2 SOCPs with ellipsoidal uncertainty

A variation of the uniform interval constraints in the preceding section is to consider ellipsoidal uncertainty on each of the rows of the matrix A; see also [BV04, Chapter 6.4]. In this case, we assume there are matrices $P_1, \ldots, P_m \in \mathbf{R}^{n \times n}$, and we represent the uncertainty in the matrix A as $a_i + P_i u$ for vectors u such that $\|u\|_2 \le 1$. Our robust (uncertain) inequality is

$$\left(\sum_{i=1}^m [(a_i + P_i u)^T x + b_i]^2 \right)^{\frac{1}{2}} \le t \quad \text{for all } u \text{ s.t. } \|u\|_2 \le 1.$$

By inspection, we may rewrite this with variables $z_i \ge \sup_{\|u\|_2 \le 1} |a_i^T x + b_i + u^T P_i^T x|$, which is equivalent to $z_i \ge |a_i^T x + b_i| + \|P_i^T x\|_2$, giving the equivalent robust inequality

$$\|z\|_2 \le t, \quad z_i \ge |a_i^T x + b_i| + \left\|P_i^T x\right\|_2 \text{ for } i = 1, \ldots, m.$$

In particular, we can take a second order cone inequality with row-wise ellipsoidal uncertainty and represent it as a collection of $O(m)$ second order cone and linear inequalities. In particular, aside from additional inequalities, there is no increase in complexity for robustness in this case.

3.3 SOCPs with matrix uncertainty

We now consider a slightly more complicated situation than that outlined in the previous section, instead looking at uncertainty in the matrix A given by a matrix Δ with an operator norm (the induced ℓ_2-norm) bound on Δ. In this case, we require substantially more complex duality arguments to give an efficient representation of our robust problem. First, our uncertain inequality in this case is specified by a matrix $P \in \mathbf{R}^{m \times n}$ and scalar $\delta > 0$, and would like to guarantee that

$$\|(A + P\Delta)x + b\|_2 \le t, \quad \text{for } \Delta \in \mathbf{R}^{n \times n} \text{ s.t. } \|\Delta\| \le \delta, \tag{8}$$

where $\|\Delta\| = \sup_{\|u\|_2 = \|v\|_2 = 1} u^T \Delta v$ is the maximum singular value (operator norm) of Δ.

To derive a dual representation of the inequality (8), we require two results: the *Schur complement representation* of a the second order cone and the *S-lemma*.[1] The first is that the following two inequalities are equivalent:

$$\|x\|_2 \le t \quad \text{and} \quad \begin{bmatrix} t & x^T \\ x & tI_n \end{bmatrix} \succeq 0.$$

The S-lemma is used to show that strong duality still obtains for non-convex quadratic problems as long as there are at most *two* quadratics, though we do not prove this here. The

[1] The formal Schur complement lemma is that if $X = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$, then if $A \succ 0$, we have $X \succeq 0$ if and only if $C - B^T A^{-1} B \succeq 0$ (*cf.* [BV04, Appendix A.5.5]). We only use the homogeneous S-lemma here; there are more general versions, *cf.* [BV04, Appendix B].

S-lemma is the following:

$$x^T A x \geq 0 \text{ implies } x^T B x \geq 0 \quad \text{if and only if} \quad \exists\, \lambda \geq 0 \text{ s.t. } B \succeq \lambda A. \tag{9}$$

For a proof of this fact using semidefinite programming duality, see Appendix A.

Now, we give a characterization of inequality (8) by combining these two results. Indeed, by Schur complements we have that the infinite inequalities (8) are equivalent to

$$\begin{bmatrix} t & ((A+P\Delta)x+b)^T \\ (A+P\Delta)x+b & tI_m \end{bmatrix} \succeq 0 \;\text{ for }\; \|\Delta\| \leq 1.$$

By left and right multiplying the positive definite inequality by $[s\ v^T]$ and $[s\ v^T]^T$ for $s \in \mathbf{R}$ and $v \in \mathbf{R}^m$, this is equivalent to

$$ts^2 + 2s((A+P\Delta)x+b)^T v + t\|v\|_2^2 \geq 0 \;\text{ for all } s \in \mathbf{R}, v \in \mathbf{R}^m, \;\; \|\Delta\| \leq 1.$$

Taking an infimum over $\|\Delta\| \leq 1$ in the preceding display, noting that $\sup_{\|\Delta\|\leq 1} u^T \Delta v = \|uv^T\| = \|u\|_2 \|v\|_2$, this is equivalent to

$$ts^2 + 2s(Ax+b)^T v + t\|v\|_2^2 - 2\|sx\|_2 \|P^T v\|_2 \geq 0 \;\text{ for all } s \in \mathbf{R}, v \in \mathbf{R}^m.$$

We now give a second order representation for the final negative term in the above inequality, which, combined with the Schur complement representation of an SOCP, gives us a linear matrix inequality that is equivalent to the robust inequality (8). Indeed, note that

$$\sup\left\{u^T x \mid \|u\|_2^2 \leq \|P^T v\|_2^2\right\} = \left\|P^T v\right\|_2 \|x\|_2,$$

so we have that inequality (8) is equivalent to

$$ts^2 + 2s(Ax+b)^T v + t\|v\|_2^2 + 2su^T x \;\text{ for all } s \in \mathbf{R}, v \in \mathbf{R}^m, u \in \mathbf{R}^n \text{ s.t. } \|u\|_2^2 \leq \left\|P^T v\right\|_2^2.$$

That is, we must have that

$$\begin{bmatrix} s \\ v \\ u \end{bmatrix}^T \begin{bmatrix} 0 & 0 & 0 \\ 0 & PP^T & 0 \\ 0 & 0 & -I_n \end{bmatrix} \begin{bmatrix} s \\ v \\ u \end{bmatrix} \geq 0 \;\text{ implies }\; \begin{bmatrix} s \\ v \\ u \end{bmatrix}^T \begin{bmatrix} t & (Ax+b)^T & x^T \\ Ax+b & t & 0 \\ x & 0 & 0 \end{bmatrix} \begin{bmatrix} s \\ v \\ u \end{bmatrix} \geq 0.$$

By the homogeneous S-lemma (9), we thus see that inequality (8) is equivalent to finding $x \in \mathbf{R}^n$ and λ such that

$$\begin{bmatrix} t & (Ax+b)^T & x^T \\ Ax+b & t-\lambda PP^T & 0 \\ x & 0 & \lambda I_n \end{bmatrix} \succeq 0.$$

3.4 Example: robust regression

We now consider an example application of the results in the preceding three sections, showing how different choices of uncertainty sets can give different robustness results. We make use of some random matrix theory, whose results we simply cite, to make our calculations somewhat more precise and our results compelling. We would like to solve

$$\underset{x}{\text{minimize}} \quad \|Ax - b\|_2 , \tag{10}$$

but we assume that $A \in \mathbf{R}^{m\times n}$ has been corrupted by a matrix $\Delta \in \mathbf{R}^{m\times n}$ of i.i.d. standard Gaussian random variables. We consider three uncertainty sets: the first based on bounds on $|\Delta_{ij}|$ for all pairs i, j with high probability, the second bounds that hold on the $\|\Delta_i\|_2$ (where Δ_i denotes the ith row of Δ) for each $i = 1, \ldots, m$ with high probability, and the last based on bounds on the operator norm $\|\Delta\|$ that hold with high probability. To do this, we require a theorem whose results are well-known in probability theory (*cf.* [Ver12, §3.1]; we note also that these results are all sharp as t gets large, meaning that they cannot be improved by any factors except outside the exponent).

Theorem 1 *For each fixed i, j pair and any $t \geq 0$,*

$$\mathbf{Prob}(|\Delta_{ij}| \geq t) \leq 2\exp\left(-\frac{t^2}{2}\right).$$

For each fixed i and any $t \geq 0$,

$$\mathbf{Prob}(\|\Delta_i\|_2 \geq \sqrt{n} + t) \leq \exp\left(-\frac{t^2}{2}\right).$$

Lastly, for the operator norm,

$$\mathbf{Prob}(\|\Delta\| \geq \sqrt{m} + \sqrt{n} + t) \leq \exp\left(-\frac{t^2}{2}\right).$$

Now we use Theorem 1 to give robust formulations similar in spirit to the preceding sections on robust SOCPs. In particular, by using a union bound, Theorem 1 implies that taking $t_\infty(\delta)^2 = 2\log(2mn/\delta)$, we have

$$\mathbf{Prob}(\max_{i,j} |\Delta_{ij}| \geq t_\infty(\delta)) \leq 2mn\exp\left(-\frac{t_\infty(\delta)^2}{2}\right) = \delta,$$

taking $t_2(\delta)^2 = 2\log(m/\delta)$ we have

$$\mathbf{Prob}(\max_i \|\Delta_i\|_2 \geq \sqrt{n} + t_2(\delta)) \leq m\exp\left(-\frac{t_2(\delta)^2}{2}\right) = \delta,$$

while taking $t_{\text{op}}(\delta)^2 = 2\log\frac{1}{\delta}$ gives

$$\mathbf{Prob}(\|\Delta\| \geq \sqrt{n} + \sqrt{m} + t_{\text{op}}(\delta)) \leq \exp\left(-\frac{t_{\text{op}}(\delta)^2}{2}\right) = \delta.$$

In particular, we consider solving the robust least squares approximation

$$\underset{x}{\text{minimize}} \quad \sup_{\Delta \in \mathcal{U}} \|(A + \Delta)x - b\|_2 \tag{11}$$

where $\mathcal{U}$ is one of the three uncertainty sets

$$\begin{aligned}
\mathcal{U}_\infty &= \{\Delta \mid \|\Delta\|_\infty \leq t_\infty(\delta)\}, \\
\mathcal{U}_2 &= \{\Delta \mid \|\Delta_i\|_2 \leq \sqrt{n} + t_2(\delta) \text{ for } i = 1, \ldots m\}, \\
\mathcal{U}_{\text{op}} &= \{\Delta \mid \|\Delta\| \leq \sqrt{n} + \sqrt{m} + t_{\text{op}}(\delta)\}.
\end{aligned}$$

Each of these guarantees that our deviations have probability no more than δ, and each satisfies $\mathbf{Prob}(\Delta \in \mathcal{U})/\delta \to 1$ as $\delta \to 0$.

In Figure 3, we plot the gap between the value $\|Ax_{\text{rob}} - b\|_2$ for the robust solution x_{rob} evaluated on the nominal data and nominal solution's value, that is, $\|Ax^\star - b\|_2 = \inf_x \|Ax - b\|_2$, as a function of the uncertainty level δ. (Small δ indicates robustness against lower probability—more extreme—deviations). From the plot, we see that while each of the uncertainty sets provides protection against deviations, the sets based on the ℓ_2 norm of the columns of Δ_i and $\|\Delta\|_\infty$ are far too conservative. In Figure 4, we plot histograms of the actual objective value $\|(A + \Delta)x - b\|_2$ for the different robust solutions (as well as the nominal solution) for several realizations of Δ. We see that the ellipsoidal uncertainty set $\mathcal{U}_2$ and interval set $\mathcal{U}_\infty$ have less variability than the others—they are quite robust!—but their solutions have very low quality.

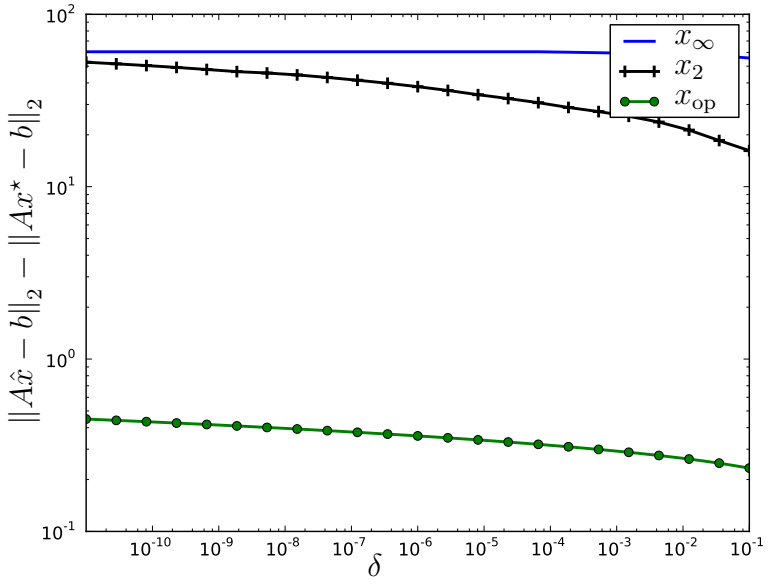

Figure 3: Objective value $\|A\hat{x} - b\|_2 - \|Ax^\star - b\|_2$ versus δ for different robust solutions, where $x^\star$ minimizes the nominal objective and $\hat{x}$ denotes a robust solution.

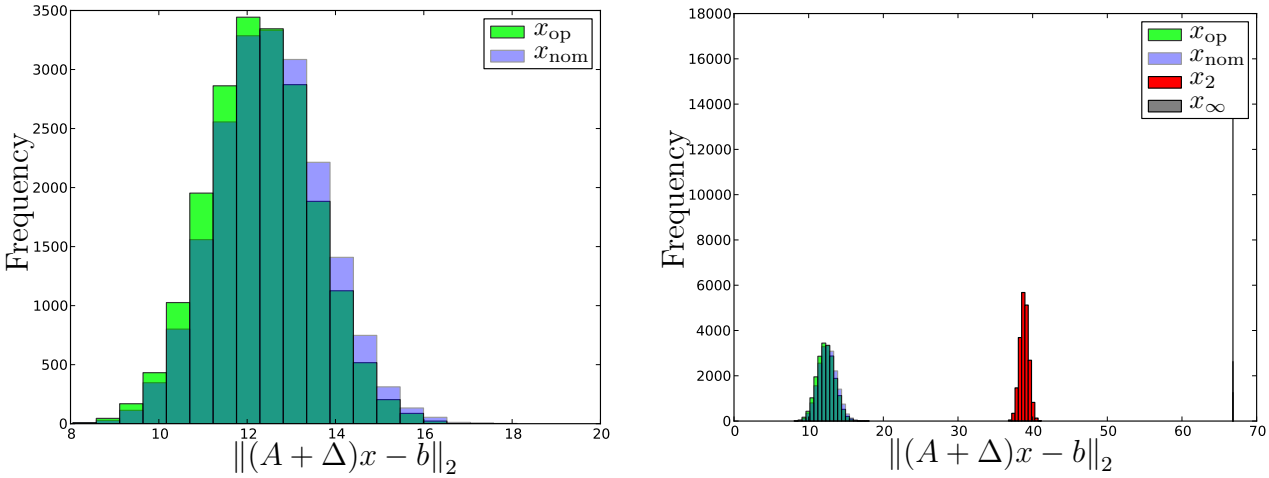

Figure 4: Distribution of residuals for the robust least squares problem (11) with different choices of uncertainty set ($\mathcal{U}_{\text{nom}} = \{0\}$ and $\mathcal{U}_\infty$, $\mathcal{U}_2$, and $\mathcal{U}_{\text{op}}$). The histograms are obtained from 10^5 random Gaussian matrices Δ generated with variance 2.25. We see that the nominal solution is somewhat more disperse than the robust solution, but the more conservative (larger) uncertainty sets $\mathcal{U}_2$ and $\mathcal{U}_\infty$ are far too large to be useful.

4 Robust semidefinite programs

TODO...

5 General robust optimization problems

Probably coming too...

6 Chance constraints and the choice of uncertainty sets

To this point, we have assumed simply that we are given an uncertainty set $\mathcal{U}$, and we would like to protect against all types of variability for every possible $u \in \mathcal{U}$. There are two major considerations in the choice of the uncertainty set $\mathcal{U}$: first, whether it is tractable to represent, meaning that our convex problem is still efficiently solvable, and second, whether it is too conservative and how well it reflects the actual variability in our problem. As the first question is generally handled on a problem-by-problem basis, we focus here on the second question, studying how (1) we may use robust optimization-like tools to represent uncertainty from randomness, and (2) how probabilistic tools allow us to efficiently represent uncertainty sets.

With this in mind, we shift our focus now to the broad problem of *chance-constrained* optimization, where we assume there exists a random variable U with (known or unknown)

probability distribution, and we would like to solve problems of the form

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & \mathbf{Prob}(f_i(x,U) > 0) \le \epsilon, \quad i = 1, \ldots, m. \end{array} \tag{12}$$

Here ϵ is some fixed probability of failing to satisfy the (random) constraint $f_i(x,U) \le 0$, and one usually chooses ϵ to be something like .1 or .01. This problem is generally non-convex, and a standard approach is to give *safe* approximations to the chance constraints (12), that is, to find (convex) functions g_i so that $g_i(x) \le 0$ guarantees that $\mathbf{Prob}(f_i(x,U) > 0) \le \epsilon$.

One immediate approach to handling these chance constraints is to choose any set $\mathcal{U}$ containing U with high probability, that is, such that $\mathbf{Prob}(U \in \mathcal{U}) \ge 1 - \epsilon$, in which case the robust constraint

$$f_i(x,u) \le 0 \;\; \text{for } u \in \mathcal{U}$$

guarantees that $\mathbf{Prob}(f_i(x,U) > 0) \le \epsilon$. Using this approach, we may apply any of the robust optimization ideas in the preceding part of this note, and this is the approach in Example 2. Section 3.4 also gives three different such constructions, each of which gave quite different performance; it is important to choose the set $\mathcal{U}$ to be as small as possible while still guaranteeing $\mathbf{Prob}(U \in \mathcal{U}) \ge 1-\epsilon$. Often, however, rather than directly choosing an uncertainty set $\mathcal{U}$, it is useful to approximate (or directly represent) the probability of error in the our chance constraints, which we now turn to.

6.1 Value at risk

We begin by looking at *value-at-risk*.[2] For a random variable Z, the value at risk ϵ is

$$\mathbf{VaR}(Z;\epsilon) = \inf\left\{\gamma \mid \mathbf{Prob}(Z \le \gamma) \ge 1 - \epsilon\right\} = \inf\left\{\gamma \mid \mathbf{Prob}(Z > \gamma) \le \epsilon\right\}.$$

That is, it is the smallest value γ such that $Z \le \gamma$ with probability at least $1 - \epsilon$, or the probability that Z exceeds γ is at most ϵ. In financial applications, one usually replaces $1 - \epsilon$ with η, in which case the value at risk involves the left (lower) quantile of a collection of assets, and Z represents a return on investment; we wish to maximize the $\mathbf{VaR}(Z;\eta)$ for some small η, say $\eta \in \{.1, .05, .01\}$, which means that the lowest η quantile of the returns Z is no smaller than $\mathbf{VaR}(Z;\eta)$. For convex optimization problems, where $Z = f(x,U)$ represents a functional whose value we wish to satisfy $f(x,U) \le 0$ with high probability, then

$$\mathbf{VaR}(Z;\epsilon) \le 0 \;\; \text{if and only if} \;\; \mathbf{Prob}(Z > 0) \le \epsilon.$$

That is, the value at risk of the random variable $Z = f_i(x,U)$ at level ϵ being non-positive is equivalent to the chance constraints (12) holding. Unfortunately, except in very special cases, the value at risk function yields non-convex constraints. Thus, as we see in the coming sections

[2] There are multiple symmetric definitions of this quantity that flip ϵ and $1 - \epsilon$. Compare the books [BTGN09], [SDR09, Chapter 6.2.4] and the paper [RU00]. We focus on the one convenient for our purposes, though we caution the reader to carefully check the definition used in any paper he or she reads.

Example 4 *Normal distributions and value at risk.* In the case of normally distributed data U and $f(x,U)$ bilinear in x and U, the value at risk yields convex constraints. Indeed, suppose that $U \in \mathbf{R}^n$ is distributed normally with mean μ and scaled identity covariance matrix $\sigma^2 I$. Then $x^T U - \gamma \sim \mathsf{N}(\mu^T x - \gamma, \sigma^2 \|x\|_2^2)$, and we have that for $Z = x^T U$, then

$$\mathbf{Prob}(x^T U \leq \gamma) = \mathbf{Prob}(Z \leq \gamma) = \Phi\left(\frac{\gamma - x^T U}{\sigma \|x\|_2}\right)$$

where Φ is the normal c.d.f., and thus

$$\mathbf{VaR}(U^T x - \gamma; \epsilon) \leq 0 \;\text{ if and only if }\; \gamma \geq \mu^T x + \sigma \Phi^{-1}(1-\epsilon) \|x\|_2 ,$$

which is a second order cone constraint if $\epsilon \leq 1/2$.

6.2 Safe convex approximations for chance constraints

Instead of directly using **VaR** or choosing an uncertainty set $\mathcal{U}$ based on the random variable U, we turn to directly constructing convex functions that upper bound the probability of error. For simplicity, let Z be any random variable (where generally, we will have $Z = f(x,U)$). Then if $\phi : \mathbf{R} \to \mathbf{R}_+$ is non-negative and non-decreasing, we have

$$1\,(z \geq 0) \leq \phi(z)$$

for any $z \in \mathbf{R}$, where $1\,(z \geq 0) = 1$ if $z \geq 0$ and 0 otherwise. In particular, we find that for any $\alpha > 0$, we have

$$\mathbf{Prob}(Z \geq 0) \leq \mathbf{E}\,\phi(\alpha^{-1} Z),$$

and thus, if we have

$$\mathbf{E}\,\phi(\alpha^{-1} Z) \leq \epsilon,$$

we are guaranteed that $\mathbf{Prob}(Z \geq 0) \leq \epsilon$. This makes ϕ a *safe* (or conservative) approximation to the probability of $Z \geq 0$, meaning it certifies that $\mathbf{Prob}(Z \geq 0) \leq \epsilon$.

Returning to the convex case, if ϕ is convex and $f(x,U)$ is convex in x, then for all $\alpha > 0$, the function $\phi(\alpha^{-1} f(x,U))$ is convex in x (an increasing convex function of a convex function). In addition, we may pose finding the best such $\alpha > 0$ as a convex problem as well. In particular, for any non-decreasing convex function $\phi : \mathbf{R} \to \mathbf{R}_+$, the constraint

$$\alpha\,\mathbf{E}\,\phi\left(\frac{f(x,U)}{\alpha}\right) \leq \alpha\epsilon$$

ensures that $\mathbf{Prob}(f(x,U) \geq 0) \leq \epsilon$ whenever $\alpha > 0$. Moreover, the perspective function $(w,\alpha) \mapsto \alpha\phi(w/\alpha)$ is jointly convex in w and α for $\alpha > 0$, so that we can optimize over both x and α using convex optimization. We arrive at the convex constraint—which is tightest for our approximating function ϕ—of

$$\inf_{\alpha \geq 0} \left\{ \alpha\,\mathbf{E}\,\phi\left(\frac{f(x,U)}{\alpha}\right) - \alpha\epsilon \right\} \leq 0. \tag{13}$$

We now look at a few special cases, including an extension of these ideas to use moment generating (or cumulant generating) functions to give more analytically tractable procedures.

6.3 Tightest convex bounds and conditional value at risk

We first look at the convex function $\phi(z) = [1+z]_+$, where $[x]_+ = \max\{x, 0\}$, which certainly satisfies $1(z \geq 0) \leq [1+z]_+$. In this case, directly applying the perspective function bound (13), we have that

$$\inf_{\alpha \geq 0} \left\{ \alpha \, \mathbf{E} \left[\frac{f(x,U)}{\alpha} + 1 \right]_+ - \alpha\epsilon \right\} = \inf_{\alpha \geq 0} \left\{ \mathbf{E} \left[f(x,U) + \alpha \right]_+ - \alpha\epsilon \right\} \leq 0 \tag{14}$$

is a safe and convex approximation to the chance constraint $\mathbf{Prob}(f(x,U) > 0) \leq \epsilon$. Often, we replace α with $-\alpha$ above, and use $\mathbf{E}\left[f(x,U) - \alpha\right]_+ + \alpha\epsilon \leq 0$, and we do not need to constrain $\alpha \leq 0$. In fact, functions of the form $[1 + \alpha^{-1} z]_+$, as α ranges between 0 and ∞, are in a sense the tightest safe convex approximations to the chance constraint (see the exercises for more on this).

The constraint (14) is related to the value at risk, and if we define the *conditional value at risk* (sometimes called the *average value at risk*)

$$\mathbf{CVaR}(Z; \epsilon) = \inf_{\alpha} \left\{ \frac{1}{\epsilon} \mathbf{E} \left[Z - \alpha \right]_+ + \alpha \right\},$$

then the constraint (14) is equivalent to $\mathbf{CVaR}(f(x,U);\epsilon) \leq 0$. Moreover, for any convex f, we have $\mathbf{CVaR}(f(x,U);\epsilon)$ is convex in x, and so we have replaced the constraint $\mathbf{VaR}(f(x,U);\epsilon) \leq 0$ with the convex constraint $\mathbf{CVaR}(f(x,U);\epsilon) \leq 0$.

Let us interpret **CVaR** a bit. First, note that we may (at least implicitly) minimize out α in the definition; taking derivatives, we have

$$0 = \frac{\partial}{\partial \alpha} \left\{ \alpha + \frac{1}{\epsilon} \mathbf{E} \left[Z - \alpha \right]_+ \right\} = 1 - \frac{1}{\epsilon} \mathbf{E} \, 1(Z \geq \alpha) = 1 - \frac{1}{\epsilon} \mathbf{Prob}(Z \geq \alpha).$$

In particular, if we set $\alpha^\star$ to be such that $\epsilon = \mathbf{Prob}(Z \geq \alpha^\star)$,[3] then $0 = 1 - \frac{1}{\epsilon} \mathbf{Prob}(Z \geq \alpha^\star)$, and we have

$$\mathbf{CVaR}(Z;\epsilon) = \frac{1}{\epsilon} \mathbf{E} \left[Z - \alpha^\star \right]_+ + \alpha^\star = \frac{1}{\epsilon} \mathbf{E} \left[Z - \alpha^\star \right]_+ + \mathbf{VaR}(Z;\epsilon).$$

So we see that conditional value at risk always upper bounds the value at risk, and gives an additional penalty for deviations above the specified level $\alpha^\star$; minimizing conditional value at risk consists of minimizing the upper ϵ quantile of the random Z plus a deviation penalty. Additionally, we also have that the *conditional tail expectation*

$$\begin{aligned} \mathbf{E}[Z \mid Z \geq \alpha^\star] &= \mathbf{E}[\alpha^\star + (Z - \alpha^\star) \mid Z \geq \alpha^\star] \\ &= \alpha^\star + \frac{\mathbf{E}\left[Z - \alpha^\star\right]_+}{\mathbf{Prob}(Z \geq \alpha^\star)} = \alpha^\star + \frac{1}{\epsilon} \mathbf{E} \left[Z - \alpha^\star \right]_+ = \mathbf{CVaR}(Z;\epsilon). \end{aligned}$$

[3] More generally, we set $\alpha^\star = \inf\{\alpha \mid \mathbf{Prob}(Z \geq \alpha) \leq \epsilon\} = \mathbf{VaR}(Z;\epsilon)$, which guarantees that for any $\alpha' > \alpha^\star$, we have $\epsilon - \mathbf{E}\, 1(Z \geq \alpha') > \epsilon - \epsilon > 0$ and for $\alpha' < \alpha^\star$ we have $\epsilon - \mathbf{E}\, 1(Z \geq \alpha') < \epsilon - \epsilon < 0$, so that $\alpha^\star$ is the minimizer in the definition of **CVaR**.

One difficulty of performing optimization using conditional value at risk is that, while $\mathbf{CVaR}(f(x,U);\epsilon)$ is convex in x whenever f is, it is very rare that we can give an analytic formula for $\mathbf{E}\left[f(x,U)-\alpha\right]_+$. For example, even if $f(x,U)=x^TU$ and the vector $U\in\{-1,1\}^n$ consists of independent coordinates taking only two values, $\mathbf{E}\left[U^Tx-\alpha\right]_+$ is a combinatorial sum. In practice, one usually performs a Monte Carlo simulation to approximate the expectation $\mathbf{E}\left[f(x,U)-\alpha\right]_+$, using the empirical expectation as a proxy for the truth. While simple, this too can become somewhat expensive for large problems, which justifies other types of approximation.

6.4 Analytic approximation using moment generating functions

The function $\phi(z)=[1+z]_+$ in the previous section gives the sharpest convex bounds on $\mathbf{Prob}(Z>0)=\mathbf{E}\,1\,(Z>0)$, but in some cases different upper bounds are useful. Here we consider using $\phi(z)=e^z$, which clearly satisfies $1\,(z>0)\le\phi(z)$. This gives the usual Chernoff bound, that is, $\mathbf{Prob}(Z>t)\le\mathbf{E}\exp(\lambda Z-\lambda t)$ for all $\lambda\ge0$, and in this case, we see that

$$\mathbf{E}\exp\left(\alpha^{-1}f(x,U)\right)\le\epsilon$$

is a safe approximation to the chance constraint $\mathbf{Prob}(f(x,U)>0)\le\epsilon$. This approximation is somewhat numerically unstable, though, so we explore bounds involving the moment generating function a bit more. We first do so abstractly, then specialize to a few more concrete examples using the probabilistic tools outlined in § 7.

If the random function $f(x,U)$ is convex in x for all U, then the generalized log-sum-exp function $\log\mathbf{E}\exp(f(x,U))$ is convex in x. Indeed, for any $\lambda\in[0,1]$ we have

$$\begin{aligned}
\log\mathbf{E}\exp\left(f(\lambda x+(1-\lambda)y,U)\right) &\le \log\mathbf{E}\exp\left(\lambda f(x,U)+(1-\lambda)f(y,U)\right)\\
&= \log\mathbf{E}\left[\exp(f(x,U))^{\lambda}\exp(f(y,U))^{1-\lambda}\right]\\
&\le \log\left[\left(\mathbf{E}\exp(f(x,U))\right)^{\lambda}\left(\mathbf{E}\exp(f(y,U))\right)^{1-\lambda}\right] \qquad (15)\\
&= \lambda\log\mathbf{E}\,e^{f(x,U)}+(1-\lambda)\log\mathbf{E}\,e^{f(y,U)}.
\end{aligned}$$

Here inequality (15) is Hölder's inequality, that is, that $a^Tb\le\|a\|_p\|b\|_q$ when $\frac{1}{p}+\frac{1}{q}=1$ taken with $p=1/\lambda$ and $q=1/(1-\lambda)$. Thus, we find that the cumulant generating function (log of the moment generating function)

$$\log\mathbf{E}\exp\left(\frac{f(x,U)}{\alpha}\right)\le\log\epsilon$$

is a safe approximation for the chance constraint $\mathbf{Prob}(f(x,U)>0)\le\epsilon$, for any $\alpha>0$. Choosing the best possible α using the perspective function, we have the conservative approximation

$$\inf_{\alpha\ge0}\left\{\alpha\log\mathbf{E}\exp\left(\frac{f(x,U)}{\alpha}\right)+\alpha\log\frac{1}{\epsilon}\right\}\le0. \qquad (16)$$

The formulation (16) may not be useful if the moment generating function $\mathbf{E}\exp(f(x,U))$ is hard to compute exactly, but if we have any convex upper bounds $\psi(x) \geq \log \mathbf{E}\exp(f(x,U))$ convex in x, we can still have useful results. We illustrate a few examples, relying on so-called *sub-Gaussian* random variables and focusing on uncertain linear constraints, that is, when $f(x,U) = U^T x$ for some random vector U. In particular, let us assume that $U = (U_1, \ldots, U_n)$ is mean-zero with independent components, and for any $\lambda \in \mathbf{R}$, we have

$$\mathbf{E}\exp(\lambda U_i) \leq \exp\left(\frac{\lambda^2\sigma_i^2}{2}\right)$$

for some $\sigma_i \geq 0$. This holds with equality for normally distributed variables $U_i \sim \mathsf{N}(0, \sigma_i^2)$, and this inequality also holds for bounded random variables U_i, that is, if $\mathbf{E}\,U_i = 0$ and $U_i \in [-\sigma_i, \sigma_i]$, we have $\log \mathbf{E}\, e^{\lambda U_i} \leq \frac{\lambda^2\sigma_i^2}{2}$. (See inequality (17) in Example 10 to follow.) In particular, if we have such an inequality, we obtain that if $f(x,t;U) = x^T U - t$, then

$$\log \mathbf{E}\exp\left(\frac{f(x,t;U)}{\alpha}\right) = \log \mathbf{E}\exp\left(\frac{x^T U}{\alpha} - \frac{t}{\alpha}\right) \leq \sum_{i=1}^n \frac{x_i^2\sigma_i^2}{2\alpha^2} - \frac{t}{\alpha}.$$

Rewriting, we see that inequality (16) is satisfied if

$$\inf_{\alpha \geq 0}\left\{\frac{1}{2\alpha}\sum_{i=1}^n x_i^2\sigma_i^2 - t + \alpha\log\frac{1}{\epsilon}\right\} = \sqrt{2\log\frac{1}{\epsilon}}\left\|\mathbf{diag}(\sigma)x\right\|_2 - t \leq 0.$$

We now give two examples of this approach.

Example 5 *Safe moment-generating function approximations with bounded random variables.* Suppose we have $f(x,U) = U^T x$, where the random variables U_i are independent and satisfy $\mathbf{E}\,U_i = \mu_i$ and $U_i \in [a_i, b_i]$. Then a safe approximation to $\mathbf{Prob}(U^T x > 0) \leq \epsilon$ is given by

$$\mu^T x + \sqrt{\frac{1}{2}\log\frac{1}{\epsilon}}\left\|\mathbf{diag}(b-a)x\right\|_2 \leq 0,$$

because $\log \mathbf{E}\exp(x^T U) = \log \mathbf{E}\exp(x^T(U-\mu) + x^T\mu) \leq \frac{1}{8}\sum_{i=1}^n (b_i - a_i)^2 x_i^2 + \mu^T x$.

Example 6 *Analytic versus Monte Carlo approximations.* Let us revisit the portfolio problem, where we set the returns $R \in \mathbf{R}^n$ to have means $\mu \in \mathbf{R}^n_+$ with $\mu_i = 1.05 + .3/i$, where the random values $R_i \in [\mu_i - .6/i, \mu_i + .6/i]$, so that the riskiest assets have the highest expected returns. Letting $\sigma_i = 1.2/i$ be the range of the R_i, we compare the analytic approximation from the moment generating function as in Example 5 with a Monte Carlo approximation to the **CVaR** constraint, that is, we maximize t subject to the constraints $x \succeq 0$, $\mathbf{1}^T x = 1$, and

$$\inf_{\alpha \geq 0}\mathbf{E}\left[t - R^T x + \alpha\right]_+ - \alpha\epsilon \leq 0 \quad \text{or} \quad t - \mu^T x + \sqrt{\frac{1}{2}\log\frac{1}{\epsilon}}\left\|\mathbf{diag}(\sigma)x\right\|_2 \leq 0,$$

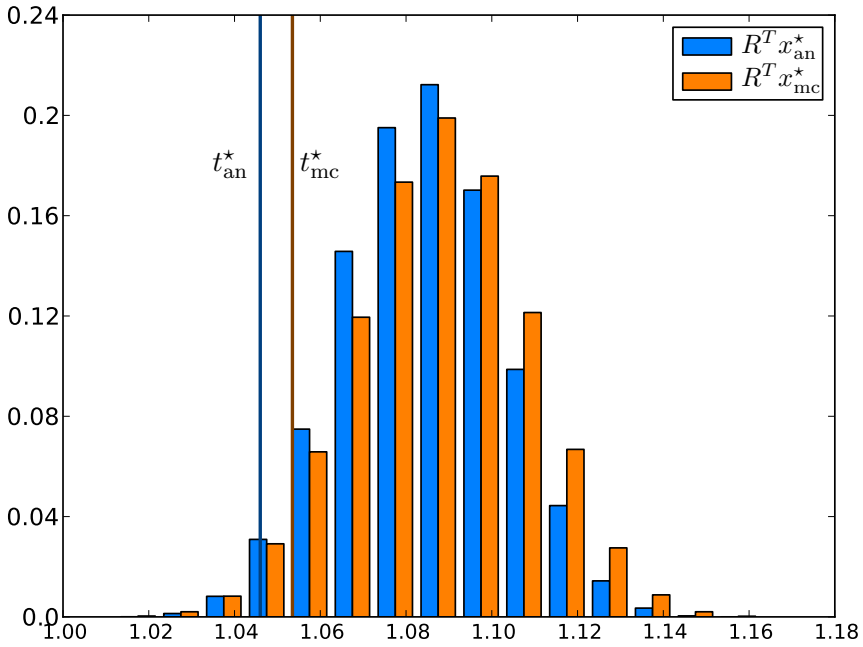

Figure 5: Performance on a portfolio optimization problem comparing Monte Carlo approximation of a chance constraint using conditional value at risk to that using the analytic SOCP derived from moment generating function bounds.

using a Monte Carlo approximation to the leftmost expectation. In Figure 5, we plot the results of this approach in $n = 15$ dimensions and using $m = 2 \cdot 10^3$ random samples R to approximate the **CVaR** constraint, where we use $\epsilon = .1$ The vertical lines represent the values $t^\star_{\text{mc}}$ and $t^\star_{\text{an}}$, the Monte Carlo and analytic lower bounds on the portfolio returns. The resulting solutions violated the constraints $R^T x^\star \leq t^\star$ with probability approximately .02 with the analytic solution and with probability approximately .06 for the Monte Carlo solution. We see that the analytic approximation with moment generating functions is more conservative than the Monte Carlo-based solution using the tighter conditional value at risk.

Example 7 *Safe moment-generating function approximations with normal random variables.* We revisit Example 4, which was exact for the normal distribution with variables $U_i \sim \mathsf{N}(\mu_i, \sigma_i^2)$. The value at risk guarantee is that

$$\mathbf{Prob}(U^T x > 0) \leq \epsilon \;\text{ iff }\; \mathbf{VaR}(U^T x; \epsilon) \leq 0 \;\text{ iff }\; 0 \geq \mu^T x + \Phi^{-1}(1-\epsilon) \left\|\mathbf{diag}(\sigma) x\right\|_2 .$$

The moment generating function approximation (16) gives constraint

$$\mu^T x + \sqrt{2 \log \frac{1}{\epsilon}} \left\|\mathbf{diag}(\sigma) x\right\|_2 \leq 0.$$

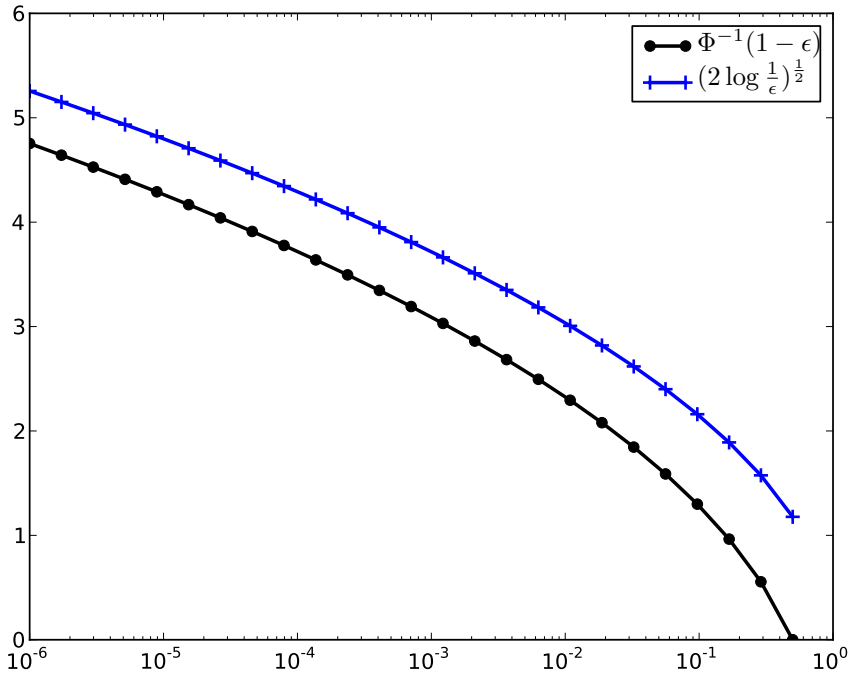

Figure 6: Approximation to value at risk using moment generating function bound for normal distribution.

In Figure 6, we plot $\Phi^{-1}(1-\epsilon)$ against $\sqrt{2\log\frac{1}{\epsilon}}$. While $\sqrt{2\log\frac{1}{\epsilon}}$ broadly tracks the true inverse CDF for the normal, we see that for reasonable ϵ (say $\epsilon \geq 10^{-2}$), the moment generating function bound is substantially more conservative than the exact bound.

---

7 Probability and tail bounds

In this section, we review several bounds on moment generating functions and tail inequalities for probability distributions that prove useful for construction of uncertainty sets and for approximation of chance constraints.

The quantity of our interest is (for the most part) a probability of the form

$$\mathbf{Prob}\left[\sum_{i=1}^{n} z_i U_i \geq t\right] \leq \epsilon,$$

where U_i are random variables and z_i are fixed parameters (that generally depend on our optimization variables). We investigate a few different techniques for controlling this probability, including choosing a set $\mathcal{U}$ such that $(U_1, \ldots, U_n) \in \mathcal{U}$ with probability at least $1-\epsilon$ or

finding a t directly such that $U^T z < t$ with probability at least $1 - \epsilon$. To this end, a variety of tail bounds may prove useful.

The simplest tail bound for random variables is Markov's inequality, that is, for a non-negative random variable U and $a \geq 0$,

$$\mathbf{Prob}(U \geq a) \leq \frac{\mathbf{E}[U]}{a}.$$

From this, we obtain the familiar Chebyshev bound, which is that

$$\mathbf{Prob}(U \geq \mathbf{E}\, U + t) \leq \mathbf{Prob}((U - \mathbf{E}\, U)^2 \geq t^2) \leq \frac{\mathbf{E}(U - \mathbf{E}\, U)^2}{t^2}.$$

For obtaining bounds on extremely low probability events, a common technique is the Chernoff bound, which also follows from Markov's inequality: for any random variable U, and t and any $\lambda > 0$, we have

$$\mathbf{Prob}(U \geq t) = \mathbf{Prob}(e^{\lambda U} \geq e^{\lambda t}) \leq \frac{\mathbf{E}\, e^{\lambda U}}{e^{\lambda t}}, \quad \text{or} \quad \mathbf{Prob}(U \geq t) \leq \inf_{\lambda \geq 0} \mathbf{E}\, e^{\lambda U} e^{-\lambda t}.$$

The key in this technique is to carefully upper bound the moment generating function $\mathbf{E}\, e^{\lambda U}$ of U, though it is very effective for bounding probabilities of deviation for sums of random variables.

Example 8 Let Z be a mean-zero Gaussian random variable with variance σ^2. Then

$$\mathbf{E} \exp(\lambda Z) = \exp\left(\frac{\lambda^2 \sigma^2}{2}\right),$$

and consequently for $t \geq 0$, we have

$$\mathbf{Prob}(Z \geq t) \leq \inf_{\lambda \geq 0} \exp\left(\frac{\lambda^2 \sigma^2}{2} - \lambda t\right) = \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

Whenever there is a parameter $\sigma \geq 0$ such that a mean-zero random variable satisfies $\mathbf{E}\, e^{\lambda U} \leq \exp(\frac{\lambda^2 \sigma^2}{2})$, we call the random variable *sub-Gaussian* with parameter σ^2; this is an extremely useful concept.

Let us give a few more examples of sub-Gaussian random variables, and show how they give useful tail bounds over sums of random variables. Indeed, we have Hoeffding's inequality, which gives sharp bounds on sums of sub-Gaussian random variables.

Example 9 *Hoeffding's inequality.* Let $U_1, \ldots, U_n$ be independent mean-zero sub-Gaussian random variables with parameters $\sigma = (\sigma_1, \ldots, \sigma_n)$. Then

$$\mathbf{Prob}\left(\sum_{i=1}^n U_i \geq t\right) \leq \exp\left(-\frac{t^2}{2\left\|\sigma\right\|_2^2}\right).$$

To see this, we use the Chernoff bound: letting $Z = \mathbf{1}^T U$, we have

$$\mathbf{Prob}(Z \geq t) \leq \exp\left(\frac{\lambda^2 \sum_{i=1}^n \sigma_i^2}{2} - \lambda t\right),$$

and minimizing over $\lambda \geq 0$ gives the tail bound.

Another useful result is that bounded random variables are sub-Gaussian, which means they have good concentration properties.

Example 10 *Bounded random variables (Hoeffding's lemma).* Let $U \in [a, b]$ be a mean zero random variable, so that $a \leq 0 \leq b$. Then U is sub-Gaussian, and moreover,

$$\mathbf{E}[\exp(\lambda U)] \leq \exp\left(\frac{\lambda^2 (b-a)^2}{8}\right). \tag{17}$$

With inequality (17) in hand, we see that if $U_1, \ldots, U_n$ are independent random variables, each bounded in $[a_i, b_i]$, then by the previous example we have

$$\mathbf{Prob}\left(\sum_{i=1}^n U_i \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right),$$

which is the classical Hoeffding inequality.

To see inequality (17), by the convexity of the exponential, we know that for $x \in [a, b]$, we have

$$e^{\lambda x} \leq e^{\lambda a} \frac{b-x}{b-a} + e^{\lambda b} \frac{x-a}{b-a}.$$

Taking expectations over the random variable U, which has mean zero, we thus find that

$$\mathbf{E}[e^{\lambda U}] \leq \frac{b}{b-a} e^{\lambda a} + \frac{-a}{b-a} e^{\lambda b}.$$

Defining $z = \lambda(b-a)$ and $p = \frac{b}{b-a}$, we obtain that $pz = \lambda b$ and $(1-p)z = -\lambda a$, so that

$$\mathbf{E}[e^{\lambda U}] \leq p e^{\lambda a} + (1-p) e^{\lambda b} = e^{(p-1)z} \left[p + (1-p) e^z\right].$$

Let $f(z) = (p-1)z + \log(p + (1-p)e^z)$. Then

$$f'(z) = (p-1) + \frac{1-p}{pe^{-z} + 1 - p} \quad \text{and} \quad f''(z) = \frac{p(1-p)e^{-z}}{(pe^{-z} + 1 - p)^2} \leq \frac{1}{4},$$

so that $f(0) = 0$, $f'(0) = 0$, by Taylor's theorem we have $f(z) \leq \frac{1}{8} z^2$. Substitute $z = \lambda(b-a)$ for the result.

In the case that $b = -a$, we can give a much simpler proof: we have

$$\begin{aligned}\mathbf{E}[e^{\lambda U}] \leq \frac{b}{2b} e^{-\lambda b} + \frac{b}{2b} e^{\lambda b} &= \sum_{k \in 2\mathbf{Z}_+} \frac{\lambda^k b^k}{k!} = 1 + \sum_{k \geq 1} \frac{\lambda^{2k} b^{2k}}{(2k)!} \\ &\leq 1 + \sum_{k \geq 1} \left(\frac{\lambda b}{2}\right)^2 \frac{1}{k!} = \exp\left(\frac{\lambda^2 b^2}{2}\right).\end{aligned}$$

For sub-Gaussian random variables, it is very natural to use these tail bounds to give uncertainty sets, or approximations to chance/uncertainty constraints. Indeed, expanding on Example 2, suppose that we have mean-zero random variables U_i and would like to guarantee that for parameters $x \in \mathbf{R}^n$ and some parameter t,

$$\mathbf{Prob}(U^T x > t) \le \epsilon.$$

That is, the probability of failing to satisfy the constraint that $U^T x \le t$ is at most ϵ. If the U_i are σ_i^2-sub-Gaussian, then the standard Hoeffding bounds guarantee that

$$\mathbf{Prob}(U^T x > t) \le \exp\left(-\frac{t^2}{2\sum_{i=1}^n \sigma_i^2 x_i^2}\right)$$

as $U_i x_i$ is mean-zero and satisfies $\log \mathbf{E}\, e^{\lambda U_i x_i} \le \frac{\lambda^2 x_i^2 \sigma_i^2}{2}$. In particular, if we replace t with $\sqrt{2\log\frac{1}{\epsilon}}\left\|\mathbf{diag}(\sigma)x\right\|_2$, then

$$U^T x \le \sqrt{2\log\frac{1}{\epsilon}}\left\|\mathbf{diag}(\sigma)x\right\|_2 \quad \text{with probability } \ge 1-\epsilon.$$

That is, second order cone constraints arise naturally out of our probabilistic considerations. If we were to directly guarantee that $U_i \in \mathcal{U}_i$ for each i and uncertainty sets $\mathcal{U}_i$, we would necessarilty choose substantially more conservative bounds, as in § 3.4.

A The S Lemma

The homogeneous S lemma is the following result: given matrices $A, B \in \mathbf{S}^n$ and assuming there exists some $x \in \mathbf{R}^n$ such that $x^T A x > 0$, we have

$$x^T A x \ge 0 \text{ implies } x^T B x \ge 0 \quad \text{if and only if} \quad \exists\, \lambda \ge 0 \text{ s.t. } B \succeq \lambda A. \tag{18}$$

A proof of this fact is actually not too difficult. For one direction, suppose that $B \succeq \lambda A$ for some $\lambda \ge 0$. Then it is clear that $x^T B x \ge \lambda x^T A x \ge 0$ any time $x^T A x \ge 0$. The converse is substantially more challenging, and is a consequence of semidefinite duality. We give an argument using a simple randomization idea (*cf.* [BTGN09]) to prove this fact. Consider the semidefinite program

$$\begin{array}{ll} \text{minimize} & \mathbf{Tr}(BX) \\ \text{subject to} & \mathbf{Tr}(AX) \ge 0,\ X \succeq 0,\ \mathbf{Tr}(X) = 1. \end{array}$$

Strong duality obtains for this problem, as it is strictly feasible (by assumption, we may take $X = xx^T/\left\|x\right\|_2^2$ for the vector x such that $x^T A x > 0$). Moreover, there exists an optimal solution $X^\star$. As a consequence, writing the Lagrangian in variables $X, Z \succeq 0$, $\lambda \ge 0$, and $\theta \in \mathbf{R}$, we have

$$L(X, Z, \lambda, \theta) = \mathbf{Tr}(BX) - \lambda\,\mathbf{Tr}(AX) - \mathbf{Tr}(ZX) + \theta(\mathbf{Tr}(X) - 1),$$

and so the dual function

$$\inf_X L(X, Z, \lambda, \theta) = \begin{cases} -\infty & \text{if } B - \lambda A - Z + \theta I \neq 0 \\ -\theta & \text{otherwise.} \end{cases}$$

In particular, we see that if the optimal value $\mathbf{Tr}(BX^\star) = -\theta^\star \geq 0$, then at optimum, we have $B = \lambda A + Z - \theta^\star I \succeq \lambda A$, which proves the implication (18).

Let $\bar{A} = (X^\star)^{\frac{1}{2}} A (X^\star)^{\frac{1}{2}}$ and $\bar{B} = (X^\star)^{\frac{1}{2}} B (X^\star)^{\frac{1}{2}}$, and let $\bar{A} = U \Lambda U^T$ be the eigendecomposition of $\bar{A}$. In addition, let $Z \in \{-1, 1\}^n$ be a mean-zero random vector of independent signs. Then

$$(UZ)^T \bar{A} (UZ) = (UZ)^T U \Lambda U^T U Z = Z^T \Lambda Z = \sum_{i=1}^n \Lambda_{ii} = \mathbf{Tr}(\bar{A}) = \mathbf{Tr}(AX^\star) \geq 0.$$

As a consequence, for any vector $Z \in \{-1, 1\}^n$, we have by the implication $x^T A x \geq 0 \Rightarrow x^T B x \geq 0$ that $(UZ)^T \bar{B} (UZ) \geq 0$. In particular, taking expectations,

$$\begin{aligned} 0 \leq \mathbf{E}(UZ)^T \bar{B} UZ &= \mathbf{E}\,\mathbf{Tr}(Z^T U^T \bar{B} U Z) = \mathbf{E}\,\mathbf{Tr}(U^T \bar{B} U Z Z^T) \\ &= \mathbf{Tr}(\bar{B} U U^T) = \mathbf{Tr}(\bar{B}) = \mathbf{Tr}(BX^\star), \end{aligned}$$

where we have used that $\mathbf{E}\, ZZ^T = I$. We see that $\mathbf{Tr}(BX^\star) \geq 0$, giving the result.

References

[BTGN09] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[Hås01] Johan Håstad. Some optimal inapproximability results. *Journal of the Association for Computing Machinery*, 48(4):798–859, 2001.

[RU00] R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000.

[SDR09] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM and Mathematical Programming Society, 2009.

[Ver12] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications*, chapter 5, pages 210–268. Cambridge University Press, 2012.