Bounding the Duality Gap for Problems with Separable Objective

Madeleine Udell and Stephen Boyd

March 8, 2014

Abstract

We consider the problem of minimizing a sum of non-convex functions over a compact domain, subject to linear inequality and equality constraints. We consider approximate solutions obtained by solving a convexified problem, in which each function in the objective is replaced by its convex envelope. We propose a randomized algorithm to solve the convexified problem which finds an ϵ-suboptimal solution to the original problem. With probability 1, ϵ is bounded by a term proportional to the number of constraints in the problem. The bound does not depend on the number of variables in the problem or the number of terms in the objective. In contrast to previous related work, our proof is constructive, self-contained, and gives a bound that is tight.

1 Problem and results

The problem. We consider the optimization problem

$$\begin{array}{ll} \text{minimize} & f(x) = \sum_{i=1}^n f_i(x_i) \\ \text{subject to} & Ax \leq b \\ & Gx = h, \end{array} \qquad (\mathcal{P})$$

with variable $x = (x_1, \ldots, x_n) \in \mathbf{R}^N$, where $x_i \in \mathbf{R}^{n_i}$, with $\sum_{i=1}^n n_i = N$. There are m_1 linear inequality constraints, so $A \in \mathbf{R}^{m_1 \times N}$, and m_2 linear equality constraints, so $G \in \mathbf{R}^{m_2 \times N}$. The optimal value of $\mathcal{P}$ is denoted $p^\star$. The objective function terms are lower semi-continuous on their domains:

$f_i : S_i \to \mathbf{R}$, where $S_i \subset \mathbf{R}^{n_i}$ is a compact set. We say that a point x is *feasible* (for $\mathcal{P}$) if $Ax \leq b$, $Gx = h$, and $x_i \in S_i$, $i = 1, \ldots, n$. We say that $\mathcal{P}$ is feasible if there is at least one feasible point. In what follows, we assume that $\mathcal{P}$ is feasible.

Linear inequality or equality constraints that pertain only to a single block of variables x_i can be expressed implicitly by modifying S_i, so that $x_i \notin S_i$ when the constraint is violated. Without loss of generality, we assume that this transformation has been carried out, so that each of the remaining linear equality or inequality constraints involves at least two blocks of variables. This reduces the total number of constraints $m = m_1 + m_2$; we will see later why this is advantageous. Since each of the linear equality or inequality constraints involves at least two blocks of variables, they are called *complicating constraints*. Thus m represents the number of complicating constraints, and can be interpreted as a measure of difficulty for the problem.

We make no assumptions about the convexity of the functions f_i or the convexity of their domains S_i, so that in general the problem is hard to solve (and even NP-hard to approximate [UB13]).

Convex envelope. For each f_i, we let $\hat{f}_i$ denote its *convex envelope*. The convex envelope $\hat{f}_i : \mathbf{conv}(S_i) \to \mathbf{R}$ is the largest closed convex function majorized by f_i, *i.e.*, $f_i(x) \geq \hat{f}_i(x)$ for all x [Roc97, Theorem 17.2]. In §5, we give a number of examples in which we compute $\hat{f}_i$ explicitly.

Nonconvexity of a function. Define the *nonconvexity* $\rho(f)$ of a function $f : S \to \mathbf{R}$ to be

$$\rho(f) = \sup_x (f(x) - \hat{f}(x)),$$

where for convenience we define a function to be infinite outside of its domain and interpret $\infty - \infty$ as 0. Evidently $\rho(f) \geq 0$, and $\rho(f) = 0$ if and only if f is convex and closed. The nonconvexity ρ is finite if f is bounded and lower semi-continuous and S is compact and convex. For later use, we define $\rho_{[i]}$ to be the ith largest of the nonconvexities $\rho(f_1), \ldots, \rho(f_n)$.

Convexified problem. Now, consider replacing each f_i by $\hat{f}_i$ to form a convex problem,

$$\begin{array}{ll} \text{minimize} & \hat{f}(x) = \sum_{i=1}^n \hat{f}_i(x_i) \\ \text{subject to} & Ax \leq b \\ & Gx = h, \end{array} \qquad (\hat{\mathcal{P}})$$

with optimal value $\hat{p}$. This problem is convex and is easy to solve, assuming we can evaluate $\hat{f}$ and a subgradient (or derivative, if $\hat{f}$ is differentiable). Furthermore, $\hat{\mathcal{P}}$ is feasible as long as $\mathcal{P}$ is feasible. Evidently $\hat{p} \leq p^\star$; that is, the optimal value of the convexified problem is a lower bound on the optimal value of the original problem. We would like to know when a solution to $\hat{\mathcal{P}}$ approximately solves $\mathcal{P}$.

Our main result is the following.

Theorem 1. *There exists a solution $\hat{x}$ of $\hat{\mathcal{P}}$ such that*

$$f(\hat{x}) \leq \hat{p} + \sum_{i=1}^{\min(m,n)} \rho_{[i]}.$$

Since $p^\star \leq f(\hat{x})$ and $\hat{p} \leq p^\star$, Theorem 1 implies that

$$p^\star \leq f(\hat{x}) \leq p^\star + \sum_{i=1}^{\min(m,n)} \rho_{[i]}.$$

In other words, there is a solution of the convexified problem that is ϵ-suboptimal for the original problem, with $\epsilon = \sum_{i=1}^{\min(m,n)} \rho_{[i]}$. It is not true (as we show in §2.1) that all solutions of the convexified problem are ϵ-suboptimal.

Theorem 1 shows that if the objective function terms are not too nonconvex, and the number of constraints is not too large, then the convexified problem has a solution that is not too suboptimal for the original problem. This theorem is similar to a number of results previously in the literature; for example, it can be derived from the well-known Shapley-Folkman theorem [Sta69]. A slightly looser version of this theorem may be obtained from the bound on the duality gap given in [AE76].

Theorem 1 also implies a bound on the duality gap for problems with separable objectives. Define the dual problem to $\mathcal{P}$,

$$\begin{array}{ll} \text{maximize} & \inf_x \mathcal{L}(x, \lambda) = \left(\sum_{i=1}^n f_i(x_i) + \lambda^T(Ax - b) + \mu^T(Gx - h)\right) \\ \text{subject to} & \lambda \geq 0, \end{array} \tag{$\hat{\mathcal{D}}$}$$

with optimal value $g^\star$. The convexified problem $\hat{\mathcal{P}}$ is the dual of $\hat{\mathcal{D}}$. Since $\hat{\mathcal{P}}$ is convex and feasible, with only linear constraints, strong duality holds by Slater's constraint qualification, and the maximum of the dual problem is attained, *i.e.*, $g^\star = \hat{p}$ and $\inf_x \mathcal{L}(x, \lambda^\star) = g^\star$ for some $\lambda^\star > 0$ [BV04]. The bound from Theorem 1 thus implies

$$p^\star - g^\star \leq \sum_{i=1}^{\min(m,n)} \rho_{[i]}.$$

What is not clear in other related work is how to construct a feasible solution that satisfies this bound. This observation leads us to the main contribution of this paper.

Theorem 2. *Let $w \in \mathbf{R}^N$ be a random variable with uniform distribution on the unit sphere, and let $(x^\star, \lambda^\star)$ denote any primal-dual optimal pair for the convexified problem $\hat{\mathcal{P}}$. Now consider the feasible convex problem*

$$\begin{array}{ll} \textit{minimize} & w^T x \\ \textit{subject to} & Ax = Ax^\star \\ & Gx = h \\ & \hat{f}(x) + \lambda^{\star T} Ax \leq \hat{f}(x^\star) + \lambda^{\star T} Ax^\star. \end{array} \qquad (\mathcal{R})$$

Then with probability one, $\mathcal{R}$ has a unique solution $\hat{x}$ which satisfies the inequality of Theorem 1,

$$f(\hat{x}) \leq \hat{p} + \sum_{i=1}^{\min(m,n)} \rho_{[i]},$$

i.e., $\hat{x}$ is ϵ-suboptimal for the orignal problem.

The problem $\mathcal{R}$ has a simple interpretation. Any feasible point x for $\mathcal{R}$ is optimal for $\hat{\mathcal{P}}$, and the constraint $\hat{f}(x) \leq \hat{p}$ is satisfied with equality. In problem $\mathcal{R}$ we minimize a random linear function over the optimal set of $\hat{\mathcal{P}}$. Theorem 2 tells us that this construction yields (almost surely) an ϵ-suboptimal solution of $\mathcal{P}$.

We give a self-contained proof of both of these theorems in §6.2.

2 Discussion

2.1 Mathematical examples

In this section, the results of which we will not use in the sequel, we show that the bound in Theorem 1 is tight, and that randomization is essential to achieving the bound in Theorem 1.

The bound is tight. Consider the problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n g(x_i) \\ \text{subject to} & \sum_{i=1}^n x_i \leq B \end{array} \tag{1}$$

with $g : [0, 1] \to \mathbf{R}$ defined as

$$g(x) = \begin{cases} 1 & 0 \leq x < 1 \\ 0 & x = 1 \end{cases}.$$

The convex envelope $\hat{g} : [0, 1] \to \mathbf{R}$ of g is given by $\hat{g}(x) = 1 - x$, with $\rho(g) = 1$. The convexified problem $\hat{\mathcal{P}}$ corresponding to (1) is

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n \hat{g}(x_i) \\ \text{subject to} & \sum_{i=1}^n x_i \leq B \\ & 0 \leq x. \end{array} \tag{2}$$

Any $\hat{x}$ satisfying $0 \leq \hat{x} \leq 1$ and $\sum_{i=1}^n \hat{x}_i = B$ is optimal for the convexified problem (2), with value $\hat{p} = n - B$. If $B < 1$, then the optimal value of (1) is $p^\star = n$. Since (1) has only one constraint, the bound from Theorem 1 applied to this problem gives

$$n = p^\star \leq \sum_{i=1}^n g(\hat{x}_i) \leq \hat{p} + \rho(g) = n - B + 1.$$

Letting $B \to 1$, we see that the bound cannot be improved.

Find the extreme points. Not all solutions to the convexified problem satisfy the bound from Theorem 1. As we show in §6, the value of the convex envelope at the extreme points of the optimal set for the convexified problem will be provably close to the value of the original function, whereas the difference between these values on the interior of the optimal set may be arbitrarily large.

For example, suppose $n - 1 < B < n$ in the problem defined above. As before, the optimal set for the convexified problem (2) is

$$M = \{x : \textstyle\sum_{i=1}^n x_i = B,\ x_i \geq 0,\ i = 1, \ldots, n\}.$$

Consider $\tilde{x} \in M$ with $\tilde{x}_i = B/n$, $i = 1, \ldots, n$, which is optimal for the convexified problem (2). This $\tilde{x}$ does *not* obey the bound in Theorem 1. With this $\tilde{x}$, the left hand side of the inequality in Theorem 1 is $\sum_{i=1}^{n} g(\tilde{x}_i) = n$, while the right hand side $\hat{p} + \rho(g) = n - B + 1 < 2$ is much smaller. On the other hand, $\hat{x} \in M$ defined by

$$\hat{x}_i = \begin{cases} 1 & i = 1, \dots, n-1 \\ B - (n-1) & i = n \end{cases},$$

which is an extreme point of the optimal set for the convexified problem, is optimal for the original problem as well. That is, $\hat{x}$, an extreme point of M, obeys Theorem 1 with equality.

For an even simpler example, consider the following univariate problem with no constraints. Let $S = \{0\} \cup \{1\}$ with $f(x) = 0$ for $x \in S$. Then $\hat{f} : [0, 1] \to \{0\}$, so the optimal set for the convexified problem consists of the entire interval $[0, 1]$. But $\tilde{x} = 1/2 \in M$ is not feasible for the original problem; its value according to the original objective is thus infinitely worse than the value guaranteed by Theorem 1. On the other hand, $x = 0$ and $x = 1$, the extreme points of the optimal set for the convexified problem, are indeed optimal for the original problem.

3 Related work

Our proof is very closely related to the Shapley-Folkman theorem [Sta69], which states, roughly, that the nonconvexity of the average of a number of nonconvex sets decreases with the number of sets. In optimization, the analogous statement is that optimizing the average of a number of functions is not too different from optimizing the average of the convex envelopes of those functions, and the difference decreases with the number of functions. However, we note that using the Shapley-Folkman theorem directly, rather than its optimization analogue, results in a bound that is slightly worse. For example, the Shapley-Folkman theorem has previously been used by Aubin and Ekeland in [AE76] to prove a bound on the duality gap. The bound they present,

$$p^\star - d^\star \leq \min(m + 1, n)\rho_{[1]},$$

is not tight; our bound is smaller by a factor of $(m + 1)/m$. The Shapley-Folkman theorem has also been used by Bertsekas et al [BLNP83] to solve a unit commitment problem in electrical power system scheduling, in which case the terms in the objective are univariate. Many authors [LR01, Ber82, Ber99] have also studied convexifications of the constraints as well as of the objective in separable problems.

The use of randomization to find approximate solutions to nonconvex problems is widespread, and often startlingly successful [Mot95, GW95]. The usual approach is to solve a convex problem to find an optimal probability distribution over possible solutions; sampling from the distribution and rounding yields the desired result. By contrast, our approach uses randomization only to explore the geometry of the optimal set [SB10]. We rely on the insight that extremal points of the epigraph of the convex envelope are likely to be closer in value to the original function, and use randomization simply to reach these points. Randomization allows us to find "simplex-style" corner points of the optimal set as solutions, rather than accepting interior points of the set.

4 Constructing the convex envelope

In this section, we give a few examples illustrating how to construct the convex envelope for a number of interesting functions and types of functions.

Sigmoidal functions. A continuous function $f : [l, u] \to \mathbf{R}$ is defined to be *sigmoidal* if it is either convex, concave, or convex for $x \le z \in [l, u]$ and concave for $x \ge z$. For a sigmoidal function, the convex envelope is particularly easy to calculate. We can write $\hat{f}$ of f piecewise as

$$\hat{f}(x) = \begin{cases} f(x) & l \le x \le w \\ f(w) + \frac{f(u)-f(w)}{u-w}(x-w) & w \le x \le u \end{cases}$$

for some appropriate $w \le z$. If f is differentiable, then $f'(w) = \frac{f(u)-f(w)}{u-w}$; in general, $\frac{f(u)-f(w)}{u-w}$ is a subgradient of f at w. The point w can easily be found by bisection: if $x > w$, then the line from $(x, f(x))$ to $(u, f(u))$ crosses the graph of f at x; if $x < w$, it crosses in the opposite direction.

Univariate functions. If the inflection points of the univariate function are known, then the convex envelope may be calculated by iterating the construction given above for the case of sigmoidal functions.

Analytically. Occasionally the convex envelope may be calculated analytically. For example, convex envelopes of multilinear functions on the unit cube are polyhedral (piecewise linear), and can be calculated using an analytical formula given in [Rik97]. A few examples of analytically tractable convex envelopes are presented in Table 4. In the table, $\hat{f} : \mathbf{conv}\, S \to \mathbf{R}$ is the convex envelope of $f : S \to \mathbf{R}$, and $\rho(f)$ gives the nonconvexity of f. We employ the following standard notation: $\mathbf{card}(x)$ denotes the cardinality (number of nonzeros) of the vector x; the spectral norm (maximum singular value) is written as $\|M\|$, and its dual, the nuclear norm (sum of singular values) is written as $\|M\|_*$.

Via differential equations. The convex envelope of a function can also be written as the solution to a certain nonlinear partial differential equation [Obe07], and hence may be calculated numerically using the standard machinery of numerical partial differential equations [Obe08].

Table 1: Examples of convex envelopes.

S	$f(x)$	$\hat{f}(x)$	$\rho(f)$
$[0,1]^2$	$\min(x,y)$	$(x+y-1)_+$	$1/2$
$[0,1]^2$	xy	$(x+y-1)_+$	$1/4$
$[0,1]^n$	$\min(x)$	$(\sum_{i=1}^n x_i - (n-1))_+$	$\frac{n-1}{n}$
$[0,1]^n$	$\prod_{i=1}^n x_i$	$(\sum_{i=1}^n x_i - (n-1))_+$	$(\frac{n-1}{n})^n$
$[-1,1]^n$	$\mathbf{card}(x)$	$\|x\|_1$	n
$\{M \in \mathbf{R}^{k\times n} : \|M\| \le 1\}$	$\mathbf{rank}(M)$	$\|M\|_*$	n

5 Examples

Resource allocation. An agent wishes to allocate resources to a collection of projects $i = 1, \ldots, n$. For example, the agent might be bidding on a number of different auctions, or allocating human and capital resources to a number of risky projects. There are m different resources to be allocated to the projects, with each project i receiving a non-negative quantity x_{ij} of resource j. The probability that project i will succeed is modeled as $f_i(x_i)$, and its value to the agent, if the project is successful, is given by v_i. The agent has access to a quantity c_j of resource j, $j = 1, \ldots, m$. An allocation is feasible if $\sum_{i=1}^n x_{ij} \le c_j$, $j = 1, \ldots, m$. The agent seeks to maximize the expected value of the successful projects by solving

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^n v_i f_i(x_i) \\ \text{subject to} & \sum_{i=1}^n x_{ij} \le c_j, \quad j = 1, \ldots, m \\ & x \ge 0. \end{array}$$

To conform to our notation in the rest of this paper, we write this as a minimization problem,

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n -v_i f_i(x_i) \\ \text{subject to} & \sum_{i=1}^n x_{ij} \le c_j, \quad j = 1, \ldots, m \\ & x \ge 0. \end{array}$$

Here, there are m complicating constraint connecting the variables. Hence the bound from Theorem 1 guarantees that $|\hat{p} - p^\star| \le \sum_{i=1}^{\min(m,n)} \rho_{[i]}$. If $p_i : \mathbf{R} \to [0, 1]$ is a probability, then $\rho(-v_i p_i) \le v_i$. For example, if there is only one resource ($m = 1$), the bound tells us that we can find a solution x by solving the convex problem $\mathcal{R}$ whose value differs from the true optimum $p^\star$ by no more than $\max_i v_i$, regardless of the number of projects n.

Flow and admission control. A set of flows pass through a network over given paths of links or edges; the goal is to maximize the total utility while respecting the capacity of the links. Let x_i denote the level of each flow $i = 1, \ldots, n$ and $u_i(x_i)$ the utility derived from that flow. Each link j, $j = 1, \ldots, m$, is shared by the flows $i \in S_j$, and can accomodate up to a total of c_j units of flow. The flow routes are defined by a matrix $A \in \mathbf{R}^{m \times n}$ mapping flows onto links, with entries a_{ji}, $i = 1, \ldots, n$, $j = 1, \ldots, m$. When flows are not split, *i.e.*, they follow simple paths, we have $a_{ij} = 1$ when flow i pass over link j, and $a_{ij} = 0$ otherwise. But it is also possible to split a flow across multiple edges, in which case the entries a_{ij} can take other values. The goal is to maximize the total utility of the flows, subject to the resource constraint,

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^n u_i(x_i) \\ \text{subject to} & Ax \le c \\ & x \ge 0. \end{array} \tag{3}$$

The utility function is often modelled by a bounded function, such as a sigmoidal function [UB13, FC05]. As an extreme case, we can consider utilities of the form

$$u(x) = \begin{cases} 0 & x < 1 \\ 1 & x \geq 1 \end{cases}.$$

Thus each flow has value 1 when its level is at least 1, and no value otherwise. In this case, the problem is to determine choose the subset of flows, of maximum cardinality, that the network can handle. (This problem is also called admission control, since we are deciding which flows to admit to the network.)

We can replace this problem with an equivalent minimization problem to facilitate the use of Theorem 1. Let $f_i(x) = -u_i(x)$. Then we minimize the negative utility of the flows by solving

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n u_i(x_i) \\ \text{subject to} & Ax \leq c \\ & x \geq 0. \end{array}$$

Suppose f_i is bounded for every i, so that $\max_i \rho(f_i) \leq R$. Then the bound from Theorem 1 guarantees that we can quickly find a solution $p^\star - \hat{p} \leq mR$. In a situation with many flows but only a modest number of links, the solution given by solving $\mathcal{R}$ may be very close to optimal.

6 Proofs

In this section, we suppose without loss of generality that the problem has only inequality constraints; the mathematical argument with equality constraints is exactly the same. We let $A = [A_1 \cdots A_n]$ with $A_i \in \mathbf{R}^{m \times n_i}$, so $Ax = \sum_i A_i x_i$. As before, $N = \sum_{i=1}^n n_i$.

6.1 Definitions

First, we review some basic definitions from convex analysis (see [Roc97, LR01] for more details). The *epigraph* of a function f is the set of points lying above the graph of f,

$$\mathbf{epi}(f) = \{(x,t) : t \geq f(x)\}.$$

The *convex hull* of a set C is the set of points that can be written as a convex combination of other points in the set,

$$\mathbf{conv}\, S = \{\textstyle\sum_j \theta_j x_j : \theta_j \geq 0,\ x_j \in S,\ \sum_j \theta_j = 1\}.$$

An *extreme point* of a set is a point that cannot be written as a convex combination of other points in the set. A *face* of a set is a set of points optimizing a linear functional over that set. It is easy to see that a zero-dimensional face of a set is an extreme point, and that any extreme point defines a zero-dimensional face of a set [Roc97].

6.2 Main lemmas

Our analysis relies on two main lemmas. Lemma 1 tells us that at the extreme points of a face of $\mathbf{epi}(\hat{f})$, the values of f and $\hat{f}$ are the same. Lemma 2 tells us that (with probability 1) we can find a point that is extreme in $\mathbf{epi}(\hat{f}_i)$ for most i, and feasible, by solving a randomized convex program. We then combine these two lemmas to prove Theorem 2 and, as a consequence, Theorem 1.

We use two other technical lemmas as ingredients in the proofs of the two main lemmas. Lemma 3 gives conditions under which the convex hull of the epigraph of a function is closed, and Corollary 1 states that the maximum of a random linear functional over a compact set is unique with probability 1. Their statements and proofs can be found in appendix B and appendix A respectively.

We begin by finding a set of points where f and $\hat{f}$ agree.

Lemma 1. *Let $S \subset \mathbf{R}^n$ be a compact set, and let $f : S \to \mathbf{R}$ be lower semi-continuous on S, with convex envelope $\hat{f} : \mathbf{conv}\, S \to \mathbf{R}$. Let $c \in \mathbf{R}^n$ be a given vector. If x is extreme in* $\operatorname{argmin}(\hat{f}(x) + c^T x)$*, then $x \in S$ and $f(x) = \hat{f}(x)$.*

Proof. The vector c defines a face $\{(y, \hat{f}(y)) \mid y \in \operatorname{argmin}(\hat{f}(x) + c^T x)\}$ of $\mathbf{epi}(\hat{f})$. If x is extreme in $\operatorname{argmin}(\hat{f}(x) + c^T x)$, then $(x, \hat{f}(x))$ is extreme in $\mathbf{epi}(\hat{f})$ [Roc97, p. 163]. But

$$\mathbf{epi}(\hat{f}) = \mathbf{cl}(\mathbf{conv}(\mathbf{epi}(f))) = \mathbf{conv}(\mathbf{epi}(f)),$$

where the first equality follows from the definition of the convex envelope, and the second from the requirement that f be lower semi-continuous, and that S is compact, using Lemma 3. Thus every extreme point of $\mathbf{epi}(\hat{f})$ is a point in $\mathbf{epi}(f)$ [Roc97, cor. 18.3.1]. So $(x, \hat{f}(x)) \in \mathbf{epi}(f)$, and hence $x \in S$ and $\hat{f}(x) \geq f(x)$. But $\hat{f}$ is the convex envelope of f, so $\hat{f}(x) \leq f(x)$. Thus $\hat{f}(x) = f(x)$. □

Now we show that a solution to a randomized convex program finds a point that is extreme for most subvectors x_i of x.

Lemma 2. *Let $M_i \in \mathbf{R}^{n_i}$, $i = 1, \dots, n$, be given convex sets, and let $A \in R^{m \times N}$ with $N = \sum_{i=1}^{n} n_i$. Choose $w \in \mathbf{R}^N$ uniformly at random on the unit sphere. Let $x^\star$ be a solution to the convex program*

$$\begin{array}{ll} \textit{minimize} & w^T x \\ \textit{subject to} & Ax = b \\ & x_i \in M_i, \quad i = 1, \dots, n. \end{array} \tag{4}$$

Then almost surely, the solution $x^\star$ is unique, and for all but m indices i, $x_i^\star$ is an extreme point of M_i.

To prove the main theorems of this paper, we'll use an optimal Lagrange multiplier vector $\lambda^\star$ to define M. We will pick $M = \prod_i M_i$, where

$$M_i = \operatorname{argmin}(\hat{f}_i(x_i) + \lambda^{\star T} Ax).$$

Then $\lambda^\star$ defines a face of $\mathbf{epi}(\hat{f}_i)$ for $i = 1, \dots, n$, which allows us to use Lemma 1. This choice also makes it easy to solve Problem (4). For example, if each f_i is univariate, M_i is simply an interval, so Problem (4) is a linear program.

Proof. By Corollary 1, the maximum of a random linear functional over a compact set is unique with probability 1. Hence we may suppose Problem (4) has a unique solution, which we call x, with probability 1. Let $B(x)$ be the intersection of an open ball around x with the set of supporting hyperplanes to M at x. Equivalently, $B(x)$ is a d-dimesional open ball around x contained in M, of maximal dimension d. If x is on the boundary of M, the dimension of B will be the dimension of the lowest-dimensional face of M containing x, and $d < N$.

Now, every $y \in B(x) \cap \mathbf{nullspace}(A)$ is feasible for Problem (4). The random vector w must be orthogonal to $y - x$ for every $y \in B(x) \cap \mathbf{nullspace}(A)$, for otherwise the solution to Problem (4) could not occur at the center x of a feasible ball, but would have to occur on the boundary. On the other hand, if w is orthogonal to this feasible subspace, then any point $y \in B(x) \cap \mathbf{nullspace}(A)$ is a solution to Problem (4). But the solution x is unique, so it must be that $S(x) \cap \mathbf{nullspace}(A) = \emptyset$. Recalling a few simple identities from linear algebra, we see that $\dim(\text{range}(A)) + \dim(\textbf{nullspace}(A)) = N$ and $\dim(B(x)) + \dim(\textbf{nullspace}(A) \leq N$, so $\dim(B(x)) \leq m$.

Furthermore, the dimension of $B(x)$ bounds the number of subvectors x_l of x that are not extreme in M_l. If x_l is not extreme, it lies on a face of M_l with dimension at least 1, so $B(x)$ contains a vector y_l that differs from x only on the lth coordinate block. This is true for any subvector of x that is not extreme. For $l \neq l'$, $y_l - x$ is orthogonal to $y_{l'} - x$, so $\dim(B(x))$ is at least as large as the number of subvectors that are not extreme. But we have already bounded $\dim(B(x)) \leq m$, and so can similarly bound the number of subvectors that are not extreme.

Thus almost surely, the solution to Problem (4) is unique, and no more than m subvectors x_l of the solution x are not at extreme points. □

6.3 Main theorems

We are now ready to prove the main theorems, using the previous lemmas.

Proof of Theorem 2. Suppose $(x^\star, \lambda^\star)$ form an optimal primal-dual pair for $\hat{\mathcal{P}}$. Any such pair satisfies the KKT conditions:

1. *The point $x^\star$ minimizes the Lagrangian at $\lambda^\star$.*

$$x^\star \in \underset{x}{\operatorname{argmin}}\left(\textstyle\sum_{i=1}^n \hat{f}_i(x_i) + \lambda^{\star T}(Ax - b)\right).$$

2. *Primal feasibility.* $Ax^\star \leq b$.

3. *Dual feasibility.* $\lambda^\star \geq 0$.

4. *Complementary slackness.* $\lambda_i^\star (Ax^\star - b)_i = 0$.

Any other x satisfying the same conditions is also a solution to $\hat{\mathcal{P}}$.

Define $M = \operatorname{argmin} \mathcal{L}(x, \lambda^\star)$, *i.e.*,

$$M = \underset{x}{\operatorname{argmin}}\left(\textstyle\sum_{i=1}^n \hat{f}_i(x_i) + \lambda^{\star T}(Ax - b)\right).$$

The set M is bounded, since the domain S_i of f_i (and hence $\textbf{conv}\, S_i$, the domain of $\hat{f}_i$) is bounded, $i = 1, \ldots, n$. Furthermore, the function defining the set M is separable. Hence $M = M_1 \times \cdots \times M_n$, where

$$M_i = \underset{x_i}{\operatorname{argmin}}\left(\hat{f}_i(x_i) + \lambda^{\star T} A_i x_i\right).$$

In other words, each M_i is a face of $\mathbf{epi}(\hat{f})$ with normal vector $A_i^T\lambda^\star$. Note that M_i is convex, since it is a sublevel set of the convex function $\hat{f}_i(x_i) + \lambda^{\star T} A_i x_i$, and it is bounded, since M is bounded.

Then using the KKT conditions, we see x is optimal for $\hat{\mathcal{P}}$ if $x \in M$, $Ax \le b$, and $\lambda_i^\star (Ax - b)_i = 0$. Writing these conditions in terms of the previous primal solution $x^\star$, we see x is optimal for $\hat{\mathcal{P}}$ if $x_i \in M_i$, $i = 1, \dots, n$, and $Ax = Ax^\star$. (Here, the condition $x_i \in M_i$, $i = 1, \dots, m$, guarantees that x minimizes the Lagrangian, while $Ax^\star = Ax$ guarantees that x preserves primal feasibility and complementary slackness.) We can find a random $\hat{x}$ satisfying these conditions by solving Problem $\mathcal{R}$.

By Lemma 2, the solution $\hat{x}_i$ to Problem $\mathcal{R}$ lies at an extreme point of M_i for all but (at most) m of the coordinate blocks i (with probability 1). By Lemma 1, extreme points x_i of the face M_i satisfy $f_i(x_i) = \hat{f}_i(x_i)$, so $f_i(\hat{x}_i) > \hat{f}_i(\hat{x}_i)$ for no more than m of the coordinate blocks i. On those blocks i where $\hat{x}_i$ is not extreme, it is still true that $f_i(\hat{x}_i) - \hat{f}_i(\hat{x}_i) \le \rho(f_i)$. Hence

$$0 \le \sum_{i=1}^{n} f_i(\hat{x}_i) - p^\star = \sum_{i=1}^{n} (f_i(\hat{x}_i) - \hat{f}_i(\hat{x}_i)) \le \sum_{i=1}^{\min(m,n)} \rho(f_{[i]}).$$

□

Proof of Theorem 1. Since a point satisfying the bound in Theorem 1 can be found almost surely by minimizing a random linear function over M, it follows that such a point exists. □

7 Numerical experiments

We now present a numerical experiment to demonstrate the utility of finding an extreme point of the solution set of the convexified problem, rather than an arbitrary solution.

Investment problem. Consider the following investment problem. Each variable $x_i \in \mathbf{R}$ represents the allocation of capital to project i. The probability that a project will fail is given by $f(x_i)$.

A sector of the economy j is given by the nonzero entries a_{ij} in the matrix $A \in \mathbf{R}^{m \times n}$, and the budget for projects in each sector is given by the vector

$b \in \mathbf{R}^m$. The constraint $Ax \leq b$ then prevents overexposure to any given sector.

The problem of minimizing the expected number of failed projects subject to these constraints can be written as

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n f(x_i) \\ \text{subject to} & Ax \leq b \\ & 0 \leq x. \end{array} \qquad (5)$$

The results of our numerical experiments are presented in Table 2. We let

$$f(x) = \begin{cases} 1 & 0 \leq x < 1 \\ 0 & x \geq 1 \end{cases} .$$

Random instances of the investment problem are generated with $n = 50$, $m = 10$. Random sector constraints are generated by choosing entries of A to be 0 or 1 uniformly at random with probability 1/2, and let $b = 1/2A\mathbf{1}$, where $\mathbf{1}$ is the vector of all ones, in order to ensure the constraints are binding. In the table, $\hat{x}$ is the solution to

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n \hat{f}(x_i) \\ \text{subject to} & Ax \leq b \\ & 0 \leq x \end{array} \qquad (6)$$

returned by an interior point solver, whereas $\tilde{x}$ is the minimizer of a random linear functional over the solution set of (6). Improvement is calculated as $\frac{f(\hat{x})-f(\tilde{x})}{f(\hat{x})-p^\star}$. Solving the random LP usually gives a substantial improvement in the value of the solution of the convexified problem according to the original objective. The observed difference between $f(\tilde{x})$ and $p^\star$ is always substantially smaller than the theoretical bound of $m\rho(f) = 10$.

Table 2: Investment problem.

$f(\hat{x})$	$f(\tilde{x})$	$p^\star$	$\hat{p}$	% improved
43.01	23.01	22.00	20.25	0.95
29.02	26.00	22.00	20.36	0.43
30.09	24.00	21.00	19.92	0.67
26.32	25.00	22.00	20.27	0.31
24.68	24.00	22.00	20.33	0.25
26.01	25.00	21.00	19.26	0.20
26.46	24.00	20.00	19.40	0.38
28.24	25.00	23.00	20.65	0.62
29.04	24.00	21.00	20.21	0.63
27.01	23.01	21.00	19.70	0.67

A Well-posedness

The following theorem, which may be of interest in its own right, characterizes the set of vectors in the dual space for which linear optimization over a compact set S is well-posed. [1]

Theorem 3 (Well-posedness of linear optimization)**.** *Suppose S is a compact set in $\mathbf{R}^n$. Then the set of $w \in \mathbf{R}^n$ for which the maximizer of $w^T x$ over S is not unique has (Lebesgue) measure zero.*

Before proceeding to a proof, we make some remarks to show why the theorem is intuitive, and why the proof is not trivial. By definition, the maximizer of a linear functional over a set S is a face R of S. The maximizer is unique if and only if R is a zero-dimensional face (*i.e.*, an extreme point). Only an outward normal to a face will be maximized on that face.

It is easy to see that the theorem is true for polyhedral sets S. For each face of the polyhedron that is not extreme, the set of vectors maximized by that face (the set of outward normals to the face, *i.e.*, the normal cone) will have dimension *smaller than n*. A polyhedron has only a bounded number of faces, so the union of these sets still has measure zero.

On the opposite extreme, consider the unit sphere. A sphere has an infinite number of faces. But every face is extreme, and every vector w has

[1]The authors are indebted to Jon Borwein for his unflagging help on the proof of this theorem, and to Julian Revalski for his insightful comments on the connection of this theorem to other related work.

a unique maximizer.

The difficulty comes when we consider cylindrical sets: those constructed as the Cartesian product of a sphere and a cube. Here, every outward normal to the "sides" of the cylinder is a vector whose maximum over the set is not extreme. That is, we find an *uncountably infinite* number of faces (parametrized by the boundary of the sphere) that are not extreme points.

Proof. Let $I_S : \mathbf{R}^n \to \mathbf{R}$ be the indicator function of S. S is compact, so the convex conjugate $I_S^*(y) = \sup_x y^T x - I_S(x)$ of I_S is finite for every $y \in \mathbf{R}^n$. Rachemacher's Theorem [BV10, Theorem 2.5.1] states that a convex function $g : \mathbf{R}^n \to \mathbf{R}$ is differentiable almost everywhere with respect to Lebesgue measure on $\mathbf{R}^n$. Furthermore, if I_S^* is differentiable at y with $\nabla I_S^*(y) = x$, then $y^T x - I_S(x)$ attains a strong maximum at x [BV10, Theorem 5.2.3]; that is, there is a unique maximizer of $y^T x$ over S. □

Clearly, the statement also holds for the minimizers, rather than maximizers, of $w^T x$.

The following corollary will be used in the proof of the main theorem of this paper.

Corollary 1. *Suppose S is a compact set in $\mathbf{R}^n$, and w is a uniform random variable on the unit sphere in $\mathbf{R}^n$. Then with probability one, there is a unique minimizer of $w^T x$ over S.*

Proof. The property of having a unique minimizer exhibits a symmetry along radial lines: there is a unique minimizer of $w^T x$ over S if and only if there is a unique minimizer of $(w/\|w\|_2)^T x$ over S. A uniform random vector on the unit sphere may be generated by taking a uniform random vector on the unit ball, and normalizing it to lie on the unit sphere. Since the set of directions whose maximizers are not unique has Lebesgue measure zero, the vectors on the unit sphere generated in this manner have maximizers that are unique with probability 1. □

We give one last corollary, which may be of mathematical interest, but is not used elsewhere in this paper.

Corollary 2. *Suppose S is a compact set in $\mathbf{R}^n$. The union of the normal cones $N(x)$ of all points $x \in S$ that are not extreme has measure zero.*

Proof. A point x minimizes $y^T x$ over S if and only if $y \in N(x)$. A point x is the only minimizer of $y^T x$ over S if and only if x is exposed, and hence extreme. Hence no y with a unique minimizer over S lies in the normal cone of a point that is not extreme. Thus the union of the normal cones $N(x)$ of all points $x \in S$ that are not extreme is a subset of the vectors which do not have a unique maximizer over S, and hence has measure zero. □

B Closure

Lemma 3. *Let $S \subset \mathbf{R}^n$ be a compact set, and let $f : S \to \mathbf{R}$ be lower semi-continuous on S. Then $\mathbf{conv}(\mathbf{epi}\, f)$ is closed.*

Proof. Every point $(x, t) \in \mathbf{cl}(\mathbf{conv}(\mathbf{epi}\, f))$ is a limit of points (x^k, t^k) in $\mathbf{conv}(\mathbf{epi}\, f)$. These points can be written as

$$(x^k, t^k) = \sum_{i=1}^{n+2} \lambda_i^k (a_i^k, b_i^k)$$

with $\sum_{i=1}^{n+2} \lambda_i^k = 1$, $0 \le \lambda_i^k \le 1$, and $(a_i^k, b_i^k) \in \mathbf{epi}(f)$. Since $[0,1]$ and S are compact, we can find a subsequence along which each sequence a_i^k converges to a limit $a_i \in S$, and each sequence λ_i^k converges to a limit $\lambda_i \in [0,1]$.

Let $P = \{i : \lambda_i > 0\}$. Note that P is not empty, since $\sum_{i=1}^{n+2} \lambda_i^k = 1$ for every k. If $l \in P$, then because the limit t exists, $\limsup_k b_i^k$ is bounded above. Recall that a lower semi-continuous function is bounded below on a compact domain, so b_i^k is also bounded below. This shows that for $i \in P$, every subsequence of b_i^k has a subsequence that converges to a limit b_i. In particular, we can pick a subsequence k_j such that simultaneously, for $i = 1, \ldots, n+2$, $a_i^{k_j}$, $b_i^{k_j}$, and $\lambda_i^{k_j}$ converge along the subsequence k_j to a_i, b_i, and λ_i, respectively.

Define $S_P = \sum_{i \in P} \lambda_i b_i$. Then along the subsequence k_j, $\lim_{j \to \infty} \sum_{i \notin P} \lambda_i^{k_j} b_i^{k_j} = t - S_P$ also exists. Since f is bounded below, b_i^k are all bounded below, and for $i \notin P$, $\lambda_i^k \to 0$, so $t - S_P \ge 0$. Therefore (x, t) can be written as $\sum_{i \in P} \lambda_i (a_i, b_i) + (0, t - S_P)$.

Recall that a function is lower semi-continuous if and only if its epigraph is closed. Hence $(a_i, b_i) \in \mathbf{epi}\, f$ for $i \in P$. Without loss of generality, suppose $1 \in P$, and note that $(a_1, b_1 + t - S_P) \in \mathbf{epi}\, f$, since $t - S_P$ is non-negative.

Armed with these facts, we see we can write (x, t) as a convex combination of points in $\mathbf{epi}\, f$,

$$(x, t) = \lambda_1 (a_1, b_1 + t - S_P) + \sum_{i \in S, i \ne 1} \lambda_i (a_i, b_i).$$

Thus every $(x, t) \in \mathbf{cl}(\mathbf{conv}(\mathbf{epi}\, f))$ can be written as a convex combination of points in $\mathbf{epi}\, f$, so $\mathbf{conv}(\mathbf{epi}\, f)$ is closed. □

Corollary 3. *Let $S \subset \mathbf{R}^{n+2}$ be a compact set, and let $f : S \to \mathbf{R}$ be lower semi-continuous on S. Then $\mathbf{epi}(\hat{f}) = \mathbf{cl}(\mathbf{conv}(\mathbf{epi}\, f)) = \mathbf{conv}(\mathbf{epi}\, f)$.*

References

[AE76] J. Aubin and I. Ekeland. Estimates of the duality gap in nonconvex optimization. *Mathematics of Operations Research*, 1(3):225–245, 1976.

[Ber82] D. P. Bertsekas. *Constrained Optimization and Lagrange Multipler Methods*. Athena Scientific, 1982.

[Ber99] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, 1999.

[BLNP83] D. Bertsekas, G. Lauer, Sandell J. N., and T. Posbergh. Optimal short-term scheduling of large-scale power systems. *Automatic Control, IEEE Transactions on*, 28(1):1–11, 1983.

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

[BV10] J. M. Borwein and J. D. Vanderwerff. *Convex functions: constructions, characterizations and counterexamples*, volume 109. Cambridge University Press, 2010.

[FC05] M. Fazel and M. Chiang. Network utility maximization with nonconcave utilities using sum-of-squares method. In *Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC '05*, pages 1867 – 1874, 2005.

[GW95] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42:1115–1145, 1995.

[LR01] C. Lemaréchal and A. Renaud. A geometric study of duality gaps, with applications. *Mathematical Programming*, 90(3):399–427, 2001.

[Mot95] R. Motwani. *Randomized algorithms*. Cambridge University Press, 1995.

[Obe07] A. M. Oberman. The convex envelope is the solution of a nonlinear obstacle problem. *Proceedings of the American Mathematical Society*, 135(6):1689–1694, 2007.

[Obe08] A. M. Oberman. Computing the convex envelope using a nonlinear partial differential equation. *Mathematical Models and Methods in Applied Sciences*, 18(05):759–780, 2008.

[Rik97] A. D. Rikun. A convex envelope formula for multilinear functions. *Journal of Global Optimization*, 10(4):425–437, 1997.

[Roc97] R. T. Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.

[SB10] J. Skaf and S. Boyd. Techniques for exploring the suboptimal set. *Optimization and Engineering*, 11(2):319–337, 2010.

[Sta69] R. M. Starr. Quasi-equilibria in markets with non-convex preferences. *Econometrica: Journal of the Econometric Society*, pages 25–38, 1969.

[UB13] M. Udell and S. Boyd. Maximizing a sum of sigmoids, 2013.