Bounding Duality Gap for Separable Problems with Linear Constraints

Madeleine Udell and Stephen Boyd

January 12, 2016

Abstract

We consider the problem of minimizing a sum of non-convex functions over a compact domain, subject to linear inequality and equality constraints. Approximate solutions can be found by solving a convexified version of the 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 one, ϵ is bounded by a term proportional to the maximal number of active 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 will state our results in terms of a (possibly) smaller quantity $\tilde{m} \leq m$, which provides a (sometimes) tighter estimate of the number of complicating constraints in the problem. Define the *active set* of inequality constraints at x to be $J(x) = \{j : (Ax - b)_j = 0\}$, let $\tilde{m}_1 = \max_x |J(x)|$ be the maximal number of inequality constraints that can be simultaneously active, and let $\tilde{m} = \tilde{m}_1 + m_2$ be the number of (equality and inequality) constraints that can be simultaneously active.

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 [Roc70, Theorem 17.2]. When f_i is lower semi-continuous and S_i is compact and nonempty, then $\mathbf{conv}(S_i)$ is compact and convex, and $\hat{f}_i$ is closed, proper, and convex [Roc70]. 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 convenience, we assume that the functions f_i are sorted in order of decreasing nonconvexity, so $\rho(f_1) \geq \cdots \geq \rho(f_n)$.

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; if we can efficiently evaluate $\hat{f}$ and a subgradient (or derivative, if $\hat{f}$ is differentiable), then the problem is easily solved using standard methods for nonlinear convex optimization. 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 first result is the following:

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

$$\hat{p} = \hat{f}(x^\star) \leq f(x^\star) \leq \hat{p} + \sum_{i=1}^{\min(\tilde{m},n)} \rho(f_i).$$

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

$$p^\star \leq f(x^\star) \leq p^\star + \sum_{i=1}^{\min(\tilde{m},n)} \rho(f_i).$$

In other words, there is a solution of the convexified problem that is ϵ-suboptimal for the original problem, with $\epsilon = \sum_{i=1}^{\min(\tilde{m},n)} \rho(f_i)$. It is not true (as we show in §2) 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 (active) 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 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. Let

$$L(x, \lambda, \mu) = \sum_{i=1}^{n} f_i(x_i) + \lambda^T(Ax - b) + \mu^T(Gx - h)$$

be the Lagrangian of $\mathcal{P}$ with dual variables λ and μ, and define the (Lagrange) dual problem to $\mathcal{P}$,

$$\begin{array}{ll} \text{maximize} & \inf_x \mathcal{L}(x, \lambda, \mu) \\ \text{subject to} & \lambda \geq 0, \end{array} \qquad (\mathcal{D})$$

with optimal value $g^\star$. The convexified problem $\hat{\mathcal{P}}$ is the dual of $\mathcal{D}$. (See Appendix A for a derivation.) Since $\hat{\mathcal{P}}$ is convex and feasible, with only linear constraints, strong duality holds by the refined Slater's constraint qualification [BV04, §5.2.3]. (For a proof, see [Roc70, p. 277].) Hence 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 $\lambda^\star \geq 0$. The bound from Theorem 1 thus implies

$$p^\star - g^\star \leq \sum_{i=1}^{\min(\tilde{m},n)} \rho(f_i).$$

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: a constructive version of Theorem 1.

Theorem 2. *Let $w \in \mathbf{R}^N$ be a random variable with uniform distribution on the unit sphere. Now consider the feasible convex problem*

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

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

$$f(x^\star) \leq \hat{p} + \sum_{i=1}^{\min(m,n)} \rho(f_i),$$

i.e., $x^\star$ is ϵ-suboptimal for the original problem $\mathcal{P}$.

The randomized problem $\mathcal{R}$ has a simple interpretation. Any feasible point x for $\mathcal{R}$ is feasible for $\hat{\mathcal{P}}$, and the constraint $\hat{f}(x) \leq \hat{p}$ is satisfied with equality. That is, $\mathcal{R}$ minimizes 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

In this section we show that the bound in Theorem 1 is tight, and that finding extreme points of the optimal set is essential to achieving the bound. In these examples, $\tilde{m} = m$.

Example 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 $x^\star$ satisfying $0 \leq x^\star \leq 1$ and $\sum_{i=1}^n x_i^\star = B$ 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

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

Letting $B \to 1$, we see that the bound is tight.

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

On the other hand, $x^\star \in M$ defined by

$$x_i^\star = \begin{cases} 1 & i = 1, \ldots, n-1 \\ B - (n-1) & i = n, \end{cases}$$

which is an extreme point of the optimal set for the convexified problem, is optimal for the original problem as well. That is, $x^\star$ is an extreme point of M that satisfies Theorem 1 with equality.

Example 3 (Nonconvex feasible set.). 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 $\hat{x} = 1/2 \in M$ 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,

$$p^\star - d^\star \leq \min(m+1, n)\rho(f_1),$$

is not tight; our bound, which is tight, is smaller by a factor of $\tilde{m}/(m+1)$.

The Shapley-Folkman theorem has found uses in a number of applications within optimization. For example, Bertsekas et al. [BLNP83] used the theorem to solve a unit commitment problem in electrical power system scheduling, in which case the terms in the objective are univariate. The Shapley-Folkman theorem and its relation to a bound on the duality gap also have found applications in integer programming [VEG+14]. While we restrict ourselves here only to nonconvex *objectives*, many authors [Ber82, Ber99, LR01] have studied convexifications of separable constraints as well. A more modern treatment, in the case of linear programs, is given in [Ber09].

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.

Our procedure for finding an extreme point is closely related to the idea of *purifying* a solution returned by, *e.g.*, an interior point solver to obtain an extremal solution. One fixes an active set of inequality constraints that hold with equality at a given point, and solves ($\mathcal{R}$) subject to the additional constraint that all inequality constraints in the active set continue to hold with equality, and then iterates this procedure until the set of active constraints completely determines the solution. It is easy to see that at each iteration

at least one constraint is added to the active set. Hence the procedure converges to an extreme point in no more than $\tilde{m}$ iterations. In contrast, our proof shows that the method finds an extreme point with probability 1 in a single iteration, without fixing an active set beforehand.

The notion that extreme points of the solution set of a convex problem have particularly nice properties is pervasive in the literature. The extreme points produced by solving $\mathcal{R}$ are simply *basic feasible solutions*, familiar from the analysis of the simplex method, whenever the functions f_i are univariate, *i.e.*, $n_i = 1$, $i = 1, \ldots, n$. Other uses of extreme points abound: for example, Anderson and Lewis [AL89] propose a simplex-style method for semi-infinite programming that proceeds by finding extreme points of the feasible set; and Barvinok [Bar95, Bar02] and Pataki [Pat96, Pat98] examine the extreme points of an affine section of the semidefinite cone to provide bounds on the rank of solutions to semidefinite programs.

4 Constructing the convex envelope

In general, the convex envelope of a function can be hard to compute. But in many special cases, we can efficiently construct the convex envelope or a close approximation to it. The problem of computing convex lower bounds on general nonconvex functions has been extensively studied in the global optimization community: see, eg, [HPN00] for a general introduction and [TS02] for a more sophisticated treatment. In this section, we give a few examples illustrating how to construct the convex envelope for a number of interesting functions and classes 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 $x \leq z \in [l, u]$ and concave for $x \geq z$. For a sigmoidal function, the convex envelope is particularly easy to calculate [UB13]. We can write $\hat{f}$ of f piecewise as

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

for some appropriate $w \leq z$. 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
$\{M \in \mathbf{R}^{k\times n} : \|M\| \leq 1\}$	$\mathrm{Rank}(M)$	$\|M\|_*$	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 $\sum_{i=1}^n x_{ij} \leq c_j$, $j = 1, \ldots, m$. The agent seeks to maximize the expected value of the successful projects by solving

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

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

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

Here, there are m complicating constraint connecting the variables. Hence the bound from Theorem 1 guarantees that $|\hat{p} - p^\star| \leq \sum_{i=1}^{\min(m,n)} \rho(f_i)$. If $p_i : \mathbf{R} \to [0, 1]$ is a probability, then $\rho(-v_i p_i) \leq v_i$. 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,

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

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

To simplify the 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. Merely note below in Lemma 3 that equality constraints are always active. 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 [Roc70, 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 S is the set of points that can be written as a convex combination of other points in the set,

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

An *exposed face* F of a convex set C is a set of points optimizing a linear functional over that set,

$$F = \operatorname*{argmin}_{x \in C} c^T x,$$

for some $c \in \mathbf{R}^n$. The vector c is called a normal vector to the face. We will use the fact that every exposed face is a *face*: a convex set $F \subset C$ for which every (closed) line segment in C with a relative interior point in F has both endpoints in F. Not all faces are exposed; but our analysis will not make use of this distinction.

An *extreme point* of a convex set is a point that cannot be written as a convex combination of other points in the set. It is easy to see that a zero-dimensional exposed face of a convex set is an extreme point, and that any extreme point defines a zero-dimensional exposed face of a convex set [Roc70].

6.2 Main lemmas

Our analysis relies on two main lemmas. Lemma 1 tells us that at the extreme points of an exposed face of $\mathbf{epi}(\hat{f})$, the values of f and $\hat{f}$ are the same. Lemma 2 tells us that (with probability one) 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 4 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 one. Their statements and proofs can be found in Appendix C and Appendix B 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 the set $\operatorname{argmin}(\hat{f}(x) + c^T x)$*, then $x \in S$ and $f(x) = \hat{f}(x)$.*

Proof. The vector c defines an (exposed) 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 $\mathbf{epi}(\hat{f})$ [Roc70, p. 163].

It is easy to see geometrically that every extreme point of $\mathbf{epi}(\hat{f})$ is a point in $\mathbf{epi}(f)$. Formally, recall that the convex envelope satisfies $\mathbf{epi}(\hat{f}) = \mathbf{cl}(\mathbf{conv}(\mathbf{epi}(f)))$ [Roc70, cor. 12.1.1]. Then use Lemma 4 (see Appendix C), which states that the $\mathbf{conv}(\mathbf{epi}(f))$ is closed if S is compact and f is lower semi-continuous, to see that $\mathbf{cl}(\mathbf{conv}(\mathbf{epi}(f))) = \mathbf{conv}(\mathbf{epi}(f))$. Thus every extreme point of $\mathbf{epi}(\hat{f})$ is a point in $\mathbf{epi}(f)$ [Roc70, 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 $M_i \in \mathbf{R}^{n_i}$, $i = 1, \ldots, n$, be given compact 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, and consider the convex program*

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

Almost surely, the solution x to problem (4) is unique. For all but at most $\tilde{m}$ indices i, x_i is an extreme point of M_i.

To prove Lemma 2, we will prove the following stronger lemma. Lemma 2 follows as a corollary, since $\tilde{m}$ bounds the number of simultaneously active constraints.

Lemma 3. *Let $M_i \in \mathbf{R}^{n_i}$, $i = 1, \ldots, n$, be given compact 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, and consider the convex program*

$$\begin{array}{ll} \textit{minimize} & w^T x \\ \textit{subject to} & Ax \leq b \\ & x_i \in M_i, \quad i = 1, \ldots, n. \end{array} \tag{5}$$

Almost surely, the solution x to problem (5) is unique. Let $J = \{j : (Ax - b)_j = 0\}$ be the set of active constraints at x. For all but at most $|J|$ indices i, x_i is an extreme point of M_i.

Proof of Lemma 3. By Corollary 1 (see Appendix B), the minimum of a random linear functional over a compact set is unique with probability one. Hence we may suppose problem (4) has a unique solution, which we call x, with probability one. Define $M = M_1 \times \cdots \times M_n$ to be the Cartesian product of the sets M_i. Let F be a minimal face of M containing x, and let $B \subset F \subseteq M$ be a ball in its relative interior. If x is on the boundary of M, then $\dim(B) < N$.

Let A_J be a matrix consisting of those rows of A with indices in J, and define the minimal distance to any non-active constraint

$$\delta = \inf_{j \in J^C} \inf_{y:(Ay-b)_j=0} \|x - y\|.$$

Let $D = (x + \mathbf{nullspace}(A_J)) \cap \mathcal{B}(x, \delta)$ where $\mathcal{B}(x, \delta)$ is an open ball around x with radius δ. With this definition, any $y \in D$ satisfies the constraints $Ay - b$ with the same active set J: $(Ay - b)_j = 0$ for every $j \in J$, and $(Ay - b)_j > 0$ for every $j \in J^C$. Note that $\dim(D) = \dim(\mathbf{nullspace}(A_J)) = N - |J|$.

Now we will show $B \cap D = \{x\}$. By way of contradiction, consider $y \in B \cap D$, $y \neq x$. Every such y is feasible for problem (4). The random vector w must be orthogonal to $y - x$, for otherwise the solution to problem (4) could not occur at the center x of the feasible ball B. On the other hand, if w is orthogonal to $y - x$, then y is a solution to problem (4). But the solution x is unique, so it must be that $B \cap D = \{x\}$. That is, B intersects the $(N - |J|)$-dimensional set D at a single point. This bounds the dimension of B: $\dim(B) + \dim(D) \leq N$, so $\dim(B) \leq |J|$.

Furthermore, $\dim(B)$ bounds the number of subvectors x_i of x that are not extreme in M_i. Let

$$\Omega = \{i \in \{1, \ldots, n\} : x_i \text{ is not extreme in } M_i\}.$$

For $i \in \Omega$, x_i lies on a face of M_i with dimension greater than zero. Hence B contains a point y^i that differs from x only on the ith coordinate block. Consider the set $Y = \{y^i : i \in \Omega\} \subset B$. The vectors $y^i - x$ for $i \in \Omega$ are mutually orthogonal, so $|\Omega| = \dim(\mathbf{conv}(Y)) \leq \dim(B)$. The argument in the last paragraph showed $\dim(B) \leq |J|$, and so we can bound the number of subvectors that are not extreme $|\Omega| \leq |J|$.

Thus almost surely, the solution to problem (4) is unique, and no more than $|J|$ subvectors x_i of the solution x are not at extreme points. $\square$

6.3 Main theorems

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

Proof of Theorem 2. By Lemma 2, the solution $x^\star$ to $\mathcal{R}$ is unique with probability 1. Every point in the feasible set for $\mathcal{R}$ is optimal for $\hat{\mathcal{P}}$, so in particular, $x^\star$ solves $\hat{\mathcal{P}}$. Pick $\lambda^\star \geq 0$ so that $(x^\star, \lambda^\star)$ form an optimal primal-dual pair for the primal-dual pair $(\hat{\mathcal{P}}, \mathcal{D})$. Note that by complementary slackness, any optimal point x for $\hat{\mathcal{P}}$ (and so any feasible point for $\mathcal{R}$) satisfies $\lambda^{\star T}(Ax - b) = 0$.

Now consider the problem

$$\begin{array}{ll} \text{minimize} & w^T x \\ \text{subject to} & Ax \leq b \\ & \hat{f}(x) - \lambda^{\star T} Ax \leq \hat{p} - \lambda^{\star T} Ax^\star, \end{array} \tag{6}$$

where, compared to $\mathcal{R}$, we have subtracted $\lambda^{\star T} Ax$ and $\lambda^{\star T} Ax^\star$ from the two sides of the inequality $\hat{f}(x) \leq \hat{p}$.

In fact, the feasible set of $\mathcal{R}$ is the same as that of problem (6). By complementary slackness, $\lambda^{\star T} Ax^\star = \lambda^{\star T} b$, so the last inequality constraint in problem (6) can be rewritten as

$$\hat{f}(x) - \lambda^{\star T}(Ax - b) \leq \hat{p}.$$

Since $\lambda^\star \geq 0$, and $Ax - b \leq 0$ on the feasible set of problem (6), we have $-\lambda^{\star T}(Ax - b) \geq 0$. Hence any x feasible for problem (6) satisfies

$$\hat{f}(x) \leq \hat{p},$$

and so satisfies the constraints of $\mathcal{R}$. Conversely, any feasible point for $\mathcal{R}$ has $\lambda^{\star T}(Ax - b) = 0$ by complementary slackness, so it is also feasible for problem (6). Since the feasible sets are the same and the objectives are the same, the solution to $\mathcal{R}$ must also be the same as that of problem (6).

Define

$$\begin{aligned} M &= \operatorname*{argmin}_x \left(\textstyle\sum_{i=1}^n \hat{f}_i(x_i) - \lambda^{\star T}(Ax - b) \right) \\ &= \operatorname*{argmin}_x \textstyle\sum_{i=1}^n \left(\hat{f}_i(x_i) - \lambda^{\star T} A_i x_i \right) - \lambda^{\star T} b. \end{aligned}$$

The function defining the set M is separable. Hence $M = M_1 \times \cdots \times M_n$, where

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

The set M_i is compact and convex: it is bounded, since the domain of $\hat{f}_i$, $\mathbf{conv}(S_i)$, is bounded; it is closed, since $\mathbf{epi}(\hat{f}_i)$ is closed; and it is convex, since $\mathbf{epi}(\hat{f}_i)$ is convex. So the M_i satisfy the conditions for Lemma 2.

By Lemma 2, the solution $x^\star$ to problem (6) is unique and lies at an extreme point of M_i for all but (at most) $\tilde{m}$ of the coordinate blocks i (with probability one). By Lemma 1, extreme points x_i of 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 $\tilde{m}$ of the coordinate blocks i. On those blocks i where $\hat{x}_i$ is not extreme, it is still true that $f_i(\hat{x}_i) - \hat{f}_i(\hat{x}_i) \leq \rho(f_i)$. Hence

$$0 \leq \sum_{i=1}^{n} f_i(x_i^\star) - p^\star = \sum_{i=1}^{n} \left(f_i(x_i^\star) - \hat{f}_i(x_i^\star) \right) \leq \sum_{i=1}^{\min(\tilde{m},n)} \rho(f_i).$$

□

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 example

We now present a numerical example to demonstrate the performance of the algorithm implied by the proof; namely, of finding an extreme point of the convexified problem to serve as an approximate solution to the original problem. This problem is not large, and is easy to solve using many methods. Our purpose in presenting the example is merely to give some intuition for 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 $f(x_i)$.

Entry a_{ij} of the matrix $A \in \mathbf{R}^{m \times n}$ represents the exposure of project i to sector j of the economy. 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

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n f(x_i) \\ \text{subject to} & Ax \leq b \\ & 0 \leq x. \end{array} \tag{7}$$

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 variables and m constraints. 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.

The results of our numerical experiments are presented in Table 2 and Figure 1. In the table, we choose $n = 50$, $m = 10$, let $\hat{x}$ be the solution to the problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n \hat{f}(x_i) \\ \text{subject to} & Ax \leq b \\ & 0 \leq x \end{array} \tag{8}$$

returned by an interior point solver, and let $x^\star$ be the solution to the random LP $\mathcal{R}$. The observed difference between $f(x^\star)$ and $p^\star$ is always substantially smaller than the theoretical bound of $m\rho(f) = 10$.

Figure 1 shows the improvement from solving $\mathcal{R}$, calculated as $\frac{f(x^\star)-f(\hat{x})}{f(x^\star)-p^\star}$, as a function of the number of variables n and constraints m, averaged over 10 random instances of the problem. Solving the random LP $\mathcal{R}$ gives a substantial improvement when $m < n$.

7.1 Solution via ADMM

Here we demonstrate how to use ADMM, a framework for distributed optimization, to find $\hat{x}$ satisfying the bound on the duality gap. This shows that a solution satisfying the bound may be found even for very large scale problems, so long as the proximal operators of the functions f_i can be evaluated efficiently.

Table 2: Investment problem.

$f(x^\star)$	$f(\hat{x})$	$p^\star$	$\hat{p}$	% 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

ADMM. The Alternating Directions Method of Multipliers (ADMM) was introduced in 1975 by Glowinski and Marocco [GM75] and Gabay and Mercier [GM76], and is closely related to a number of classical operator-splitting methods such as Douglas-Rachford and Peaceman-Rachford [Gab83, PR55, LM79, Glo14]. ADMM has recently received renewed interest as a method for solving distributed optimization problems due both to its ease of implementation and its robust convergence in practice and in theory on convex problems [Gab83, FG83b, FG83a, GT87, Tse91, Fuk92, EB92, EF93, CT94, HL12, HY12]. For an introduction to ADMM, we refer the reader to the survey [BPC+11] and references therein.

ADMM is not guaranteed to converge to the global solution when applied to a nonconvex problem [Zha10, MWF14]. However, its computational advantages still make ADMM a popular method for nonconvex optimization [MWF14, DBEY13, CW13, Cha12, BTP13, GZ13, KT12] even in the absence of convergence guarantees. In constrast to this previous work, here we use ADMM to find a feasible point for the nonconvex problem which obeys the global error bound of Theorem 1.

ADMM for the convexified problem. A generalized consensus ADMM iteration can be used to solve the convexified problem. (See [BPC+11] for details.) We rewrite the problem as

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n \hat{f}_i(x_i) + \mathbf{1}_{Ax\leq b, Gx=h}(z) \\ \text{subject to} & x = z, \end{array} \tag{9}$$

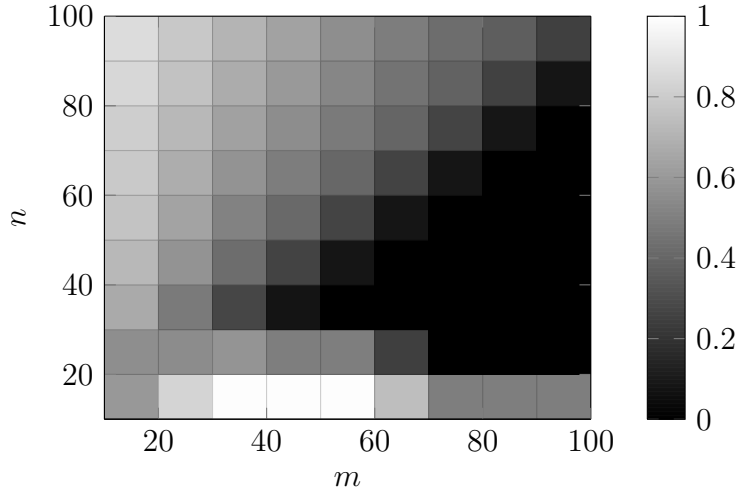

Figure 1: Improvement $\frac{f(x^\star)-f(\hat{x})}{f(x^\star)-p^\star}$ on investment problem.

where $\mathbf{1}_{\mathcal{C}}$ denotes the indicator function of the set $\mathcal{C}$. An ADMM iteration solving the above problem is given by

$$\begin{aligned} x_i^k &= \operatorname{argmin} \hat{f}_i(x) + \rho/2\|x - z_i^{k-1} + y_i^{k-1}\|_2^2 \\ z^k &= \Pi_{Ax\leq b, Gx=h}(x^k) \\ y_i^k &= y_i^{k-1} + 1/\rho(x_i^k - z_i^k). \end{aligned}$$

Here, $\Pi_{\mathcal{C}}$ denotes projection onto the set $\mathcal{C}$, and $\rho > 0$ is a parameter. Under some mild conditions [HL12], the iterates z^k and x^k both converge linearly to a primal optimal solution $x^\star$ for the convexified problem; y^k converges to a dual optimal solution $\lambda^\star$ for the convexified problem.

This iteration requires very little communication between nodes in a distributed system. This property may be very useful if it is expensive to compute or to optimize the convex envelopes $\hat{f}_i$. Each processor in the distributed architecture may perform the x update for one block i in parallel, with no need to communicate with other processors. The only centralized computation is the projection of x^k onto the constraints.

However, we have already seen in §2 that projecting a solution to the dual problem onto the constraint set can work very poorly for nonconvex separable problems. To understand this phenomenon better, consider a *symmetric* problem, which has the same f_i for every $i = 1, \ldots, n$, and constraint matrices A and G whose columns are identical. ADMM will not break the symmetry

between different coordinate blocks, since the iteration above is completely symmetric, resulting in a symmetric solution to the convexified problem. But we have seen in §2 that a symmetric solution is the worst sort of solution; it can have an error that grows linearly with n.

ADMM for the randomized problem. We want a solution at an extreme point of the optimal set for the convexified problem. Fortunately, it is also easy to compute the solution to the randomized problem $\mathcal{R}$ using distributed optimization, which allows us to find a point $\hat{x}$ satisfying the bound in Theorem 1.

Taking the primal and dual optimal pair $(x^\star, \lambda^\star)$ for (9) computed by the first round of ADMM iterations, we can rewrite problem (6) in ADMM consensus form. Let

$$M = \left\{ x : \hat{f}(x) + \lambda^{\star T} A x \leq \hat{f}(x^\star) + \lambda^{\star T} A x^\star \right\}.$$

We saw in §6 that M is separable, and can be written as $M = M_1 \times \cdots \times M_n$. Hence we can rewrite $\mathcal{R}$ as

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n (w_i^T x_i + \mathbf{1}_{M_i}(x_i)) + \mathbf{1}_{Ax \leq b, Gx = h}(z) \\ \text{subject to} & x = z, \end{array}$$

which gives rise to the ADMM consensus iteration

$$\begin{aligned} x_i^k &= \operatorname*{argmin}_{x \in M_i} w_i^T x + \rho/2 \|x - z_i^{k-1} + y_i^{k-1}\|_2^2 \\ z^k &= \Pi_{Ax \leq b, Gx = h}(x^k) \\ y_i^k &= y_i^{k-1} + 1/\rho (x_i^k - z_i^k). \end{aligned}$$

The solution z produced by this distributed iteration will satisfy Theorem 1.

Acknowledgements

The authors thank Haitham Hindi, Ernest Ryu and the anonymous reviewers for their very careful readings of and comments on early drafts of this paper, and Jon Borwein and Julian Revalski for their generous advice on the technical lemmas in the appendix.

A The dual of the dual is the convexified problem

In this appendix, we prove that the dual of the dual of $\mathcal{P}$ is the convexified problem $\hat{\mathcal{P}}$.

Before we begin, note that the convex envelope has a close connection to duality. Let $f^*(y) = \sup(y^T x - f(x)) = -\inf(f(x) - y^T x)$ be the (Fenchel) *conjugate* of f. Then $\hat{f}(x) = f^{**}(x)$ is the *biconjugate* of f [Roc70]. The conjugate function arises naturally when taking the dual of a problem, as we show below. Hence it should come as no suprise that the biconjugate appears upon taking the dual twice.

Below, we refer to the dual of the dual problem as the dual dual problem, the dual function of the dual problem as the dual dual function, and the variables in the dual dual problem as the dual dual variables.

Recall the primal problem, which we write as

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

We can write the Lagrangian of the primal problem as

$$L(x, \lambda, \mu) = \sum_{i=1}^n f_i(x_i) + \lambda^T (Ax - b) + \mu^T (Gx - h),$$

with dual variables $\lambda \geq 0$ and μ. The dual function $g(\lambda, \mu)$ is the minimum of the Lagrangian over x,

$$\begin{aligned} g(\lambda, \mu) &= \inf_x L(x, \lambda, \mu) \\ &= \inf_x \sum_{i=1}^n f_i(x_i) + \lambda^T (Ax - b) + \mu^T (Gx - h) \\ &= \sum_{i=1}^n \inf_{x_i} (f_i(x_i) - \gamma_i x_i) - \lambda^T b - \mu^T h \\ &= \sum_{i=1}^n -f_i^*(\gamma_i) - \lambda^T b - \mu^T h, \end{aligned}$$

where we have defined $\gamma = -A^T\lambda - G^T\mu$ in the second to last equality and used the relation $f^*(y) = -\inf(f(x) - y^T x)$ in the last.

The dual problem is to maximize the dual function over μ and λ with $\lambda \geq 0$:

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^n -f_i^*(\gamma_i) - \lambda^T b - \mu^T h \\ \text{subject to} & \gamma = -A^T\lambda - G^T\mu \\ & \lambda \geq 0. \end{array}$$

The conjugate function f_i^* is a pointwise supremum of affine functions, and so is always convex even if f_i is not. Hence the dual problem is a concave maximization problem.

To take the dual of the dual, we perform exactly the same computations again on the dual problem now instead of the primal. The dual Lagrangian is

$$L_D(\lambda, \mu, \gamma, x, y) = \sum_{i=1}^n -f_i^*(\gamma_i) - \lambda^T b - \mu^T h + x^T(\gamma + A^T\lambda + G^T) + s^T\lambda,$$

with dual dual variables $s \geq 0$ and x. We maximize the dual Lagrangian over the dual variables λ, μ, and γ to form the dual dual function

$$\begin{aligned} g_D(x, s) &= \sup_{\lambda \geq 0, \mu, \gamma} L_D(\lambda, \mu, \gamma, x, y) \\ &= \sup_{\lambda \geq 0, \mu, \gamma} \sum_{i=1}^n -f_i^*(\gamma_i) - \lambda^T b - \mu^T h + x^T(\gamma + A^T\lambda + G^T) + s^T\lambda \\ &= \sup_{\lambda \geq 0, \mu} \sum_{i=1}^n f_i^{**}(x_i) + \lambda^T(Ax + s - b) + \mu^T(Gx - h), \end{aligned}$$

using now the relation $f^*(y) = \sup(y^T x - f(x))$. This is finite only if $Ax + s - b \leq 0$ and $Gx - h = 0$. So we see

$$g_D(x, s) = \sum_{i=1}^n f_i^{**}(x_i)$$

so long as these equalities are satisfied.

To form the dual dual problem, we minimize the dual dual function over x and $s \geq 0$:

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

where we have solved for $s = b - Ax$. Hence we see that we have recovered the convexified problem by dualizing the primal twice.

B Well-posedness

The following theorem characterizes the set of vectors in the dual space for which linear optimization over a compact set S is well-posed.

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.*

This result is well-known; for example, it follows from [BDL11, §2], taking into account that if $S \subseteq \mathbf{R}^n$ is compact, then so is its convex hull $K = \mathbf{conv}(S)$ and the set of extreme points of S and K coincide. In fact, one can derive much stronger results using, for example, Alexandrov's theorem for convex functions to show quadratic decay, or finite identifiability in the case of semialgebraic functions. However, our purpose here is more modest; we merely prove the weaker result stated as Theorem 3 so that this paper may be self-contained.

Before proceeding to a proof, however, let us make sense of the statement of the theorem. 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 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 one. □

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. □

C Closure

The following lemma techical lemma will be useful in the main body of the paper.

Lemma 4. *Let $S \subset \mathbf{R}^n$ be a nonempty compact set, and let $f : S \to \mathbf{R}$ be lower semi-continuous on S. Then* $\mathbf{conv}(\mathbf{epi}\, f)$ *is closed.*

This result follows from [BHU96, Thm. 4.6], since every function defined on a compact set is in particular 1-coercive. The earliest proof known to the authors can be found in [Val70, p. 69]; for a simpler exposition, see [HUL96, Ch. X, §1.5]. Here, we provide a self-contained elementary proof for the curious reader.

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 $\mathbf{epi}\, f$,

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

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 $S \subset \mathbf{R}^n$ 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.

[AL89] E. Anderson and A. Lewis. An extension of the simplex algorithm for semi-infinite linear programming. *Mathematical Programming*, 44(1-3):247–269, 1989.

[Bar95] Alexander I. Barvinok. Problems of distance geometry and convex properties of quadratic maps. *Discrete & Computational Geometry*, 13(1):189–202, 1995.

[Bar02] Alexander Barvinok. *A course in convexity*, volume 54. American Mathematical Society Providence, 2002.

[BDL11] J. Bolte, A. Daniilidis, and A. S. Lewis. Generic optimality conditions for semialgebraic convex programs. *Mathematics of Operations Research*, 36(1):55–70, 2011.

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

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

[Ber09] D Bertsekas. *Convex optimization theory*. Athena Scientific, 2009.

[BHU96] J. Benoist and J.-B. Hiriart-Urruty. What is the subdifferential of the closed convex hull of a function? *SIAM Journal on Mathematical Analysis*, 27(6):1661–1679, 1996.

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

[BPC+11] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[BTP13] S. Bouaziz, A. Tagliasacchi, and M. Pauly. Sparse iterative closest point. In *Computer Graphics Forum*, volume 32, pages 113–123. Wiley Online Library, 2013.

[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*. Cambridge University Press, 2010.

[Cha12] R. Chartrand. Nonconvex splitting for regularized low-rank + sparse decomposition. *IEEE Transactions on Signal Processing*, 60(11):5810–5819, 2012.

[CT94] G. Chen and M. Teboulle. A proximal-based decomposition method for convex minimization problems. *Mathematical Programming*, 64:81–101, 1994.

[CW13] R. Chartrand and B. Wohlberg. A nonconvex admm algorithm for group sparsity with sparse groups. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6009–6013. IEEE, 2013.

[DBEY13] N. Derbinsky, J. Bento, V. Elser, and J. S. Yedidia. An improved three-weight message-passing algorithm. *arXiv preprint arXiv:1305.1961*, 2013.

[EB92] J. Eckstein and D. P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55:293–318, 1992.

[EF93] J. Eckstein and M. Fukushima. Some reformulations and applications of the alternating direction method of multipliers. *Large Scale Optimization: State of the Art*, pages 119–138, 1993.

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

[FG83a] M. Fortin and R. Glowinski. *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*. North-Holland: Amsterdam, 1983.

[FG83b] M. Fortin and R. Glowinski. On decomposition-coordination methods using an augmented Lagrangian. In M. Fortin and R. Glowinski, editors, *Augmented Lagrangian Methods: Applications to the Solution of Boundary-Value Problems*. North-Holland: Amsterdam, 1983.

[Fuk92] M. Fukushima. Application of the alternating direction method of multipliers to separable convex programming problems. *Computational Optimization and Applications*, 1:93–111, 1992.

[Gab83] D. Gabay. Applications of the method of multipliers to variational inequalities. In M. Fortin and R. Glowinski, editors, *Augmented Lagrangian Methods: Applications to the Solution of Boundary-Value Problems*. North-Holland: Amsterdam, 1983.

[Glo14] Roland Glowinski. On alternating direction methods of multipliers: A historical perspective. In William Fitzgibbon, Yuri A. Kuznetsov, Pekka Neittaanmki, and Olivier Pironneau, editors, *Modeling, Simulation and Optimization for Science and Technology*, volume 34 of *Computational Methods in Applied Sciences*, pages 59–82. Springer Netherlands, 2014.

[GM75] R. Glowinski and A. Marrocco. Sur l'approximation, par elements finis d'ordre un, et la resolution, par penalisation-dualité, d'une

classe de problems de Dirichlet non lineares. *Revue Française d'Automatique, Informatique, et Recherche Opérationelle*, 9:41–76, 1975.

[GM76] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximations. *Computers and Mathematics with Applications*, 2:17–40, 1976.

[GT87] R. Glowinski and P. Le Tallec. Augmented Lagrangian methods for the solution of variational problems. Technical Report 2965, University of Wisconsin-Madison, 1987.

[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.

[GZ13] R. Gilimyanov and H. Zhuang. Power allocation in OFDMA networks: An ADMM approach. In *5th Traditional Youth Summer School on Control, Information, and Optimization*, 2013.

[HL12] M. Hong and Z.-Q. Luo. On the linear convergence of the alternating direction method of multipliers. *arXiv preprint arXiv:1208.3922*, 2012.

[HPN00] R. Horst, P. M. Pardalos, and Van T. N. *Introduction to global optimization*. Kluwer Academic Pub, 2000.

[HUL96] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms I: Part 1: Fundamentals*, volume 305. Springer, 1996.

[HY12] B. He and X. Yuan. On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2):700–709, 2012.

[KT12] T. Kanamori and A. Takeda. Non-convex optimization on Stiefel manifold and applications to machine learning. In T. Huang, Z. Zeng, C. Li, and C. Leung, editors, *Neural Information Processing*, volume 7663 of *Lecture Notes in Computer Science*, pages 109–116. Springer, 2012.

[LM79] P. L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16:964–979, 1979.

[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.

[MWF14] S. Magnusson, P. C. Weeraddana, and C. Fischione. A distributed approach for the optimal power flow problem based on ADMM and sequential convex approximations. *arXiv preprint arXiv:1401.4621*, 2014.

[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.

[Pat96] G. Pataki. Cone-LP's and semidefinite programs: Geometry and a simplex-type method. In *Integer programming and combinatorial optimization*, pages 162–174. Springer, 1996.

[Pat98] G. Pataki. On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Mathematics of Operations Research*, 23(2):339–358, 1998.

[PR55] D. W. Peaceman and H. H. Rachford. The numerical solution of parabolic and elliptic differential equations. *Journal of the Society for Industrial and Applied Mathematics*, 3:28–41, 1955.

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

[Roc70] R. Rockafellar. *Convex analysis*. Princeton University Press, 1970.

[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 nonconvex preferences. *Econometrica: Journal of the Econometric Society*, pages 25–38, 1969.

[TS02] M. Tawarmalani and N. Sahinidis. *Convexification and global optimization in continuous and mixed-integer nonlinear programming: theory, algorithms, software, and applications*, volume 65. Springer Science & Business Media, 2002.

[Tse91] P. Tseng. Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM Journal on Control and Optimization*, 29(1):119–138, 1991.

[UB13] M. Udell and S. Boyd. Maximizing a sum of sigmoids. Available at `http://www.stanford.edu/~boyd/papers/max_sum_sigmoids.html`, 2013.

[Val70] M. Valadier. Intégration de convexes fermés notamment d'épigraphes inf-convolution continue. *ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique*, 4(R2):57–73, 1970.

[VEG+14] R. Vujanic, P. M. Esfahani, P. Goulart, S. Mariethoz, and M. Morari. Vanishing duality gap in large scale mixed-integer optimization: a solution method with power system applications. *submitted to Journal of Mathematical Programming*, 2014.

[Zha10] Y. Zhang. Recent advances in alternating direction methods: Practice and theory. In *IPAM Workshop on Continuous Optimization*, 2010.