Radial Subgradient Descent

Benjamin Grimmer*

Abstract

We present a subgradient method for minimizing non-smooth, non-Lipschitz convex optimization problems. The only structure assumed is that a strictly feasible point is known. We extend the work of Renegar [1] by taking a different perspective, leading to an algorithm which is conceptually more natural, has notably improved convergence rates, and for which the analysis is surprisingly simple. At each iteration, the algorithm takes a subgradient step and then performs a line search to move radially towards (or away from) the known feasible point. Our convergence results have striking similarities to those of traditional methods that require Lipschitz continuity. Costly orthogonal projections typical of subgradient methods are entirely avoided.

1 Introduction

We consider the convex optimization problem of minimizing a lower-semicontinuous convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ when a point x_0 in the interior of the domain of f is known. Importantly, it is not assumed that f is either smooth or Lipschitz. Note that any constrained convex optimization problem fits under this model by setting the function value of all infeasible points to infinity (provided the feasible region is closed and convex, and a point exists in the interior of both the feasible region and the domain of f). This model can easily be extended to allow affine constraints by considering the relative interior of the domain instead.

Without loss of generality, we have $x_0 = \vec{0} \in \text{int dom} f$ and $f(\vec{0}) < 0$, which can be achieved by replacing $f(x)$ by $f(x + x_0) - f(x_0) - h$ (for any positive constant $h > 0$). Let $f^* = \inf\{f(x)\}$ denote the function's minimum value (which equals $-\infty$ if the function is unbounded below).

This general minimization problem has been the focus of a recent paper by Renegar [1]. Renegar develops a framework for converting the original non-Lipschitz problem into an equivalent Lipschitz problem, in a slightly lifted space. This transformation is geometric in nature and applied to a conic reformulation of the original problem. By applying a subgradient approach to the reformulated problem, Renegar achieves convergence bounds analogous to those of traditional methods that assume Lipschitz continuity.

One difference in Renegar's approach is that it guarantees relative accuracy of a solution rather than absolute accuracy. A point x has absolute accuracy of ϵ if $f(x) - f^* < \epsilon$.

*bdg79@cornell.edu; Cornell University, Ithaca NY

In contrast, a point x has relative accuracy of ϵ if $(f(x) - f^*)/(0 - f^*) < \epsilon$. Note that here we measure error relative to an objective value of 0 since we assume $f(x_0) < 0$. This relative accuracy can be interpreted as a multiplicative accuracy through simple algebraic rearrangement:

$$\frac{f(x) - f^*}{0 - f^*} < \epsilon \iff f(x) < f^*(1 - \epsilon).$$

From this framework, Renegar presents two algorithms addressing when the optimal objective value is known and when it is unknown (referred to below as Algorithms B and A in [1], respectively). The two (paraphrased) convergence bounds stated in Theorems 1 and 2 are then proven.

Renegar's analysis requires two additional assumptions on f beyond the basic assumptions we make of lower-semicontinuity and convexity. In particular, they further assume that the minimum objective value is finite and attained at some point, and that the set of optimal solutions is bounded.

Let $\langle \cdot, \cdot \rangle$ denote an inner product on $\mathbb{R}^n$ with associated norm $\|\cdot\|$. Define D to be the diameter of the sublevel set $\{x \mid f(x) \leq f(\vec{0})\}$ and $R = \sup\{r \in \mathbb{R} \mid f(x) \leq 0 \text{ for all } x \text{ with } \|x\| \leq r\}$. These scalars appear only in convergence bounds, but are never assumed to be known. Note that the set of optimal solutions must be bounded for D to be finite.

Theorem 1 (Renegar [1], Theorem 1.1)**.** *If $\{x_i\}$ is the sequence generated by Algorithm B in [1] (which is given the value of f^*), then for any $\epsilon > 0$, some iteration*

$$i \leq \left\lceil 8\left(\frac{D}{R}\right)^2 \left(\frac{1}{\epsilon^2} + \frac{1}{\epsilon} \log_{4/3}\left(1 + \frac{D}{R}\right)\right)\right\rceil \quad \textit{has} \quad \frac{f(x_i) - f^*}{0 - f^*} \leq \epsilon.$$

Theorem 2 (Renegar [1], Theorem 1.2)**.** *Consider any $0 < \epsilon < 1$. If $\{x_i\}$ is the sequence generated by Algorithm A in [1] (which is given the value of ϵ), then some iteration*

$$i \leq \left\lceil 4\left(\frac{D}{R}\right)^2 \left(\frac{4}{3}\left(\frac{1-\epsilon}{\epsilon}\right)^2 + 4\frac{1-\epsilon}{\epsilon} + \log_2\left(\frac{1-\epsilon}{\epsilon}\right) + \log_2\left(\frac{D}{R}\right) + 1\right)\right\rceil \textit{ has } \frac{f(x_i) - f^*}{0 - f^*} \leq \epsilon.$$

These bounds are remarkable in that they attain the same rate of growth with respect to ϵ as traditional methods that assume Lipschitz continuity. However, the constants involved could be very large. In particular, the value of D can be enormous if a small perturbation of the problem would have an unbounded set of optimal solutions.

Another notable property of Renegar's algorithms is that they completely avoid computing orthogonal projections. Instead only radial projections are done, which only require a line search to compute. This can lead to substantial improvements in runtime over projected subgradient methods, which require an orthogonal projection every iteration.

We take a different perspective on the transformation underlying Renegar's framework. Instead of first converting the convex problem into conic form in a lifted space, we define an equivalent transformation directly in terms of the original function.

Before stating our function-oriented transformation, we define the *perspective function* of f for any $\gamma > 0$ as $f^p(x, \gamma) = \gamma f(x/\gamma)$. Note that perspective functions have been studied

in a variety of other contexts. Two interesting recent works have considered optimization problems where perspective functions occur naturally. In [2], Combettes and Müller investigate properties of the proximity operator of a perspective function. In [3], Aravkin et al. construct a duality framework based on gauges, which extends to perspective functions.

Based on the perspective function of f, our functional version of Renegar's transformation is defined as follows.

Definition 1. *The radial reformulation of f of level $z \in \mathbb{R}$ is given by*

$$\gamma_z(x) = \inf\{\gamma > 0 \mid f^p(x, \gamma) \leq z\}.$$

As a simple consequence of Renegar's transformation converting non-Lipschitz problems into Lipschitz ones, the radial reformulation of any convex function is both convex and Lipschitz (shown in Lemma 1 and Proposition 2). As a result, the radial reformulation is amenable to the application of traditional subgradient descent methods, even if f is not.

Using this functional version of Renegar's transformation, we present a modified version of Renegar's algorithms which is both conceptually simpler and achieves notably improved convergence bounds. We let $\partial\gamma_z(x)$ denote the set of subgradients of $\gamma_z(\cdot)$ at x. Then our algorithm is stated in Algorithm 1 with step sizes given by a positive sequence $\{\alpha_i\}$.

Algorithm 1 Radial Subgradient Descent

$x_0 = \vec{0}$, $z_0 = f(x_0)$
for $i = 0, 1, \ldots$ **do**
  Select $\zeta_i \in \partial\gamma_{z_i}(x_i)$ {Pick a subgradient of $\gamma_{z_i}(\cdot)$}
  $\tilde{x}_{i+1} = x_i - \alpha_i\zeta_i$ {Move in the subgradient direction}
  if $\gamma_{z_i}(\tilde{x}_{i+1}) = 0$ **then** report unbounded objective and terminate.
  $x_{i+1} = \dfrac{\tilde{x}_{i+1}}{\gamma_{z_i}(\tilde{x}_{i+1})}$, $z_{i+1} = \dfrac{z_i}{\gamma_{z_i}(\tilde{x}_{i+1})}$ {Radially update current solution}
end for

Critical to the strength of our results is the specification of the initial iterate, $x_0 = \vec{0}$. Each iteration of this algorithm takes a subgradient step (with respect to a radial reformulation $\gamma_{z_i}(\cdot)$ of changing level) and then moves the resulting point radially towards (or away from) the origin. This radial movement ensures that each iterate x_i lies in the domain of f.

This algorithm differs from Renegar's approach by doing a radial update at every step. Algorithm A of [1] only does an update when the threshold $\gamma_z(x) \leq 3/4$ is met, and Algorithm B of [1] never does radial updates. Recently in [4], Freund and Lu presented an interesting approach to first-order optimization with similarities to Renegar's approach. Their method utilizes a similar thresholding condition for doing periodic updates.

We first give a general convergence guarantee for the Radial Subgradient Descent algorithm, where $\text{dist}(x_0, X)$ denotes the minimum distance from $x_0 = \vec{0}$ to a set X. As a consequence, we find that proper selection of the step size will guarantee a subsequence of the iterates has objective values converging to the optimal value.

Theorem 3. *Let $\{x_i\}$ be the sequence generated by Algorithm 1 with steps sizes α_i. Consider any $\hat{f} < 0$ with nonempty level set $\hat{X} = \{x \mid f(x) = \hat{f}\}$. Then for any $k \geq 0$, some iteration*

$i \leq k$ either identifies that f is unbounded on the ray $\{t\tilde{x}_{i+1} \mid t > 0\}$ or has

$$\frac{f(x_i) - \hat{f}}{0 - f(x_i)} \leq \frac{\operatorname{dist}(x_0, \hat{X})^2 + \frac{1}{R^2}\sum_{j=0}^{k} \alpha_j^2 \left(\frac{\hat{f}}{z_j}\right)^2}{2\sum_{j=0}^{k} \alpha_j \frac{\hat{f}}{z_j}}.$$

Corollary 4. *Consider any positive sequence $\{\beta_i\}$ with $\sum_{i=0}^{\infty} \beta_i = \infty$ and $\sum_{i=0}^{\infty} \beta_i^2 < \infty$. Let $\{x_i\}$ be the sequence generated by Algorithm 1 with $\alpha_i = -z_i\beta_i$. Then either some iteration identifies that f is unbounded on the ray $\{t\tilde{x}_{i+1} \mid t > 0\}$ or*

$$\lim_{k\to\infty} \min_{i\leq k}\{f(x_i)\} = f^*.$$

Although Corollary 4 proves the convergence of our algorithm, it does not provide any guarantees on the rate of convergence. As done in Renegar's analysis, we bound the convergence rate of our algorithm when the optimal objective f^* is known and when it is unknown. The resulting convergence bounds for two particular choices of step sizes are given in the following theorems.

Theorem 5. *Suppose f^* is finite and attained at some set of points X^*. Let $\{x_i\}$ be the sequence generated by Algorithm 1 with $\alpha_i = \dfrac{z_i - f^*}{0 - f^*}\dfrac{1}{\|\zeta_i\|^2}$. Then for any $\epsilon > 0$, some iteration*

$$i \leq \left\lceil \frac{\operatorname{dist}(x_0, X^*)^2}{R^2}\frac{1}{\epsilon^2} \right\rceil \quad \text{has} \quad \frac{f(x_i) - f^*}{0 - f^*} \leq \epsilon.$$

Theorem 6. *Consider any $\epsilon > 0$ and $\hat{f} < 0$ with nonempty level set $\hat{X} = \{x \mid f(x) = \hat{f}\}$. Let $\{x_i\}$ be the sequence generated by Algorithm 1 with $\alpha_i = \dfrac{\epsilon}{2\|\zeta_i\|^2}$. Then some iteration*

$$i \leq \left\lceil \frac{4}{3}\frac{\operatorname{dist}(x_0, \hat{X})^2}{R^2}\frac{1}{\epsilon^2} \right\rceil$$

either identifies that f is unbounded on the ray $\{t\tilde{x}_{i+1} \mid t > 0\}$ or has

$$\frac{f(x_i) - \hat{f}}{0 - \hat{f}} \leq \epsilon.$$

These convergence bounds are remarkably simple and should have fairly small constants in practice. Importantly, the bounds have no dependence on D, which means (unlike [1]) we do not need to assume the problem has a bounded set of optimal solutions.

It is also worth noting that this algorithm, like Renegar's algorithms, completely avoids computing orthogonal projections. As a result, its per iteration cost can be drastically less than that of standard projected subgradient descent methods. We defer a deeper discussion of this improvement in iteration cost to [1].

In Section 2, we establish a number of relevant properties of our radial reformulation, which follow from the equivalent structures in Renegar's framework. Then in Section 3, we prove our main results on the convergence of the Radial Subgradient Descent algorithm.

2 Preliminaries

Renegar's framework begins by converting the problem of minimizing f into conic form. Let $\mathcal{K}$ be the closure of $\{(xs, s, ts) \mid s > 0, (x, t) \in \text{epi} f\}$, which is the conic extension of the epigraph of f. Note that the restriction of $\mathcal{K}$ to $s = 1$ is exactly the epigraph of f. Then Renegar considers the following equivalent conic form problem

$$\begin{cases} \min & t \\ \text{s.t.} & s = 1, (x, s, t) \in \mathcal{K}. \end{cases} \tag{1}$$

Observe that $(\vec{0}, 1, 0)$ lies in the interior of $\mathcal{K}$. Then Renegar defines the following function, which lies at the heart of the framework. Although this function was originally developed for any conic program, we state it specifically in terms the above program.

$$\lambda(x, s, t) = \inf\{\lambda \mid (x, s, t) - \lambda(\vec{0}, 1, 0) \notin \mathcal{K}\}.$$

At this point, it is easy to show the connection between this function when restricted to $t = z$ and our radial reformulation of level z.

Lemma 1. *For any $x \in \mathbb{R}^n$ and $z \in \mathbb{R}$, $\gamma_z(x) = 1 - \lambda(x, 1, z)$.*

Proof. Follows directly from the definitions of $\mathcal{K}$ and $\gamma_z(\cdot)$:

$$\begin{aligned} \lambda(x, 1, z) &= \sup\Big\{\lambda \mid (x, 1, z) - \lambda(\vec{0}, 1, 0) \in \mathcal{K}\Big\} \\ &= \sup\left\{\lambda \mid 1 - \lambda > 0, (\frac{x}{1-\lambda}, \frac{z}{1-\lambda}) \in \text{epi} f\right\} \\ &= \sup\left\{\lambda < 1 \mid f\left(\frac{x}{1-\lambda}\right) \leq \frac{z}{1-\lambda}\right\} \\ &= 1 - \inf\left\{(1-\lambda) > 0 \mid (1-\lambda) f\left(\frac{x}{1-\lambda}\right) \leq z\right\} \\ &= 1 - \gamma_z(x). \end{aligned}$$

□

As a consequence of Propositions 2.1 and 3.2 of [1], we know that the radial reformulation is both convex and Lipschitz.

Proposition 2. *For any $z \in \mathbb{R}$, the radial reformulation of level z, $\gamma_z(\cdot)$, is convex and Lipschitz with constant $1/R$ (independent of the level z).*

We now see that the radial reformulation is notably more well-behaved than the original function f. However, to justify it as a meaningful proxy for the original function in an optimization setting, we need to relate their minimum values. In the following proposition, we establish such a connection between f and any radial reformulation with a negative level.

Note that f^p is strictly decreasing in γ. To see this, observe that, for any $x \in \mathbb{R}^n$, all sufficiently large γ have $f(x/\gamma) < f(\vec{0})/2 < 0$. Then it follows that $\lim_{\gamma \to \infty} f^p(x, \gamma) \leq \lim_{\gamma \to \infty} \gamma f(\vec{0})/2 = -\infty$. Since perspective functions are convex (see [5] for an elementary proof of this fact), we conclude that f^p is strictly decreasing in γ.

Proposition 3. *For any $z < 0$, the minimum value of $\gamma_z(\cdot)$ is z/f^* (where $z/-\infty := 0$). Further, if $\gamma_z(x) = 0$, then $\lim_{t\to\infty} f(tx) = -\infty$ (that is, x is an unbounded direction).*

Proof. First we show that this minimum value lower bounds $\gamma_z(x)$ for all $x \in \mathbb{R}^n$. Since $f((f^*/z)x) \geq f^*$, we know that $\gamma = z/f^*$ has $f^p(x,\gamma) \geq z$. Thus $\gamma_z(x) \geq z/f^*$ since f^p is strictly decreasing in γ.

Consider any sequence $\{x_i\}$ with $f(x_i) < 0$ and $\lim_{i\to\infty} f(x_i) = f^*$. Observe that

$$f^p\left(\frac{z}{f(x_i)}x_i, \frac{z}{f(x_i)}\right) = \frac{z}{f(x_i)}f(x_i) = z.$$

Since f^p is strictly decreasing in γ, we have $\gamma_z((z/f(x_i))x_i) = z/f(x_i)$. It follows that $\lim_{i\to\infty} \gamma_z((z/f(x_i))x_i) = z/f^*$. Then our lower bound is indeed the minimum value of the radial reformulation.

Our second observation follows from the definition of the radial reformulation (using the change of variables $t = 1/\gamma$):

$$\begin{aligned}\gamma_z(x) = 0 &\Leftrightarrow \inf\{\gamma > 0 \mid f(x/\gamma) < z/\gamma\} = 0\\ &\Leftrightarrow \sup\{t > 0 \mid f(tx) < tz\} = \infty\\ &\Rightarrow \lim_{t\to\infty} f(tx) = -\infty.\end{aligned}$$

□

Finally, we give a characterization of the subgradients of our radial reformulation. Although this description is not necessary for our analysis, it helps establish the practicality of the Radial Subgradient Descent algorithm. We see that the subgradients of $\gamma_z(\cdot)$ can be computed easily from normal vectors of the epigraph of the original function. This result follows as a direct consequence of Proposition 7.1 of [1].

Proposition 4. *For any $z < 0$, the subgradients of the radial reformulation are given by*

$$\partial\gamma_z(x) = \left\{\frac{\gamma_z(x)}{\langle\zeta, x\rangle + \delta z}\zeta \mid (\vec{0}, 0) \neq (\zeta, \delta) \in N_{\text{epi}f}\left(\frac{x}{\gamma_z(x)}, \frac{z}{\gamma_z(x)}\right)\right\}.$$

3 Analysis of Convergence

First, we observe the following two properties hold at each iteration of our algorithm.

Lemma 5. *At any iteration $k \geq 0$ of Algorithm 1, $\gamma_{z_k}(x_k) = 1$.*

Proof. When $k = 0$, we have $f^p(x_0, 1) = z_0$, and so the result follows from f^p being strictly decreasing in γ. The general case follows from simple algebraic manipulation:

$$\begin{aligned}\gamma_{z_{k+1}}(x_{k+1}) &= \inf\{\gamma > 0 \mid \gamma f(x_{k+1}/\gamma) \leq z_{k+1}\}\\ &= \inf\left\{\gamma > 0 \mid \gamma f\left(\frac{\tilde{x}_{k+1}}{\gamma_{z_k}(\tilde{x}_{k+1})\gamma}\right) \leq z_{k+1}\right\}\\ &= \inf\left\{\gamma > 0 \mid \gamma_{z_k}(\tilde{x}_{k+1})\gamma f\left(\frac{\tilde{x}_{k+1}}{\gamma_{z_k}(\tilde{x}_{k+1})\gamma}\right) \leq z_k\right\}\\ &= \frac{1}{\gamma_{z_k}(\tilde{x}_{k+1})}\gamma_{z_k}(\tilde{x}_{k+1}) = 1.\end{aligned}$$

□

Lemma 6. *At any iteration $k \geq 0$ of Algorithm 1, the following ordering holds*

$$f^* \leq f(x_k) \leq z_k < 0.$$

Proof. The first inequality follows from f^* being the minimum value of f. The second inequality is trivially true when $k = 0$. From the lower-semicontinuity of $f^p(x, \gamma)$, we know $f^p(\tilde{x}_{k+1}, \gamma_{z_k}(\tilde{x}_{k+1})) \leq z_k$. Then we have the second inequality in general since $f(x_{k+1}) = f(\tilde{x}_{k+1}/\gamma_{z_k}(\tilde{x}_{k+1})) \leq z_k/\gamma_{z_k}(\tilde{x}_{k+1}) = z_{k+1}$. The third inequality follows inductively since $z_0 < 0$ and $z_{k+1} = z_k/\gamma_{z_k}(\tilde{x}_{k+1}) < 0$. □

The traditional analysis of subgradient descent, assuming Lipschitz continuity, is based on an elementary inequality, which is stated in the following lemma.

Lemma 7. *Consider any convex function $g : \mathbb{R}^n \to \mathbb{R}$, $x, y \in \mathbb{R}^n$, and $\zeta \in \partial g(x)$. Then for any $\alpha > 0$,*

$$\|(x - \alpha\zeta) - y\|^2 \leq \|x - y\|^2 - 2\alpha(g(x) - g(y)) + \alpha^2\|\zeta\|^2.$$

Proof. Follows directly from applying the subgradient inequality, $g(y) \geq g(x) + \langle \zeta, y - x \rangle$:

$$\begin{aligned}\|(x - \alpha\zeta) - y\|^2 &= \|x - y\|^2 - 2\alpha\langle \zeta, x - y\rangle + \alpha^2\|\zeta\|^2 \\ &\leq \|x - y\|^2 - 2\alpha(g(x) - g(y)) + \alpha^2\|\zeta\|^2.\end{aligned}$$ □

The core of proving the traditional subgradient descent bounds is to inductively apply this lemma at each iteration. However, such an approach cannot be applied directly to Algorithm 1 since the iterates are rescaled every iteration and the underlying function changes every iteration. The key to proving our convergence bounds is setting up a modified inequality that can be applied inductively. This is done in the following lemma.

Lemma 8. *Consider any y with $f(y) < 0$. Then at any iteration $k \geq 0$ of Algorithm 1,*

$$\|\frac{f(y)}{z_{k+1}}x_{k+1} - y\|^2 \leq \|\frac{f(y)}{z_k}x_k - y\|^2 - 2\alpha_k\frac{f(y)}{z_k}\frac{z_k - f(y)}{0 - z_k} + \alpha_k^2\left(\frac{f(y)}{z_k}\right)^2\|\zeta_k\|^2.$$

Proof. Applying Lemma 7 on $\gamma_{z_k}(\cdot)$ with x_k and $\dfrac{z_k}{f(y)}y$ implies

$$\|\tilde{x}_{k+1} - \frac{z_k}{f(y)}y\|^2 \leq \|x_k - \frac{z_k}{f(y)}y\|^2 - 2\alpha_k\left(\gamma_{z_k}(x_k) - \gamma_{z_k}\left(\frac{z_k}{f(y)}y\right)\right) + \alpha_k^2\|\zeta_k\|^2. \tag{2}$$

The value of $\gamma_{z_k}\left(\dfrac{z_k}{f(y)}y\right)$ can be derived directly. Observe that

$$f^p\left(\frac{z_k}{f(y)}y, \frac{z_k}{f(y)}\right) = \frac{z_k}{f(y)}f(y) = z_k.$$

Then $\gamma_{z_k}\left(\dfrac{z_k}{f(y)}y\right) = \dfrac{z_k}{f(y)}$ since f^p is strictly decreasing in γ. Combining this with Lemma 5 allows us to restate our inequality as

$$\|\tilde{x}_{k+1} - \frac{z_k}{f(y)}y\|^2 \leq \|x_k - \frac{z_k}{f(y)}y\|^2 - 2\alpha_k\left(1 - \frac{z_k}{f(y)}\right) + \alpha_k^2\|\zeta_k\|^2. \tag{3}$$

Multiplying through by $(f(y)/z_k)^2$ yields

$$\|\frac{f(y)}{z_k}\tilde{x}_{k+1} - y\|^2 \leq \|\frac{f(y)}{z_k}x_k - y\|^2 - 2\alpha_k\frac{f(y)}{z_k}\frac{z_k - f(y)}{0 - z_k} + \alpha_k^2\left(\frac{f(y)}{z_k}\right)^2\|\zeta_k\|^2. \tag{4}$$

Noting that $\tilde{x}_{k+1} = \dfrac{z_k}{z_{k+1}}x_{k+1}$ completes the proof. □

3.1 Proof of Theorem 3

We assume that for all $i \leq k$, we have $\gamma_{z_i}(\tilde{x}_{i+1}) > 0$ (otherwise the theorem immediately holds by Proposition 3). Then the first k iterates of Algorithm 1 are well-defined. Consider any $\hat{x} \in \hat{X}$. Inductively applying Lemma 8 with $y = \hat{x}$ produces

$$\|\frac{\hat{f}}{z_{k+1}}x_{k+1} - \hat{x}\|^2 \leq \|\frac{\hat{f}}{z_0}x_0 - \hat{x}\|^2 - \sum_{i=0}^{k}\left(2\alpha_i\frac{\hat{f}}{z_i}\frac{z_i - \hat{f}}{0 - z_i} - \alpha_i^2\left(\frac{\hat{f}}{z_i}\right)^2\|\zeta_i\|^2\right). \tag{5}$$

Noting that $x_0 = \vec{0}$, this implies

$$2\sum_{i=0}^{k}\alpha_i\frac{\hat{f}}{z_i}\frac{z_i - \hat{f}}{0 - z_i} \leq \|\hat{x}\|^2 + \sum_{i=0}^{k}\alpha_i^2\left(\frac{\hat{f}}{z_i}\right)^2\|\zeta_i\|^2. \tag{6}$$

Minimizing over all $\hat{x} \in \hat{X}$, we have

$$2\sum_{i=0}^{k}\alpha_i\frac{\hat{f}}{z_i}\frac{z_i - \hat{f}}{0 - z_i} \leq \operatorname{dist}(x_0, \hat{X})^2 + \sum_{i=0}^{k}\alpha_i^2\left(\frac{\hat{f}}{z_i}\right)^2\|\zeta_i\|^2. \tag{7}$$

From Proposition 2, we have $\|\zeta_i\| \leq 1/R$. Then rearrangement of this inequality gives

$$\min_{i\leq k}\left\{\frac{z_i - \hat{f}}{0 - z_i}\right\} \leq \frac{\operatorname{dist}(x_0, \hat{X})^2 + \frac{1}{R^2}\sum_{i=0}^{k}\alpha_i^2\left(\frac{\hat{f}}{z_i}\right)^2}{2\sum_{i=0}^{k}\alpha_i\frac{\hat{f}}{z_i}}. \tag{8}$$

Then Theorem 3 follows from the fact that $f(x_i) \leq z_i < 0$, as shown in Lemma 6.

3.2 Proof of Corollary 4

Corollary 4 follows directly from Theorem 3. By Proposition 3, the algorithm will correctly report unbounded objective if it ever encounters $\gamma_{z_i}(\tilde{x}_{i+1}) = 0$ (and thus the corollary holds). So we assume this never occurs, which implies the sequence of iterates $\{x_i\}$ is well-defined. By selecting $\alpha_i = -z_i\beta_i$, the upper bound in Theorem 3 converges to zero:

$$\lim_{k\to\infty}\frac{\operatorname{dist}(x_0, \hat{X})^2 + \frac{1}{R^2}\sum_{i=0}^{k}\alpha_i^2\left(\frac{\hat{f}}{z_i}\right)^2}{2\sum_{i=0}^{k}\alpha_i\frac{\hat{f}}{z_i}} = \lim_{k\to\infty}\frac{\operatorname{dist}(x_0, \hat{X})^2 + \hat{f}^2\frac{1}{R^2}\sum_{i=0}^{k}\beta_i^2}{-2\hat{f}\sum_{i=0}^{k}\beta_i} = 0.$$

Then we have the following convergence result

$$\lim_{k\to\infty} \min_{i\leq k} \left\{ \frac{f(x_i) - \hat{f}}{0 - f(x_i)} \right\} \leq 0.$$

This implies $\lim_{k\to\infty} \min_{i\leq k}\{f(x_i)\} \leq \hat{f}$. Considering a sequence of $\hat{f}$ approaching f^* gives the desired result.

3.3 Proof of Theorem 5

Since we assume f^* is finite, we know all $\gamma_{z_i}(\tilde{x}_{i+1}) \geq z_i/f^* > 0$ (by Proposition 3). Thus the sequence of iterates $\{x_i\}$ is well-defined. Consider any $k \geq 0$. Then taking (7) with $\hat{f} = f^*$ gives

$$2\sum_{i=0}^{k} \alpha_i \frac{f^*}{z_i} \frac{z_i - f^*}{0 - z_i} \leq \operatorname{dist}(x_0, X^*)^2 + \sum_{i=0}^{k} \alpha_i^2 \left(\frac{f^*}{z_i}\right)^2 \|\zeta_i\|^2. \tag{9}$$

Combining this with our choice of step size $\alpha_i = \dfrac{z_i - f^*}{0 - f^*} \dfrac{1}{\|\zeta_i\|^2}$ yields

$$2\sum_{i=0}^{k} \frac{1}{\|\zeta_i\|^2} \left(\frac{z_i - f^*}{0 - z_i}\right)^2 \leq \operatorname{dist}(x_0, X^*)^2 + \sum_{i=0}^{k} \frac{1}{\|\zeta_i\|^2} \left(\frac{z_i - f^*}{0 - z_i}\right)^2 \tag{10}$$

$$\implies \sum_{i=0}^{k} \frac{1}{\|\zeta_i\|^2} \left(\frac{z_i - f^*}{0 - z_i}\right)^2 \leq \operatorname{dist}(x_0, X^*)^2 \tag{11}$$

$$\implies (k+1) \min_{i\leq k} \left\{ \left(\frac{1}{\|\zeta_i\|^2} \frac{z_i - f^*}{0 - z_i}\right)^2 \right\} \leq \operatorname{dist}(x_0, X^*)^2. \tag{12}$$

From Proposition 2, we know $\|\zeta_i\| \leq 1/R$, and thus

$$\min_{i\leq k} \left\{ \left(\frac{z_i - f^*}{0 - z_i}\right)^2 \right\} \leq \frac{\operatorname{dist}(x_0, X^*)^2}{R^2(k+1)}. \tag{13}$$

From Lemma 6, we have $f^* \leq f(x_i) \leq z_i$, which completes our proof by implying

$$\min_{i\leq k} \left\{ \frac{f(x_i) - f^*}{0 - f^*} \right\} \leq \frac{\operatorname{dist}(x_0, X^*)}{R\sqrt{k+1}}. \tag{14}$$

3.4 Proof of Theorem 6

Let $k = \left\lceil \frac{4}{3} \frac{\operatorname{dist}(x_0, \hat{X})^2}{R^2} \frac{1}{\epsilon^2} \right\rceil$. We assume $\gamma_{z_i}(\tilde{x}_{i+1}) > 0$ for all $i \leq k$ (otherwise the theorem immediately holds by Proposition 3). Then the first k iterates of Algorithm 1 are well-defined. Combining (7) with our choice of step size $\alpha_i = \dfrac{\epsilon}{2\|\zeta_i\|^2}$ yields

$$\sum_{i=0}^{k} \frac{\epsilon}{\|\zeta_i\|^2} \frac{\hat{f}}{z_i} \frac{z_i - \hat{f}}{0 - z_i} \leq \operatorname{dist}(x_0, \hat{X})^2 + \sum_{i=0}^{k} \left(\frac{\hat{f}}{z_i}\right)^2 \frac{\epsilon^2}{4\|\zeta_i\|^2} \tag{15}$$

$$\implies \sum_{i=0}^{k} \frac{\epsilon}{\|\zeta_i\|^2} \left(\frac{\hat{f}}{z_i}\right)^2 \left(\frac{z_i - \hat{f}}{0 - \hat{f}} - \frac{\epsilon}{4}\right) \leq \text{dist}(x_0, \hat{X})^2 \tag{16}$$

$$\implies \epsilon(k+1) \min_{i \leq k} \left\{ \frac{1}{\|\zeta_i\|^2} \left(\frac{\hat{f}}{z_i}\right)^2 \left(\frac{z_i - \hat{f}}{0 - \hat{f}} - \frac{\epsilon}{4}\right) \right\} \leq \text{dist}(x_0, \hat{X})^2. \tag{17}$$

If any $i \leq k$ has $\hat{f} > z_i$, the theorem holds (as this would imply $f(x_i) - \hat{f} \leq z_i - \hat{f} < 0$). So we now assume $\hat{f}/z_i \geq 1$. Then, noting that $\|\zeta_i\| \leq 1/R$, we can simplify our inequality to

$$\min_{i \leq k} \left\{ \frac{z_i - \hat{f}}{0 - \hat{f}} - \frac{\epsilon}{4} \right\} \leq \frac{\text{dist}(x_0, \hat{X})^2}{\epsilon R^2 (k+1)}. \tag{18}$$

Since $f(x_i) \leq z_i$ from Lemma 6, we have the following, completing our proof of Theorem 6,

$$\min_{i \leq k} \left\{ \frac{f(x_i) - \hat{f}}{0 - \hat{f}} \right\} \leq \frac{\text{dist}(x_0, \hat{X})^2}{\epsilon R^2 (k+1)} + \frac{\epsilon}{4}. \tag{19}$$

Acknowledgments. The author wishes to express his deepest gratitude to Jim Renegar for numerous insightful discussions helping motivate this work, and for advising on both the presentation and positioning of this paper.

References

[1] James Renegar. "Efficient" subgradient methods for general convex optimization. *SIAM Journal on Optimization*, 26(4):2649–2676, 2016.

[2] Patrick L. Combettes and Christian L. Müller. Perspective functions: Proximal calculus and applications in high-dimensional statistics. *Journal of Mathematical Analysis and Applications*, 2016.

[3] A. Y. Aravkin, J. V. Burke, D. Drusvyatskiy, M. P. Friedlander, and K. MacPhee. Foundations of gauge and perspective duality. February 2017. preprint, arXiv:1702.08649.

[4] Robert M. Freund and Haihao Lu. New Computational Guarantees for Solving Convex Optimization Problems with First Order Methods, via a Function Growth Condition Measure. November 2016. preprint, arXiv:1511.02974.

[5] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004. p.89.