Projected Gradient Descent Efficiently Solves the Trust Region Subproblem

Mark Nishimura Reese Pathak

June 12, 2018

Abstract

We show that projected gradient descent asymptotically converges to a global minimizer of the trust region subproblem. We then show that iterates shortly hit the boundary, after which consecutive iterates remain on the boundary. Conditional on a single conjectured inequality from empirical evidence we are able to show that projected gradient descent achieves the typical $O(\log(1/\varepsilon))$ rate enjoyed by smooth convex functions.

1 Introduction

Trust region methods are sequential programming procedures in which heuristics are used to approximately solve a general optimization problem through multiple constrained quadratic programs. As a subroutine, these methods formulate and solve many instances of the following *trust region subproblem*

$$\begin{array}{ll} \text{minimize} & (1/2)x^T Ax + b^T x \\ \text{subject to} & \|x\| \leq R \end{array} \tag{1}$$

with variable $x \in \mathbf{R}^n$. The problem data are a symmetric matrix $A \in \mathbf{R}^{n \times n}$, a vector $b \in \mathbf{R}^n$, and a radius parameter $R > 0$. Crucially, the matrix A is possibly indefinite.

1.1 Previous works

The trust region subproblem is well-studied, and thus there many previous works worth mentioning. In earlier papers, the problem was solved either via subspace methods such as Steihaug-Toint (where no global convergence guarantees have been proven, to our knowledge), or using fast eigenvector and eigenvalue computation procedures like the Lanczos method [CGT00, EG09, GLRT99, GRT10]. More recently, however, some authors have provided convergence guarantees for this problem. For example, by reducing the trust region subproblem to a sequence of approximate eigenvector computations, Hazan and Koren [HK16] demonstrate that $\tilde{O}(1/\sqrt{\varepsilon})$[1] matrix-vector multiplies are enough to guarantee an ε-suboptimal point. In [HK17], Nguyen and Kilinç-Karzan reduce the trust region problem to a convex QCQP using eigenvector calculations, where first-order methods apply.

However, perhaps the most obvious algorithm to solve (1), is the *projected gradient method*, which we study in this paper. To our knowledge, the only previous work that analyzes the convergence properties of this procedure on (1) is [TA98], where Tao and An augment this procedure

[1] We use the $\tilde{O}(\cdot)$ notation to hide logarithmic factors.

by a restarting scheme, requiring possibly $O(d)$ restarts, which could scale poorly for large-scale problems. We also mention a recent work by Carmon and Duchi [CD16], studying the closely related problem

$$\text{minimize } (1/2)x^T Ax + b^T x + (\rho/3)\|x\|_2^3, \tag{2}$$

in variable $x \in \mathbf{R}^n$, again with A symmetric, possibly indefinite, and parameter $\rho > 0$. The authors analyze gradient descent, proving that $\tilde{O}(1/\varepsilon)$ gradient steps are enough to output an ε-suboptimal point.

In this paper we demonstrate that the projected gradient method on (1) asymptotically converges to a global minimizer on the trust region subproblem.

1.2 Notation and classical results

In the sequel, we refer to the objective function as $f : \mathbf{R}^n \to \mathbf{R}$, given by $f(x) = (1/2)x^T Ax + 2b^T x$. Additionally, the constraint set is the closed ball $\mathcal{B}(R) \triangleq \{x \in \mathbf{R}^n \mid \|x\| \le R\}$, where $\|\cdot\|$ denotes the Euclidean norm. We use the notation $x^\star$ to denote the global minimum of f when it is unique, so that $x^\star = \operatorname{argmin}_{x \in \mathcal{B}(R)} f(x)$. We use $f^\star$ to denote the optimal value of f, so that $f^\star = \inf_{x \in \mathcal{B}(R)} f(x)$. Hence, when $x^\star$ exists, $f^\star = f(x^\star)$.

We fix the eigendecomposition of $A = UDU^T$, where $D = \mathbf{diag}(\lambda_1, \dots, \lambda_n)$, and U has orthonormal columns u_i. We impose without loss that $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$. By $\|\cdot\|_{\text{op}}$, we denote the ℓ_2-operator norm $\|M\|_{\text{op}} = \sup_{\|x\|=1} \|Mx\|$, for any $M \in \mathbf{R}^{n \times n}$. A useful identity is that $\|M\|_{\text{op}} = \max_i |\lambda_i(M)|$ when M is a symmetric $n \times n$ matrix. We will put $\beta \triangleq \|A\|_{\text{op}}$.

Additionally, say a differentiable function $g : \mathbf{R}^n \to \mathbf{R}$ is L-smooth on convex set $C \subset \mathbf{R}^n$, provided that

$$\|\nabla g(x) - \nabla g(y)\| \le L\|x - y\| \qquad \text{for any } x, y \in C.$$

It is well known that this implies

$$g(x) - g(y) \le \nabla g(y)^T (x - y) + \frac{L}{2}\|x - y\|^2 \qquad \text{for any } x, y \in C. \tag{3}$$

Equivalently, $\|g(x)\|_{\text{op}} \le L$, for Lebesgue almost every $x \in C$. For nonempty, closed, convex sets $C \subset \mathbf{R}^n$, associate the projection operator $\Pi_C : \mathbf{R}^n \to C$ given by

$$\Pi_C(x) = \operatorname*{argmin}_{y \in C} \left(\frac{1}{2}\|x - y\|^2 \right),$$

for any $x \in \mathbf{R}^n$. In the sequel we denote by $I : \mathbf{R}^n \to \mathbf{R}^n$ the identity operator on $\mathbf{R}^n$.

2 Asymptotic convergence to a global minimizer

2.1 Projected gradient descent

Projected gradient descent (PGD) begins at an initialization $x^{(0)} \in \mathbf{R}^n$ and generates iterates

$$y^{(k+1)} = x^{(k)} - \eta \nabla f(x^{(k)}) \tag{4}$$

$$x^{(k+1)} = \Pi_{\mathcal{B}(R)}(y^{(k+1)}), \tag{5}$$

for nonnegative integer k and step size η. We make the following assumptions about this procedure.

Assumption 2.1. In (4), the step size η satisfies $0 < \eta < \frac{1}{\beta}$.

Assumption 2.2. The initial point satisfies $x^{(0)} = 0$.

2.2 Asymptotic convergence to a global minimizer

We begin by providing a few results, which characterize the iterates of projected gradient descent.

Lemma 2.3. *Let Assumptions 2.1 and 2.2 hold. Then the iterates of gradient descent satisfy* $(u_i^T x^{(k)})(u_i^T b) \leq 0$ *for all* $i = 1, \ldots, n$ *and every* $k \geq 0$. *0*

Proof. Evidently, the claim holds due to Assumption 2.2 when $k = 0$. Thus, inductively assume that for some k

$$(u_i^T x^{(k)})(u_i^T b) \leq 0 \qquad \text{for all } i = 1, \ldots, n. \tag{6}$$

By definition, $x^{(k+1)} = cy^{(k+1)}$ for some $c \in (0, 1]$, so it suffices to ensure $(u_i^T y^{(k+1)})(u_i^T b) \leq 0$. Using (6) along with Assumption 2.1,

$$(u_i^T y^{(k+1)})(u_i^T b) = (1 - \eta\lambda_i)(u_i^T x^{(k)})(u_i^T b) - \eta(u_i^T b)^2 \leq 0,$$

since $\eta < \beta^{-1} \leq \lambda_i^{-1}$, for all $i = 1, \ldots, n$. This proves the result. □

The following result shows projected gradient descent is a descent method for (1).

Lemma 2.4. *Let Assumption 2.1 hold. Then for any* $k > 0$*,*

$$f(x^{(k+1)}) - f(x^{(k)}) \leq \left(\frac{\beta}{2} - \frac{1}{2\eta}\right) \|x^{(k+1)} - x^{(k)}\|^2.$$

Proof. Basic manipulations imply

$$\nabla f(x^{(k)})^T(x - x^{(k)}) + \frac{1}{2\eta}\|x - x^{(k)}\|^2 = \frac{1}{2\eta}\|x - (x^{(k)} - \eta\nabla f(x^{(k)}))\|^2 - \frac{\eta}{2}\|\nabla f(x^{(k)})\|^2.$$

Thus, as $\eta > 0$ it follows that

$$\operatorname*{argmin}_{x\in\mathcal{B}(R)} \left(\nabla f(x^{(k)})^T(x - x^{(k)}) + \frac{1}{2\eta}\|x - x^{(k)}\|^2\right) = \operatorname*{argmin}_{x\in\mathcal{B}(R)} \left(\frac{1}{2}\|x - (x^{(k)} - \eta\nabla f(x^{(k)}))\|^2\right).$$

Comparing the display above to (4), (5), and the definition of $\Pi_{\mathcal{B}(R)}$,

$$x^{(k+1)} = \operatorname*{argmin}_{x\in\mathcal{B}(R)} \left(\nabla f(x^{(k)})^T(x - x^{(k)}) + \frac{1}{2\eta}\|x - x^{(k)}\|^2\right). \tag{7}$$

Appealing to the β-smoothness of f and evaluating (7) at $x^{(k)} \in \mathcal{B}(R)$,

$$f(x^{(k+1)}) - f(x^{(k)}) \leq \nabla f(x^{(k)})^T(x^{(k+1)} - x^{(k)}) + \frac{\beta}{2}\|x^{(k+1)} - x^{(k)}\|^2 \leq \left(\frac{\beta}{2} - \frac{1}{2\eta}\right)\|x^{(k+1)} - x^{(k)}\|^2.$$

□

The following result provides a useful optimality criterion for the trust region subproblem (1).

Theorem 2.5 ([CGT00], Corollary 7.2.2.). *A point $x \in \mathcal{B}(R)$ is a global minimizer of f subject to $\|x\| \le R$ if and only if for some $z \ge 0$,*

$$(A + zI)x = -b \qquad A + zI \succeq 0 \qquad z(\|x\| - R) = 0.$$

Furthermore, x is unique if and only if $A + zI \succ 0$. In this case, we write $x = x^\star$.

An important special case from Theorem 2.5 is that when $\|x^\star\| < R$, then $\nabla f(x^\star) = 0$. Furthermore, with a simplifying assumption, we can provide a set of simpler optimality criterion.

Corollary 2.6. *Suppose that $b^T u_1 \neq 0$. Then if for some $\tilde{x} \in \mathcal{B}(R)$ and $z \ge 0$, it holds that*

$$(A + zI)\tilde{x} = -b \qquad z(\|\tilde{x}\| - R) = 0 \qquad (u_1^T \tilde{x})(u_1^T b) \le 0 \tag{8}$$

then $\tilde{x}$ is the unique global minimizer to f over $\mathcal{B}(R)$, i.e., $\tilde{x} = x^\star$.

Proof. Focusing on the first condition, $b^T u_1 = -(z + \lambda_1)(u_1^T \tilde{x})$. Thus, $b^T u_1 \neq 0$ implies that $(u_1^T \tilde{x}) \neq 0$ and $z + \lambda_1 \neq 0$, strengthening the third condition to $(u_1^T \tilde{x})(u_1^T b) < 0$. But this implies that $z + \lambda_1 = -(u_1^T b)(u_1^T \tilde{x})/(u_1^T \tilde{x})^2 > 0$, which implies that $z > \lambda_i$ for all i, whence $A + zI \succ 0$, establishing the result. □

The assumptions along with Corollary 2.6 and Lemmas 2.3 and 2.4 give us our desired asymptotic convergence gaurantee.

Proposition 2.7 (Asymptotic convergence). *Let Assumptions 2.1 and 2.2 hold, and suppose $b^T u_1 \neq 0$. Then as $k \to \infty$, the iterates of projected gradient descent satisfy $x^{(k)} \to x^\star$ and $f(x^{(k)}) \downarrow f(x^\star)$.*

Proof. It suffices to demonstrate that $x^{(k)} \to x^\star$, because then the conclusion follows via continuity of f and To that end, Lemma 2.4. Lemma 2.4 and Assumption 2.1 yield the following bound for any integer $T \ge 1$,

$$\left(\frac{1}{2\eta} - \frac{\beta}{2}\right) \sum_{k=0}^{T-1} \|x^{(k+1)} - x^{(k)}\|^2 \le f(x^{(0)}) - f(x^{(T)}) \le f(x^{(0)}) - f^\star. \tag{9}$$

Now, define $\phi : \mathcal{B}(R) \to \mathbf{R}^n$ by $\phi(x) = \Pi_{\mathcal{B}(R)}(x - \eta \nabla f(x)) - x$, for points $x \in \mathcal{B}(R)$. The bound in (9) implies that the displayed series is convergent as $T \to \infty$ and thus $\phi(x^{(k)}) \to 0$. Note also that the map ϕ is evidently continuous, as ∇f is β-Lipschitz and $\Pi_{\mathcal{B}(R)}$ is non-expansive, thus 1-Lipschitz.

Suppose now that $\tilde{x} \in \mathcal{B}(R)$ is a subsequential limit of $(x^{(k)})$ (indeed, one exists since this sequence is bounded), and observe by continuity $\phi(\tilde{x}) = 0$. To show that $\tilde{x} = x^\star$, by Corollary 2.6, it suffices to establish the first two conditions of (8), as the third immediately holds by Lemma 2.3. Observe first that $\phi(\tilde{x}) = 0$ implies that for some $c \ge 1$,

$$\tilde{x} - \eta \nabla f(\tilde{x}) = \tilde{x} - \eta(A\tilde{x} - b) = c\tilde{x}. \tag{10}$$

Indeed, setting $z = (c-1)\eta^{-1}$, this implies that $(A + zI)\tilde{x} = -b$. If $\tilde{x}$ lies on the boundary of $\mathcal{B}(R)$, so that $\|\tilde{x}\| = R$, then as $z \ge 0$, this establishes (8) and hence $\tilde{x} = x^\star$. On the other hand, if $\tilde{x}$ is in the interior of $\mathcal{B}(R)$, so that $\|\tilde{x}\| < R$, then $\phi(\tilde{x}) = 0$ implies that $c = 1$ in (10), and thus $z = 0$, once again establishing (8), hence also that $\tilde{x} = x^\star$. As this analysis applies to any such subsequential limit $\tilde{x}$ of the bounded sequence $(x^{(k)})$, the claim is now proven (since the iterates lie in $\mathcal{B}(R)$, which is compact). □

We provide some numerical evidence demonstrating the effect of Proposition 2.7 in Figure 1.

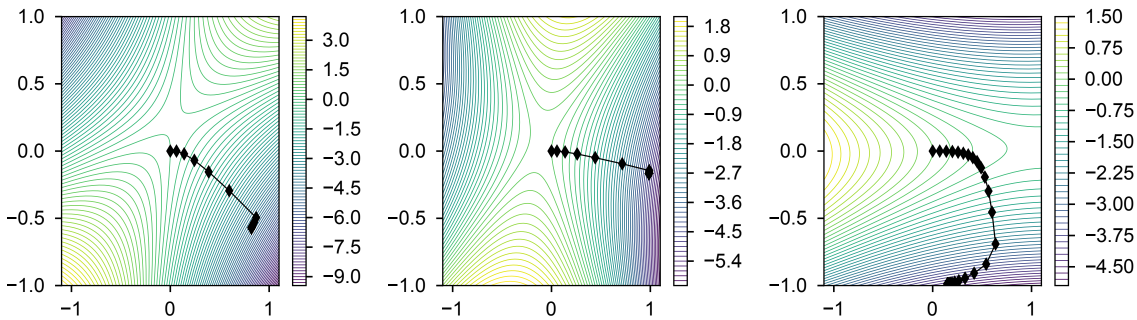

Figure 1: Three random indefinite instances of the the trust region subproblem (1), with $R = 1$, $\eta = 1/(2\|A\|_{\mathrm{op}})$ and $x^{(0)} = 0$. From left to right, the eigenvalues are $\lambda = (-8, 3)$, $\lambda = (-9, 3)$, and $\lambda = (-7, 1)$. The dots indicate iterates of projected gradient descent and the lines indicate the process $\dot{x} = -\nabla f(x)$.

3 Non-asymptotic convergence guarantees

In this section, we use the notation $\pi^{(k)} \in (0, 1]$ to denote a constant such that $x^{(k)} = \pi^{(k)} y^{(k)}$. Additionally we tacitly assume that Assumptions 2.1 and 2.2 hold.

We first prove a technical result about the signs of an iterative process.

Lemma 3.1. *Let $\kappa \in \mathbf{R}^n$ satisfy $\kappa_1 \le \kappa_2 \le \cdots \le \kappa_n \le 1$, and let $c^{(t)}$ denote a non-negative sequence. If $z_i^{(0)} = 0$ for all i, and*

$$z_i^{(k+1)} = c^{(k)}(1 - \kappa_i) z^{(k)} + 1$$

for all k, then the following three statements hold

1. *If $z_j^{(k)} \le c^{(k-1)} z_j^{(k-1)}$ then $z_j^{(k')} \le c^{(k'-1)} z_j^{(k'-1)}$ for all $k' > k$.*
2. *If $z_i^{(k+1)} \le c^{(k)} z_i^{(k)}$ then for all $j \ge i$, $z_j^{(k+1)} \le c^{(k)} z_j^{(k)}$.*

Proof. To see (1.), it suffices to show (by induction) that the claim holds for $k' = k + 1$. Thus,

$$z_j^{(k+1)} - c^{(k)} z_j^{(k)} = c^{(k)}(1 - \kappa_j) \left(z_j^{(k)} - c^{(k-1)} z_j^{(k-1)} \right) \le 0,$$

by assumption and $c^{(k)} \ge 0$, $1 - \kappa_j \ge 0$.

To see (2.), fix $j \ge i$. Note that evidently $z_i^{(k)} \ge 0$ for all i, k thus

$$\frac{c^{(k)} z_j^{(k)}}{z_j^{(k+1)}} - \frac{c^{(k)} z_i^{(k)}}{z_i^{(k+1)}} = \frac{(c^{(k)})^2 (\kappa_j - \kappa_i) z_j^{(k)} z_i^{(k)}}{z_i^{(k+1)} z_j^{(k+1)}} \ge 0.$$

Above, we use that $\kappa_j \ge \kappa_i$ when $j \ge i$. Hence, $c^{(k)} z_j^{(k)} / z_j^{(k+1)} \ge c^{(k)} z_i^{(k)} / z_i^{(k+1)} \ge 1$. The claim follows. ☐

Lemma 3.2. *Fix $k \in \mathbf{N}$ such that $\nabla f(x^{(k)})^T x^{(k)} \le 0$. Then $x^{(k)^T} A \nabla f(x^{(k)}) \ge \beta x^{(k)^T} \nabla f(x^{(k)})$.*

Proof. Note first that $x^{(k')} = \pi^{(k')} y^{(k')}$, thus, for all $k' \leq k$,

$$\sum_{i=1}^{n} (u_i^T y^{(k')})(u_i^T (x^{(k')} - y^{(k'+1)})) \leq 0.$$

Define the following sets for $k' \leq k$

$$I_+^{(k')} \triangleq \{i \in [n] : (u_i^T y^{(k')})(u_i^T (x^{(k')} - y^{(k'+1)})) \geq 0\}$$
$$I_-^{(k')} \triangleq \{i \in [n] : (u_i^T y^{(k')})(u_i^T (x^{(k')} - y^{(k'+1)})) \leq 0\}.$$

Associated to these sets, define $\lambda_+^{(k')} = \lambda_i$ and $\lambda_-^{(k')} = \lambda_j$ for $i = \min I_+^{(k')}$, and $j = \max I_-^{(k')}$. Then now observe that, expanding ${y^{(k)}}^T A \nabla f(x^{(k)})$ in the eigenbasis of A,

$$\begin{aligned}
{y^{(k)}}^T A \nabla f(x^{(k)}) &= \frac{1}{\eta}\left(\sum_{i \in I_+^{(k)}} \lambda_i (u_i^T y^{(k)})(u_i^T (x^{(k)} - y^{(k+1)})) + \sum_{i \in I_-^{(k)}} \lambda_i (u_i^T y^{(k)})(u_i^T (x^{(k)} - y^{(k+1)}))\right) \\
&\geq \frac{1}{\eta}\left(\lambda_+^{(k)} \sum_{i \in I_+^{(k)}} (u_i^T y^{(k)})(u_i^T (x^{(k)} - y^{(k+1)})) + \lambda_-^{(k)} \sum_{i \in I_-^{(k)}} (u_i^T y^{(k)})(u_i^T (x^{(k)} - y^{(k+1)}))\right) \\
&\geq \lambda_-^{(k)} {y^{(k)}}^T \nabla f(x^{(k)}) \geq \beta y^{(k)} \nabla f(x^{(k)}).
\end{aligned}$$

Recalling that $x^{(k)} = \pi^{(k)} y^{(k)}$ with $\pi^{(k)} \in (0, 1]$, this proves the claim. The last inequality was obtained by assumption that ${x^{(k)}}^T \nabla f(x^{(k)}) \leq 0$, and that $\lambda_i \leq \lambda_n \leq \beta$ for all $i \leq n$. The penultimate inequality is due to $\lambda_-^{(k)} \leq \lambda_+^{(k)}$. To see this, it suffices to show that

$$(u_i^T y^{(k)})(u_i^T (x^{(k)} - y^{(k+1)})) \geq 0 \quad \text{implies} \quad (u_j^T y^{(k)})(u_j^T (x^{(k)} - y^{(k+1)})) \geq 0 \quad \text{for all } j \geq i \tag{11}$$

We prove this using Lemma 3.1. Indeed, $z_j^{(k)} = (u_j^T y^{(k)})/(-\eta u_j^T b)$ for all $k \in \mathbf{N}$. Then

$$z_j^{(k+1)} = \frac{u_j^T ((I - \eta A) x^{(k)} - \eta b)}{-\eta u_j^T b} = \underbrace{\pi^{(k)}}_{\triangleq c^{(k)}} (1 - \underbrace{\eta \lambda_i}_{\triangleq \kappa_i}) z_j^{(k)} + 1.$$

Note that $z_j^{(0)} = 0$, and additionally $\eta \lambda_i = \kappa_i$ is non-decreasing in i, and bounded above by 1 since $\eta \leq 1/\beta$. Additionally, by assumption we have $c^{(k')} = \pi^{(k')} = R/\|y^{(k')}\|$, which is evidently non-negative. Thus (2.) of Lemma 3.1 implies that

$$\pi^{(k)} z_i^{(k)} - z_i^{(k+1)} \geq 0 \quad \text{implies} \quad \pi^{(k)} z_i^{(k)} - z_i^{(k+1)} \geq 0 \quad \text{for all } j \geq i.$$

Note that as $z_i^{(k)} \geq 0$, this is equivalent to the display in (11). The result is now proven. $\square$

Lemma 3.3. *For all $k \geq 0$, the iterates of projected gradient descent satisfy $b^T x^{(k)} \leq 0$.*

Proof. We inductively estlabish the following, stronger, result.

$$\text{for all } i \in [n], \qquad (u_i^T x^{(k)})(u_i^T b) \le 0 \qquad \text{for all } k \in \mathbf{N}. \tag{12}$$

Claim (12) evidently holds when $k = 0$, so now suppose it holds for $k' \le k$. Fix $i \in [n]$. Note

$$\begin{aligned}\operatorname{sign}(u_i^T x^{(k+1)}) &= \operatorname{sign}\left(\pi^{(k+1)}((1-\eta\lambda_i)u_i^T x^{(k)} - \eta u_i^T b)\right) \\ &= \operatorname{sign}((1-\eta\lambda_i)u_i^T x^{(k)} + \eta(-u_i^T b)) = -\operatorname{sign}(u_i^T b)\end{aligned}$$

The final equality holds since $\pi^{(k+1)} \in (0, 1]$, and the final inequality holds due to the inductive assumption. This proves Claim (12), and the result follows by summing these inequalites: $b^T x^{(k)} = \sum_{i=1}^n (x^{(k)^T} u_i)(u_i^T b) \le 0$. □

Lemma 3.4. *For all $k \ge 0$, the iterates of projected gradient descent satisfy $x^{(k)^T} \nabla f(x^{(k)}) \le 0$. Furthermore, for all k, $\|y^{(k+1)}\| \ge \|x^{(k)}\|$.*

Proof. By definition of the projected gradient descent iteration, we have

$$\|y^{(k+1)}\|^2 = \|x^{(k)}\|^2 + \eta\|\nabla\nabla f(x^{(k)})\|^2 - 2\eta\nabla f(x^{(k)})^T x^{(k)}.$$

Thus, to prove the claim it would be sufficient to show inductively that $x^{(k)^T} \nabla f(x^{(k)}) \le 0$. The basis of induction is clear as the statement trivially holds at $x^{(0)} = 0$. Suppose the claim holds for k, and note it suffices to demonstrate $y^{(k+1)^T} \nabla f(x^{(k+1)}) \le 0$. For all k, denote $0 < \pi^{(k)} \le 1$ such that $\pi^{(k)} y^{(k)} = x^{(k)}$. Lemma 3.3 implies

$$\begin{aligned}y^{(k+1)^T} \nabla f(x^{(k+1)}) &\le \pi^{(k)} \left(x^{(k)^T} \nabla f(x^{(k)}) - \eta x^{(k)^T} A \nabla f(x^{(k)}) - \eta\|\nabla f(x^{(k)})\|^2 + \eta^2 \nabla f(x^{(k)})^T A \nabla f(x^{(k)})\right) \\ &\le -\pi^{(k)}(\eta - \eta^2\beta)\|f(x^{(k)})\|^2 + \pi^{(k)}(1-\eta\beta)x^{(k)^T}\nabla f(x^{(k)}) \le 0\end{aligned}$$

The penultimate inequality is due Lemma 3.2, and the final inequality is because $\eta \le 1/\beta$. □

Lemma 3.4 immediately implies the following claim.

Corollary 3.5. *Suppose an iterate of projected gradient descent satisfies $x^{(\tau)} \in \partial\mathcal{B}(R)$ for some $\tau \in \mathbf{N}$. Then $x^{(t)} \in \partial\mathcal{B}(R)$ for all $t \ge \tau$.*

Proof. Lemma 3.4 implies $\|y^{(\tau+1)}\|^2 \ge \|x^{(\tau)}\|^2 \ge R^2$, thus $x^{(\tau+1)} \in \partial B(R)$. The claim now follows via induction. □

In words, once an iterate hits the boundary, all subsequent iterates remain on the boundary. We believe this result is relevant. Figure 2 demonstrates the effect of iterates remaining on the boundary; empirically we observe exponential convergence, characteristic of gradient descent on smooth, convex functions.

The following result bounds the time to the boundary.

Claim 3.6. *Suppose that $\lambda_1 < 0$, and $b^T u_1 \ne 0$. Then after $O\left(\frac{R|\lambda_1|}{|b^T u_1| \log(1-\eta\lambda_1)}\right)$ iterations, the iterates of projected gradient descent lie on the boundary.*

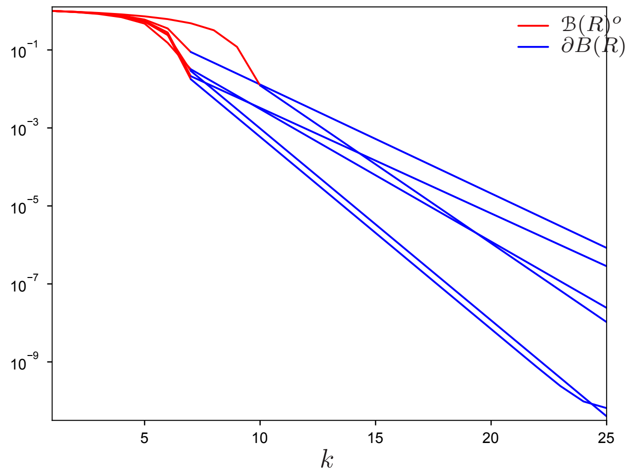

Figure 2: Two regimes of convergence, before and after hitting the boundary.

Proof. Suppose that for k iterations, $\|y^{(k)}\|^2 < R^2$. Then, as $x^{(0)} = 0$, we have

$$y^{(k)} = -\eta \sum_{t=0}^{k-1} (I - \eta A)^t b = -\eta \sum_{i=1}^{n} \left(\sum_{t=0}^{k-1} (1 - \eta \lambda_i)^t \right) b^T u_i u_i$$

Hence,

$$\|y^{(k)}\|^2 = \sum_{i=1}^{n} \frac{(b^T u_i)^2}{\lambda_i^2} (1 - (1 - \eta\lambda_i)^k)^2 \geq \frac{(b^T u_1)^2}{\lambda_1^2} (1 - (1 - \eta\lambda_1)^k)^2.$$

Setting the right-hand side to R^2, one obtains that if $k \geq 1 + \frac{R|\lambda_1|}{|b^T u_1|(\log(1-\eta\lambda_1)}$, then the display above is larger than R^2. Thus, the claim now follows via Corollary 3.5. $\square$

Lemma 3.7. *Let $\tau^{\mathrm{bd}} = \min\{k : x^{(k)} \in \partial\mathcal{B}(R)\}$. If for some $\delta > 0$, and for all $k \geq \tau^{\mathrm{bd}}$, $(x^{(k+1)} - \eta\delta x^{(k)})^T x^\star \geq (1 - \eta\delta) R^2$, then $f(x^{(t)})$ is $\varepsilon > 0$-suboptimal provided that $t \geq \tau^{\mathrm{bd}} + \frac{1}{\eta\delta} \log\left(\frac{2R^2(\beta+z)}{\varepsilon}\right)$.*

Proof. By hypothesis, we have $\|x^{(k+1)} - x^\star\|^2 \leq (1 - \eta\delta)^2 \|x^{(k)} - x^\star\|^2$, provided $k \geq \tau^{\mathrm{bd}}$. Thus, as $\delta < 1/\eta$, and with $\tau := \tau^{\mathrm{bd}}$,

$$\|x^{(k+\tau)} - x^\star\|^2 \leq (1 - \eta\delta)^{2k} \|x^\tau - x^\star\|^2 \leq 4R^2 (1 - \eta\delta)^k \leq 4R^2 e^{-2\eta\delta k},$$

as $1 + \alpha \leq e^\alpha$ for all $\alpha \in \mathbf{R}$. Now appealing to the β-smoothness of f,

$$f(x^{(k+\tau)}) - f^\star \leq 2(\beta + z) R^2 e^{-\eta\delta k} \leq \varepsilon,$$

provided that $k \geq \frac{1}{\eta\delta} \log((2R^2(\beta + z))/\varepsilon)$. Hence if

$$t \geq \tau^{\mathrm{bd}} + \frac{1}{\eta\delta} \log\left(\frac{2R^2(\beta + z)}{\varepsilon}\right),$$

then $f(x^{(t)}) - f^\star \leq \varepsilon$. $\square$

Conjecture 3.8. *In the notation of Lemma 3.7, for all* $k \geq \tau^{\mathrm{bd}}$, $(x^{(k+1)} - \eta\delta x^{(k)})^T x^\star \geq (1-\eta\delta)R^2$ *holds with* $\delta = 1/(z+\lambda_1)$.

Conditional on Conjecture 3.8, this implies the $\tilde{O}(\log(1/\varepsilon))$ convergence rate enjoyed by smooth, convex functions under the gradient descent iteration. Of course, the most obvious open problem is to prove (or disprove) Conjecture 3.8.

References

[CD16] Yair Carmon and John C. Duchi. Gradient descent efficiently finds the cubic-regularized non-convex Newton step. *CoRR*, abs/1612.00547, 2016.

[CGT00] Andrew R. Conn, Nicholas I. M. Gould, and Philippe L. Toint. *Trust-region methods*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2000.

[EG09] Jennifer B. Erway and Philip E. Gill. A subspace minimization method for the trust-region step. *SIAM Journal on Optimization*, 20(3):1439–1461, 2009.

[GLRT99] Nicholas I. M. Gould, Stefano Lucidi, Massimo Roma, and Philippe L. Toint. Solving the trust-region subproblem using the Lanczos method. *SIAM Journal on Optimization*, 9(2):504–525, 1999.

[GRT10] Nicholas I. M. Gould, Daniel P. Robinson, and H. Sue Thorne. On solving trust-region and other regularized subproblems in optimization. *Math. Program. Comput.*, 2(1):21–57, 2010.

[HK16] Elad Hazan and Tomer Koren. A linear-time algorithm for trust region problems. *Math. Program.*, 158(1-2):363–381, 2016.

[HK17] Nam Ho-Nguyen and Fatma Kilinç-Karzan. A second-order cone based approach for solving the trust-region subproblem and its variants. *SIAM Journal on Optimization*, 27(3):1485–1512, 2017.

[TA98] Pham Dinh Tao and Le Thi Hoai An. A D.C. optimization algorithm for solving the trust-region subproblem. *SIAM Journal on Optimization*, 8(2):476–505, 1998.