Projected Gradient Descent Efficiently* Solves the Trust Region Subproblem

Mark Nishimura Reese Pathak

May 19, 2018

Abstract

We show that projected gradient descent asymptotically converges to a global minimizer of the trust region subproblem. We remark on next steps in this project at the end.

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 by a restarting scheme, requiring possibly $O(d)$ restarts, which could scale poorly for large-scale

*Technically, a conjecture.

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

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 $\mathcal{B}(R) \triangleq \{x \in \mathbf{R}^n \mid \|x\| \leq R\}$, 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 $D = \mathbf{diag}(\lambda_1, \ldots, \lambda_n)$, and U has orthonormal columns u_i. We impose without loss that $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. 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

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

It is well known that this implies

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

Equivalently, $\|g(x)\|_{\text{op}} \leq L$, 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* $\|x\| \leq R$ *if and only if for some* $z \geq 0$,

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

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 $z \geq 0$, it holds that*

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

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 $T \geq 1$,

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

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 $c \geq 1$,

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

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 $z \geq 0$, 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.

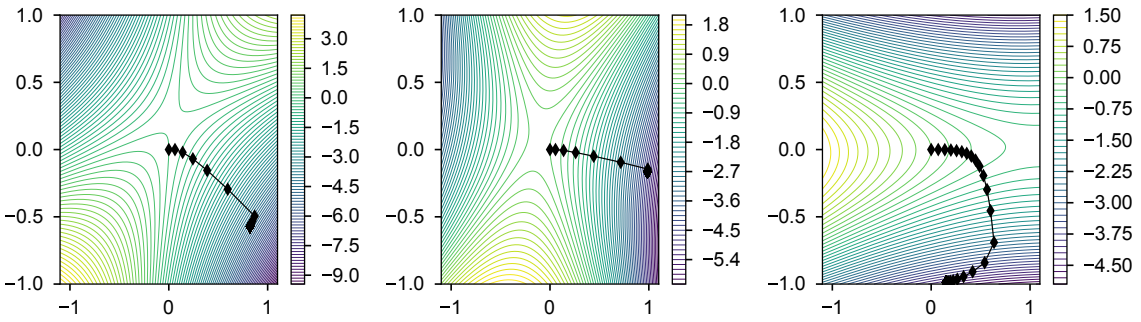

Figure 1: Three random indefinite instances of the the trust region subproblem (1), with $R = 1$, $\eta = 1/(2\|A\|_{\mathrm{op}})$ and $x^{(0)} = 0$. From left to right, the eigenvalues are $\lambda = (-8, 3)$, $\lambda = (-9, 3)$, and $\lambda = (-7, 1)$. The dots indicate iterates of projected gradient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

3 Non-asymptotic convergence guarantees

We plan to continue this work by providing convergence rates for this problem. In particular, our current goal is to obtain a non-asympototic convergence rate, similar to Theorem 3.1, [CD16].

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.

Theo Diamandis: 2/2
Your introduction explains the problem you wish to solve in a clear, concise manner. The previous works section provided a good background, which made clear your contribution to the literature. I was surprised that few previous works looked at the convergence of projected GD. It sounds like you all found a great problem to tackle. The following section proving your convergence result was organized, clear, and concise. Looking forward to seeing your future work on convergence guarantees.

I list more specific comments and suggestions below.

First, three minor points. In Section 1, I noticed that you used the ball B(R) notation before it is introduced in section 1.2. In Section 1.1, should the second sentence say “have been proven” rather than “have proven”? In Section 1.2, you don't mention your Otilde(n) notation. If you use this notation in Section 3, it might be useful to put in Section 1.2 as a reference to the reader.

In Section 2, I found the figure difficult to read, especially when printed out. I would make the lines thicker and darker, possibly using markers rather than color to differentiate. In addition, since the figure was not referenced in your paper, I was unsure if it served to support a particular point or simply show GD steps. Also, since the example is low dimensional, I would have liked to know the matrix A and vector b used to create the instance of (1).

My theory background is lacking, so take what follows with a grain of salt. I found two parts of Section 2 tricky to follow: the beginning of the Lemma 2.5 proof, which invokes Lemma 2.4, was not obvious to me; and I did not follow the ||x_tilde|| = R case in your final proof. In the latter case, I was not sure where the gradient expression of (1-c)*eta^-1*x came from.

--
Rahul Trivedi: 2/2
The authors mention that they exclude Lemma 2.3 and 2.4 due to space constraints. It might be a good idea to provide a sketch of the proof here since both of them seem crucial to the proof in Lemma 2.5 and proposition 2.7.
The statement that 'The bound in (6) implies that the displayed series is convergent as T → ∞ and thus φ(x(k)) → 0' is a bit unclear. It might be a good idea to expand on this a bit more. In particular, on reading the proof it seems that the fact that \sum ||x^(k) - x^(k+1)|| ^2 goes is bounded by f(x^(0))-f(x*) implies that x^(k) -> x* as k-> infty. I am not quite sure why this is necessarily true, unless some specific properties of the projected gradient descent are exploited

Other than that, the paper was very well written and the proofs were very connected and understandable. I look forward to seeing the

convergence guarantees that you come up with.

--
Pulkit Tandon: 2/2
Problem is very clearly motivated and the idea seems neat. Paper is well written.
Figure 1 is unclear to me. What is the experiment? It is not explained in the text. Also, it will be nice to have a clearer figure: what exactly are the ‘orange’ and ‘green’ lines and the dots on them (caption of the figure is unclear).
In proposition 2.7, convergence of \phi is unclear to me.

Hope this helps and I would love to see the complete proofs of Lemma 2.3, 2.4 and the convergence rate results in your final report.

--
Georgia Murray: 2/2
1) Can you provide some intuition into the assumptions in section 2.1? The first seems like an actually necessary assumption (although I haven't thought about why it is strictly necessary), but the second just seems like it is for convenience. Is this the case, or is initializing at the origin strictly necessary for some reason?

2) Just to clarify, you have already proven all the lemmas in section 2.2? You imply that you have, but I wasn't entirely sure.

3) Since (if I understand correctly) the point of your work is that you can use gradient descent to solve the trust-region problem despite its nonconvexity, I think it would make more sense to have Figure 1 show a nonconvex example.

Sorry, I wish I had more over-arching feedback, but great job!

--
Akshay Rajagopal: 2/2
For the problem instance used for Figure 1, is A an indefinite matrix? If not, since projected gradient descent is able to solve the trust region problem even with A indefinite, it might be nice to use an indefinite A to illustrate that.
I know there's not much extra space, but if you can include it, maybe an exact statement of the result you will prove going forward ("the equivalent of Theorem 3.1, [CD16].") could help.
I was curious how the results in section 2.2 would fail if assumptions 2.1 and 2.2 did not hold. That's probably not necessary for this midterm report, but a brief mention of it could be good for the final report.

Otherwise, I think it looks good. Looking forward to seeing how it progresses!