Projected Gradient Descent Solves the Trust Region Problem

Mark Nishimura, Reese Pathak

EE364b, Convex Optimization II Project

Introduction

Trust region methods are sequential programming procedures which formulate and solve many instances of the following **trust region problem**

$$\begin{array}{ll} \text{minimize} & (1/2)x^TAx + b^Tx \\ \text{subject to} & \|x\| \le R \end{array} \qquad (1)$$

with variable x. Do **not** assume A is definite.

Recall... If $A \in \mathbf{R}^{n\times n}$ is symmetric then for Λ, orthonormal U,

$$A = U\Lambda U^T \quad \Lambda = \mathbf{diag}(\lambda) \quad \lambda_1 \le \cdots \le \lambda_n \quad U = [u_1 \mid \cdots \mid u_n]$$

Also have for $f : \mathbf{R}^n \to \mathbf{R}$ with L-Lipschitz gradient

$$f(x) - f(y) \le \nabla f(y)^T(x-y) + \frac{L}{2}\|x-y\|^2 \qquad (x, y \in \mathbf{dom}\, f)$$

Projected Gradient Descent

We investigate the behavior of **projected gradient descent** (PGD) which begins at an initialization $x^{(0)} \in \mathbf{R}^n$ and generates iterates

$$\begin{aligned} y^{(k+1)} &= x^{(k)} - \eta\nabla f(x^{(k)}) \\ x^{(k+1)} &= \Pi_{\mathcal{B}(R)}(y^{(k+1)}). \end{aligned}$$

We make the following assumptions about this procedure:

- step size satisfies $0 < \eta < 1/\|A\|_{\text{op}}$
- initialize at $x^{(0)} = 0 \in \mathbf{R}^n$.

Variational interpretation

Complete the square to verify that PGD iterates satisfy (cf. Nesterov)

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

(Essentially) immediately implies that PGD is a descent method:

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

Optimality criterion

The following result provides a useful optimality condition.

Theorem 1 ([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 iff $A + zI \succ 0$. In this case, we write $x = x^\star$.

We show (and make use of) the following weaker statement.

Corollary 2. *Suppose that $b^Tu_1 \ne 0$. Then if at $\tilde{x} \in \mathcal{B}(R)$, $z \ge 0$, have*

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

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

Asymptotic result

Under mild assumptions we obtain the following result.

Proposition 3 (Asymptotic convergence). *Let the step-size and initialization assumptions hold, and assume further that suppose $b^Tu_1 \ne 0$. Then as $k \to \infty$, the iterates of projected gradient descent satisfy $x^{(k)} \to x^\star$ and $f(x^{(k)}) \downarrow f(x^\star)$, where $x^\star$ is the unique global minimizer to f over $\mathcal{B}(R)$.*

- In English, unless you have a pretty sick trust region problem, PGD eventually gets to the global minimizer of f.
- *Disclaimer:* this statement (nor its proof) admit an obvious convergence rate. This means, you could get to opt quite slowly (though not in practice, ...)

Convergence proof

Idea: Show that $\|x^{(k+1)} - x^{(k)}\|_2^2 \to 0$ and hope that's enough.

- Use descent method inequality to show that
  $$x^{(k+1)} - x^{(k)} \to \infty. \qquad (2)$$
- Introduce continuous map $g : \mathcal{B}(R) \to \mathbf{R}^n$,
  $$x \mapsto \Pi_{\mathcal{B}(R)}(x - \eta\nabla f(x)) - x$$
  - this is just a single PGD step
  - can rewrite Eq. (2) as $g(x^{(k)}) \to 0$
- Consider a subsequential limit (one exists since $x^{(k)}$ lie in compact set) L
  - Use continuity to conclude that limits satisfy $g(L) = 0$
  - Use the optimality criterion, projection map, and case work to analyze $\ker g$
  - Conclude that $L = x^\star$
- The analysis above applies to any limit point, so we're done (here we use the fact that $\{x^{(k)}\} \subset K = \mathcal{B}(R)$)

Numerical example

Example below has $A \in \mathbf{R}^{2\times 2}$ with $\lambda_1(A) = -8$, $\lambda_2(A) = 3$. We took $R = 1$, $\eta = 1/16$. Note that $b^Tu_1 = 0.790857$ and $\|x^\star\| = 1$.

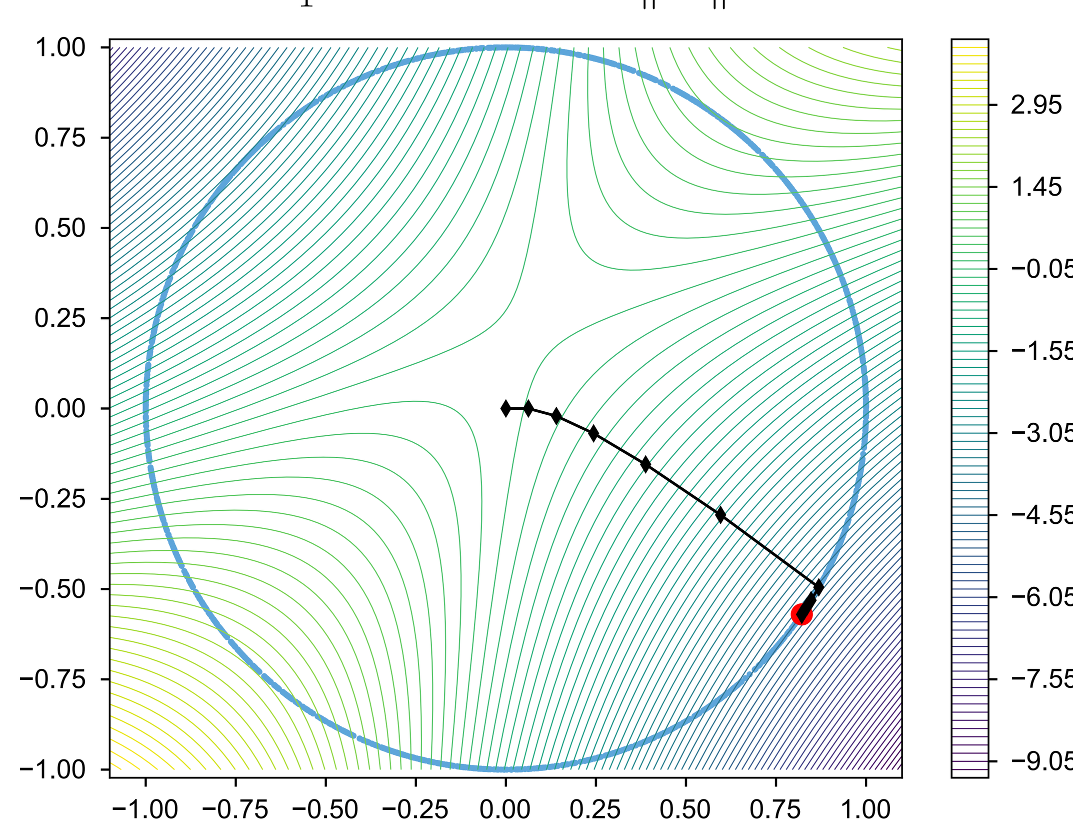

Figure 1: Trust region problem. Black diamonds are $x^{(k)}$, red dot is $x^\star$, blue circle is $\partial\mathcal{B}(R)$.

Ideas for non-asymptotics

Basically, it looks like a non-asymptotic proof of convergence could follow the following lines (roughly)

1. Show that there is a τ^{bd} such that when $t \ge \tau^{\text{bd}}$, we can guarantee that you've used projection at least once (*i.e.*, you've hit the boundary)
   $$\|y^{(t)}\|^2 = \eta^2 \left\| \sum_{k=0}^{t-1}(I - \eta A)b \right\|^2 = \sum_{i=1}^n \left(\frac{b^Tu_i}{\lambda_i}\right)^2 (1 - (1-\eta\lambda_i)^t)^2 > R^2$$
2. Show that once the boundary is reached, successive iterates remain on the boundary. In other words, $\|y^{(t+1)}\| > R$ for all $t \ge \tau^{\text{bd}}$.

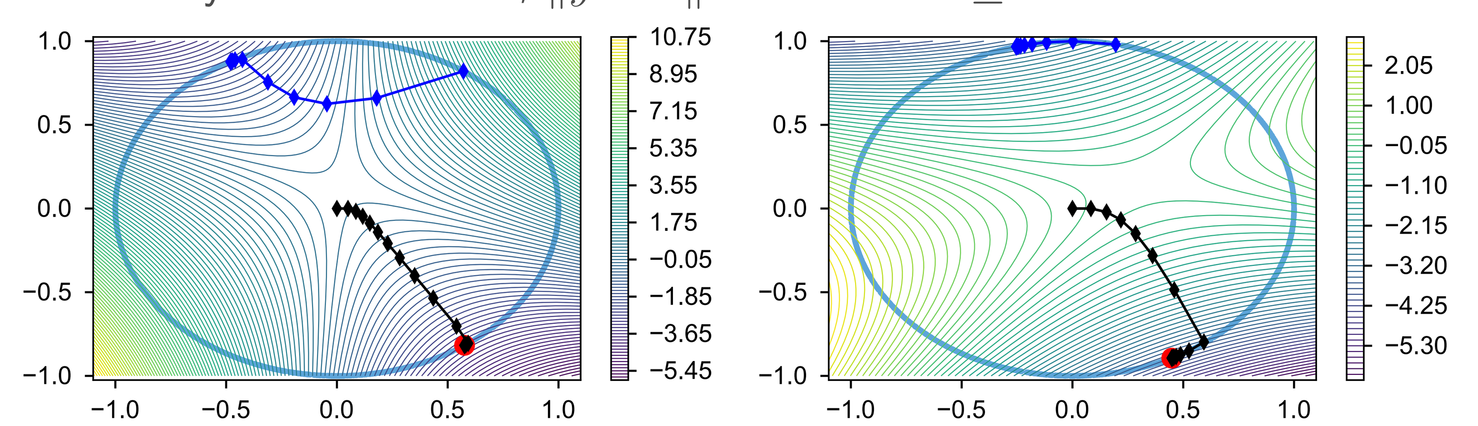

Figure 2: Initializing randomly on the boundary of $\mathcal{B}(R)$ doesn't always work!

- Importantly, property of remaining on boundary for all successive iterates is not true of all points $x \in \partial\mathcal{B}(R)$, ...

3. Show a contraction inequality like (for $k \ge \tau^{\text{bd}}$)
   $$\|x^{(k+1)} - x^\star\| \le (1-\epsilon)\|x^{(k)} - x^\star\| \qquad (\epsilon > 0)$$

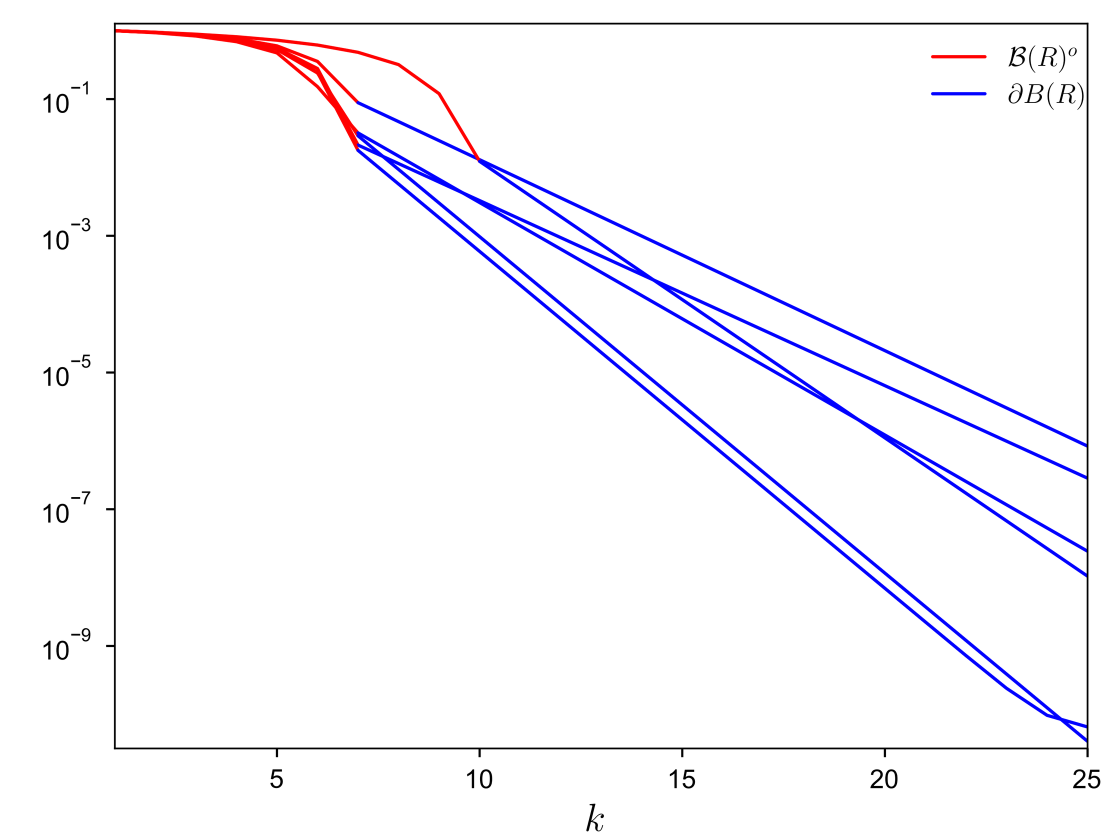

Figure 3: Two regimes of convergence

4. Conclude via smoothness, standard GD analysis for smooth problems.

References & Acknowledgements

[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.

Thanks to Yair and John for putting up with our (mostly stupid) questions.