EECS 559
Optimization Methods for SIPML

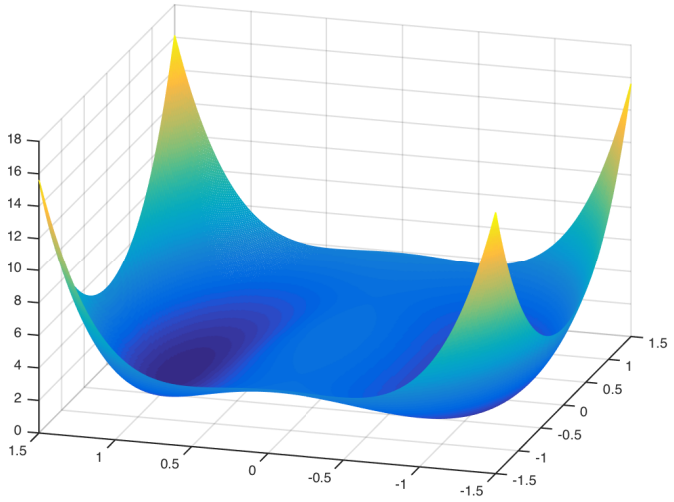
Lecture 12 – Trust-Region and Cubic Regularization

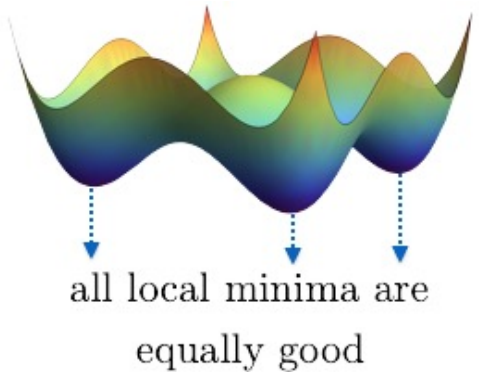
Instructor: Prof. Qing Qu (qingqu@umich.edu)

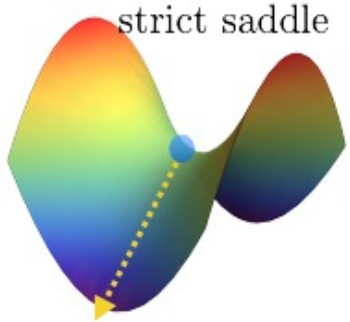
Lecture Agenda

- **Trust-Region Method**
  - **Algorithmic Introduction**
  - Solving the TRM Subproblem
  - Convergence for Strict Saddle Function
- Cubic Regularization of Newton's Method

Example: Generalized Phase Retrieval



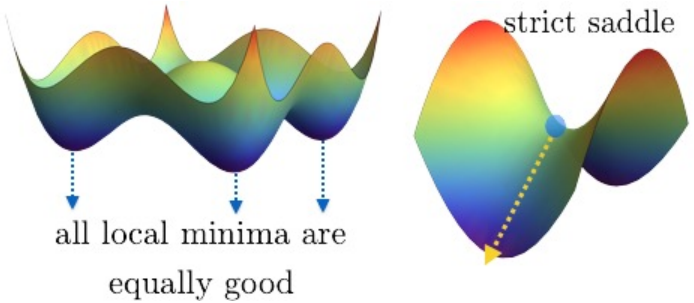




$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) \;=\; \frac{1}{4m}\sum_{i=1}^{m}\left(y_i - (\boldsymbol{a}_i^\top \boldsymbol{x})^2\right)^2$$

Unconstraint Nonconvex Problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \; f(\boldsymbol{x})$$



- Assume the function $f(\boldsymbol{x}) : \mathbb{R}^n \mapsto \mathbb{R}$ is twice continuously differentiable.
- We aim to solve nonconvex/nonlinear problems, e.g. strict saddle functions.
- We need to *exploit the negative curvature* of functions to escape saddle points.

Idea: Optimizing Local Approximation

At every iteration $\boldsymbol{x}_k$, approximate $f(\boldsymbol{x}_k + \boldsymbol{d})$ by

- **Linear model:**

$$f_L(\boldsymbol{d}; \boldsymbol{x}_k) \;=\; f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d} \rangle$$

- **Quadratic model:**

$$f_Q(\boldsymbol{d}; \boldsymbol{x}_k) \;=\; f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d} \rangle + \frac{1}{2} \boldsymbol{d}^\top \boldsymbol{B}_k \boldsymbol{d}$$

for some symmetric matrix $\boldsymbol{B}_k \approx \nabla^2 f(\boldsymbol{x}_k)$
(e.g., quasi-Newton SR-1/SR-2 update).

Idea: Optimizing Local Approximation

- **We want to optimize**

$$f_Q(\boldsymbol{d}; \boldsymbol{x}_k) \;=\; f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d} \rangle + \frac{1}{2} \boldsymbol{d}^\top \boldsymbol{B}_k \boldsymbol{d}$$

for some symmetric matrix
(e.g., quasi-Newton SR-1/SR-2 update).

- **Difficulties for optimization:**
  - the approximation error can be large when $\|\boldsymbol{d}\|_2$ is large.
  - minimizing the model might not be possible
    (e.g., $f_Q(\boldsymbol{d}; \boldsymbol{x}_k)$ is unbounded when $\boldsymbol{B}_k$ is *indefinite*)

Trust-region Methods (TRM)

Idea: add a trust-region constraint

$$\|\boldsymbol{d}\|_{\square} \leq \Delta_k,$$

for some "suitable" trust-region radius $\Delta_k > 0$ and solve

$$\boldsymbol{d}_k = \arg\min_{\boldsymbol{d}\in\mathbb{R}^n} f_Q(\boldsymbol{d}; \boldsymbol{x}_k), \quad \text{s.t.} \quad \|\boldsymbol{d}\|_{\square} \leq \Delta_k.$$

- We usually choose $\|\cdot\|_{\square} = \|\cdot\|_2$. Other common norms are $\|\cdot\|_1$ and $\|\cdot\|_\infty$.
- The global convergence does not depend on the choice of $\|\cdot\|_{\square}$, but the empirical performance do.

TRM Subproblem

Definition. (Trust-region subproblem)
The trust-region subproblem at the k-th iterate is

$$\min_{\boldsymbol{d}\in\mathbb{R}^n} f_Q(\boldsymbol{d};\boldsymbol{x}_k) \;=\; f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}\rangle + \frac{1}{2}\boldsymbol{d}^\top \boldsymbol{B}_k \boldsymbol{d}, \text{ s.t. } \|\boldsymbol{d}\|_\square \le \Delta_k,$$

where $\boldsymbol{B}_k$ is a symmetric matrix and $\Delta_k > 0$ is the trust-region radius.

- Once the step $\boldsymbol{d}_k$ is computed, update

$$\boldsymbol{x}_{k+1} \;=\; \boldsymbol{x}_k + \boldsymbol{d}_k$$

Linesearch vs Trust-Region Methods

- **Linesearch methods:**
  - Descent direction $\boldsymbol{d}_k$ first (e.g., steepest descent, Newton)
  - Then compute a stepsize τ_k based on $\boldsymbol{d}_k$ to reduce $f(\boldsymbol{x}_k + \tau \boldsymbol{d}_k)$
  - Update $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \tau_k \boldsymbol{d}_k$
- **Trust-region methods:**
  - Length first (trust-region radius Δ_k), direction second ("solve" subproblem for $\boldsymbol{d}_k$)
  - pick $\boldsymbol{d}_k$ to reduce some (quadratic) model of $f(\boldsymbol{x}_k + \boldsymbol{d})$
  - Update $\boldsymbol{x}_{k+1} \leftarrow \begin{cases} \boldsymbol{x}_k + \boldsymbol{d}_k & \text{if } f(\boldsymbol{x}_k + \boldsymbol{d}_k) < f(\boldsymbol{x}_k) \\ \boldsymbol{x}_k & \text{otherwise} \end{cases}$

Choice of the Trust-Region Radius

How do we choose the trust-region radius to guarantee sufficient decrease $f_Q(\boldsymbol{d}_k; \boldsymbol{x}_k) < f_Q(\boldsymbol{0}; \boldsymbol{x}_k)$?

- **Measure of predicted model decrease:**

$$f_Q(\boldsymbol{0}; \boldsymbol{x}_k) - f_Q(\boldsymbol{d}; \boldsymbol{x}_k) \;=\; f(\boldsymbol{x}_k) - f_Q(\boldsymbol{d}; \boldsymbol{x}_k)$$

- **Measure of actual function decrease:**

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + \boldsymbol{d}_k)$$

- **Measure of the quality of reduction:**

$$\rho_k \;=\; \frac{f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + \boldsymbol{d}_k)}{f(\boldsymbol{x}_k) - f_Q(\boldsymbol{d}; \boldsymbol{x}_k)},$$

characterizes how well $f_Q(\boldsymbol{d}; \boldsymbol{x}_k)$ models $f(\boldsymbol{x}_k + \boldsymbol{d})$.

Choice of the Trust-Region Radius

Solve the trust-region subproblem for $\boldsymbol{d}_k$ so that we have $f_Q(\boldsymbol{d}_k; \boldsymbol{x}_k) < f_Q(\boldsymbol{0}; \boldsymbol{x}_k)$. Set

$$\rho_k = \frac{f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + \boldsymbol{d}_k)}{f_Q(\boldsymbol{0}; \boldsymbol{x}_k) - f_Q(\boldsymbol{d}; \boldsymbol{x}_k)},$$

- If $\rho_k \geq \eta_{vs}$, set $\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k + \boldsymbol{d}_k, \quad \Delta_{k+1} \leftarrow \gamma_i \Delta_k$
  increase the radius
- If $\eta_s \leq \rho_k < \eta_{vs}$, set $\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k + \boldsymbol{d}_k, \quad \Delta_{k+1} \leftarrow \Delta_k$
- If $0 < \rho_k < \eta_s$, set $\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k, \quad \Delta_{k+1} \leftarrow \gamma_d \Delta_k$
  decrease the radius

Algorithm 1 Trust-region Algorithm

Input: $0 < \gamma_d < 1 < \gamma_i$, and $0 < \eta_s \leq \eta_{vs} < 1$

Initialize $\boldsymbol{x}_0$ and $\Delta_0 = 1$.

for $k = 0, 1, 2, \cdots$ **do**

Build the 2nd-order model $f_Q(\boldsymbol{d})$ of $f(\boldsymbol{x}_k + \boldsymbol{d})$;

Solve TRM subproblem for $\boldsymbol{d}_k$ such that $f_Q(\boldsymbol{d}_k; \boldsymbol{x}_k) < f(\boldsymbol{x}_k)$

Set $\rho_k \leftarrow \frac{f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + \boldsymbol{d}_k)}{f(\boldsymbol{x}_k) - f_Q(\boldsymbol{d}_k; \boldsymbol{x}_k)}$

if $\rho_k > \eta_{vs}$, **then**

set $\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k + \boldsymbol{d}_k$ and $\Delta_{k+1} \leftarrow \gamma_i \Delta_k$ (very successful)

else if $\rho_k \geq \eta_s$, **then**

set $\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k + \boldsymbol{d}_k$ and $\Delta_{k+1} \leftarrow \Delta_k$ (successful)

else, set $\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k$ and $\Delta_{k+1} \leftarrow \gamma_d \Delta_k$ (unsuccessful)

end for

Extra Details: Trust-Region Methods

- **Termination criteria:** in practice, we often stop when

$$\|\nabla f(\boldsymbol{x}_k)\|_2 \;\leq\; 10^{-6} \max\{1, \|\nabla f(\boldsymbol{x}_0)\|_2\}$$

- **Typical parameter values:**

$$\eta_s \;=\; 0.1, \quad \eta_{vs} \;=\; 0.9, \quad \gamma_d \;=\; \frac{1}{2}, \quad \gamma_i \;=\; 2.$$

Further Readings

- *Numerical Optimization*. Jorge Nocedal and Stephen J. Wright. (**Chapter 4**)

- *Trust Region Methods*. Andrew R. Conn, Nicholas I.M. Gould, and Philippe L. Toint. MOS-SIAM Series on Optimization, 2000.

Lecture Agenda

- **Trust-Region Method**
  - Algorithmic Introduction
  - **Solving the TRM Subproblem**
  - Convergence for Strict Saddle Function

- Cubic Regularization of Newton's Method

TRM Subproblem

Definition. (Trust-region subproblem)
The trust-region subproblem at the k-th iterate is

$$\min_{\boldsymbol{d}\in\mathbb{R}^n} f_Q(\boldsymbol{d};\boldsymbol{x}_k) \;=\; f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}\rangle + \frac{1}{2}\boldsymbol{d}^\top \boldsymbol{B}_k \boldsymbol{d}, \text{ s.t. } \|\boldsymbol{d}\|_\square \le \Delta_k,$$

where $\boldsymbol{B}_k$ is a symmetric matrix and $\Delta_k > 0$ is the trust-region radius.

- Once the step $\boldsymbol{d}_k$ is computed, update

$$\boldsymbol{x}_{k+1} \;=\; \boldsymbol{x}_k + \boldsymbol{d}_k$$

The Effect of Different Norms

The choice of $\|\cdot\|_\square$ determines the behavior of $\boldsymbol{d}_k$ as $\Delta_k \to 0$
If $\|\boldsymbol{d}\|_\square \ll 1$, then

$$\begin{aligned} f_Q(\boldsymbol{d}; \boldsymbol{x}_k) &= f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d} \rangle + \frac{1}{2} \boldsymbol{d}^\top \boldsymbol{B}_k \boldsymbol{d} \\ &\approx f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d} \rangle \end{aligned}$$

so that for $\Delta_k \ll 1$, we have

$$\min_{\|\boldsymbol{d}\|_\square \leq \Delta_k} f_Q(\boldsymbol{d}; \boldsymbol{x}_k) \approx \min_{\|\boldsymbol{d}\|_\square \leq \Delta_k} f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d} \rangle$$

The Effect of Different Norms

The solution approaches the *steepest-descent* direction of length Δ_k as $\Delta_k \to 0$

- ℓ_2 **-norm:**

$$\lim_{\Delta_k \to 0} \boldsymbol{d}_k \;\to\; -\frac{\Delta_k}{\|\nabla f(\boldsymbol{x}_k)\|_2} \nabla f(\boldsymbol{x}_k)$$

- ℓ_∞**-norm:**

$$\lim_{\Delta_k \to 0} \boldsymbol{d}_k \;\to\; -\Delta_k \widehat{\boldsymbol{e}}, \quad \text{with} \quad \widehat{e}_j \;=\; \text{sign}\left([\nabla f(\boldsymbol{x}_k)]_j\right)$$

TRM Subproblem with ℓ_∞-norm

For ℓ_∞-norm, the subproblem is

$$\min_{\boldsymbol{d}\in\mathbb{R}^n} f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}\rangle + \frac{1}{2}\boldsymbol{d}^\top \boldsymbol{B}_k \boldsymbol{d},$$
$$\text{s.t. } -\Delta_k \mathbf{1} \leq \boldsymbol{d} \leq \Delta_k \mathbf{1},$$

- This is a quadratic program (QP) and possibly nonconvex (i.e., when $\boldsymbol{B}_k$ is indefinite);
- Finding the global minimizer could be NP-hard.

TRM Subproblem with ℓ_∞-norm

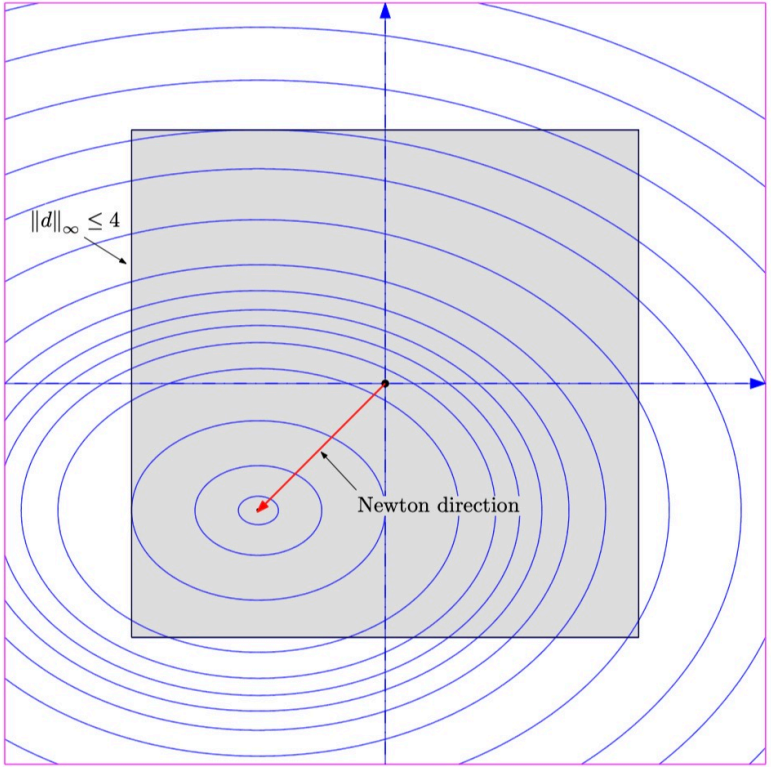
Example: Consider the following TRM subproblem

$$\min_{\boldsymbol{d}} f_Q(\boldsymbol{d}) = f + \boldsymbol{g}^\top \boldsymbol{d} + \frac{1}{2}\boldsymbol{d}^\top \boldsymbol{B}\boldsymbol{d}, \quad \text{s.t. } \|\boldsymbol{d}\|_\infty \leq 4,$$

with $f = 0$, $\boldsymbol{g} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$, and $\boldsymbol{B} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$.

The unique global minimizer $\boldsymbol{d}_\star = \begin{bmatrix} -2 \\ -2 \end{bmatrix}$ lies inside the trust region with $f_Q(\boldsymbol{d}_\star) = -6$.

TRM Subproblem with ℓ_∞-norm



TRM Subproblem with ℓ_∞-norm

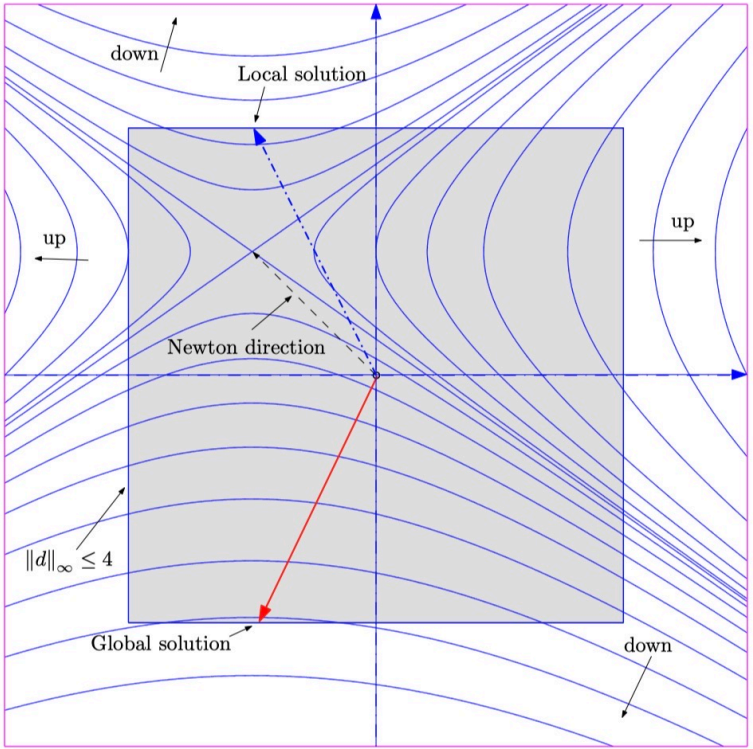
Example: Consider the following TRM subproblem

$$\min_{\boldsymbol{d}} f_Q(\boldsymbol{d}) = f + \boldsymbol{g}^\top \boldsymbol{d} + \frac{1}{2}\boldsymbol{d}^\top \boldsymbol{B}\boldsymbol{d}, \quad \text{s.t. } \|\boldsymbol{d}\|_\infty \leq 4,$$

with $f = 0$, $\boldsymbol{g} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$, and $\boldsymbol{B} = \begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix}$.

Then $f_Q(\boldsymbol{d})$ is unbounded below and $\boldsymbol{d}_N = \begin{bmatrix} -2 \\ 2 \end{bmatrix}$ is the step to a saddle point of $f_Q(\boldsymbol{d})$.

TRM Subproblem with ℓ_∞-norm



- the unique global minimizer lies on the boundary of the trust-region;
- there are two local solutions.

TRM Subproblem with ℓ_2 -norm

For ℓ_2-norm, the subproblem is

$$\min_{\boldsymbol{d}\in\mathbb{R}^n} \; f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}\rangle + \frac{1}{2}\boldsymbol{d}^\top \boldsymbol{B}_k \boldsymbol{d},$$
$$\text{s.t. } \|\boldsymbol{d}\|_2 \leq \Delta_k,$$

This is a nonlinearly constrained optimization problem

- the global minimizer can be computed *efficiently*
- we will focus on the ℓ_2-norm trust-region subproblem

The Cauchy point of TRM Subproblem

Definition (Cauchy Point) At point $\boldsymbol{x}$, we call $\boldsymbol{y}_C := \boldsymbol{x} + \boldsymbol{d}_C$, the Cauchy point if $\boldsymbol{d}_C = -\tau_C \nabla f(\boldsymbol{x})$ with

$$\tau_C := \arg\min_{\tau > 0} f_Q(-\tau \nabla f(\boldsymbol{x}); \boldsymbol{x}), \quad \text{s.t.} \quad \|\tau \nabla f(\boldsymbol{x})\|_2 \leq \Delta$$

- Cauchy point requires minimal condition of sufficient decrease in the model;
- Using Cauchy direction can only guarantee convergence to a critical point $\nabla f(\boldsymbol{x}) \to 0$;
- We need a direction that can escape saddle point.

TRM Subproblem with ℓ_2 -norm

For ℓ_2-norm, the subproblem is

$$\min_{\boldsymbol{d}\in\mathbb{R}^n} f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}\rangle + \frac{1}{2}\boldsymbol{d}^\top \boldsymbol{B}_k \boldsymbol{d},$$
$$\text{s.t. } \|\boldsymbol{d}\|_2 \leq \Delta_k,$$

This is a nonlinearly constrained optimization problem

- Cauchy step $\boldsymbol{d}_C$ implies a steepest descent like method;
- exact solution implies Newton-like method

Exact Solution of TRM Subproblem

Case 1: $\|\boldsymbol{d}_\star\|_2 \leq \Delta$

$$\nabla f_Q(\boldsymbol{d}; \boldsymbol{x}) \;=\; \mathbf{0}$$

This gives the following optimality condition

$$\nabla f(\boldsymbol{x}) + \boldsymbol{B}\boldsymbol{d} \;=\; \mathbf{0} \quad \Longrightarrow \quad \boldsymbol{B}\boldsymbol{d}_\star = -\nabla f(\boldsymbol{x})$$

so that $\boldsymbol{d}_\star$ is the Newton direction.

Exact Solution of TRM Subproblem

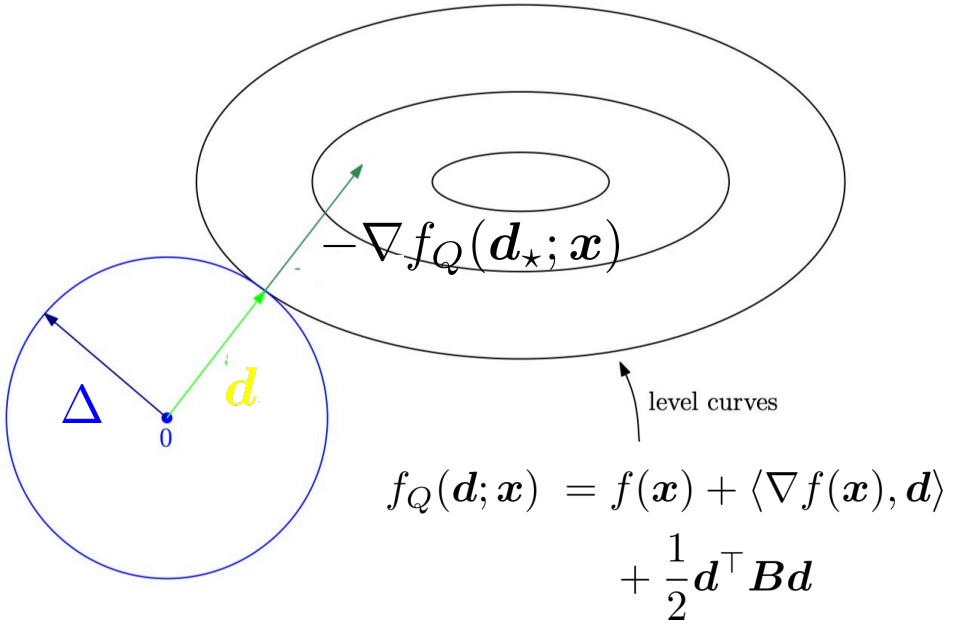
Case 2: $\|\boldsymbol{d}_\star\|_2 = \Delta$

- $\boldsymbol{d}_\star$ and $-\nabla f_Q(\boldsymbol{d}_\star; \boldsymbol{x})$ point in the same direction
- There exists $\lambda_\star > 0$ such that

$$\lambda_\star \boldsymbol{d}_\star = -\nabla f_Q(\boldsymbol{d}; \boldsymbol{x})$$

which is equivalent to

$$(\boldsymbol{B} + \lambda_\star \boldsymbol{I})\boldsymbol{d}_\star = -\nabla f(\boldsymbol{x})$$



Exact Solution of TRM Subproblem

Theorem. A vector $\boldsymbol{d}_\star$ is a global minimizer of

$$\min_{\boldsymbol{d}\in\mathbb{R}^n} \; f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{d}\rangle + \frac{1}{2}\boldsymbol{d}^\top \boldsymbol{B}\boldsymbol{d}, \quad \text{s.t. } \|\boldsymbol{d}\|_2 \leq \Delta,$$

if and only if $\|\boldsymbol{d}_\star\|_2 \leq \Delta$ and there exists a scalar $\lambda_\star$ such that

- $\lambda_\star \geq 0$;
- $(\boldsymbol{B} + \lambda_\star \boldsymbol{I})\,\boldsymbol{d}_\star = -\nabla f(\boldsymbol{x})$
- $\boldsymbol{B} + \lambda_\star \boldsymbol{I}$ *is positive semi-definite (PSD)*
- $\lambda_\star(\|\boldsymbol{d}_\star\|_2 - \Delta) = 0.$

Moreover, if $\boldsymbol{B} + \lambda_\star \boldsymbol{I}$ is positive definite, then $\boldsymbol{d}_\star$ is unique.

Approach I: Newton's Method

Thus, we need to consider **three** cases:

1. $\boldsymbol{B}$ is PSD and

$$\boldsymbol{B}\boldsymbol{d}_\star = -\nabla f(\boldsymbol{x}), \quad \|\boldsymbol{d}_\star\|_2 \leq \Delta. \qquad \text{(easy)}$$

2. $\boldsymbol{B}$ is PSD and $\lambda_\star > 0$ that

$$(\boldsymbol{B} + \lambda_\star \boldsymbol{I})\boldsymbol{d}_\star = -\nabla f(\boldsymbol{x}), \quad \|\boldsymbol{d}_\star\|_2 = \Delta. \qquad \text{(challenging)}$$

3. $\boldsymbol{B}$ is *not* PSD (e.g., indefinite) and $\lambda_\star > 0$ that

$$(\boldsymbol{B} + \lambda_\star \boldsymbol{I})\boldsymbol{d}_\star = -\nabla f(\boldsymbol{x}), \quad \|\boldsymbol{d}_\star\|_2 = \Delta. \qquad \text{(challenging)}$$

We only need to consider Case 2 & Case 3 with $\|\boldsymbol{d}_\star\|_2 = \Delta$, solving nonlinear system in $\boldsymbol{d}$ and λ.

Approach I: Newton's Method

We only need to consider Case 2 & Case 3 with $\|\boldsymbol{d}_\star\|_2 = \Delta$, solving nonlinear system in $\boldsymbol{d}$ and λ such that

$$(\boldsymbol{B} + \lambda \boldsymbol{I})\boldsymbol{d} = -\nabla f(\boldsymbol{x}), \quad \|\boldsymbol{d}\|_2 = \Delta,$$
$$\boldsymbol{B} + \lambda \boldsymbol{I} \succeq \boldsymbol{0}, \quad \lambda \geq 0.$$

Goal. Find a scalar $\lambda_\star \geq \max\{0, -\lambda_n\}$ and a vector $\boldsymbol{d}_\star$ such that

$$(\boldsymbol{B} + \lambda \boldsymbol{I})\boldsymbol{d}_\star = -\nabla f(\boldsymbol{x}), \quad \|\boldsymbol{d}_\star\|_2 = \Delta.$$

Approach I: Newton's Method

We only need to consider Case 2 & Case 3 with $\|\boldsymbol{d}_\star\|_2 = \Delta$, solving nonlinear system in $\boldsymbol{d}$ and λ such that

$$(\boldsymbol{B} + \lambda \boldsymbol{I})\boldsymbol{d} \;=\; -\nabla f(\boldsymbol{x}), \quad \|\boldsymbol{d}\|_2 \;=\; \Delta,$$
$$\boldsymbol{B} + \lambda \boldsymbol{I} \;\succeq\; \boldsymbol{0}, \quad \lambda \;\geq\; 0.$$

- Given $\boldsymbol{B}$ is symmetric, consider the eigen decomposition

$$\boldsymbol{B} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top, \quad \boldsymbol{\Lambda} \;=\; \operatorname{diag}\left(\lambda_1, \lambda_2, \cdots, \lambda_n\right),$$

so that $\boldsymbol{d}(\lambda) \;=\; -\left(\boldsymbol{B} + \lambda \boldsymbol{I}\right)^{-1} \nabla f(\boldsymbol{x}) \;=\; -\sum_{i=1}^{n} \dfrac{\boldsymbol{v}_i^\top \nabla f(\boldsymbol{x})}{\lambda_i + \lambda} \boldsymbol{v}_i$

Approach I: Newton's Method

$$\boldsymbol{d}(\lambda) \;=\; -\left(\boldsymbol{B}+\lambda \boldsymbol{I}\right)^{-1} \nabla f(\boldsymbol{x}) \;=\; -\sum_{i=1}^{n} \frac{\boldsymbol{v}_i^{\top} \nabla f(\boldsymbol{x})}{\lambda_i + \lambda} \boldsymbol{v}_i$$

Thus, we would like

$$\|\boldsymbol{d}(\lambda)\|_2^2 \;=\; \sum_{i=1}^{n} \frac{\left(\boldsymbol{v}_i^{\top} \nabla f(\boldsymbol{x})\right)^2}{(\lambda_i + \lambda)^2} \;=\; \Delta^2$$

- This implies that $\psi(\lambda) \;:=\; \|\boldsymbol{d}(\lambda)\|_2$ has poles that $\lambda = -\lambda_i$ if $\boldsymbol{v}_i^{\top} \nabla f(\boldsymbol{x}) \neq 0$.

Approach I: Newton's Method

Question: How do we actually solve $\|\boldsymbol{d}(\lambda)\|_2 = \Delta$?

A: Solve the following secular equation

$$\phi(\lambda) := \frac{1}{\|\boldsymbol{d}(\lambda)\|_2} - \frac{1}{\Delta} = 0$$

using Newton's method (safeguarded).

- The method works except for the "hard" case.

Approach I: Newton's Method

Example: Consider the following TRM subproblem

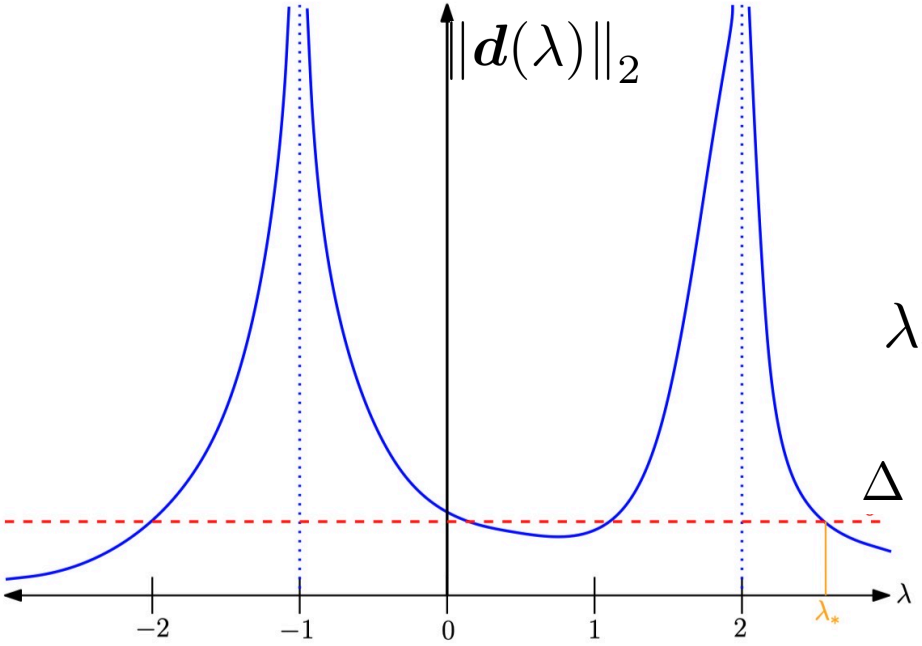
$$\min_{\boldsymbol{d}} f_Q(\boldsymbol{d}) = f + \boldsymbol{g}^\top \boldsymbol{d} + \frac{1}{2}\boldsymbol{d}^\top \boldsymbol{B}\boldsymbol{d}, \quad \text{s.t. } \|\boldsymbol{d}\|_2 \leq 4,$$

with $f = 0$, $\boldsymbol{g} = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$, and $\boldsymbol{B} = \begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix}$.

$$\boldsymbol{d}(\lambda) = \begin{bmatrix} -2/(\lambda+1) \\ -4/(\lambda-2) \end{bmatrix} \implies \|\boldsymbol{d}(\lambda)\|_2^2 = \frac{4}{(\lambda+1)^2} + \frac{16}{(\lambda-2)^2}$$

- It follows that $\psi(\lambda) = \|\boldsymbol{d}(\lambda)\|_2$ has two poles at $\lambda_1 = -1$ and $\lambda_2 = 2$.

Approach I: Newton's Method

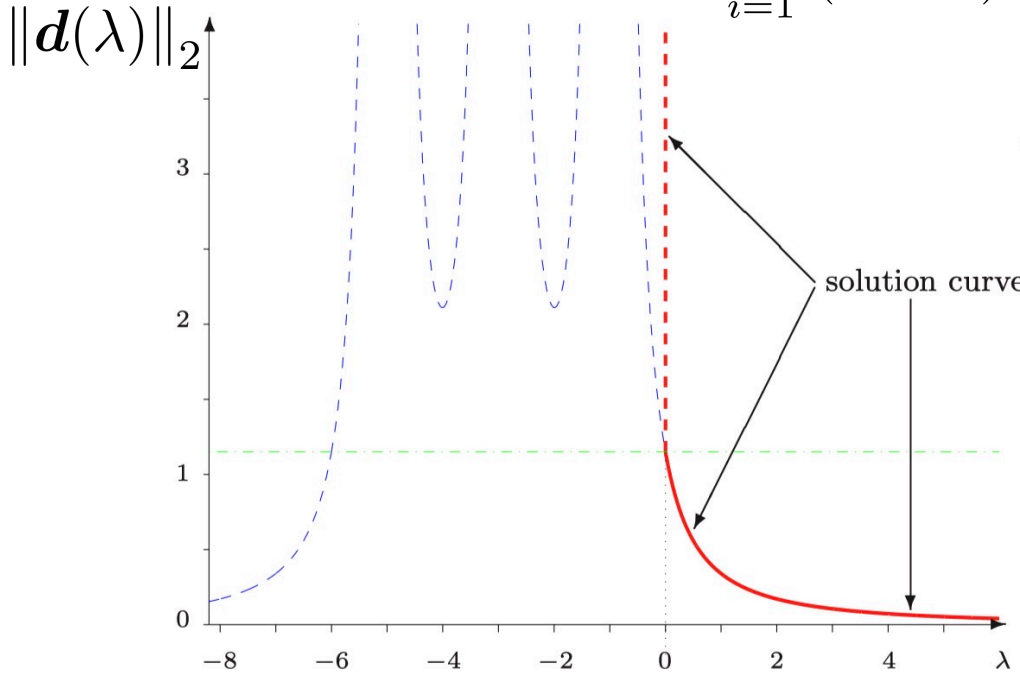


$$\lambda_\star \geq \max\{0, -\lambda_n\}$$

It is clear that for any $\Delta > 0$ there is a $\lambda_\star > 2$ such that $\psi(\lambda_\star) = \Delta$

Approach I: Newton's Method

- **convex case:** $\|\boldsymbol{d}(\lambda)\|_2^2 = \sum_{i=1}^{n} \frac{\left(\boldsymbol{v}_i^\top \boldsymbol{g}\right)^2}{(\lambda_i + \lambda)^2} = \Delta^2$



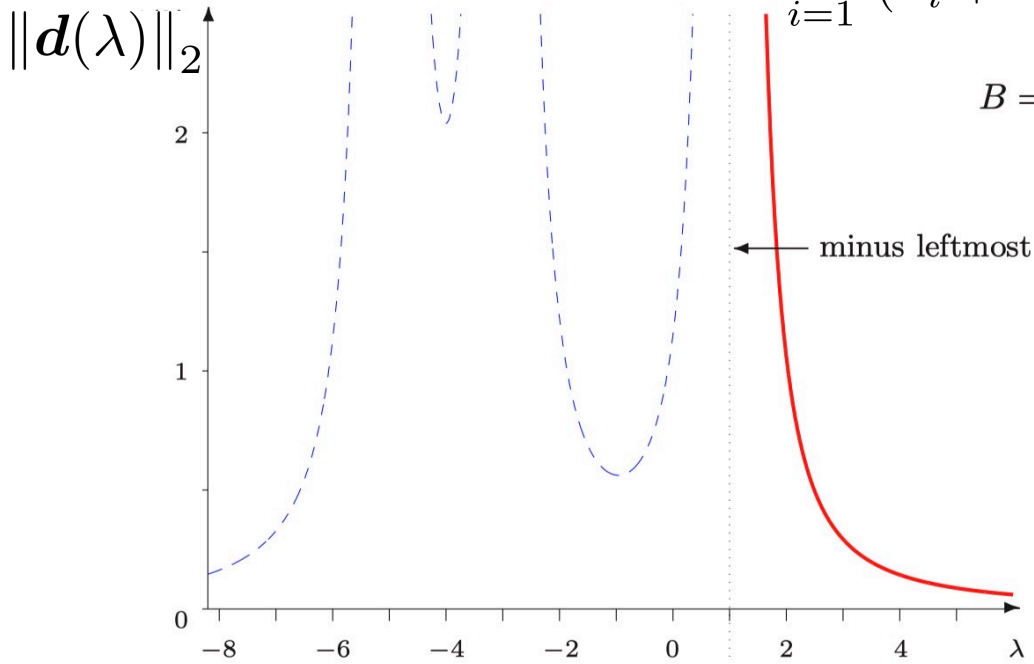
$$B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{pmatrix}$$

$$\Delta = 1.151$$

$$g = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Approach I: Newton's Method

- **nonconvex case:** $\|\boldsymbol{d}(\lambda)\|_2^2 = \sum_{i=1}^{n} \frac{\left(\boldsymbol{v}_i^\top \boldsymbol{g}\right)^2}{(\lambda_i + \lambda)^2} = \Delta^2$

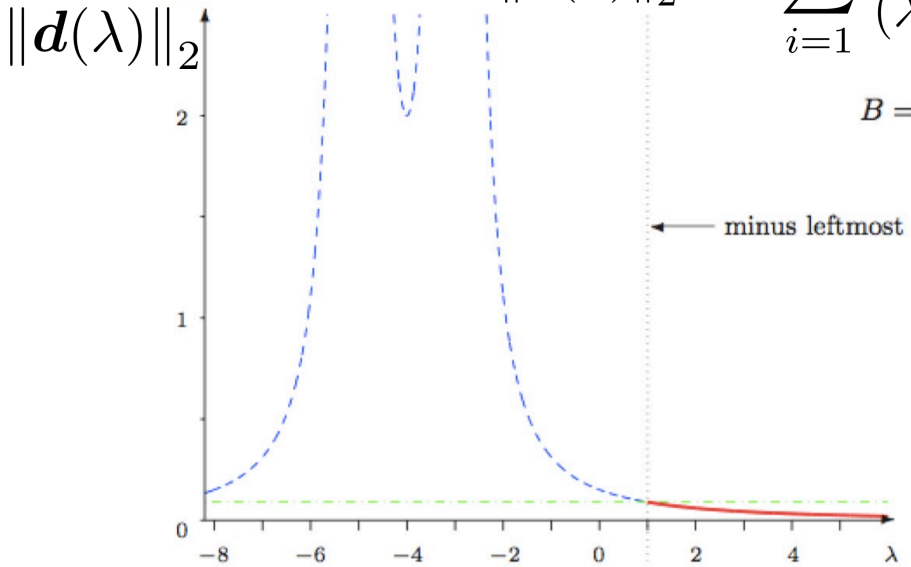


$$B = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{pmatrix}$$

$$g = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Approach I: Newton's Method

- **the "hard" case:** $\|\boldsymbol{d}(\lambda)\|_2^2 = \sum_{i=1}^{n} \frac{\left(\boldsymbol{v}_i^\top \boldsymbol{g}\right)^2}{(\lambda_i + \lambda)^2} = \Delta^2$



$$B = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{pmatrix}$$

$$g = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$

$\Delta = 0.0903$

- $\boldsymbol{g}^\top \boldsymbol{v}_3 = 0$ so that no pole at $\lambda = -\lambda_3 = 1$.
- No obvious solution for $\Delta > 0.0903$. (of course, there is one!)

Approach I: Newton's Method

- **the "hard" case:** $\|\boldsymbol{d}(\lambda)\|_2^2 = \sum_{i=1}^{n} \frac{\left(\boldsymbol{v}_i^\top \boldsymbol{g}\right)^2}{(\lambda_i + \lambda)^2} = \Delta^2$

For indefinite $\boldsymbol{B}$, the hard case occurs when $\boldsymbol{g}$ (i.e., $\nabla f(\boldsymbol{x})$) is orthogonal to the eigenvector $\boldsymbol{v}_n$ associated with λ_n

- "Okay" if the radius Δ is "small enough";
- No "obvious" solution when Δ is big, but in fact a solution is of the form

$$\boldsymbol{d}_{\lim} + \sigma \boldsymbol{v}_n$$

where $\boldsymbol{d}_{\lim} = \lim_{\lambda \to -\lambda_n} \boldsymbol{d}(\lambda), \quad \|\boldsymbol{d}_{\lim} + \sigma \boldsymbol{v}_n\|_2 = \Delta$

Approach II: SDP Relaxation

$$\min_{\boldsymbol{d}\in\mathbb{R}^n} \; f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}\rangle + \frac{1}{2}\boldsymbol{d}^\top \boldsymbol{B}_k \boldsymbol{d},$$
$$\text{s.t.} \; \|\boldsymbol{d}\|_2 \leq \Delta_k,$$

which can be *equivalently* rewritten as

$$\min_{\boldsymbol{d}\in\mathbb{R}^n} \; \frac{1}{2}\begin{bmatrix}\boldsymbol{d}\\ 1\end{bmatrix}^\top \begin{bmatrix}\boldsymbol{B}_k & \nabla f(\boldsymbol{x}_k)\\ \nabla f(\boldsymbol{x}_k)^\top & 2f(\boldsymbol{x}_k)\end{bmatrix}\begin{bmatrix}\boldsymbol{d}\\ 1\end{bmatrix}$$
$$\text{s.t.} \; \begin{bmatrix}\boldsymbol{d}\\ 1\end{bmatrix}^\top \begin{bmatrix}\boldsymbol{d}\\ 1\end{bmatrix} \leq \Delta_k^2 + 1,$$

Approach II: SDP Relaxation

$$\min_{\boldsymbol{d}\in\mathbb{R}^n} \frac{1}{2}\begin{bmatrix}\boldsymbol{d}\\1\end{bmatrix}^\top \begin{bmatrix}\boldsymbol{B}_k & \nabla f(\boldsymbol{x}_k)\\ \nabla f(\boldsymbol{x}_k)^\top & 2f(\boldsymbol{x}_k)\end{bmatrix}\begin{bmatrix}\boldsymbol{d}\\1\end{bmatrix}$$

$$\text{s.t.} \begin{bmatrix}\boldsymbol{d}\\1\end{bmatrix}^\top \begin{bmatrix}\boldsymbol{d}\\1\end{bmatrix} \le \Delta_k^2 + 1,$$

Let $\boldsymbol{D} = \begin{bmatrix}\boldsymbol{d}\\1\end{bmatrix}\begin{bmatrix}\boldsymbol{d}\\1\end{bmatrix}^\top$, then consider the convex relaxation via *lifting*

$$\min_{\boldsymbol{D}\in\mathbb{R}^{(n+1)\times(n+1)}} \left\langle \begin{bmatrix}\boldsymbol{B}_k & \nabla f(\boldsymbol{x}_k)\\ \nabla f(\boldsymbol{x}_k)^\top & 2f(\boldsymbol{x}_k)\end{bmatrix}, \boldsymbol{D}\right\rangle$$

$$\text{s.t. } \langle \boldsymbol{D}, \boldsymbol{I}\rangle \le \Delta_k^2 + 1, \quad D_{n+1,n+1} = 1, \quad \boldsymbol{D} \succeq \boldsymbol{0}.$$

Approach II: SDP Relaxation

$$\min_{\boldsymbol{D}\in\mathbb{R}^{(n+1)\times(n+1)}} \left\langle \begin{bmatrix} \boldsymbol{B}_k & \nabla f(\boldsymbol{x}_k) \\ \nabla f(\boldsymbol{x}_k)^\top & 2f(\boldsymbol{x}_k) \end{bmatrix}, \boldsymbol{D} \right\rangle$$
$$\text{s.t. } \langle \boldsymbol{D}, \boldsymbol{I} \rangle \leq \Delta_k^2 + 1, \quad D_{n+1,n+1} = 1, \quad \boldsymbol{D} \succeq \boldsymbol{0}.$$

- Once the optimal $\boldsymbol{D}_\star$ is obtained, find the direction $\boldsymbol{d}_\star$ via computing the leading eigenvector of $\boldsymbol{D}_\star$;
- However, solving the semidefinite programming (SDP) problem can be very *expensive* !

Further Readings

- *Numerical Optimization*. Jorge Nocedal and Stephen J. Wright. (**Chapter 4**)

- *Trust Region Methods*. Andrew R. Conn, Nicholas I.M. Gould, and Philippe L. Toint. MOS-SIAM Series on Optimization, 2000.

Lecture Agenda

- **Trust-Region Method**
  - Algorithmic Introduction
  - Solving the TRM Subproblem
  - **Convergence for Strict Saddle Function**

- Cubic Regularization of Newton's Method

Assumptions

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \; f(\boldsymbol{x})$$

- The nonlinear function $f(\boldsymbol{x}) : \mathbb{R}^n \mapsto \mathbb{R}$ is twice continuously differentiable;
- The gradient is L_1-Lipschitz continuous

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}')\|_2 \;\leq\; L_1 \|\boldsymbol{x} - \boldsymbol{x}'\|_2 , \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^n.$$

- The Hessian is L_2-Lipschitz continuous

$$\left\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x}')\right\|_2 \;\leq\; L_2 \|\boldsymbol{x} - \boldsymbol{x}'\|_2 , \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^n.$$

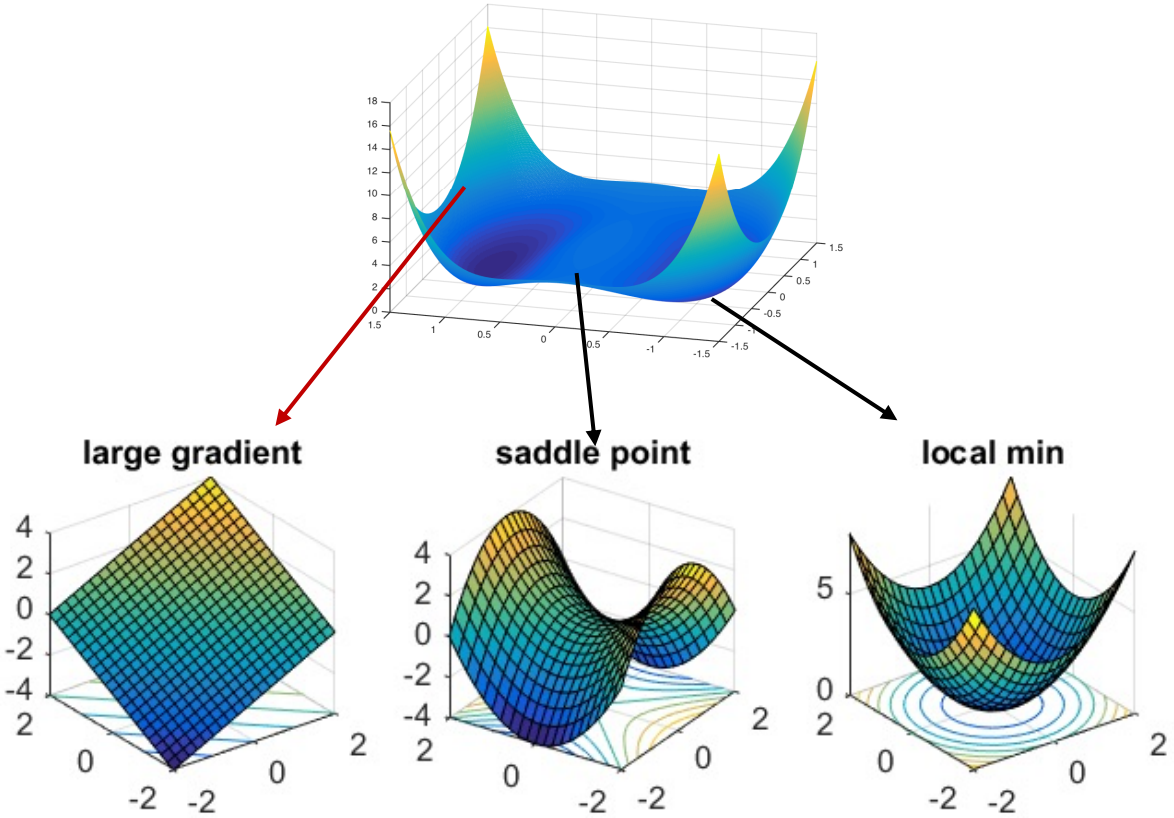
Strict Saddle Function in $\mathbb{R}^n$

Definition. (Strict Saddle Function in $\mathbb{R}^n$, Ge et al.'15) A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is $(\alpha, \beta, \gamma, \delta)$-strict saddle, if $\forall \boldsymbol{x} \in \mathbb{R}^n$ obeys *at least one* of the following:

- **[Large gradient]** $\|\nabla f(\boldsymbol{x})\|_2 \geq \beta$;
- **[Negative curvature]** $\exists\, \boldsymbol{v} \in \mathbb{S}^{n-1}$, such that
$$\boldsymbol{v}^\top \nabla^2 f(\boldsymbol{x}) \boldsymbol{v} \leq -\alpha;$$
- **[Strong convexity around minimizers]** $\exists \boldsymbol{x}_\star$ such that $\|\boldsymbol{x} - \boldsymbol{x}_\star\|_2 \leq \delta$, and for all $\boldsymbol{y} \in \mathcal{B}(\boldsymbol{x}_\star, 2\delta)$, we have $\nabla^2 f(\boldsymbol{y}) \succeq \gamma \boldsymbol{I}$.

Escaping From Saddle Points --- Online Stochastic Gradient for Tensor Decomposition. Rong Ge, Furong Huang, Chi Jin, Yang Yuan, COLT'15, 2015.

Strict Saddle Function in $\mathbb{R}^n$



Convergence to 2nd Order Critical Points

- For strict saddle function, we show that the iterates produced by TRM converges to a point $\boldsymbol{x}_\star$ that satisfies

$$\nabla f(\boldsymbol{x}_\star) \;=\; \mathbf{0}, \quad \nabla^2 f(\boldsymbol{x}_\star) \;\succeq\; \mathbf{0}.$$

- For analysis, we divide $\mathbb{R}^n$ into three regions

$$\begin{aligned}
\mathcal{R}_g &:= \left\{\boldsymbol{x} \in \mathbb{R}^n \;\middle|\; \|\nabla f(\boldsymbol{x})\|_2 \geq \beta\right\} \\
\mathcal{R}_n &:= \left\{\boldsymbol{x} \in \mathbb{R}^n \;\middle|\; \lambda_{\min}\left(\nabla^2 f(\boldsymbol{x})\right) \leq -\alpha\right\} \\
\mathcal{R}_s &:= \left\{\boldsymbol{x} \in \mathbb{R}^n \;\middle|\; \nabla^2 f(\boldsymbol{x}) \succeq \gamma \boldsymbol{I}\right\}
\end{aligned}$$

Constant Decrease in Gradient Region

Proposition 1 (Constant Decrease in $\mathcal{R}_g$)
Suppose $\boldsymbol{x}_k \in \mathcal{R}_g := \{\boldsymbol{x} \in \mathbb{R}^n \mid \|\nabla f(\boldsymbol{x})\|_2 \geq \beta\}$, and the radius of TRM subproblem satisfies

$$\Delta \leq \min\left\{\frac{\beta}{L_1}, \sqrt{\frac{3\beta}{4L_2}}\right\}.$$

Then for $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{d}_k$, we have

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \frac{\beta\Delta}{4}.$$

Decrease in Negative Curvature Region

Proposition 2 (Constant Decrease in $\mathcal{R}_n$)
Suppose $\boldsymbol{x}_k \in \mathcal{R}_n := \left\{\boldsymbol{x} \in \mathbb{R}^n \mid \lambda_{\min}\left(\nabla^2 f(\boldsymbol{x})\right) \leq -\alpha\right\}$, and the radius of TRM subproblem satisfies

$$\Delta \leq \frac{3}{8}\frac{\alpha}{L_2}.$$

Then for $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{d}_k$, we have

$$f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k) - \frac{\alpha\Delta^2}{8}.$$

Quadratic Convergence Near Local Minimizer

Proposition 3 (Local Quadratic Convergence in $\mathcal{R}_s$)
Suppose $\boldsymbol{x}_k \in \mathcal{R}_s := \left\{\boldsymbol{x} \in \mathbb{R}^n \mid \nabla^2 f(\boldsymbol{x}) \geq \gamma \boldsymbol{I}\right\}$, and the radius of TRM subproblem is sufficiently small.
Suppose there exists a local minimizer $\boldsymbol{x}_\star \in \mathcal{R}_s$, then for $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{d}_k$, we have

$$\|\nabla f(\boldsymbol{x}_{k+1})\|_2 \leq c_1 \|\nabla f(\boldsymbol{x}_k)\|_2^2$$

$$\|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2 \leq c_2 2^{-2^k}$$

Further Readings

- *Numerical Optimization*. Jorge Nocedal and Stephen J. Wright. (**Chapter 4**)

- *Trust Region Methods*. Andrew R. Conn, Nicholas I.M. Gould, and Philippe L. Toint. MOS-SIAM Series on Optimization, 2000.

Lecture Agenda

- Trust-Region Method
  - Algorithmic Introduction
  - Solving the TRM Subproblem
  - Convergence for Strict Saddle Function

- **Cubic Regularization of Newton's Method**

Assumptions

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \; f(\boldsymbol{x})$$

- The *nonlinear* function $f(\boldsymbol{x}) : \mathbb{R}^n \mapsto \mathbb{R}$ is twice continuously differentiable;
- The gradient is L_1-Lipschitz continuous

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{x}')\|_2 \;\leq\; L_1 \|\boldsymbol{x} - \boldsymbol{x}'\|_2 , \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^n.$$

- The Hessian is L_2-Lipschitz continuous

$$\left\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x}')\right\|_2 \;\leq\; L_2 \|\boldsymbol{x} - \boldsymbol{x}'\|_2 , \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^n.$$

Recall from Proximal Method

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \; F(\boldsymbol{x}) \;=\; f(\boldsymbol{x}) \;+\; g(\boldsymbol{x})$$

Minimize the quadratic upper bound:

$$\widehat{F}_\mu(\boldsymbol{x},\boldsymbol{x}_k) \;=\; \underbrace{f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}-\boldsymbol{x}_k\rangle + g(\boldsymbol{x})}_{=:\,\ell_F(\boldsymbol{x},\boldsymbol{x}_k)} + \frac{1}{2\mu}\|\boldsymbol{x}-\boldsymbol{x}_k\|_2^2$$

that for any $\mu \in (0, 1/L_1)$,

$$F(\boldsymbol{x}) \;\le\; \widehat{F}_\mu(\boldsymbol{x},\boldsymbol{x}_k)$$

Recall from Proximal Method

$$\widehat{F}_\mu(\boldsymbol{x}, \boldsymbol{x}_k) = g(\boldsymbol{x}) + \frac{1}{2\mu} \|\boldsymbol{x} - (\boldsymbol{x}_k - \mu \nabla f(\boldsymbol{x}_k))\|_2^2 + f(\boldsymbol{x}_k) - \frac{\mu}{2} \|\nabla f(\boldsymbol{x}_k)\|_2^2$$

Let $\boldsymbol{w}_k = \boldsymbol{x}_k - \frac{1}{\mu} \nabla f(\boldsymbol{x}_k)$,

$$\begin{aligned} \boldsymbol{x}_{k+1} &= \arg\min_{\boldsymbol{x}} \left\{ g(\boldsymbol{x}) + \frac{1}{2\mu} \|\boldsymbol{x} - \boldsymbol{w}_k\|_2^2 \right\} \\ &= \operatorname{prox}_{\mu g}(\boldsymbol{w}_k). \end{aligned}$$

Recall from Proximal Method

$$G_\mu(\boldsymbol{x}) := \frac{1}{\mu}\left(\boldsymbol{x} - \mathrm{prox}_{\mu g}(\boldsymbol{x} - \mu\nabla f(\boldsymbol{x}))\right), \text{ then}$$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \mu \cdot \underbrace{G_\mu(\boldsymbol{x}_k)}_{\text{proximal gradient}}.$$

- However, minimizing quadratic upper bound cannot help us escape saddle point.
- It is built upon a linear approximation, with no Hessian (negative curvature) information utilized.
- To escape strict saddle, we need second-order approximation.

Idea: Minimize the Cubic Upper Bound?

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \; f(\boldsymbol{x})$$

From Lecture 3, when $\nabla^2 f$ is L_2-Lipschitz, we have

$$f(\boldsymbol{x}+\boldsymbol{d}) \;\leq\; \underbrace{f_Q(\boldsymbol{d};\boldsymbol{x}) + \frac{L_2}{6}\|\boldsymbol{d}\|_2^3}_{\widehat{f}_Q(\boldsymbol{d};\boldsymbol{x})},$$

with that

$$f_Q(\boldsymbol{d};\boldsymbol{x}) \;:= f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{d}\rangle + \frac{1}{2}\boldsymbol{d}^\top \nabla^2 f(\boldsymbol{x})\boldsymbol{d}.$$

Minimize the Cubic Upper Bound?

Idea: instead of minimizing the quadratic upper bound, minimize the *cubic upper bound*:

$$\boldsymbol{d}_k \;=\; \arg\min_{\boldsymbol{d}} \widehat{f}_Q(\boldsymbol{d}; \boldsymbol{x}_k)$$

With

$$\widehat{f}_Q(\boldsymbol{d}; \boldsymbol{x}) \;:= f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{d} \rangle + \frac{1}{2}\boldsymbol{d}^\top \nabla^2 f(\boldsymbol{x})\boldsymbol{d} + \frac{L_2}{6}\|\boldsymbol{d}\|_2^3 .$$

The resulting method is known as the *Cubic Regularized Newton's Method*.

Cubic Regularized Newton's Method

Problem Class:

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^n,$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is nonconvex, and is twice continuously differentiable, with both gradient and Hessian Lipschitz continuous. We have access to the second order oracle: $\nabla f(\boldsymbol{x}) \in \mathbb{R}^n$ and $\nabla^2 f(\boldsymbol{x}) \in \mathbb{R}^{n \times n}$.

Setup: Let $\hat{f}(\boldsymbol{y}, \boldsymbol{x})$ be defined similarly as in (9.2.4):

$$\hat{f}(\boldsymbol{y}, \boldsymbol{x}) \doteq \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{x})^* \nabla^2 f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) + \frac{L_2}{6} \|\boldsymbol{y} - \boldsymbol{x}\|_2^3 .$$

Initialization: Set $\boldsymbol{x}_0 \in \mathbb{R}^n$,

Iteration: For $k = 0, 1, 2, \ldots$

$$\boldsymbol{x}_{k+1} = \arg\min_{\boldsymbol{y}} \hat{f}(\boldsymbol{y}, \boldsymbol{x}_k).$$

Convergence Guarantee: $\boldsymbol{x}_k$ converges with $\lim_{k \to \infty} \mu(\boldsymbol{x}_k) = 0$.

Relationship with Trust-Region Method

- **Subproblem of cubic regularized Newton:**

$$\min_{\boldsymbol{d}\in\mathbb{R}^n} f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}\rangle + \frac{1}{2}\boldsymbol{d}^\top \nabla^2 f(\boldsymbol{x}_k)\boldsymbol{d} + \frac{L_2}{6}\|\boldsymbol{d}\|_2^3.$$

- **Subproblem of trust-region method:**

$$\min_{\boldsymbol{d}\in\mathbb{R}^n} f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}\rangle + \frac{1}{2}\boldsymbol{d}^\top \nabla^2 f(\boldsymbol{x}_k)\boldsymbol{d}, \text{ s.t. } \|\boldsymbol{d}\|_2 \le \Delta_k,$$

The cubic regularization method can be viewed as replacing the constraint $\|\boldsymbol{d}\|_2 \le \Delta_k$ in the trust region method by a cubic regularization $\frac{L_2}{6}\|\boldsymbol{d}\|_2^3$.

Relationship with Trust-Region Method

- **Subproblem of cubic regularized Newton:**

$$\min_{\boldsymbol{d}\in\mathbb{R}^n} f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}\rangle + \frac{1}{2}\boldsymbol{d}^\top \nabla^2 f(\boldsymbol{x}_k)\boldsymbol{d} + \frac{L_2}{6}\|\boldsymbol{d}\|_2^3.$$

- **Subproblem of trust-region method:**

$$\min_{\boldsymbol{d}\in\mathbb{R}^n} f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}\rangle + \frac{1}{2}\boldsymbol{d}^\top \nabla^2 f(\boldsymbol{x}_k)\boldsymbol{d}, \text{ s.t. } \|\boldsymbol{d}\|_2 \leq \Delta_k,$$

If L_2 is not known, we can estimate it using backtracking linesearch similar to the proximal method

Solving the Subproblem of Cubic Regularization

- **Subproblem of cubic regularized Newton:**

$$\min_{\boldsymbol{d}\in\mathbb{R}^n} f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}\rangle + \frac{1}{2}\boldsymbol{d}^\top \nabla^2 f(\boldsymbol{x}_k)\boldsymbol{d} + \frac{L_2}{6}\|\boldsymbol{d}\|_2^3.$$

  - When $\nabla^2 f(\boldsymbol{x}_k)$ is not PSD (i.e., $\nabla^2 f(\boldsymbol{x}_k) \not\succeq \mathbf{0}$), the subproblem is still *nonconvex*.
  - The subproblem shares a lot of similarities to TRM subproblem, where we might potentially run into "hard cases", similar to TRM subproblem.

Solving the Subproblem of Cubic Regularization

- **Subproblem of cubic regularized Newton:**

$$\min_{\boldsymbol{d}\in\mathbb{R}^n} \psi(\boldsymbol{d}) = f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{d}\rangle + \frac{1}{2}\boldsymbol{d}^\top \nabla^2 f(\boldsymbol{x}_k)\boldsymbol{d} + \frac{L_2}{6}\|\boldsymbol{d}\|_2^3.$$

  - Nonetheless, we can still run gradient descent, with

$$\nabla\psi(\boldsymbol{d}) = \nabla f(\boldsymbol{x}_k) + \nabla^2 f(\boldsymbol{x}_k)\boldsymbol{d} + \frac{L_2}{6}\nabla\|\boldsymbol{d}\|_2^3$$

$$\nabla^2 f(\boldsymbol{x}_k)\boldsymbol{d} \approx \frac{\nabla f(\boldsymbol{x}_k + t\boldsymbol{d}) - \nabla f(\boldsymbol{x}_k)}{t}$$

  - Recent work showed that gradient descent with small noise can globally optimize the problem with $O(\varepsilon^{-1}\log(1/\varepsilon))$.

Convergence to 2nd Order Critical Points

Similar to TRM, the cubic regularized Newton's method converges to a 2nd-order critical point $\boldsymbol{x}_\star$ with

$$\nabla f(\boldsymbol{x}_\star) \;=\; \mathbf{0}, \quad \nabla^2 f(\boldsymbol{x}_\star) \;\succeq\; \mathbf{0}.$$

- We measure our progress by the following quantity

$$\mu(\boldsymbol{x}) \;:=\; \max\left\{\sqrt{\frac{1}{L_2}\left\|\nabla f(\boldsymbol{x})\right\|_2},\; -\frac{2}{3L_2}\lambda_{\min}\left(\nabla^2 f(\boldsymbol{x})\right)\right\}$$

- If $\mu(\boldsymbol{x}) \to 0$, then $\boldsymbol{x}_k \to \boldsymbol{x}_\star$.

Convergence to 2nd Order Critical Points

Theorem (Convergence Rate of Cubic Newton's Method) Suppose $f(\boldsymbol{x})$ is bounded from below. The sequence $\{\boldsymbol{x}_k\}$ that

$$\boldsymbol{x}_{k+1} = \arg\min_{\boldsymbol{d}} \widehat{f}_Q(\boldsymbol{d}; \boldsymbol{x}_k)$$

converges to a non-empty set of limit points $\mathcal{X}$. Let $\boldsymbol{x}_\star \in \mathcal{X}$, then we further have $\lim_{k\to\infty} \mu(\boldsymbol{x}_k) = 0$ and for any $k \geq 1$,

$$\min_{1\leq i\leq k} \mu(\boldsymbol{x}_i) \leq C \left(\frac{f(\boldsymbol{x}_0) - f(\boldsymbol{x}_\star)}{k \cdot L_2} \right)^{1/3}$$

for some constant $C > 0$.

Convergence to 2nd Order Critical Points

- The fact $\mu(\boldsymbol{x}_k) \to 0$ implies that $\{\boldsymbol{x}_k\}_{k\geq 1}$ converges to a 2nd order critical point with

$$\nabla f(\boldsymbol{x}_\star) = \mathbf{0}, \quad \nabla^2 f(\boldsymbol{x}_\star) \succeq \mathbf{0}$$

- The bound on $\mu(\boldsymbol{x}_k)$ implies that

$$\min_{1\leq i\leq k} \|\nabla f(\boldsymbol{x}_i)\|_2 \;\leq\; O(k^{-2/3})$$

which improves over $O(k^{-1/2})$ for 1st-order gradient descent method, and it is *tight* for methods with access to 2nd-order oracle.

Further Readings

- *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications*. John Wright, Yi Ma. (**Chapter 9.2**)

- *Cubic Regularization of Newton Method and Its Global Performance*. Yurii Nesterov, and B.T. Polyak. Mathematical Programming, 2006.

- *Gradient Descent Finds the Cubic-regularized Nonconvex Newton Step*. Yair Carmon and John Duchi, SIAM Journal on Optimization, 2019.