Stochastic Second-Order Optimization Methods

Part I: Convex

Fred Roosta

School of Mathematics and Physics
University of Queensland

Disclaimer

Disclaimer:

- To preserve brevity and flow, many cited results are simplified, slightly-modified, or generalized...please see the cited work for the details.
- Unfortunately, due to time constraints, many relevant and interesting works are not mentioned in this presentation.

Problem Statement

Problem

$$\min_{\mathbf{x} \in \mathcal{X} \subseteq \mathcal{R}^d} F(\mathbf{x}) = \mathbb{E}_{\mathbb{w}} f(\mathbf{x}; \mathbb{w}) = \overbrace{\int_{\Omega} \underbrace{f(\mathbf{x}; \mathbb{w}(\omega))}_{\substack{\text{e.g., Loss/Penalty} \\ + \\ \text{prediction function}}} \, \mathrm{d}P(\mathbb{w}(\omega))}^{\text{Risk}}$$

- F: (non-)convex/(non-)smooth
- High-dimensional: $d \gg 1$
- With empirical (counting) measure over $\{\mathbf{w}_i\}_{i=1}^{n} \subset \mathcal{R}^p$, i.e.,

$$\int_{\mathcal{A}} \mathrm{d}P(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{\mathbf{w}_i \in \mathcal{A}\}}, \ \forall \mathcal{A} \subseteq \Omega \Longrightarrow \underbrace{F(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{x}; \mathbf{w}_i)}_{\text{finite-sum/empirical risk}}$$

- "Big data": $n \gg 1$


SAY BIG DATA
ONE MORE TIME
memegenerator.net


WE'VE DECIDED TO TAKE BIG DATA TO THE NEXT LEVEL...
HUMONGOUS DATA
© D.Fletcher for CloudTweaks.com

Humongous Data / High Dimension

Classical algorithms $\Longrightarrow$ High per-iteration costs

Humongous Data / High Dimension

Modern variants:

1. Low per-iteration costs
2. Maintain original convergence rate

First Order Methods

- Use only gradient information
  - E.g. : Gradient Descent

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha^{(k)} \nabla F(\mathbf{x}^{(k)})$$

- Smooth Convex $F \Rightarrow$ Sublinear, $\mathcal{O}(1/k)$
- Smooth Strongly Convex $F \Rightarrow$ Linear, $\mathcal{O}(\rho^k)$, $\rho < 1$
- However, iteration cost is high!

First Order Methods

- Stochastic variants e.g., (mini-batch) SGD
  - For some s small, chosen at random $\{\mathbf{w}_i\}_{i=1}^{s}$ and

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \frac{\alpha^{(k)}}{s} \sum_{i=1}^{s} \nabla f(\mathbf{x}^{(k)}, \mathbf{w}_i)$$

  - Cheap per-iteration costs!
  - However slower to converge:
    - Smooth Convex $F \Rightarrow \mathcal{O}(1/\sqrt{k})$
    - Smooth Strongly Convex $F \Rightarrow \mathcal{O}(1/k)$

- Modifications...
  - Achieve the convergence rate of the full GD
  - Preserve the per-iteration cost of SGD
- E.g.: SAG, SDCA, SVRG, Prox-SVRG, Acc-Prox-SVRG, Acc-Prox-SDCA, S2GD, mS2GD, MISO, SAGA, AMSVRG, ...

Second Order Methods

- Use both gradient and Hessian information
  - E.g. : Classical Newton's method

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha^{(k)} \underbrace{[\nabla^2 F(\mathbf{x}^{(k)})]^{-1} \nabla F(\mathbf{x}^{(k)})}_{\text{Non-uniformly Scaled Gradient}}$$

- Smooth Convex $F \Rightarrow$ Locally Superlinear
- Smooth Strongly Convex $F \Rightarrow$ Locally Quadratic and Globally Linear
- However, per-iteration cost is much higher!

Outline

- Part I: Convex
  - Smooth
    - Newton-CG
  - Non-Smooth
    - Proximal Newton
    - Semi-smooth Newton
- Part II: Non-Convex
  - Line-Search Based Methods
    - L-BFGS
    - Gauss-Newton
    - Natural Gradient
  - Trust-Region Based Methods
    - Trust-Region
    - Cubic Regularization
- Part III: Discussion and Examples

Outline

- Part I: Convex
  - Smooth
    - Newton-CG
  - Non-Smooth
    - Proximal Newton
    - Semi-smooth Newton
- Part II: Non-Convex
  - Line-Search Based Methods
    - L-BFGS
    - Gauss-Newton
    - Natural Gradient
  - Trust-Region Based Methods
    - Trust-Region
    - Cubic Regularization
- Part III: Discussion and Examples

Finite Sum / Empirical Risk Minimization

FSM/ERM

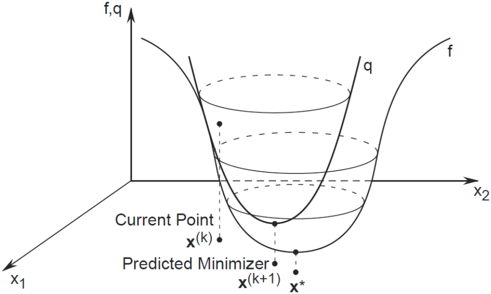
$$\min_{\mathbf{x}\in\mathcal{X}\subseteq\mathcal{R}^d} F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x};\mathbf{w}_i) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})$$

- f_i: Convex and Smooth
- F: Strongly Convex $\Longrightarrow$ Unique minimizer $\mathbf{x}^\star$
- $n \gg 1$ and/or $d \gg 1$

Iterative Scheme

$$\mathbf{y}^{(k)} = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}} \left\{ (\mathbf{x}-\mathbf{x}^{(k)})^T \mathbf{g}^{(k)} + \frac{1}{2}(\mathbf{x}-\mathbf{x}^{(k)})^T \mathbf{H}^{(k)} (\mathbf{x}-\mathbf{x}^{(k)}) \right\}$$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \left(\mathbf{y}^{(k)} - \mathbf{x}^{(k)} \right)$$



Iterative Scheme

$$\mathbf{y}^{(k)} = \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmin}} \left\{ (\mathbf{x}-\mathbf{x}^{(k)})^T \mathbf{g}^{(k)} + \frac{1}{2}(\mathbf{x}-\mathbf{x}^{(k)})^T \mathbf{H}^{(k)} (\mathbf{x}-\mathbf{x}^{(k)}) \right\}, \quad \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha_k \left(\mathbf{y}^{(k)} - \mathbf{x}^{(k)}\right)$$

- Newton: $\mathbf{g}^{(k)} = \nabla F(\mathbf{x}^{(k)}) \quad \& \quad \mathbf{H}^{(k)} = \nabla^2 F(\mathbf{x}^{(k)})$
- Gradient Descent: $\mathbf{g}^{(k)} = \nabla F(\mathbf{x}^{(k)}) \quad \& \quad \mathbf{H}^{(k)} = \mathbf{I}$
- Frank-Wolfe: $\mathbf{g}^{(k)} = \nabla F(\mathbf{x}^{(k)}) \quad \& \quad \mathbf{H}^{(k)} = 0$
- (mini-batch) SGD: $\mathcal{S}_{\mathbf{g}} \subset \{1,2,\ldots,n\} \implies \mathbf{g}^{(k)} = \frac{1}{|\mathcal{S}_{\mathbf{g}}|}\sum_{j\in\mathcal{S}_{\mathbf{g}}} \nabla f_j(\mathbf{x}^{(k)}) \quad \& \quad \mathbf{H}^{(k)} = \mathbf{I}$
- Sub-Sampled Newton:

Hessian Sub-Sampling

$$\mathbf{g}^{(k)} = \nabla F(\mathbf{x}^{(k)})$$

$$\mathcal{S}_{\mathbf{g}} \subset \{1,2,\ldots,n\} \implies \mathbf{H}^{(k)} = \frac{1}{|\mathcal{S}_H|}\sum_{j\in\mathcal{S}_H} \nabla^2 f_j(\mathbf{x}^{(k)})$$

Gradient and Hessian Sub-Sampling

$$\mathcal{S}_{\mathbf{g}} \subset \{1,2,\ldots,n\} \implies \mathbf{g}^{(k)} = \frac{1}{|\mathcal{S}_{\mathbf{g}}|}\sum_{j\in\mathcal{S}_{\mathbf{g}}} \nabla f_j(\mathbf{x}^{(k)})$$

$$\mathcal{S}_{\mathbf{H}} \subset \{1,2,\ldots,n\} \implies \mathbf{H}^{(k)} = \frac{1}{|\mathcal{S}_H|}\sum_{j\in\mathcal{S}_H} \nabla^2 f_j(\mathbf{x}^{(k)})$$

Sub-sampled Newton's Method

Let $\mathcal{X} = \mathcal{R}^d$, i.e., unconstrained optimization

Iterative Scheme

$$\mathbf{p}^{(k)} \approx \operatorname*{argmin}_{\mathbf{p} \in \mathcal{R}^d} \left\{ \mathbf{p}^T \mathbf{g}^{(k)} + \frac{1}{2} \mathbf{p}^T \mathbf{H}^{(k)} \mathbf{p} \right\},$$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha^{(k)} \mathbf{p}^{(k)}$$

Hessian Sub-Sampling

$$\mathbf{g}^{(k)} = \nabla F(\mathbf{x}^{(k)})$$

$$\mathbf{H}^{(k)} = \frac{1}{|\mathcal{S}_H|} \sum_{j \in \mathcal{S}_H} \nabla^2 f_j(\mathbf{x}^{(k)})$$

Sub-sampled Newton's Method

Global convergence, i.e., starting from any initial point

Sub-sampled Newton's Method

Iterative Scheme

Descent Direction: $\mathbf{H}^{(k)}\mathbf{p}^{(k)} \approx -\mathbf{g}^{(k)}$,

Step Size: $\begin{cases} \alpha_k = \underset{\alpha \le 1}{\text{argmax}} \;\; \alpha \\ \qquad\qquad \text{s.t. } F(\mathbf{x}^{(k)} + \alpha \mathbf{p}_k) \le F(\mathbf{x}^{(k)}) + \alpha\beta \mathbf{p}_k^T \nabla F(\mathbf{x}^{(k)}) \end{cases}$

Update: $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha^{(k)}\mathbf{p}^{(k)}$

Sub-sampled Newton's Method

Algorithm Newton's Method with Hessian Sub-Sampling

1: **Input:** $\mathbf{x}^{(0)}$
2: **for** $k = 0, 1, 2, \cdots$ until termination **do**
3: - $\mathbf{g}^{(k)} = \nabla F(\mathbf{x}^{(k)})$
4: - $\mathcal{S}_H^{(k)} \subseteq \{1, 2, \ldots, n\}$
5: - $\mathbf{H}^{(k)} = \frac{1}{|\mathcal{S}_H^{(k)}|} \sum_{j \in \mathcal{S}_H^{(k)}} \nabla^2 f_j(\mathbf{x}^{(k)})$
6: - $\mathbf{H}^{(k)}\mathbf{p}^{(k)} \approx -\mathbf{g}^{(k)}$
7: - Find $\alpha^{(k)}$ that passes Armijo linesearch
8: - Update $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha^{(k)}\mathbf{p}^{(k)}$
9: **end for**

Sub-sampled Newton's Method

Theorem ([Byrd et al., 2011])

If each f_i is **strongly** *convex, twice continuously differentiable with Lipschitz continuous gradient, then*

$$\lim_{k \to \infty} \nabla F(\mathbf{x}^{(k)}) = 0.$$

Sub-sampled Newton's Method

Theorem ([Bollapragada et al., 2016])

If each f_i is **strongly** *convex, twice continuously differentiable with Lipschitz continuous gradient, then*

$$\mathbb{E}\left(F(\mathbf{x}^{(k)}) - F(\mathbf{x}^\star)\right) \leq \rho^k \left(F(\mathbf{x}^{(0)}) - F(\mathbf{x}^\star)\right),$$

where $\rho \leq \left(1 - 1/\kappa^2\right)$ with κ being the "condition number" of the problem.

Sub-sampled Newton's Method

What if F is **strongly** convex, but each f_i is only **weakly** convex?, e.g.,

- $F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(\mathbf{a}_i^T\mathbf{x})$
- $\mathbf{a}_i \in \mathcal{R}^d$ and $\text{Range}(\{\mathbf{a}_i\}_{i=1}^{n}) = \mathbb{R}^d$
- $\ell_i : \mathcal{R} \to \mathcal{R}$ and $\ell_i'' \geq \gamma > 0$
- Each $\nabla^2 f_i(\mathbf{x}) = \ell_i''(\mathbf{a}_i^T\mathbf{x})\mathbf{a}_i\mathbf{a}_i^T$ is rank one!
- But $\nabla^2 F(\mathbf{x}) \succeq \gamma \cdot \underbrace{\lambda_{\min}\left(\sum_{i=1}^{n} \mathbf{a}_i\mathbf{a}_i^T\right)}_{>0}$

Sub-Sampling Hessian

$$\mathbf{H} = \frac{1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} \nabla^2 f_j(\mathbf{x})$$

Lemma ([Roosta and Mahoney, 2016a])

Suppose $\nabla^2 F(\mathbf{x}) \succeq \gamma \mathbf{I}$ *and* $0 \preceq \nabla^2 f_i(\mathbf{x}) \preceq L\mathbf{I}$. *Given any* $0 < \epsilon < 1$, $0 < \delta < 1$, *if Hessian is uniformly sub-sampled with*

$$|\mathcal{S}| \geq \frac{2\kappa \log(d/\delta)}{\epsilon^2},$$

then

$$\mathbf{Pr}\Big(\mathbf{H} \succeq (1-\epsilon)\gamma\mathbf{I}\Big) \geq 1 - \delta.$$

where $\kappa = L/\gamma$.

Sub-Sampling Hessian

Inexact Update

$\mathbf{H}\mathbf{p} \approx -\mathbf{g} \Longrightarrow \|\mathbf{H}\mathbf{p} + \mathbf{g}\| \leq \theta \|\mathbf{g}\|$, $\theta < 1$, using CG.

Why CG and not other solvers (at least theoretically)?

- $\mathbf{H}$ is SPD
- $\mathbf{p}^{(t)} = \underset{\mathbf{p} \in \mathcal{K}_t}{\text{argmin}}\ \mathbf{p}^T\mathbf{g} + \frac{1}{2}\mathbf{p}^T\mathbf{H}\mathbf{p} \to \left\langle \mathbf{p}^{(t)}, \mathbf{g} \right\rangle \leq -\frac{1}{2}\left\langle \mathbf{p}^{(t)}, \mathbf{H}\mathbf{p}^{(t)} \right\rangle$

Sub-Sampling Hessian

Theorem ([Roosta and Mahoney, 2016a])

If $\theta \leq \sqrt{\frac{(1-\epsilon)}{\kappa}}$, *then, w.h.p,*

$$F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^{\star}) \leq \rho \left(F(\mathbf{x}^{(k)}) - F(\mathbf{x}^{\star}) \right),$$

where $\rho = 1 - (1-\epsilon)/\kappa^2$.

Sub-sampled Newton's Method

Local convergence, i.e., in a neighborhood of $\mathbf{x}^\star$, and with $\alpha^{(k)} = 1$

Linear-Quadratic Error Recursion

$$\left\| \mathbf{x}^{(k+1)} - \mathbf{x}^\star \right\| \leq \xi_1 \left\| \mathbf{x}^{(k)} - \mathbf{x}^\star \right\| + \xi_2 \left\| \mathbf{x}^{(k)} - \mathbf{x}^\star \right\|^2$$

Sub-sampled Newton's Method

[Erdogdu and Montanari, 2015]

$$[\mathbf{U}_{r+1}, \mathbf{\Lambda}_{r+1}] = \text{TSVD}\left(\mathbf{H}, r+1\right)$$

$$\hat{\mathbf{H}}^{-1} = \lambda_{r+1}^{-1}\mathbf{I} + \mathbf{U}_r\left(\mathbf{\Lambda}_r^{-1} - \frac{1}{\lambda_{r+1}}\mathbf{I}\right)\mathbf{U}_r^T$$

$$\mathbf{x}^{(k+1)} = \arg\min_{\mathbf{x}\in\mathcal{X}}\left\{\mathbf{x} - \left(\mathbf{x}^{(k)} - \alpha^{(k)}\hat{\mathbf{H}}^{-1}\nabla F(\mathbf{x}^{(k)})\right)\right\}$$

Theorem ([Erdogdu and Montanari, 2015])

$$\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^\star\right\| \leq \xi_1^{(k)}\left\|\mathbf{x}^{(k)} - \mathbf{x}^\star\right\| + \xi_2^{(k)}\left\|\mathbf{x}^{(k)} - \mathbf{x}^\star\right\|^2,$$

$$\xi_1^{(k)} = 1 - \frac{\lambda_{\min}(\mathbf{H}^{(k)})}{\lambda_{r+1}(\mathbf{H}^{(k)})} + \frac{L_{\mathbf{g}}}{\lambda_{r+1}(\mathbf{H}^{(k)})}\sqrt{\frac{\log d}{|\mathcal{S}_{\mathbf{H}}^{(k)}|}}, \quad \textit{and} \quad \xi_2^{(k)} = \frac{L_{\mathbf{H}}}{2\lambda_{r+1}(\mathbf{H}^{(k)})}.$$

Note: For constrained optimization, the method is based on two metric gradient projection $\Longrightarrow$ starting from an arbitrary point, the algorithm might not recognize, i.e., fail to stop at, an stationary point.

Sub-sampling Hessian

$$\mathbf{H} = \frac{1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} \nabla^2 f_j(\mathbf{x})$$

Lemma ([Roosta and Mahoney, 2016b])

Given any $0 < \epsilon < 1$, $0 < \delta < 1$, *if Hessian is uniformly sub-sampled with*

$$|\mathcal{S}| \geq \frac{2\kappa^2 \log(d/\delta)}{\epsilon^2},$$

then

$$\mathbf{Pr}\Big((1-\epsilon)\nabla^2 F(\mathbf{x}) \preceq H(\mathbf{x}) \preceq (1+\epsilon)\nabla^2 F(\mathbf{x}) \Big) \geq 1 - \delta.$$

Sub-sampled Newton's Method [Roosta and Mahoney, 2016b]

Theorem ([Roosta and Mahoney, 2016b])

With high-probability, we get

$$\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{\star}\right\| \leq \xi_1 \left\|\mathbf{x}^{(k)} - \mathbf{x}^{\star}\right\| + \xi_2 \left\|\mathbf{x}^{(k)} - \mathbf{x}^{\star}\right\|^2,$$

where

$$\xi_1 = \frac{\epsilon}{(1-\epsilon)} + \left(\sqrt{\frac{\kappa}{1-\epsilon}}\right)\theta, \quad \textit{and} \quad \xi_2 = \frac{L_{\mathbf{H}}}{2(1-\epsilon)\gamma}.$$

If $\theta = \epsilon/\sqrt{\kappa}$, then ξ_1 is problem-independent! $\Rightarrow$ Can be made arbitrarily small!

Sub-sampled Newton's Method

Theorem ([Roosta and Mahoney, 2016b])

Consider any $0 < \rho_0 < \rho < 1$ *and* $\epsilon \leq \rho_0/(1+\rho_0)$. *If* $\|\mathbf{x}^{(0)} - \mathbf{x}^*\| \leq (\rho - \rho_0)/\xi_2$, *and*

$$\theta \leq \rho_0 \sqrt{\frac{(1-\epsilon)}{\kappa}},$$

we get locally Q-linear convergence

$$\|\mathbf{x}^{(k)} - \mathbf{x}^*\| \leq \rho \|\mathbf{x}^{(k-1)} - \mathbf{x}^*\|, \quad k = 1, \ldots, k_0$$

with probability $(1-\delta)^{k_0}$.

- Problem-independent local convergence rate
- By increasing Hessian accuracy, super-linear rate is possible

Putting it all together

Theorem ([Roosta and Mahoney, 2016b])

Under certain assumptions, starting at any $\mathbf{x}^{(0)}$, *we have*

- *linear convergence*
- *after certain number of iterations, we get "problem-independent" linear convergence*
- *after certain number of iterations, the step size of* $\alpha^{(k)} = 1$ *passes Armijo rule for all subsequent iterations*

"Any optimization algorithm for which the unit step length works has some wisdom. It is too much of a fluke if the unite step length [accidentally] works."

Prof. Jorge Nocedal
IPAM Summer School, 2012

Sub-sampled Newton's Method

Theorem ([Bollapragada et al., 2016])

Suppose

$$\left\| \mathbb{E}_i \left(\nabla^2 f_i(\mathbf{x}) - \nabla^2 F(\mathbf{x}) \right)^2 \right\| \leq \sigma^2.$$

Then,

$$\mathbb{E} \left\| \mathbf{x}^{(k+1)} - \mathbf{x}^\star \right\| \leq \xi_1 \left\| \mathbf{x}^{(k)} - \mathbf{x}^\star \right\| + \xi_2 \left\| \mathbf{x}^{(k)} - \mathbf{x}^\star \right\|^2,$$

where

$$\xi_1 = \frac{\sigma}{\gamma \sqrt{|\mathcal{S}_{\mathbf{H}}^{(k)}|}} + \kappa\theta, \quad \textit{and} \quad \xi_2 = \frac{L_{\mathbf{H}}}{2\gamma}.$$

Exploiting the structure....

Example:

$$F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \ell_i(\mathbf{a}_i^T\mathbf{x}) \Longrightarrow \nabla^2 F(\mathbf{x}) = \mathbf{A}^T\mathbf{D}\mathbf{A},$$

where

$$\mathbf{A} \triangleq \begin{pmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_n^T \end{pmatrix} \in \mathbb{R}^{n\times d}, \quad \mathbf{D}_{\mathbf{x}} \triangleq \frac{1}{n}\begin{pmatrix} \ell''(\mathbf{a}_1^T\mathbf{x}) & & \\ & \ddots & \\ & & \ell''(\mathbf{a}_n^T\mathbf{x}) \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

Sketching [Pilanci and Wainwright, 2017]

$$\mathbf{x}^{(k+1)} = \underset{\mathbf{x}\in\mathcal{X}}{\text{argmin}} \left\{ (\mathbf{x}-\mathbf{x}^{(k)})^T \nabla F(\mathbf{x}^{(k)}) + \frac{1}{2}(\mathbf{x}-\mathbf{x}^{(k)})^T \nabla^2 F(\mathbf{x}^{(k)})(\mathbf{x}-\mathbf{x}^{(k)}) \right\}$$

$$\nabla^2 F(\mathbf{x}^{(k)}) = \mathbf{B}^{(k)\,T}\mathbf{B}^{(k)}, \quad \mathbf{B}^{(k)} \in \mathcal{R}^{n\times d}$$

$$\mathbf{x}^{(k+1)} = \underset{\mathbf{x}\in\mathcal{X}}{\text{argmin}} \left\{ (\mathbf{x}-\mathbf{x}^{(k)})^T \nabla F(\mathbf{x}^{(k)}) + \frac{1}{2}\| \underbrace{\mathbf{B}^{(k)}}_{n\times d} (\mathbf{x}-\mathbf{x}^{(k)})\|^2 \right\}$$

$$\nabla^2 F(\mathbf{x}^{(k)}) \approx \mathbf{B}^{(k)\,T}\mathbf{S}^{(k)\,T}\mathbf{S}^{(k)}\mathbf{B}^{(k)}, \quad \mathbf{S}^{(k)} \in \mathcal{R}^{s\times n}, \; d \leq s \ll n$$

$$\mathbf{x}^{(k+1)} = \underset{\mathbf{x}\in\mathcal{X}}{\text{argmin}} \left\{ (\mathbf{x}-\mathbf{x}^{(k)})^T \nabla F(\mathbf{x}^{(k)}) + \frac{1}{2}\| \underbrace{\mathbf{S}^{(k)}\mathbf{B}^{(k)}}_{s\times d} (\mathbf{x}-\mathbf{x}^{(k)})\|^2 \right\}$$

Sketching [Pilanci and Wainwright, 2017]

- Sub-Gaussian sketches
  - well-known concentration properties
  - involve dense/unstructured matrix operations
- Randomized orthonormal systems, e.g., Hadamard or Fourier
  - sub-optimal sample sizes
  - fast matrix multiplication
- Random row sampling
  - Uniform
  - Non-uniform (more on this later)
- Sparse JL sketches

Non-uniform Sampling [Xu et al., 2016]

- Non-uniformity among $\nabla^2 f_i \implies \mathcal{O}(n)$ uniform samples!!!
- Find $|\mathcal{S}_{\mathbf{H}}|$ independent of n
- Immune non-uniformity
- Non-Uniform sampling schemes base on

  1. Row norms
  2. Leverage scores

LiSSA [Agarwal et al., 2017]

- Key idea: use Taylor expansion (Neumann series) to construct an estimator of the Hessian inverse
- $\|\mathbf{A}\| \leq 1, \mathbf{A} \succ 0 \Longrightarrow \mathbf{A}^{-1} = \sum_{i=0}^{\infty} (\mathbf{I} - \mathbf{A})^i$
- $\mathbf{A}_j^{-1} = \sum_{i=0}^{j} (\mathbf{I} - \mathbf{A})^i = \mathbf{I} + (\mathbf{I} - \mathbf{A})\, \mathbf{A}_{j-1}^{-1}$
- $\lim_{j \to \infty} \mathbf{A}_j^{-1} = \mathbf{A}^{-1}$
- Uniformly sub-sampled Hessian + Turncated Neumann series
- Three-nested loop involving HVP, i.e., $\nabla^2 f_{i,j}(\mathbf{x}^{(k)})\mathbf{v}_{i,j}^{(k)}$
- Leverage fast multiplication by Hessian of GLMs

Name	Complexity per iteration	Reference
Newton-CG	$\mathcal{O}(\text{nnz}(\mathbf{A})\sqrt{\kappa})$	Folklore
SSN-LS	$\tilde{\mathcal{O}}(\text{nnz}(\mathbf{A}) \log n + p^2 \kappa^{3/2})$	[Xu et al., 2016]
SSN-RNS	$\tilde{\mathcal{O}}(\text{nnz}(\mathbf{A}) + \mathbf{sr}(\mathbf{A}) p \kappa^{5/2})$	[Xu et al., 2016]
SRHT	$\tilde{\mathcal{O}}(np(\log n)^4 + p^2 (\log n)^4 \kappa^{3/2})$	[Pilanci et al., 2016]
SSN-Uniform	$\tilde{\mathcal{O}}(\text{nnz}(\mathbf{A}) + p \hat{\kappa} \kappa^{3/2})$	[Roosta et al., 2016]
LiSSA	$\tilde{\mathcal{O}}(\text{nnz}(\mathbf{A}) + p \hat{\kappa} \bar{\kappa}^2)$	[Agarwal et al., 2017]

$$\left.\begin{aligned} \kappa &= \max_{\mathbf{x}} \frac{\lambda_{\max} \nabla^2 F(\mathbf{x})}{\lambda_{\min} \nabla^2 F(\mathbf{x})} \\ \hat{\kappa} &= \max_{\mathbf{x}} \frac{\max_i \lambda_{\max} \nabla^2 f_i(\mathbf{x})}{\lambda_{\min} \nabla^2 F(\mathbf{x})} \\ \bar{\kappa} &= \max_{\mathbf{x}} \frac{\max_i \lambda_{\max} \nabla^2 f_i(\mathbf{x})}{\min_i \lambda_{\min} \nabla^2 f_i(\mathbf{x})} \end{aligned}\right\} \Rightarrow \kappa \leq \hat{\kappa} \leq \bar{\kappa}$$

Sub-sampled Newton's Method

Iterative Scheme

$$\mathbf{p}^{(k)} \approx \underset{\mathbf{p} \in \mathcal{R}^d}{\operatorname{argmin}} \left\{ \mathbf{p}^T \mathbf{g}^{(k)} + \frac{1}{2} \mathbf{p}^T \mathbf{H}^{(k)} \mathbf{p} \right\},$$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha^{(k)} \mathbf{p}^{(k)}$$

Gradient and Hessian Sub-Sampling

$$\mathbf{g}^{(k)} = \frac{1}{|\mathcal{S}_{\mathbf{g}}|} \sum_{j \in \mathcal{S}_{\mathbf{g}}} \nabla f_j(\mathbf{x}^{(k)})$$

$$\mathbf{H}^{(k)} = \frac{1}{|\mathcal{S}_H|} \sum_{j \in \mathcal{S}_H} \nabla^2 f_j(\mathbf{x}^{(k)})$$

Sub-sampled Newton's Method

Algorithm Newton's Method with Hessian and Gradient Sub-Sampling

1: **Input:** $\mathbf{x}^{(0)}$
2: **for** $k = 0, 1, 2, \cdots$ until termination **do**
3: - $\mathcal{S}_{\mathbf{g}}^{(k)} \subseteq \{1, 2, \ldots, n\} \Longrightarrow \mathbf{g}^{(k)} = \frac{1}{|\mathcal{S}_{\mathbf{g}}^{(k)}|} \sum_{j \in \mathcal{S}_{\mathbf{g}}^{(k)}} \nabla f_j(\mathbf{x}^{(k)})$
4: **if** $\left\| \mathbf{g}^{(k)} \right\| \leq \sigma \epsilon_{\mathbf{g}}$ **then**
5: - STOP
6: **end if**
7: - $\mathcal{S}_{H}^{(k)} \subseteq \{1, 2, \ldots, n\} \Longrightarrow \mathbf{H}^{(k)} = \frac{1}{|\mathcal{S}_{H}^{(k)}|} \sum_{j \in \mathcal{S}_{H}^{(k)}} \nabla^2 f_j(\mathbf{x}^{(k)})$
8: - $\mathbf{H}^{(k)} \mathbf{p}^{(k)} \approx -\mathbf{g}^{(k)}$
9: - Find $\alpha^{(k)}$ that passes Armijo linesearch
10: - Update $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha^{(k)} \mathbf{p}^{(k)}$
11: **end for**

Sub-sampled Newton's Method

Global convergence, i.e., starting from any initial point

[Byrd et al., 2012, Roosta and Mahoney, 2016a, Bollapragada et al., 2016]

Sub-sampled Newton's Method

We can write $\nabla F(\mathbf{x}) = \mathbf{A}\mathbf{B}$ where

$$\mathbf{A} := \begin{pmatrix} | & | & & | \\ \nabla f_1(\mathbf{x}) & \nabla f_2(\mathbf{x}) & \cdots & \nabla f_n(\mathbf{x}) \\ | & | & & | \end{pmatrix} \quad \text{and} \quad \mathbf{B} := \begin{pmatrix} 1/n \\ 1/n \\ \vdots \\ 1/n \end{pmatrix}$$

Lemma ([Roosta and Mahoney, 2016a])

Let $\|\nabla f_i(\mathbf{x})\| \leq G(\mathbf{x}) < \infty$. *For any* $0 < \epsilon, \delta < 1$, *if*

$$|\mathcal{S}| \geq G(\mathbf{x})^2 \left(1 + \sqrt{8 \ln(1/\delta)}\right)^2 / \epsilon^2,$$

then

$$\mathbf{Pr}\left(\|\nabla F(\mathbf{x}) - \mathbf{g}(\mathbf{x})\| \leq \epsilon \right) \geq 1 - \delta.$$

Sub-sampled Newton's Method

Theorem ([Roosta and Mahoney, 2016a])

If

$$\theta \leq \sqrt{\frac{(1-\epsilon_{\mathbf{H}})}{\kappa}},$$

then, w.h.p,

$$F(\mathbf{x}^{(k+1)}) - F(\mathbf{x}^{\star}) \leq \rho \left(F(\mathbf{x}^{(k)}) - F(\mathbf{x}^{\star}) \right),$$

where $\rho = 1 - (1-\epsilon_{\mathbf{H}})/\kappa^2$, and upon "STOP", we have $\|\nabla F(\mathbf{x}^{(k)})\| < (1+\sigma)\,\epsilon_{\mathbf{g}}$.

Sub-sampled Newton's Method

Local convergence, i.e., in a neighborhood of $\mathbf{x}^\star$, and with $\alpha^{(k)} = 1$

Error Recursion

$$\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^\star\right\| \leq \xi_0 + \xi_1 \left\|\mathbf{x}^{(k)} - \mathbf{x}^\star\right\| + \xi_2 \left\|\mathbf{x}^{(k)} - \mathbf{x}^\star\right\|^2$$

[Roosta and Mahoney, 2016b, Bollapragada et al., 2016]

Sub-sampled Newton's Method

Theorem ([Roosta and Mahoney, 2016b])

With high-probability, we get

$$\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{\star}\right\| \leq \xi_0 + \xi_1 \left\|\mathbf{x}^{(k)} - \mathbf{x}^{\star}\right\| + \xi_2 \left\|\mathbf{x}^{(k)} - \mathbf{x}^{\star}\right\|^2,$$

where

$$\xi_0 = \frac{\epsilon_{\mathbf{g}}}{(1-\epsilon_{\mathbf{H}})\gamma}$$

$$\xi_1 = \frac{\epsilon_{\mathbf{H}}}{(1-\epsilon_{\mathbf{H}})} + \left(\sqrt{\frac{\kappa}{1-\epsilon_{\mathbf{H}}}}\right)\theta,$$

$$\xi_2 = \frac{L_{\mathbf{H}}}{2(1-\epsilon_{\mathbf{H}})\gamma}.$$

Sub-sampled Newton's Method

Theorem ([Roosta and Mahoney, 2016b])

Consider any $0 < \rho_0 + \rho_1 < \rho < 1$. *If* $\|\mathbf{x}^{(0)} - \mathbf{x}^*\| \leq c(\rho_0, \rho_1, \rho)$, $\epsilon_{\mathbf{g}}^{(k)} = \rho^k \epsilon_{\mathbf{g}}$ *and*

$$\theta \leq \rho_0 \sqrt{\frac{(1 - \epsilon_{\mathbf{H}})}{\kappa}},$$

we get locally R-linear convergence

$$\|\mathbf{x}^{(k)} - \mathbf{x}^*\| \leq c\rho^k$$

with probability $(1 - \delta)^{2k}$.

- Problem-independent local convergence rate

Putting it all together

Theorem ([Roosta and Mahoney, 2016b])

Under certain assumptions, starting at any $\mathbf{x}^{(0)}$*, we have*

- *linear convergence,*
- *after certain number of iterations, we get "problem-independent" linear convergence,*
- *after certain number of iterations, the step size of* $\alpha^{(k)} = 1$ *passes Armijo rule for all subsequent iterations,*
- *upon "STOP" at iteration* k*, we have* $\|\nabla F(\mathbf{x}^{(k)})\| < (1 + \sigma)\sqrt{\rho^k \epsilon_{\mathbf{g}}}$.

Outline

- Part I: Convex
  - Smooth
    - Newton-CG
  - Non-Smooth
    - Proximal Newton
    - Semi-smooth Newton
- Part II: Non-Convex
  - Line-Search Based Methods
    - L-BFGS
    - Gauss-Newton
    - Natural Gradient
  - Trust-Region Based Methods
    - Trust-Region
    - Cubic Regularization
- Part III: Discussion and Examples

Newton-type Methods for Non-Smooth Problems

Convex Composite Problem

$$\min_{\mathbf{x} \in \mathcal{X} \subseteq \mathcal{R}^d} F(\mathbf{x}) + R(\mathbf{x})$$

- F: convex and smooth
- R: convex and non-smooth

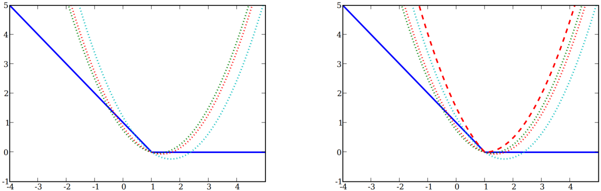
Non-Smooth Newton-type [Yu et al., 2010]

Let $\mathcal{X} = \mathcal{R}^d$, i.e., unconstrained optimization

Iterative Scheme (Non-Quadratic Sub-Problem)

$$\mathbf{p}^{(k)} \approx \underset{\mathbf{p} \in \mathcal{R}^d}{\operatorname{argmin}} \left\{ \sup_{\mathbf{g}^{(k)} \in \partial(F+R)(\mathbf{x}^{(k)})} \left\{ \mathbf{p}^T \mathbf{g}^{(k)} \right\} + \frac{1}{2} \mathbf{p}^T \underbrace{\mathbf{H}^{(k)}}_{\substack{\text{e.g.,} \\ \text{QN}}} \mathbf{p} \right\},$$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha^{(k)} \mathbf{p}^{(k)}$$



Proximal Newton-type [Lee et al., 2014, Byrd et al., 2016]

Iterative Scheme

$$\mathbf{p}^{(k)} \approx \underset{\mathbf{p} \in \mathcal{R}^d}{\operatorname{argmin}} \left\{ \mathbf{p}^T \nabla F(\mathbf{x}^{(k)}) + \frac{1}{2} \mathbf{p}^T \underbrace{\mathbf{H}^{(k)}}_{\approx \nabla^2 F(\mathbf{x}^{(k)})} \mathbf{p} + R(\mathbf{x}^{(k)} + \mathbf{p}) \right\}$$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha^{(k)} \mathbf{p}^{(k)}$$

Notable examples of proximal Newton methods:

- glmnet ([Friedman et al., 2010]): ℓ_1-regularized GLMs, sub-problems are solved using coordinate descent
- QUIC ([Hsieh et al., 2014]): Graphical Lasso problem with factorization tricks, sub-problems are solved using coordinate descent

Proximal Newton-type [Lee et al., 2014, Byrd et al., 2016]

Inexactness of sub-problem solver

$$\mathbf{p}^{(k)} \approx \underset{\mathbf{p} \in \mathcal{R}^d}{\operatorname{argmin}} \left\{ \mathbf{p}^T \nabla F(\mathbf{x}^{(k)}) + \frac{1}{2} \mathbf{p}^T \mathbf{H}^{(k)} \mathbf{p} + R(\mathbf{x}^{(k)} + \mathbf{p}) \right\}$$

$\mathbf{p}^\star$ is the minimizer of the above sub-problem iff $\mathbf{p}^\star = \mathbf{Prox}_\lambda(\mathbf{p}^\star)$, where

$$\mathbf{Prox}_\lambda^{(k)}(\mathbf{p}) \triangleq \underset{\mathbf{q} \in \mathcal{R}^d}{\operatorname{argmin}} \left\{ \left\langle \nabla F(\mathbf{x}^{(k)}) + \mathbf{H}^{(k)} \mathbf{p}, \mathbf{q} \right\rangle + R(\mathbf{x}^{(k)} + \mathbf{q}) + \frac{\lambda}{2} \|\mathbf{q} - \mathbf{p}\|^2 \right\}.$$

Inexactness

$$\left\| \mathbf{p} - \mathbf{Prox}_\lambda^{(k)}(\mathbf{p}) \right\| \leq \theta \left\| \mathbf{Prox}_\lambda^{(k)}(\mathbf{0}) \right\|,$$

and some sufficient decrease condition on the subproblem

Proximal Newton-type [Lee et al., 2014, Byrd et al., 2016]

Step-size

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \alpha^{(k)}\mathbf{p}^{(k)}$$

Line-Search

$$(F + R)(\mathbf{x}^{(k)} + \alpha\mathbf{p}^{(k)}) - (F + R)(\mathbf{x}^{(k)}) \leq \beta\left(\ell_k\left(\mathbf{x}^{(k)} + \alpha\mathbf{p}\right) - \ell_k\left(\mathbf{x}^{(k)}\right)\right)$$

$$\ell_k(\mathbf{x}) = F(\mathbf{x}^{(k)}) + (\mathbf{x} - \mathbf{x}^{(k)})^T\nabla F(\mathbf{x}^{(k)}) + R(\mathbf{x})$$

Proximal Newton-type [Lee et al., 2014, Byrd et al., 2016]

- Global convergence: $\mathbf{x}^{(k)}$ converges to an optimal solution starting at any $\mathbf{x}^{(0)}$
- Local convergence: for $\mathbf{x}^{(0)}$ close enough to $\mathbf{x}^\star$
  - $\mathbf{H}^{(k)} = \nabla^2 F(\mathbf{x}^{(k)})$
    - Exact solve: quadratic convergence
    - Approximate solve with decaying forcing term: superlinear convergence
    - Approximate solve with fixed forcing term: linear convergence
  - $\mathbf{H}^{(k)}$: Dennis-Moré $\Longrightarrow$ supelinear convergence

Sub-Sampled Proximal Newton-type [Liu et al., 2017]

FSM/ERM

$$\min_{\mathbf{x} \in \mathcal{X} \subseteq \mathcal{R}^d} F(\mathbf{x}) + R(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{a}_i^T \mathbf{x}) + R(\mathbf{x})$$

- f_i: Strongly-Convex, Smooth, and Self-Concordant
- R: Convex and Non-Smooth
- $n \gg 1$ and/or $d \gg 1$
- Dennis-Moré condition:

$$|\mathbf{p}^{(k)}{}^T \left(\mathbf{H}^{(k)} - \nabla^2 F(\mathbf{x}^{(k)}) \right) \mathbf{p}^{(k)}| \leq \eta_k \mathbf{p}^{(k)}{}^T \nabla^2 F(\mathbf{x}^{(k)}) \mathbf{p}^{(k)}$$

- Leverage score sampling to ensure Dennis-Moré
- Inexact sub-problem solver

Finite Sum / Empirical Risk Minimization

Self-Concordant

$$\left|\mathbf{v}^T \left(\nabla^3 f(\mathbf{x})[\mathbf{v}]\right) \mathbf{v}\right| \leq M \left(\mathbf{v}^T \nabla^2 f(\mathbf{x}) \mathbf{v}\right)^{3/2}, \ \forall \mathbf{x}, \mathbf{v} \in \mathcal{R}^d$$

$M = 2$ is called standard self-concordant.

Theorem ([Zhang and Lin, 2015])

Suppose there exists $\gamma > 0$ *and* $\eta \in [0, 1)$ *such that* $|f_i'''(t)| \leq \gamma \left(f_i''(t)\right)^{1-\eta}$. *Then*

$$F(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{a}_i^T \mathbf{x}) + \frac{\lambda}{2} \|\mathbf{x}\|^2$$

is self-concordant with

$$M = \frac{\max_i \|\mathbf{a}_i\|^{1+2\eta} \gamma}{\lambda^{\eta+1/2}}.$$

Proximal Newton-type Methods

- General treatment: [Fukushima and Mine, 1981, Lee et al., 2014, Byrd et al., 2016, Becker and Fadili, 2012, Schmidt et al., 2012, Schmidt et al., 2009, Shi and Liu, 2015]
- Tailored to specific problems: [Friedman et al., 2010, Hsieh et al., 2014, Yuan et al., 2012, Oztoprak et al., 2012]
- Self-Concordant: [Li et al., 2017, Kyrillidis et al., 2014, Tran-Dinh et al., 2013]

Outline

- Part I: Convex
  - Smooth
    - Newton-CG
  - Non-Smooth
    - Proximal Newton
    - Semi-smooth Newton
- Part II: Non-Convex
  - Line-Search Based Methods
    - L-BFGS
    - Gauss-Newton
    - Natural Gradient
  - Trust-Region Based Methods
    - Trust-Region
    - Cubic Regularization
- Part III: Discussion and Examples

Semi-Smooth Newton-type Methods

Convex Composite Problem

$$\min_{\mathbf{x} \in \mathcal{X} \subseteq \mathcal{R}^d} F(\mathbf{x}) + R(\mathbf{x})$$

- F: (non-)convex and (non-)smooth
- R: convex and non-smooth

Semi-Smooth Newton-type Methods

Recall:

Proximal Newton-type Methods

$$\mathbf{p}^{(k)} \approx \underset{\mathbf{p}\in\mathcal{R}^d}{\operatorname{argmin}} \left\{ \mathbf{p}^T \nabla F(\mathbf{x}^{(k)}) + \frac{1}{2}\mathbf{p}^T \mathbf{H}^{(k)} \mathbf{p} + R(\mathbf{x}^{(k)} + \mathbf{p}) \right\}$$

$$\equiv \mathbf{prox}_R^{\mathbf{H}^{(k)}} \left(\mathbf{x}^{(k)} - {\mathbf{H}^{(k)}}^{-1} \nabla F(\mathbf{x}^{(k)}) \right) - \mathbf{x}^{(k)}$$

$$\mathbf{prox}_R^{\mathbf{H}^{(k)}}(\mathbf{x}) \triangleq \underset{\mathbf{y}\in\mathcal{R}^d}{\operatorname{argmin}}\ R(\mathbf{y}) + \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|^2_{\mathbf{H}^{(k)}}$$

$$\|\mathbf{y}-\mathbf{x}\|^2_{\mathbf{H}} = \langle \mathbf{y}-\mathbf{x}, \mathbf{H}\left(\mathbf{y}-\mathbf{x}\right)\rangle$$

Semi-Smooth Newton-type Methods

Recall Newton for smooth root-finding problems:

$$\mathbf{r}(\mathbf{x}) = \mathbf{0} \Longrightarrow \mathbf{J}(\mathbf{x}^{(k)})\mathbf{p}^{(k)} = -\mathbf{r}(\mathbf{x}^{(k)}) \Longrightarrow \mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{p}^{(k)}$$

Key Idea

For solving non-linear non-smooth systems of equations, replace the Jacobian in the Newtonian iteration system by an element of Clarke's generalized Jacobian, which may be nonempty (e.g., locally Lipschitz continuous) even if the true Jacobian does not exist.

Example:

$$\mathbf{r}(\mathbf{x}) = \mathbf{prox}_R^{\mathbf{B}}\left(\mathbf{x} - \mathbf{B}^{-1}\nabla F(\mathbf{x})\right) - \mathbf{x}$$

Semi-Smooth Newton-type Methods

Clarke's generalized Jacobian

Let $\mathbf{r} : \mathcal{R}^d \rightarrow \mathcal{R}^p$ be locally Lipschitz continuous at $\mathbf{x}$[a]. Let $\mathcal{D}_{\mathbf{r}}$ be the set of differentiable points. The Bouligand-subdifferential of $\mathbf{r}$ at $\mathbf{x}$ is defined as

$$\partial_{\mathrm{B}} \triangleq \left\{ \lim_{k\to\infty} \mathbf{r}'(\mathbf{x}_k) \mid \mathbf{x}_k \in \mathcal{D}_{\mathbf{r}}, \mathbf{x}_k \rightarrow \mathbf{x} \right\}.$$

Clarke's generalized Jacobian is $\partial \mathbf{r}(\mathbf{x}) = \text{Conv-hull}\left(\partial_{\mathrm{B}} \mathbf{r}(\mathbf{x})\right)$.

[a] By Rademacher's Theorem, $\mathbf{r}$ is almost everywhere differentiable

Semi-Smooth Newton-type Methods

Semi-Smooth Map

$\mathbf{r} : \mathcal{R}^d \to \mathcal{R}^p$ is semi-smooth at $\mathbf{x}$ if

- $\mathbf{r}$ a locally Lipschitz continuous function,
- $\mathbf{r}$ is directionally differentiable at $\mathbf{x}$, and
- $\|\mathbf{r}(\mathbf{x}+\mathbf{v}) - \mathbf{r}(\mathbf{x}) - \mathbf{J}\mathbf{v}\| = o(\|\mathbf{v}\|), \ \mathbf{v} \in \mathcal{R}^d, \ \mathbf{J} \in \partial \mathbf{r}(\mathbf{x}+\mathbf{v})$

Example: Piecewise continuously differentiable functions.

A vector-valued function is (strongly) semi-smooth if and only if each of its component functions is (strongly) semi-smooth.

Semi-Smooth Newton-type Methods

- First-order stationarity is written as a fixed point-type non-linear system of equations, e.g.,

$$\mathbf{r}^{\mathbf{B}}(\mathbf{x}) \triangleq \mathbf{x} - \mathbf{prox}_R^{\mathbf{B}}\left(\mathbf{x} - \mathbf{B}^{-1}\nabla F(\mathbf{x})\right) = \mathbf{0}, \; \mathbf{B} \succ 0$$

- (Stochastic) semi-smooth Newton's method used to (approximately) solve these system of nonlinear equations , e.g., as in [Milzarek et al., 2018]

$$\hat{\mathbf{r}}^{\mathbf{B}}(\mathbf{x}) \triangleq \mathbf{x} - \mathbf{prox}_R^{\mathbf{B}}\left(\mathbf{x} - \mathbf{B}^{-1}\mathbf{g}(\mathbf{x})\right)$$

$$\mathbf{J}^{(k)}\mathbf{p}^{(k)} = -\mathbf{r}^{\mathbf{B}}(\mathbf{x}^{(k)}), \quad \mathbf{J}^{(k)} \in \mathcal{J}^{\mathbf{B}}(\mathbf{x}^{(k)})$$

$$\mathcal{J}_k^{\mathbf{B}} \triangleq \left\{\mathbf{J} \in \mathcal{R}^{d\times d} \mid \mathbf{J} = (\mathbf{I} - \mathbf{A}) + \mathbf{A}\mathbf{B}^{-1}\mathbf{H}^{(k)}, \mathbf{A} \in \partial\mathbf{prox}_R^{\mathbf{B}}\left(\mathbf{u}(\mathbf{x})\right)\right\}$$

$$\mathbf{u}(\mathbf{x}) \triangleq \mathbf{x} - \mathbf{B}^{-1}\mathbf{g}(\mathbf{x})$$

Semi-Smooth Newton-type Methods

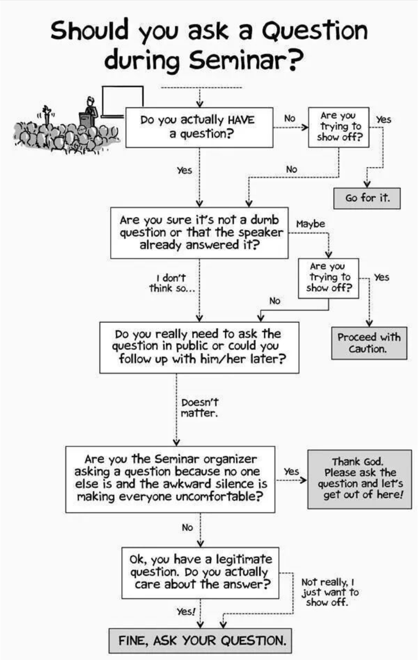
- Semi-smoothness of proximal mapping does not hold in general.
- In certain settings, these vector-valued functions are monotone and can be semi-smooth due to the properties of the proximal mappings.
- In such cases, their generalized Jacobian matrix is positive semi-definite due to monotonicity.

Lemma

For a monotone and Lipschitz continuous mapping $\mathbf{r} : \mathcal{R}^d \to \mathcal{R}^d$, *and any* $\mathbf{x} \in \mathcal{R}^d$, *each element of* $\partial_B \mathbf{r}(x)$ *is positive semi-definite.*

Semi-Smooth Newton-type Methods

- General treatment: [Patrinos et al., 2014, Milzarek and Ulbrich, 2014, Xiao et al., 2016, Patrinos and Bemporad, 2013]
- Tailored to a specific problem: [Yuan et al., 2018]
- Stochastic: [Milzarek et al., 2018]


Should you ask a Question during Seminar?
Do you actually HAVE a question?
No
Are you trying to show off?
Yes
Yes
No
Go for it.
Are you sure it's not a dumb question or that the speaker already answered it?
Maybe
Are you trying to show off?
Yes
I don't think so...
No
Do you really need to ask the question in public or could you follow up with him/her later?
Proceed with Caution.
Doesn't matter.
Are you the Seminar organizer asking a question because no one else is and the awkward silence is making everyone uncomfortable?
Yes
Thank God. Please ask the question and let's get out of here!
No
Ok, you have a legitimate question. Do you actually care about the answer?
Not really, I just want to show off.
Yes!
FINE, ASK YOUR QUESTION.

THANK YOU!

Agarwal, N., Bullins, B., and Hazan, E. (2017).
Second-order stochastic optimization for machine learning in linear time.
The Journal of Machine Learning Research, 18(1):4148–4187.

Becker, S. and Fadili, J. (2012).
A quasi-Newton proximal splitting method.
In *Advances in Neural Information Processing Systems*, pages 2618–2626.

Bollapragada, R., Byrd, R. H., and Nocedal, J. (2016).
Exact and inexact subsampled Newton methods for optimization.
IMA Journal of Numerical Analysis.

Byrd, R. H., Chin, G. M., Neveitt, W., and Nocedal, J. (2011).
On the use of stochastic Hessian information in optimization methods for machine learning.
SIAM Journal on Optimization, 21(3):977–995.

Byrd, R. H., Chin, G. M., Nocedal, J., and Wu, Y. (2012).
Sample size selection in optimization methods for machine learning.
Mathematical programming, 134(1):127–155.

Byrd, R. H., Nocedal, J., and Oztoprak, F. (2016).
An inexact successive quadratic approximation method for l-1 regularized optimization.
Mathematical Programming, 157(2):375–396.

Erdogdu, M. A. and Montanari, A. (2015).
Convergence rates of sub-sampled Newton methods.
In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, pages 3052–3060. MIT Press.

Friedman, J., Hastie, T., and Tibshirani, R. (2010).
Regularization paths for generalized linear models via coordinate descent.
Journal of statistical software, 33(1):1.

Fukushima, M. and Mine, H. (1981).
A generalized proximal point algorithm for certain non-convex minimization problems.
International Journal of Systems Science, 12(8):989–1000.

Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., and Ravikumar, P. (2014).
QUIC: quadratic approximation for sparse inverse covariance estimation.
The Journal of Machine Learning Research, 15(1):2911–2947.

Kyrillidis, A., Mahabadi, R. K., Tran-Dinh, Q., and Cevher, V. (2014).
Scalable Sparse Covariance Estimation via Self-Concordance.
In *AAAI*, pages 1946–1952.

Lee, J. D., Sun, Y., and Saunders, M. A. (2014).
Proximal Newton-type methods for minimizing composite functions.
SIAM Journal on Optimization, 24(3):1420–1443.

Li, J., Andersen, M. S., and Vandenberghe, L. (2017).
Inexact proximal Newton methods for self-concordant functions.
Mathematical Methods of Operations Research, 85(1):19–41.

Liu, X., Hsieh, C.-J., Lee, J. D., and Sun, Y. (2017).
An inexact subsampled proximal Newton-type method for large-scale machine learning.
arXiv preprint arXiv:1708.08552.

Milzarek, A. and Ulbrich, M. (2014).
A semismooth newton method with multidimensional filter globalization for l_1-optimization.
SIAM Journal on Optimization, 24(1):298–333.

Milzarek, A., Xiao, X., Cen, S., Wen, Z., and Ulbrich, M. (2018).
A Stochastic Semismooth Newton Method for Nonsmooth Nonconvex Optimization.
arXiv preprint arXiv:1803.03466.

Oztoprak, F., Nocedal, J., Rennie, S., and Olsen, P. A. (2012).
Newton-like methods for sparse inverse covariance estimation.
In *Advances in neural information processing systems*, pages 755–763.

Patrinos, P. and Bemporad, A. (2013).
Proximal Newton methods for convex composite optimization.
In *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*, pages 2358–2363. IEEE.

Patrinos, P., Stella, L., and Bemporad, A. (2014).
Forward-backward truncated Newton methods for convex composite optimization.
arXiv preprint arXiv:1402.6655.

Pilanci, M. and Wainwright, M. J. (2017).
Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence.
SIAM Journal on Optimization, 27(1):205–245.

Roosta, F. and Mahoney, M. W. (2016a).
Sub-sampled Newton methods I: globally convergent algorithms.
arXiv preprint arXiv:1601.04737.

Roosta, F. and Mahoney, M. W. (2016b).
Sub-sampled Newton methods II: Local convergence rates.
arXiv preprint arXiv:1601.04738.

Schmidt, M., Berg, E., Friedlander, M., and Murphy, K. (2009).
Optimizing costly functions with simple constraints: A limited-memory projected quasi-Newton algorithm.
In *Artificial Intelligence and Statistics*, pages 456–463.

Schmidt, M., Kim, D., and Sra, S. (2012).
Projected Newton-type Methods in Machine Learning.
Optimization for Machine Learning, (1).

Shi, Z. and Liu, R. (2015).

Large scale optimization with proximal stochastic newton-type gradient descent.
In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 691–704. Springer.

Tran-Dinh, Q., Kyrillidis, A., and Cevher, V. (2013).
Composite self-concordant minimization.
arXiv preprint arXiv:1308.2867.

Xiao, X., Li, Y., Wen, Z., and Zhang, L. (2016).
A regularized semi-smooth Newton method with projection steps for composite convex programs.
Journal of Scientific Computing, pages 1–26.

Xu, P., Yang, J., Roosta, F., Ré, C., and Mahoney, M. W. (2016).
Sub-sampled Newton methods with non-uniform sampling.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 3000–3008.

Yu, J., Vishwanathan, S., Günter, S., and Schraudolph, N. N. (2010).
A quasi-Newton approach to nonsmooth convex optimization problems in machine learning.
Journal of Machine Learning Research, 11(Mar):1145–1200.

Yuan, G.-X., Ho, C.-H., and Lin, C.-J. (2012).
An improved glmnet for l1-regularized logistic regression.
Journal of Machine Learning Research, 13(Jun):1999–2030.

Yuan, Y., Sun, D., and Toh, K.-C. (2018).
An Efficient Semismooth Newton Based Algorithm for Convex Clustering.
arXiv preprint arXiv:1802.07091.

Zhang, Y. and Lin, X. (2015).
DiSCO: Distributed optimization for self-concordant empirical loss.
In *International conference on machine learning*, pages 362–370.