Newton-MR: Newton's Method Without Smoothness or Convexity

Michael W. Mahoney

ICSI and Department of Statistics
University of California at Berkeley

Joint work with Fred Roosta, Yang Liu, and Peng Xu

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$$

Newton's Method

Classical Newton's Method

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \underbrace{\alpha_k}_{\text{step-size}} \underbrace{[\nabla^2 f(\mathbf{x}^{(k)})]^{-1} \nabla f(\mathbf{x}^{(k)})}_{\text{Newton Direction}}$$

First Order Methods

Classical Gradient Descent

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha_k \nabla f(\mathbf{x}^{(k)})$$

Machine Learning ♡ First Order Methods...

But why 1st?

Q: But Why 1st Order Methods?

- Cheap Iterations
- Easy To Implement
- "Good" Worst-Case Complexities
- Good Generalization

But why Not 2nd?

Q: But Why Not 2nd Order Methods?

- ~~Cheap~~ Expensive Iterations
- ~~Easy~~ Hard To Implement
- ~~"Good"~~ "Bad" Worst-Case Complexities
- ~~Good~~ Bad Generalization

Our Goal...

Goal: Improve 2nd Order Methods...

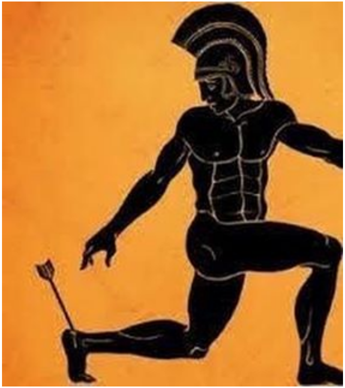
- Cheap ~~Expensive~~ Iterations
- Easy ~~Hard~~ To Use
- "Good" ~~"Bad"~~ Average(?)-Case Complexities
- Good ~~Bad~~ Generalization

Our Goal..

Any Other Advantages?

- Effective Iterations $\Rightarrow$ Less Iterations $\Rightarrow$ Less Communications
- Saddle Points For Non-Convex Problems
- Less Sensitive to Parameter Tuning
- Less Sensitive to Initialization

Achilles' heel for most 2nd-order methods is...



Achilles' heel: Solving the Sub-problems!!!

Sub-Problems

- Trust Region:

$$\mathbf{s}^{(k)} = \arg \min_{\|\mathbf{s}\| \leq \Delta_k} \left\langle \mathbf{s}, \nabla f(\mathbf{x}^{(k)}) \right\rangle + \frac{1}{2} \left\langle \mathbf{s}, \nabla^2 f(\mathbf{x}^{(k)}) \mathbf{s} \right\rangle$$

- Cubic Regularization:

$$\mathbf{s}^{(k)} = \arg \min_{\mathbf{s} \in \mathbb{R}^d} \left\langle \mathbf{s}, \nabla f(\mathbf{x}^{(k)}) \right\rangle + \frac{1}{2} \left\langle \mathbf{s}, \nabla^2 f(\mathbf{x}^{(k)}) \mathbf{s} \right\rangle + \frac{\sigma_k}{3} \|\mathbf{s}\|^3$$

Newton's Method

Recall: Classical Newton's method

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha_k \underbrace{[\nabla^2 f(\mathbf{x}^{(k)})]^{-1} \nabla f(\mathbf{x}^{(k)})}_{\text{Linear System}}$$

$$\nabla^2 f(\mathbf{x}^{(k)})\mathbf{p} = -\nabla f(\mathbf{x}^{(k)})$$

We know how to solve "$\mathbf{A}\mathbf{x} = \mathbf{b}$" very well!

Newton-CG

$$f\text{: Strongly Convex} \implies \text{Newton-CG} \implies \nabla^2 f(\mathbf{x}^{(k)})\mathbf{p} \approx -\nabla f(\mathbf{x}^{(k)})$$

$$\mathbf{p} \approx \underset{\mathbf{p}\in\mathbb{R}^d}{\operatorname{argmin}} \left\langle \mathbf{p}, \nabla f(\mathbf{x}^{(k)}) \right\rangle + \frac{1}{2} \left\langle \mathbf{p}, \nabla^2 f(\mathbf{x}^{(k)})\mathbf{p} \right\rangle$$

Why CG?

- f is strongly convex $\Longrightarrow \nabla^2 f(\mathbf{x}^{(k)})$ is SPD
- More subtly...

$$\mathbf{p}^{(t)} = \underset{\mathbf{p} \in \mathcal{K}_t}{\operatorname{argmin}} \left\langle \mathbf{p}, \nabla f(\mathbf{x}^{(k)}) \right\rangle + \frac{1}{2} \left\langle \mathbf{p}, \nabla^2 f(\mathbf{x}^{(k)}) \mathbf{p} \right\rangle$$

$\Downarrow$

$$\left\langle \mathbf{p}, \nabla f(\mathbf{x}^{(k)}) \right\rangle \leq -\frac{1}{2} \left\langle \mathbf{p}, \nabla^2 f(\mathbf{x}^{(k)}) \mathbf{p} \right\rangle < 0$$

$\Downarrow$

$\mathbf{p}^{(t)}$ is a descent direction for f for all t!

Classical Newton's Method

But...what if the Hessian is indefinite and/or singular?

- Indefinite Hessian $\Longrightarrow$ Unbounded sub-problem
- Singular Hessian and $\nabla f(\mathbf{x}) \notin \text{Range}(\nabla^2 f(\mathbf{x})) \Longrightarrow$ Unbounded sub-problem
  - $\nabla^2 f(\mathbf{x})\mathbf{p} = -\nabla f(\mathbf{x})$ has no solution

strong convexity $\Longrightarrow$ linear system sub-problems

~~strong convexity~~ $\Longrightarrow$ ~~linear system~~ sub-problems

$$\underbrace{\mathbf{A}\mathbf{x} = \mathbf{b}}_{\text{Linear System}} \quad \Longrightarrow \quad \underbrace{\|\mathbf{A}\mathbf{x} - \mathbf{b}\|}_{\text{Least Squares}}$$

$$\min_{\mathbf{p}\in\mathbb{R}^d} \| \overbrace{\nabla^2 f(\mathbf{x}_k)}^{\mathbf{A}} \overbrace{\mathbf{p}}^{\mathbf{x}} + \overbrace{\nabla f(\mathbf{x}_k)}^{-\mathbf{b}} \|$$

The underlying matrix in OLS is

- symmetric
- (possibly) indefinite
- (possibly) singular
- (possibly) ill-conditioned

MINRES-type OLS Solvers $\Longrightarrow$ MINRES-QLP [Choi et al., 2011]

Sub-problems of MINRES:

$$\mathbf{p}^{(t)} = \underset{\mathbf{p}\in\mathcal{K}_t}{\operatorname{argmin}} \frac{1}{2}\|\nabla^2 f(\mathbf{x}_k)\mathbf{p} + \nabla f(\mathbf{x}_k)\|^2$$

- There is always a solution (sometimes infinitely many)
-

$$\mathbf{p}^{(t)} = \underset{\mathbf{p}\in\mathcal{K}_t}{\operatorname{argmin}} \frac{1}{2}\|\nabla^2 f(\mathbf{x}_k)\mathbf{p} + \nabla f(\mathbf{x}_k)\|^2$$

$$\Downarrow$$

$$\left\langle \mathbf{p}^{(t)}, \nabla^2 f(\mathbf{x}^{(k)})\nabla f(\mathbf{x}^{(k)})\right\rangle \leq -\frac{1}{2}\|\nabla^2 f(\mathbf{x}_k)\mathbf{p}^{(t)}\|^2 < 0$$

$$\Downarrow$$

$\mathbf{p}^{(t)}$ is a descent direction for $\|\nabla f(\mathbf{x})\|^2$ <u>for all t!</u>

Newton-MR vs. Newton-CG

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

Newton-CG:

$$\mathbf{p}_k \approx \underset{\mathbf{p}\in\mathbb{R}^d}{\operatorname{argmin}} \langle \mathbf{g}_k, \mathbf{p} \rangle + \frac{1}{2} \langle \mathbf{p}, \mathbf{H}_k \mathbf{p} \rangle = -[\mathbf{H}_k]^{-1} \mathbf{g}_k$$

$$\alpha_k : f(\mathbf{x}_k + \alpha_k \mathbf{p}_k) \leq f(\mathbf{x}_k) + \alpha_k \beta \langle \mathbf{p}_k, \mathbf{g}_k \rangle$$

Newton-MR:

$$\mathbf{p}_k \approx \underset{\mathbf{p}\in\mathbb{R}^d}{\operatorname{argmin}} \|\mathbf{H}_k \mathbf{p} + \mathbf{g}_k\|^2 = -[\mathbf{H}_k]^{\dagger} \mathbf{g}_k$$

$$\alpha_k : \|\mathbf{g}(\mathbf{x}_k + \alpha_k \mathbf{p}_k)\|^2 \leq \|\mathbf{g}_k\|^2 + 2\alpha_k \beta \langle \mathbf{p}_k, \mathbf{H}_k \mathbf{g}_k \rangle$$

Newton-MR vs. Newton-CG

	Newton-CG	Newton-MR
Sub-problems	$\min_{\mathbf{p} \in \mathcal{K}_t} \frac{1}{2} \langle \mathbf{p}, \mathbf{H}\mathbf{p} \rangle + \langle \mathbf{p}, \mathbf{g} \rangle$	$\min_{\mathbf{p} \in \mathcal{K}_t} \Vert \mathbf{H}\mathbf{p} + \mathbf{g} \Vert^2$
Line Search	$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + \alpha\rho \langle \mathbf{p}_k, \mathbf{g}_k \rangle$	$\Vert \mathbf{g}_{k+1} \Vert^2 \leq \Vert \mathbf{g}_k \Vert^2 + 2\alpha\rho \langle \mathbf{p}_k, \mathbf{H}_k \mathbf{g}_k \rangle$

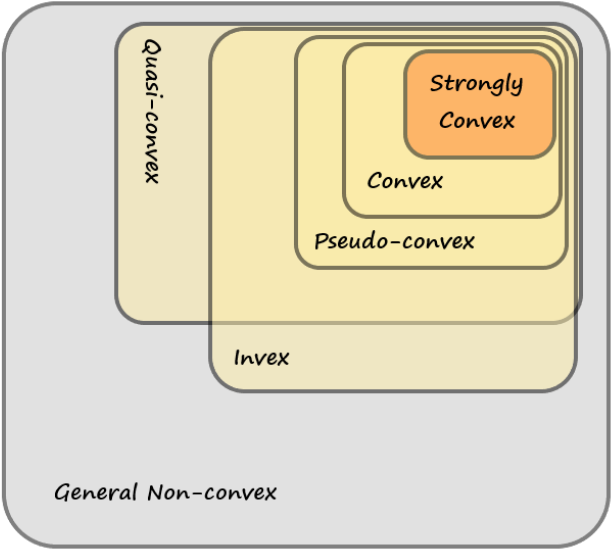
Invexity

Invexity

$$f(\mathbf{y}) - f(\mathbf{x}) \geq \langle \phi(\mathbf{y}, \mathbf{x}), \nabla f(\mathbf{x}) \rangle$$

- Necessary and sufficient for optimality: $\nabla f(\mathbf{x}) = 0$
- E.g.: Convex $\Longrightarrow \phi(\mathbf{y}, \mathbf{x}) = \mathbf{y} - \mathbf{x}$
- $$\left.\begin{array}{l} g : \mathbb{R}^p \to \mathbb{R} \text{ is differentiable and convex} \\ \mathbf{h} : \mathbb{R}^d \to \mathbb{R}^p \text{ has full-rank Jacobian } (p \leq d) \end{array}\right\} \Rightarrow g \circ \mathbf{h} \text{ is invex}$$
- "Global optimality" of stationary points in deep residual networks [Bartlett et al., 2018]

Strong Convexity $\subsetneq$ Invexity



Newton-MR vs. Newton-CG

	Newton-CG	**Newton-MR**
Sub-problems	$\min_{\mathbf{p}\in\mathcal{K}_t} \frac{1}{2}\langle\mathbf{p},\mathbf{Hp}\rangle + \langle\mathbf{p},\mathbf{g}\rangle$	$\min_{\mathbf{p}\in\mathcal{K}_t} \|\mathbf{Hp}+\mathbf{g}\|^2$
Line Search	$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + \alpha\rho\langle\mathbf{p}_k,\mathbf{g}_k\rangle$	$\|\mathbf{g}_{k+1}\|^2 \leq \|\mathbf{g}_k\|^2 + 2\alpha\rho\langle\mathbf{p}_k,\mathbf{H}_k\mathbf{g}_k\rangle$
Problem class	**Strongly Convex**	**Invex**

Moral Smoothness

(Recall) Typical Smoothness Assumptions:

Lipschitz Gradient: $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L_{\mathbf{g}} \|\mathbf{x} - \mathbf{y}\|$

Lipschitz Hessian: $\left\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\right\| \leq L_{\mathbf{H}} \|\mathbf{x} - \mathbf{y}\|$

These smoothness assumptions
are *stronger* than
what is required for first-order methods.

Moral Smoothness

Moral-Smoothness

Let $\mathcal{X}_0 \triangleq \left\{\mathbf{x} \in \mathbb{R}^d \mid \|\nabla f(\mathbf{x})\| \leq \|\nabla f(\mathbf{x}_0)\|\right\}$. For any $\mathbf{x}_0 \in \mathbb{R}^d$, there is a constant $0 < L(\mathbf{x}_0) < \infty$, such that $\forall(\mathbf{x}, \mathbf{y}) \in \mathcal{X}_0 \times \mathbb{R}^d$, we have

$$\left\|\nabla^2 f(\mathbf{y})\nabla f(\mathbf{y}) - \nabla^2 f(\mathbf{x})\nabla f(\mathbf{x})\right\| \leq L(\mathbf{x}_0)\left\|\mathbf{y} - \mathbf{x}\right\|.$$

Smoothness $\subsetneq$ Moral Smoothness

Moral Smoothness

Hessian of the quadratically smoothed hinge-loss is not continuous.

$$f(\mathbf{x}) = \frac{1}{2} \max \left\{ 0, b \langle \mathbf{a}, \mathbf{x} \rangle \right\}^2$$

But it satisfies moral-smoothness with $L = b^4 \|\mathbf{a}\|^4$.

Newton-MR vs. Newton-CG

	Newton-CG	Newton-MR
Sub-problems	$\min_{\mathbf{p}\in\mathcal{K}_t} \frac{1}{2}\langle \mathbf{p}, \mathbf{H}\mathbf{p}\rangle + \langle \mathbf{p}, \mathbf{g}\rangle$	$\min_{\mathbf{p}\in\mathcal{K}_t} \|\mathbf{H}\mathbf{p} + \mathbf{g}\|^2$
Line Search	$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + \alpha\rho\,\langle \mathbf{p}_k, \mathbf{g}_k\rangle$	$\|\mathbf{g}_{k+1}\|^2 \leq \|\mathbf{g}_k\|^2 + 2\alpha\rho\,\langle \mathbf{p}_k, \mathbf{H}_k\mathbf{g}_k\rangle$
Problem class	**Strongly Convex**	**Invex**
Smoothness	**H & g**	**Hg**

Null-Space Property

For any $\mathbf{x} \in \mathbb{R}^d$, let

- $\mathbf{U}_{\mathbf{x}}$ be an orthogonal basis for Range$(\nabla^2 f(\mathbf{x}))$
- $\mathbf{U}_{\mathbf{x}}^{\perp}$ be its orthogonal complement

Gradient-Hessian Null-Space Property

$$\left\| \left(\mathbf{U}_{\mathbf{x}}^{\perp}\right)^T \nabla f(\mathbf{x}) \right\|^2 \leq \left(\frac{1-\nu}{\nu}\right) \left\| \mathbf{U}_{\mathbf{x}}^T \nabla f(\mathbf{x}) \right\|^2, \ \forall \mathbf{x} \in \mathbb{R}^d, \ 0 < \nu \leq 1$$

- Strictly convex $f(\mathbf{x})$: $\nu = 1$
- Non-convex $f(\mathbf{x}) = \sum_{i=1}^n f_i(\mathbf{a}_i^T \mathbf{x})$: $\nu = 1$
- Some fractional programming: $\nu = 8/9$
- Some non-linear composition of functions $f(\mathbf{x}) = g(\mathbf{h}(\mathbf{x}))$

Inexactness

- Newton-CG [Roosta and Mahoney, Mathematical Programming, 2018]

$$\left\|\mathbf{H}_k \mathbf{p}_k + \mathbf{g}_k\right\| \leq \theta \left\|\mathbf{g}_k\right\| \implies \theta \leq 1/\sqrt{\kappa}$$

- Newton-MR [Roosta, Liu, Xu and Mahoney, arXiv, 2019]

$$\langle \mathbf{H}_k \mathbf{p}_k, \mathbf{g}_k \rangle \leq -(1-\theta) \left\|\mathbf{g}_k\right\|^2 \implies 1 - \nu \leq \theta < 1$$

Examples of Convergence Results

Global Linear Rate in "$\|\mathbf{g}\|$"

$$\left\|\mathbf{g}^{(k+1)}\right)\right\|^2 \leq \left(1 - \frac{4\rho(1-\rho)\gamma^2(1-\theta)^2}{L(\mathbf{x}_0)}\right)\|\mathbf{g}_k\|^2 .$$

Global Linear Rate in "$f(\mathbf{x}) - f^\star$" Under Polyak-Łojasiewicz

$$f(\mathbf{x}_k) - f^\star \leq C\zeta^k, \ \zeta < 1.$$

Error Recursion with $\alpha_k = 1$ Under Error Bound

$$\min_{\mathbf{y}\in\mathcal{X}^\star} \|\mathbf{x}_{k+1} - \mathbf{y}\| \leq c_1 \min_{\mathbf{y}\in\mathcal{X}^\star} \|\mathbf{x}_k - \mathbf{y}\|^2 + \sqrt{(1-\nu)}c_2 \min_{\mathbf{y}\in\mathcal{X}^\star} \|\mathbf{x}_k - \mathbf{y}\| .$$

Newton-MR vs. Newton-CG

	Newton-CG	**Newton-MR**
Sub-problems	$\min_{\mathbf{p}\in\mathcal{K}_t} \frac{1}{2}\langle \mathbf{p}, \mathbf{Hp}\rangle + \langle \mathbf{p}, \mathbf{g}\rangle$	$\min_{\mathbf{p}\in\mathcal{K}_t} \Vert\mathbf{Hp}+\mathbf{g}\Vert^2$
Line Search	$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) + \alpha\rho\langle \mathbf{p}_k, \mathbf{g}_k\rangle$	$\Vert\mathbf{g}_{k+1}\Vert^2 \leq \Vert\mathbf{g}_k\Vert^2 + 2\alpha\rho\langle \mathbf{p}_k, \mathbf{H}_k\mathbf{g}_k\rangle$
Problem class	**Strongly Convex**	**Invex**
Smoothness	**H & g**	**Hg**
Inexactness	$\Vert\mathbf{Hp}+\mathbf{g}\Vert \leq \theta\Vert\mathbf{g}\Vert$ $\theta < 1/\sqrt{\kappa}$	$\langle \mathbf{p}, \mathbf{Hg}\rangle \leq -(1-\theta)\Vert\mathbf{g}\Vert$ $\theta < 1$
Metric / Rate	$\Vert\mathbf{g}\Vert$: **R-linear** $f(\mathbf{x}) - f^\star$: **Q-Linear**	$\Vert\mathbf{g}\Vert$: **Q-linear** $f(\mathbf{x}) - f^\star$: **R-Linear (GPL)**

Newton-MR vs. Newton-CG for $\min f(\mathbf{x})$

&

MINRES vs. CG for $\mathbf{Ax} = \mathbf{b}$

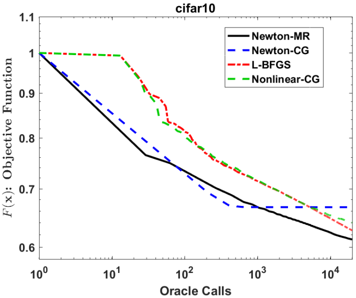
Newton-MR vs. Newton-CG

	$\min f(\mathbf{x})$	
	Newton-CG	**Newton-MR**
Sub-problems	$\min_{\mathbf{p} \in \mathcal{K}_t} \frac{1}{2} \langle \mathbf{p}, \mathbf{H}\mathbf{p} \rangle + \langle \mathbf{p}, \mathbf{g} \rangle$	$\min_{\mathbf{p} \in \mathcal{K}_t} \|\mathbf{H}\mathbf{p} + \mathbf{g}\|^2$
Problem class	**Strongly Convex**	**Invex**
Metric / Rate	$\|\mathbf{g}\|$: **R-linear** $f(\mathbf{x}) - f^\star$: **Q-Linear**	$\|\mathbf{g}\|$: **Q-linear** $f(\mathbf{x}) - f^\star$: **R-Linear (GPL)**

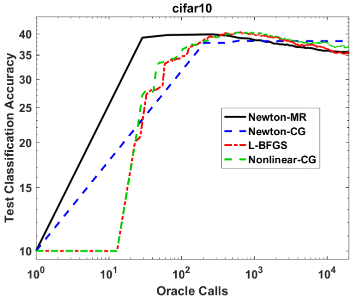
MINRES vs. CG

	$\mathbf{Ax} = \mathbf{b}$	
	CG	**MINRES**
Sub-problems	$\min_{\mathbf{x}\in\mathcal{K}_t} \frac{1}{2}\langle \mathbf{x}, \mathbf{Ax}\rangle + \langle \mathbf{x}, \mathbf{b}\rangle$	$\min_{\mathbf{x}\in\mathcal{K}_t} \Vert\mathbf{Ax} - \mathbf{b}\Vert^2$
Problem class	**Symmetric Positive Definite**	**Symmetric**
Metric / Rate	$\Vert A\mathbf{x} - \mathbf{b}\Vert$: **R-linear** $\Vert\mathbf{x} - \mathbf{x}^\star\Vert_{\mathbf{A}}$: **Q-Linear**	$\Vert A\mathbf{x} - \mathbf{b}\Vert$: **Q-linear** $\Vert\mathbf{x} - \mathbf{x}^\star\Vert_{\mathbf{A}}$: **R-Linear (SPD)**

Weakly-Convex ($n = 50,000, d = 27,648$): Softmax-Cross Entropy

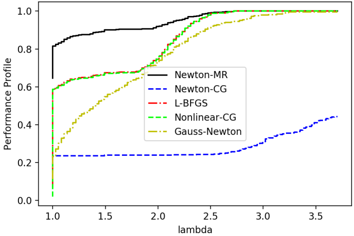


(a) $f(\mathbf{x}_k)$

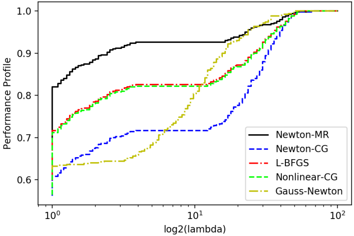


(b) Test Accuracy

Non-Convex: GMM

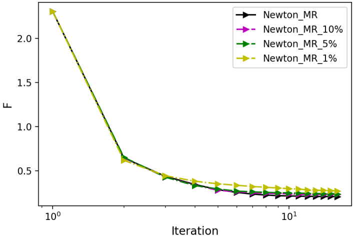


(c) $f(\mathbf{x}_k)$

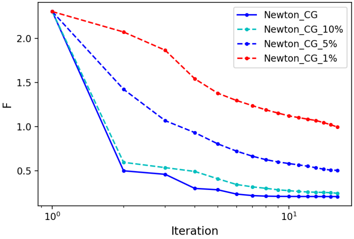


(d) Estimation error

Weakly-Convex ($n = 50,000, d = 7,056$): Softmax-Cross Entropy

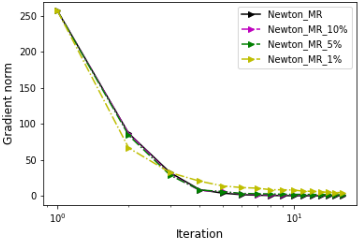


(e) $f(\mathbf{x}_k)$

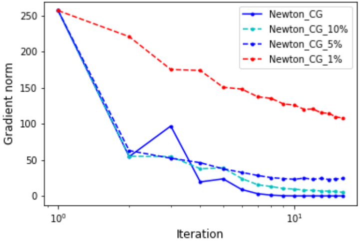


(f) $f(\mathbf{x}_k)$

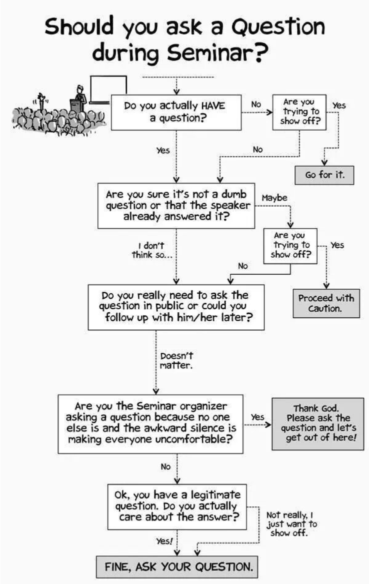
Weakly-Convex ($n = 50,000, d = 7,056$): Softmax-Cross Entropy



(g) $\|\nabla f(\mathbf{x}_k)\|$



(h) $\|\nabla f(\mathbf{x}_k)\|$


Should you ask a Question during Seminar?
Do you actually HAVE a question?
No
Are you trying to show off?
Yes
Go for it.
No
Yes
Are you sure it's not a dumb question or that the speaker already answered it?
Maybe
Are you trying to show off?
Yes
Proceed with Caution.
No
I don't think so...
Do you really need to ask the question in public or could you follow up with him/her later?
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!