Sketched Newton-Raphson

Rui Yuan[†§], Alessandro Lazaric[†], Robert M. Gower[§]
[†]Facebook AI Research and [§]LTCI, Télécom Paris

Seminar of 2nd Year PhD of EDMH
April 15 2021

FACEBOOK AI

Table of contents

Introduction

Context

- Solving nonlinear equations

$$F(x) = 0$$

with $F : \mathbb{R}^d \to \mathbb{R}^n$

Context

- Solving nonlinear equations

$$F(x) = 0$$

  with $F : \mathbb{R}^d \to \mathbb{R}^n$
- Applications: phase retrieval problems, matrix completion problems, PDE, ...

Context

- Solving nonlinear equations
$$F(x) = 0$$
with $F : \mathbb{R}^d \to \mathbb{R}^n$
- Applications: phase retrieval problems, matrix completion problems, PDE, ...
- *Main interest*: Solving finite-sum minimization problems in machine learning

Context

- Solving nonlinear equations

$$F(x) = 0$$

  with $F : \mathbb{R}^d \to \mathbb{R}^n$
- Applications: phase retrieval problems, matrix completion problems, PDE, ...
- *Main interest*: Solving finite-sum minimization problems in machine learning
- Newton-Raphson (NR) method

$$x^{k+1} = x^k - \gamma \left(DF(x^k)^\top \right)^\dagger F(x^k)$$

Context

- Solving nonlinear equations
$$F(x) = 0$$
with $F : \mathbb{R}^d \to \mathbb{R}^n$
- Applications: phase retrieval problems, matrix completion problems, PDE, ...
- *Main interest*: Solving finite-sum minimization problems in machine learning
- Newton-Raphson (NR) method

$$x^{k+1} = x^k - \gamma \left(DF(x^k)^\top\right)^\dagger F(x^k)$$

$DF(x) = [\nabla F_1(x) \cdots \nabla F_n(x)] \in \mathbb{R}^{d\times n}$: Jacobian matrix of F

$\left(DF(x^k)^\top\right)^\dagger$: Moore-Penrose pseudoinverse of $DF(x^k)^\top$

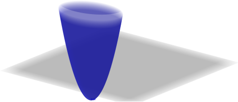
Newton-Raphson methods

$$x^{k+1} \quad = \quad x^k - \gamma \left(DF(x^k)^\top \right)^\dagger F(x^k)$$

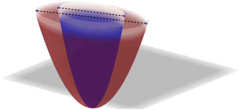
Newton-Raphson methods

$$x^{k+1} \;=\; x^k - \gamma \left(DF(x^k)^\top \right)^\dagger F(x^k)$$

- *Pros:* Scale invariant



Function F

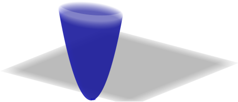


Function $C \times F$ with $C > 0$

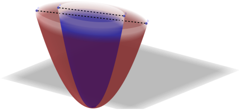
Newton-Raphson methods

$$x^{k+1} \quad = \quad x^k - \gamma \left(DF(x^k)^\top \right)^\dagger F(x^k)$$

- *Pros:* Scale invariant



Function F



Function $C \times F$ with $C > 0$

- *Cons:* Cost per iteration is $\mathcal{O}\left(\min\left(nd^2, dn^2\right)\right)$ which is prohibitive when both n and d are large

Solving Large Nonlinear Equations with Sketched Newton-Raphson

Sketch-and-project

[Gower and Richtárik, 2015]

(1)

Sketch-and-project

[Gower and Richtárik, 2015]

- Newton-Raphson (NR) method

(1)

Sketch-and-project

[Gower and Richtárik, 2015]

- Newton-Raphson (`NR`) method

$$x^{k+1} \quad = \quad x^k - \gamma \left(DF(x^k)^\top \right)^\dagger F(x^k) \tag{1}$$

Sketch-and-project

[Gower and Richtárik, 2015]

- Newton-Raphson (`NR`) method

$$
\begin{aligned}
x^{k+1} &= x^k - \gamma \left(DF(x^k)^\top \right)^\dagger F(x^k) \\
&= \arg\min_{x \in \mathbb{R}^d} \|x - x^k\|_2^2 \\
&\quad \text{subject to} \quad DF(x^k)^\top (x - x^k) = -\gamma F(x^k).
\end{aligned} \tag{1}
$$

Sketch-and-project

[Gower and Richtárik, 2015]

- Newton-Raphson (`NR`) method

$$
\begin{aligned}
x^{k+1} &= x^k - \gamma \left(DF(x^k)^\top \right)^\dagger F(x^k) \\
&= \arg\min_{x \in \mathbb{R}^d} \|x - x^k\|_2^2 \\
&\quad \text{subject to} \quad DF(x^k)^\top (x - x^k) = -\gamma F(x^k).
\end{aligned} \tag{1}
$$

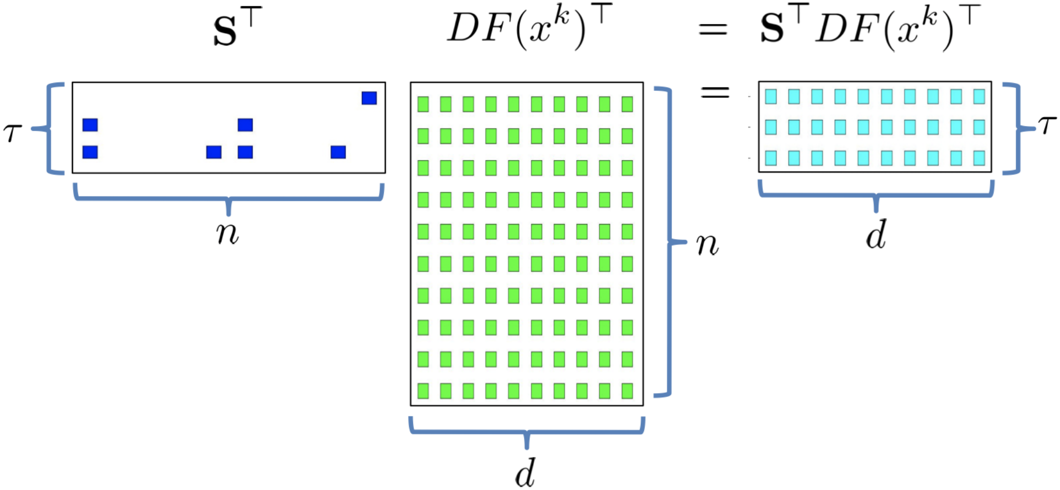
- *Sketched* Newton-Raphson (`SNR`) method

$$
\begin{aligned}
x^{k+1} &= \arg\min_{x \in \mathbb{R}^d} \|x - x^k\|_2^2 \\
&\quad \text{subject to} \quad \mathbf{S}_k^\top DF(x^k)^\top (x - x^k) = -\gamma \mathbf{S}_k^\top F(x^k)
\end{aligned} \tag{2}
$$

$\mathbf{S}_k \sim \mathcal{D}$: sketching matrix of size $n \times \tau$ with $\tau \ll n$, low rank

Decrease dimension using sketching

The sketching matrix
$\mathbf{S} \sim \mathcal{D}$ a distribution over matrices $\mathbf{S} \in \mathbb{R}^{n \times \tau}$ and $\tau \ll n$



Simple examples of sketches

- Sample

$$\mathbf{S} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} = e_j, \qquad \mathbf{S}^\top DF(x)^\top = \nabla F_j(x)^\top$$

Simple examples of sketches

- Sample

$$\mathbf{S} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} = e_j, \qquad \mathbf{S}^\top DF(x)^\top = \nabla F_j(x)^\top$$

- Average sample

$$\mathbf{S} = \begin{bmatrix} a_1 \\ 0 \\ a_3 \\ a_4 \end{bmatrix} = \sum_{i \in C} a_i e_i, \qquad \mathbf{S}^\top DF(x)^\top = \sum_{i \in C} a_i \nabla F_i(x)^\top$$

Simple examples of sketches

- Sample

$$\mathbf{S} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} = e_j, \qquad \mathbf{S}^\top DF(x)^\top = \nabla F_j(x)^\top$$

- Average sample

$$\mathbf{S} = \begin{bmatrix} a_1 \\ 0 \\ a_3 \\ a_4 \end{bmatrix} = \sum_{i \in C} a_i e_i, \qquad \mathbf{S}^\top DF(x)^\top = \sum_{i \in C} a_i \nabla F_i(x)^\top$$

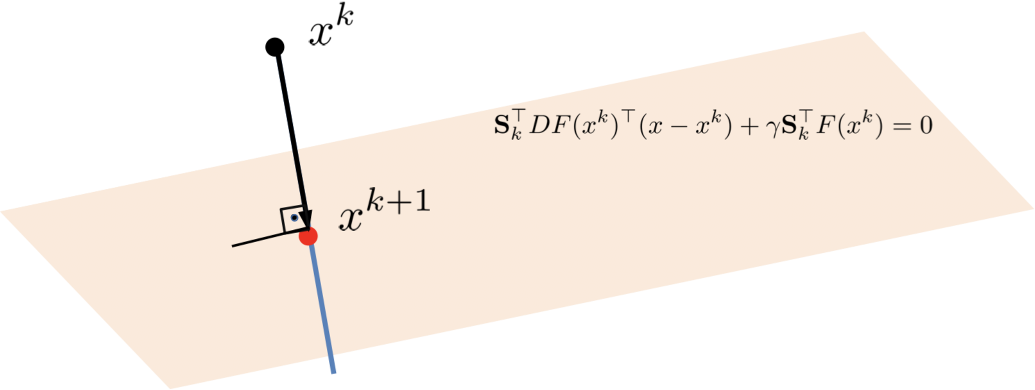
- Batch sample

$$\mathbf{S} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = [e_i \; e_j \; e_k], \qquad \mathbf{S}^\top DF(x)^\top = \begin{bmatrix} \nabla F_i(x)^\top \\ \nabla F_j(x)^\top \\ \nabla F_k(x)^\top \end{bmatrix} \in \mathbb{R}^{\tau \times d}$$

Sketched Newton-Raphson (SNR)

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^d} \|x - x^k\|_2^2$$

subject to $\mathbf{S}_k{}^\top DF(x^k)^\top (x - x^k) + \gamma \mathbf{S}_k{}^\top F(x^k) = 0$



Sketched Newton-Raphson (SNR)

Explicit update:

(3)

Sketched Newton-Raphson (SNR)

Explicit update:

$$
\begin{aligned}
x^{k+1} \quad = \quad & \underset{x \in \mathbb{R}^d}{\arg\min} \|x - x^k\|_2^2 \\
& \text{subject to} \quad \mathbf{S}_k^\top DF(x^k)^\top (x - x^k) + \gamma \mathbf{S}_k^\top F(x^k) = 0
\end{aligned} \tag{3}
$$

Sketched Newton-Raphson (SNR)

Explicit update:

$$\begin{aligned} x^{k+1} &= \underset{x\in\mathbb{R}^d}{\arg\min} \|x - x^k\|_2^2 \\ &\quad \text{subject to} \quad \mathbf{S}_k^\top DF(x^k)^\top (x - x^k) + \gamma \mathbf{S}_k^\top F(x^k) = 0 \\ &= x^k - \gamma DF(x^k)\mathbf{S}_k(\underbrace{\mathbf{S}_k^\top DF(x^k)^\top DF(x^k)\mathbf{S}_k}_{\in\mathbb{R}^{\tau\times\tau}})^\dagger \mathbf{S}_k^\top F(x^k) \end{aligned} \tag{3}$$

Sketched Newton-Raphson (SNR)

Explicit update:

$$\begin{aligned} x^{k+1} &= \arg\min_{x \in \mathbb{R}^d} \|x - x^k\|_2^2 \\ &\quad \text{subject to} \quad \mathbf{S}_k^\top DF(x^k)^\top (x - x^k) + \gamma \mathbf{S}_k^\top F(x^k) = 0 \\ &= x^k - \gamma DF(x^k)\mathbf{S}_k \underbrace{(\mathbf{S}_k^\top DF(x^k)^\top DF(x^k)\mathbf{S}_k)}_{\in \mathbb{R}^{\tau \times \tau}}{}^\dagger \mathbf{S}_k^\top F(x^k) \qquad (3) \end{aligned}$$

Complexity: Cost per iteration $\mathcal{O}(\tau^3 + \tau^2 d)$

Sketched Newton-Raphson (SNR)

Explicit update:

$$\begin{aligned} x^{k+1} &= \arg\min_{x\in\mathbb{R}^d} \|x - x^k\|_2^2 \\ &\quad \text{subject to} \quad \mathbf{S}_k^\top DF(x^k)^\top (x - x^k) + \gamma \mathbf{S}_k^\top F(x^k) = 0 \\ &= x^k - \gamma DF(x^k)\mathbf{S}_k(\underbrace{\mathbf{S}_k^\top DF(x^k)^\top DF(x^k)\mathbf{S}_k}_{\in\mathbb{R}^{\tau\times\tau}})^\dagger \mathbf{S}_k^\top F(x^k) \end{aligned} \tag{3}$$

Complexity: Cost per iteration $\mathcal{O}(\tau^3 + \tau^2 d)$

Assumptions:

- F is continuously twice differentiable
- F contains at least one solution

Algorithm

Input: $\mathcal{D} =$ distribution of sketching matrix, stepsize $\gamma > 0$
Choose $x^0 \in \mathbb{R}^d$
for $k = 0, 1, \ldots,$ **do**
  Sample a fresh sketching matrix: $\mathbf{S}_k \sim \mathcal{D}_{x^k}$
  $x^{k+1} = x^k - \gamma_k DF(x^k)\mathbf{S}_k \left(\mathbf{S}_k^\top DF(x^k)^\top DF(x^k)\mathbf{S}_k\right)^\dagger \mathbf{S}_k^\top F(x^k)$
end
Output: Last iterate x^k

Convergence Theories of Sketched Newton-Raphson

Sketched Newton-Raphson as SGD

$$F(x) \;=\; 0 \qquad \Longleftrightarrow \qquad \min_{x \in \mathbb{R}^d} \frac{1}{2} \|F(x)\|^2_{\mathbb{E}[\mathbf{H}_{\mathbf{S}}(x^k)]}$$

Sketched Newton-Raphson as SGD

$$F(x) = 0 \qquad \Longleftrightarrow \qquad \min_{x \in \mathbb{R}^d} \frac{1}{2} \|F(x)\|^2_{\mathbb{E}[\mathbf{H}_{\mathbf{S}}(x^k)]}$$

where

$$\mathbf{H}_{\mathbf{S}}(x) \overset{\text{def}}{=} \mathbf{S}\left(\mathbf{S}^\top DF(x)^\top DF(x)\mathbf{S}\right)^\dagger \mathbf{S}^\top$$

Sketched Newton-Raphson as SGD

With small technical assumption

Assumption

$F(\mathbb{R}^d) \;\cap\; \mathbf{Ker}\left(\mathbb{E}\left[\mathbf{H}_{\mathbf{S}}(x)\right]\right) \;=\; \{0\}, \quad \forall x \in \mathbb{R}^d.$

$$F(x) \;=\; 0 \qquad \Longleftrightarrow \qquad \min_{x\in\mathbb{R}^d} \frac{1}{2}\|F(x)\|^2_{\mathbb{E}[\mathbf{H}_{\mathbf{S}}(x^k)]}$$

where

$$\mathbf{H}_{\mathbf{S}}(x) \;\overset{\text{def}}{=}\; \mathbf{S}\left(\mathbf{S}^\top DF(x)^\top DF(x)\mathbf{S}\right)^\dagger \mathbf{S}^\top$$

Sketched Newton-Raphson as SGD

With small technical assumption

Assumption

$F(\mathbb{R}^d) \;\cap\; \mathbf{Ker}\left(\mathbb{E}\left[\mathbf{H}_{\mathbf{S}}(x)\right]\right) \;=\; \{0\}, \quad \forall x \in \mathbb{R}^d.$

$$F(x) \;=\; 0 \qquad \Longleftrightarrow \qquad \min_{x\in\mathbb{R}^d} \frac{1}{2}\|F(x)\|^2_{\mathbb{E}[\mathbf{H}_{\mathbf{S}}(x^k)]}$$

where

$$\mathbf{H}_{\mathbf{S}}(x) \;\overset{\text{def}}{=}\; \mathbf{S}\left(\mathbf{S}^\top DF(x)^\top DF(x)\mathbf{S}\right)^{\dagger}\mathbf{S}^\top$$

If we define

$$f_{\mathbf{S},k}(x) \;\overset{\text{def}}{=}\; \frac{1}{2}\|F(x)\|^2_{\mathbf{H}_{\mathbf{S}}(x^k)} \qquad \text{and} \qquad f_k(x) \;\overset{\text{def}}{=}\; \mathbb{E}\left[f_{\mathbf{S},k}(x)\right],$$

Sketched Newton-Raphson as SGD

With small technical assumption

Assumption

$$F(\mathbb{R}^d) \;\cap\; \mathbf{Ker}\left(\mathbb{E}\left[\mathbf{H_S}(x)\right]\right) \;=\; \{0\}, \quad \forall x \in \mathbb{R}^d.$$

$$F(x) \;=\; 0 \qquad \Longleftrightarrow \qquad \min_{x\in\mathbb{R}^d} \frac{1}{2}\|F(x)\|^2_{\mathbb{E}[\mathbf{H_S}(x^k)]}$$

where

$$\mathbf{H_S}(x) \;\overset{\text{def}}{=}\; \mathbf{S}\left(\mathbf{S}^\top DF(x)^\top DF(x)\mathbf{S}\right)^\dagger \mathbf{S}^\top$$

If we define

$$f_{\mathbf{S},k}(x) \;\overset{\text{def}}{=}\; \frac{1}{2}\|F(x)\|^2_{\mathbf{H_S}(x^k)} \qquad \text{and} \qquad f_k(x) \;\overset{\text{def}}{=}\; \mathbb{E}\left[f_{\mathbf{S},k}(x)\right],$$

then it is equivalent to solving $\min_{x\in\mathbb{R}^d} f_k(x)$.

Sketched Newton-Raphson as online SGD

Solve

$$\min_{x \in \mathbb{R}^d} \; f_k(x) \; = \; \mathbb{E}\left[f_{\mathbf{S},k}(x)\right]$$

Sketched Newton-Raphson as online SGD

Solve

$$\min_{x \in \mathbb{R}^d} f_k(x) = \mathbb{E}\left[f_{\mathbf{S},k}(x)\right]$$

At kth iteration

Sketched Newton-Raphson as online SGD

Solve

$$\min_{x \in \mathbb{R}^d} \; f_k(x) \;=\; \mathbb{E}\left[f_{\mathbf{S},k}(x)\right]$$

At kth iteration

$$x^{k+1} \;=\; x^k - \gamma \nabla f_{\mathbf{S},k}(x^k)$$

Sketched Newton-Raphson as online SGD

Solve

$$\min_{x \in \mathbb{R}^d} f_k(x) = \mathbb{E}\left[f_{\mathbf{S},k}(x)\right]$$

At kth iteration

$$\begin{aligned} x^{k+1} &= x^k - \gamma \nabla f_{\mathbf{S},k}(x^k) \\ &= x^k - \gamma DF(x^k)\mathbf{H}_{\mathbf{S}}(x^k)F(x^k) \end{aligned}$$

Sketched Newton-Raphson as online SGD

Solve

$$\min_{x\in\mathbb{R}^d} f_k(x) = \mathbb{E}\left[f_{\mathbf{S},k}(x)\right]$$

At kth iteration

$$\begin{aligned} x^{k+1} &= x^k - \gamma \nabla f_{\mathbf{S},k}(x^k) \\ &= x^k - \gamma DF(x^k)\mathbf{H}_{\mathbf{S}}(x^k)F(x^k) \\ &= x^k - \gamma DF(x^k)\mathbf{S}\left(\mathbf{S}^\top DF(x^k)^\top DF(x^k)\mathbf{S}\right)^\dagger \mathbf{S}^\top F(x^k) \end{aligned}$$

Sketched Newton-Raphson as online SGD

Solve

$$\min_{x\in\mathbb{R}^d} \; f_k(x) \;=\; \mathbb{E}\left[f_{\mathbf{S},k}(x)\right]$$

At kth iteration

$$\begin{aligned} x^{k+1} &= x^k - \gamma \nabla f_{\mathbf{S},k}(x^k) \\ &= x^k - \gamma DF(x^k)\mathbf{H}_{\mathbf{S}}(x^k)F(x^k) \\ &= x^k - \gamma DF(x^k)\mathbf{S}\left(\mathbf{S}^\top DF(x^k)^\top DF(x^k)\mathbf{S}\right)^{\dagger}\mathbf{S}^\top F(x^k) \end{aligned}$$

Satisfy *strong growth condition* and *zero noise stochastic gradient* for free!

Fits need one assumption

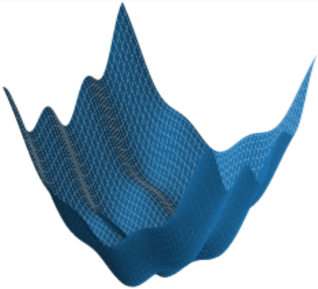
Assumption (Star-convexity)

$$f_k(x^*) \quad \geq \quad f_k(x^k) + \left\langle \nabla f_k(x^k), x^* - x^k \right\rangle$$

Fits need one assumption

Assumption (Star-convexity)

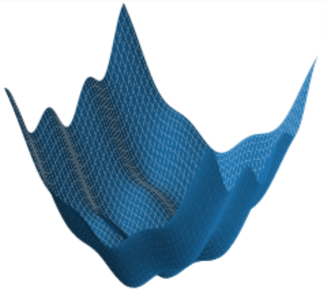
$$f_k(x^*) \quad \geq \quad f_k(x^k) + \langle \nabla f_k(x^k), x^* - x^k \rangle$$



Fits need one assumption

Assumption (Star-convexity)

$$f_k(x^*) \quad \geq \quad f_k(x^k) + \langle \nabla f_k(x^k), x^* - x^k \rangle$$



Class of non-convex functions includes:

- SGD path on DNNs [Zhou et al., 2019]
- Learning systems in control [Hardt et al., 2018]
- Non-convex generalized linear models [Lee and Valiant, 2016]

Online SGD inspired theory

(see paper for technique details and additional properties)

Theorem

Let x^k be the iterates of `SNR`*. Suppose star-convexity*

$$f_k(x^*) \;\geq\; f_k(x^k) + \left\langle \nabla f_k(x^k), x^* - x^k \right\rangle$$

and the technical assumption hold, then

$$\mathbb{E}\left[\min_{t=0,\ldots,k-1} f_t(x^t)\right] \quad\leq\quad \frac{1}{k}\sum_{t=0}^{k-1}\mathbb{E}[f_t(x^t)] \quad\leq\quad \frac{1}{k}\frac{\|x^0 - x^*\|^2}{2\gamma\,(1-\gamma)}.$$

Online SGD inspired theory

(see paper for technique details and additional properties)

Theorem

Let x^k be the iterates of `SNR`. *Suppose star-convexity*

$$f_k(x^*) \;\geq\; f_k(x^k) + \left\langle \nabla f_k(x^k), x^* - x^k \right\rangle$$

and the technical assumption hold, then

$$\mathbb{E}\left[\min_{t=0,\ldots,k-1} f_t(x^t)\right] \quad\leq\quad \frac{1}{k}\sum_{t=0}^{k-1}\mathbb{E}[f_t(x^t)] \quad\leq\quad \frac{1}{k}\frac{\|x^0 - x^*\|^2}{2\gamma\left(1-\gamma\right)}.$$

Direct consequence:

New global convergence theory for the original Newton-Raphson method under strictly weaker assumptions

Applications of Sketched Newton-Raphson

Applications in machine learning

(see paper for additional applications)

- Stochastic Newton method [Kovalev et al., 2019] (First global convergence theory)
- New method for solving generalized linear models (GLM)

Stochastic Newton method ($\texttt{SNM}$)

[Kovalev et al., 2019]

- Solving a finite-sum minimization problem

$$\min_{x\in\mathbb{R}^d}\left[f(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x)\right]$$

Stochastic Newton method (SNM)

[Kovalev et al., 2019]

- Solving a finite-sum minimization problem

$$\min_{x\in\mathbb{R}^d}\left[f(x)\overset{\text{def}}{=}\frac{1}{n}\sum_{i=1}^{n}f_i(x)\right]$$

- Finding a stationary point of the gradient of f : $\nabla f(x)=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\nabla f_i(x)=0$

Stochastic Newton method (SNM)

[Kovalev et al., 2019]

- Solving a finite-sum minimization problem

$$\min_{x\in\mathbb{R}^d}\left[f(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n f_i(x)\right]$$

- Finding a stationary point of the gradient of f : $\nabla f(x) = \frac{1}{n}\sum_{i=1}^n \nabla f_i(x) = 0$
- Re-write the problem as

$$\frac{1}{n}\sum_{i=1}^n \nabla f_i(w^i) \;=\; 0, \qquad \text{and} \qquad x \;=\; w^i, \quad \text{for } i = 1,\ldots,n \tag{4}$$

Stochastic Newton method (SNM)

[Kovalev et al., 2019]

- Solving a finite-sum minimization problem

$$\min_{x\in\mathbb{R}^d}\left[f(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x)\right]$$

- Finding a stationary point of the gradient of f : $\nabla f(x) = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x) = 0$
- Re-write the problem as

$$\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(w^i) \;=\; 0, \qquad \text{and} \qquad x \;=\; w^i, \qquad \text{for } i = 1,\ldots,n \tag{4}$$

- Sketching matrix : based on subsampling rows of (4) and the Hessian matrices of the f_i functions

Stochastic Newton method (SNM)

[Kovalev et al., 2019]

- Solving a finite-sum minimization problem

$$\min_{x\in\mathbb{R}^d}\left[f(x) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x)\right]$$

- Finding a stationary point of the gradient of f : $\nabla f(x) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\nabla f_i(x) = 0$
- Re-write the problem as

$$\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(w^i) \;=\; 0, \qquad \text{and} \qquad x \;=\; w^i, \quad \text{for } i = 1,\dots,n \tag{4}$$

- Sketching matrix : based on subsampling rows of (4) and the Hessian matrices of the f_i functions
- *SNM is a special case of SNR*

Stochastic Newton method (SNM)

[Kovalev et al., 2019]

- Solving a finite-sum minimization problem

$$\min_{x\in\mathbb{R}^d}\left[f(x) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x)\right]$$

- Finding a stationary point of the gradient of f : $\nabla f(x) = \dfrac{1}{n}\sum_{i=1}^{n}\nabla f_i(x) = 0$
- Re-write the problem as

$$\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(w^i) \;=\; 0, \qquad \text{and} \qquad x \;=\; w^i, \quad \text{for } i = 1,\ldots,n \tag{4}$$

- Sketching matrix : based on subsampling rows of (4) and the Hessian matrices of the f_i functions
- *SNM is a special case of SNR*
- Consequently, establish the first global convergence theory of SNM

Tossing-coin-sketch (TCS) for solving GLMs

Generalized linear model

$$\min_{x\in\mathbb{R}^d} f(x) = \frac{1}{n}\sum_{i=1}^{n}\phi_i(a_i^\top x) + \frac{\lambda}{2}\|x\|^2$$

Tossing-coin-sketch (TCS) for solving GLMs

Generalized linear model

$$\min_{x\in\mathbb{R}^d} f(x) = \frac{1}{n}\sum_{i=1}^{n}\phi_i(a_i^\top x) + \frac{\lambda}{2}\|x\|^2$$

We aim to solve $\nabla f(x) = 0$

$$\nabla f(x) \quad = \quad \frac{1}{n}\sum_{i=1}^{n}\underbrace{\phi_i'(a_i^\top x)}_{-\alpha_i} a_i + \lambda x = 0$$

Tossing-coin-sketch (TCS) for solving GLMs

Generalized linear model

$$\min_{x\in\mathbb{R}^d} f(x) = \frac{1}{n}\sum_{i=1}^{n}\phi_i(a_i^\top x) + \frac{\lambda}{2}\|x\|^2$$

We aim to solve $\nabla f(x) = 0$

$$\nabla f(x) \quad = \quad \frac{1}{n}\sum_{i=1}^{n}\underbrace{\phi_i'(a_i^\top x)}_{-\alpha_i} a_i + \lambda x = 0$$

Fixed point equations

$$x \quad = \quad \frac{1}{\lambda n}A\alpha, \tag{5}$$

$$\alpha_i \quad = \quad -\phi_i'(a_i^\top x), \quad \text{for } i = 1,\ldots,n, \tag{6}$$

with $A = [a_1, \cdots, a_n]$

Experiments for TCS method applied for GLM

(see paper for additional experiments)

Logistic regression for binary classification

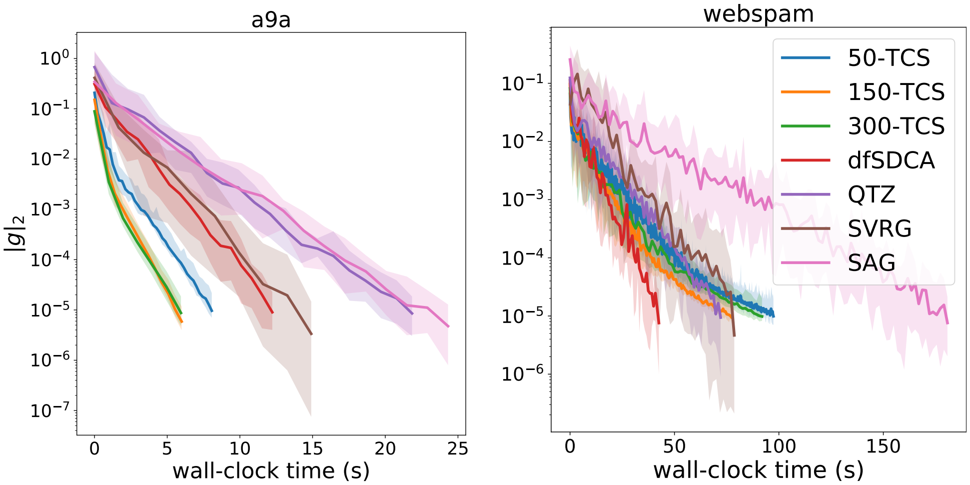


Figure: Experiments for TCS method applied for generalized linear model.

Conclusion

Conclusion

Summary

- Principled development of adaptive scale invariant methods using projected sketched Newton-Raphson
- SGD interpretation gives fast convergence theory (even for non-convex)
- Open the way to designing and analyzing a host of new stochastic second order methods

Future work

- Extend `SNR` by using matrix weighted projection
- Design and analyze more applications of `SNR`
- Develop efficient accelerated `SNR`, `SNR` with momentum or variance reduced `SNR` methods

Details are in our paper:

Sketched Newton-Raphson

https://arxiv.org/abs/2006.12120

Rui Yuan, Alessandro Lazaric, Robert M. Gower

Thank you

FACEBOOK AI

Robert Mansel Gower and Peter Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018.

Dmitry Kovalev, Konstantin Mishchenko, and Peter Richtarik. Stochastic newton and cubic newton methods with simple local linear-quadratic rates. *arxiv:1912.01597*, 2019.

Jasper C. H. Lee and Paul Valiant. Optimizing star-convex functions. In Irit Dinur, editor, *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS*, pages 603–614, 2016.

Yi Zhou, Junjie Yang, Huishuai Zhang, Yingbin Liang, and Vahid Tarokh. SGD converges to global minimum in deep learning via star-convex path. In *International Conference on Learning Representations*, 2019.