Stochastic second-order optimization: global bounds, subspaces, and momentum

Nikita Doikov

EPFL, Machine Learning and Optimization Lab (MLO)

Based on joint works:
with J. Zhao and A. Lucchi — **arXiv:2406.16666**
with E.M. Chayti and M. Jaggi — **arXiv:2410.19644**

Weierstrass Institute, Berlin

December 3, 2024

Outline

I. Introduction: regularization of Newton's method

II. Stochastic subspaces

III. Stochastic oracles and momentum

IV. Conclusions

$$\min_x f(x), \qquad x \in \mathbb{R}^n$$

f is differentiable, can be non-convex

The Gradient Method. Iterate, for $k \geq 0$:

$$\boxed{x_{k+1} := x_k - \alpha_k \nabla f(x_k), \quad \text{for some} \quad \alpha_k > 0}$$

- **+ Cheap iterations:** $\mathcal{O}(n)$
- **+ Easy to analyse**
- **+ Global convergence**
- **− Slow rate**

Let the gradient be Lipschitz: $\|\nabla f(x) - \nabla f(y)\| \leq L_1 \|x - y\|$

We have global upper model of the function:

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \tfrac{L_1}{2}\|x_{k+1} - x_k\|^2 \\ &= f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \tfrac{\alpha_k^2 L_1}{2}\|\nabla f(x_k)\|^2. \end{aligned}$$

Main Proposition. Let $\alpha_k := 1/L_1$. Then,

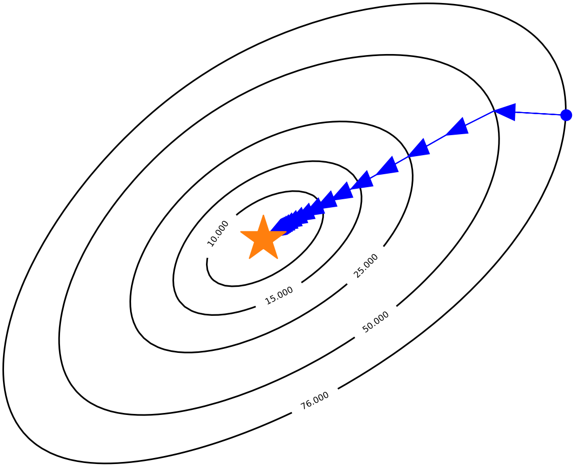
$$f(x_k) - f(x_{k+1}) \geq \tfrac{1}{2L_1}\|\nabla f(x_k)\|^2 \geq \tfrac{1}{2L_1}\varepsilon^2.$$

$\Rightarrow$ telescoping this bound, we obtain the complexity:

$$K = \tfrac{2L_1(f(x_0) - f^\star)}{\varepsilon^2}$$

to find $\|\nabla f(\bar{x}_K)\| \leq \varepsilon$.

$$x_{k+1} \quad := \quad x_k - \alpha_k \nabla f(x_k)$$



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

The Hessian $\nabla^2 f(x) \in \mathbb{R}^{n \times n}$ is the second-order information about the objective,

$$[\nabla^2 f(x)]^{(i,j)} = \frac{\partial^2 f(x)}{\partial x^{(i)} \partial x^{(j)}}, \qquad 1 \leq i, j \leq n$$

A full quadratic model of the objective, $f(y) \approx \Omega_2(x; y)$, where

$$\Omega_2(x; y) \stackrel{\text{def}}{=} f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{1}{2} \langle \nabla^2 f(x)(y - x), y - x \rangle.$$

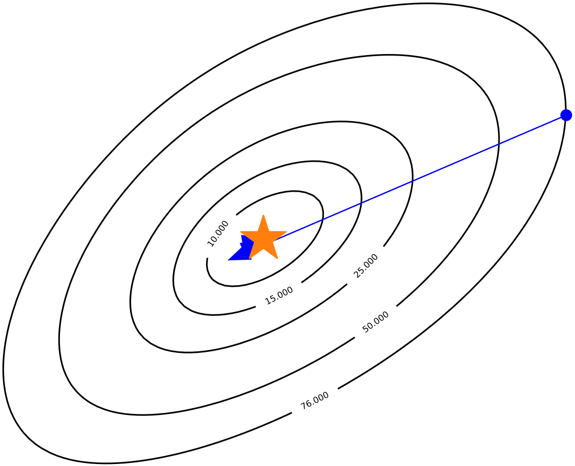
Newton's Method. Iterate, for $k \geq 0$:

$$\boxed{\begin{aligned} x_{k+1} &:= \operatorname*{argmin}_{y \in \mathbb{R}^n} \Omega_2(x_k; y) \\ &= x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k) \end{aligned}}$$

[Newton, 1669; Raphson, 1690; Fine-Bennett, 1916; Kantorovich, 1948]

Newton's Method: Trajectory

$$x_{k+1} \quad := \quad x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k)$$



Newton's Method: Analysis

$$x_{k+1} \quad := \quad x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k)$$

– **Expensive iterations:** $\mathcal{O}(n^3)$

– **More difficult to analyse**

\+ **Fast local convergence**:

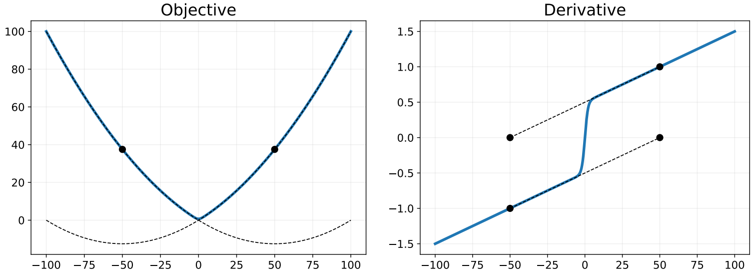
$$\mathcal{O}(\log \log \tfrac{1}{\varepsilon})$$

iterations to find an ε-solution, **when in the neighbourhood of the optimum**

- Global convergence — ?

$$\min_{x \in \mathbb{R}} \left\{ f(x) \;\; := \;\; \log(1 + \exp(x)) - \tfrac{1}{2}x + \tfrac{\mu}{2}x^2 \right\}, \qquad \mu \; := \; 10^{-2}.$$

- The objective is smooth and strongly convex; $x^* = 0$.



The method oscillates between two points!

Let the Hessian be Lipschitz: $\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L_2 \|x - y\|$
$\Rightarrow$ global upper model of the objective, for $H \geq L_2$:

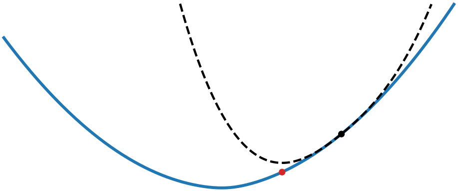
$$f(y) \quad \leq \quad \Omega_2(x; y) + \frac{H}{6}\|y - x\|^3, \qquad \forall x, y \in \mathbb{R}^n$$

where $\Omega_2(x; y) := f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}\langle \nabla^2 f(x)(y - x), y - x \rangle$

Cubic Newton. Iterate, for $k \geq 0$:

$$\boxed{x_{k+1} \quad := \quad \operatorname*{argmin}_{y \in \mathbb{R}^n} \Big[\Omega_2(x_k; y) + \frac{H}{6}\|y - x_k\|^3 \Big]}$$

[Griewank, 1981; Nesterov-Polyak, 2006; Cartis-Gould-Toint, 2011]



H = 0.1

$$x_{k+1} := \operatorname*{argmin}_{y\in\mathbb{R}^n}\Big[\, \Omega_2(x_k; y) + \frac{H}{6}\|y - x_k\|^3 \,\Big]$$

Main Lemma. Let $H := L_2$. Then

$$f(x_k) - f(x_{k+1}) \;\geq\; \frac{1}{12\sqrt{L_2}}\|\nabla f(x_{k+1})\|^{3/2} \;\geq\; \frac{1}{12\sqrt{L_2}}\varepsilon^{3/2}$$

$\Rightarrow$ telescoping this bound, we obtain the complexity:

$$K = \frac{12\sqrt{L_2}(f(x_0) - f^\star)}{\varepsilon^{3/2}}$$

iterations to find $\|\nabla f(\bar{x}_k)\| \leq \varepsilon$.

NB: for the Gradient Method we have $\frac{2L_1(f(x_0)-f^\star)}{\varepsilon^2}$

- Adaptive strategy for H: ensure

$$f(x_{k+1}) \;\leq\; \Omega_2(x_k; x_{k+1}) + \frac{H}{6}\|x_{k+1} - x_k\|^3$$

How to compute one step?

$$h^+ = \operatorname*{argmin}_{h \in \mathbb{R}^n} \Big\{ \langle g, h \rangle + \tfrac{1}{2}\langle Ah, h \rangle + \tfrac{H}{6}\|h\|^3 \Big\}$$

Step 1: compute factorization of $A = A^\top \in \mathbb{R}^{n \times n}$:

$$A = U \Lambda U^\top,$$

where $U \in \mathbb{R}^{n \times n}$ is orthonormal basis: $UU^\top = I$, and Λ is **diagonal** or **tridiagonal** — $\mathcal{O}(n^3)$ arithmetic operations

Step 2: solve

$$P_\star = \min_{h \in \mathbb{R}^n} \Big\{ \langle \bar{g}, h \rangle + \tfrac{1}{2}\langle \Lambda h, h \rangle + \tfrac{H}{6}\|h\|^3 \Big\}$$

using duality:

$$P_\star = D^\star = \max_{\substack{\tau \in \mathbb{R} \text{ s.t.} \\ \tau > [-\lambda_{\min}]_+}} \Big\{ -\tfrac{1}{2}\langle (\Lambda + \tau I)^{-1}\bar{g}, \bar{g} \rangle - \tfrac{2^4}{3H^2}\tau^3 \Big\}$$

concave maximization of univariate function — $\tilde{\mathcal{O}}(n^2)$ operations

- **Cubic Newton step:**

$$\begin{aligned} x^+ &= \operatorname*{argmin}_{y \in \mathbb{R}^n} \Big\{ \Omega_2(x;y) + \tfrac{H}{6}\|y-x\|^3 \Big\} \\ &= x - \Big(\nabla^2 f(x) + \beta I\Big)^{-1} \nabla f(x), \end{aligned}$$

where β is the solution of the dual. We have $\beta = \frac{H}{2}\|x^+ - x\|$.

- Let f be convex. Then,

$$r \stackrel{\text{def}}{=} \|x^+ - x\| = \|\big(\nabla^2 f(x) + \tfrac{Hr}{2} I\big)^{-1} \nabla f(x)\| \leq \tfrac{2}{Hr}\|\nabla f(x)\|$$

Hence, we have an upper bound: $\boxed{\beta = \frac{Hr}{2} \leq \sqrt{\frac{H\|\nabla f(x)\|}{2}}}$.

Gradient Regularization. [Ueda-Yamashita, 2014; Mishchenko, 2021; D-Nesterov, 2021]:

$$x^+ = x - \Big(\nabla^2 f(x) + \sqrt{\tfrac{H\|\nabla f(x)\|}{2}} I\Big)^{-1} \nabla f(x)$$

- One matrix inversion; fast global rates

Classic Newton's step:

$$x_{k+1} \;=\; x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$$

Three major issues:

- **No global convergence** $\Rightarrow$ **Cubic Regularization:**

$$x_{k+1} \;=\; x_k - \left[\nabla^2 f(x_k) + \beta_k I\right]^{-1} \nabla f(x_k)$$

where β_k is computed at each step by univariate maximization.

For convex functions we can use **Gradient Regularization:** $\beta_k = \sqrt{\frac{H\|\nabla f(x_k)\|}{2}}$.

- **High arithmetic cost $\mathcal{O}(n^3)$** $\Rightarrow$ stochastic subspaces $\mathcal{O}(\tau^3)$
- **Requires exact $\nabla f(x), \nabla^2 f(x)$** $\Rightarrow$ stochastic oracles

Outline

I. Introduction: regularization of Newton's method

II. Stochastic subspaces

III. Stochastic oracles and momentum

IV. Conclusions

Coordinate Subspace Model

- Fix subset of coordinates: $S \subset \{1, \ldots, n\}$
- For any $y \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$, denote by

$$y_{[S]} \quad \in \quad \mathbb{R}^n, \qquad A_{[S]} \quad \in \quad \mathbb{R}^{n \times n}$$

  the vector/matrix with zeroed $i \notin S$

Cubic subspace second-order model. For any $h \in \mathbb{R}^n$:

$$\begin{aligned} m_{x,S}(h) \quad &\stackrel{\text{def}}{=} \quad f(x) + \langle \nabla f(x), h_{[S]} \rangle + \tfrac{1}{2} \langle \nabla^2 f(x) h_{[S]}, h_{[S]} \rangle + \tfrac{H}{6} \| h_{[S]} \|^3 \\ &= \quad f(x) + \langle \nabla f(x)_{[S]}, h \rangle + \tfrac{1}{2} \langle \nabla^2 f(x)_{[S]} h, h \rangle + \tfrac{H}{6} \| h_{[S]} \|^3 \end{aligned}$$

- By smoothness, for a sufficiently large $H \geq L_2$, we have:

$$f(x+h) \quad \leq \quad m_{x,S}(h), \qquad \forall x, h \in \mathbb{R}^n$$

$\Rightarrow$ at iteration $k \geq 0$, we can compute a new point as:

$$\boxed{\begin{aligned} x_{k+1} \quad &= \quad x_k + \operatorname*{argmin}_h m_{x_k,S}(h) \\ &= \quad x_k - \Big(\nabla^2 f(x_k)_{[S]} + \beta_k I \Big)^{-1} \nabla f(x_k)_{[S]} \end{aligned}}$$

Stochastic Subspace Cubic Newton

Init: $x_0 \in \mathbb{R}^n$ and subspace size $1 \leq \tau \leq n$

Iteration, $k \geq 0$:

1. Sample $S_k \subset \{1, \dots, n\}$ of size $|S_k| = \tau$
2. Estimate regularization parameter H_k
3. Compute Subspace Cubic Step:

$$x_{k+1} = x_k + \mathrm{argmin}_h \, m_{x_k, S_k}(h)$$

$$= x_k + \underset{h}{\mathrm{argmin}} \Big\{ \langle \nabla f(x_k)_{[S_k]} h, h \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)_{[S_k]} h, h \rangle + \frac{H_k}{6} \|h\|^3 \Big\}$$

- The cost of solving the subproblem is $\mathcal{O}(\tau^3)$
- Very efficient for small $\tau \ll n$

[D-Richtárik, 2018; Cartis-Scheinberg, 2018;
Hanzely-D-Richtárik-Nesterov, 2020; Zhao-Lucchi-D, 2024]

Lemma. Let $H := L_2$. Then

$$f(x_k) - f(x_{k+1}) \quad \geq \quad \tfrac{L_2}{12}\|x_{k+1} - x_k\|^3$$

- Progress of every step
- **NB:** for the full Cubic Newton ($\tau = n$), we have

$$\tfrac{L_2}{12}\|x_{k+1} - x_k\|^3 \quad \geq \quad \tfrac{1}{12\sqrt{L_2}}\|\nabla f(x_{k+1})\|^{3/2}$$

- Difficult to analyse for stochastic step ($\tau < n$)

Lemma. For any $x \in \mathbb{R}^n$ and $|S| = \tau$, we have

$$\mathbb{E}\|\nabla f(x)_{[S]} - \nabla f(x)\| \quad \leq \quad \sqrt{1 - \tfrac{\tau}{n}}\|\nabla f(x)\|$$

$$\mathbb{E}\|\nabla^2 f(x)_{[S]} - \nabla^2 f(x)\| \quad \leq \quad \sqrt{1 - \tfrac{\tau(\tau-1)}{n(n-1)}}\|\nabla^2 f(x)\|_F$$

- The error $\to 0$ with $\tau \to n$

Complexity Result

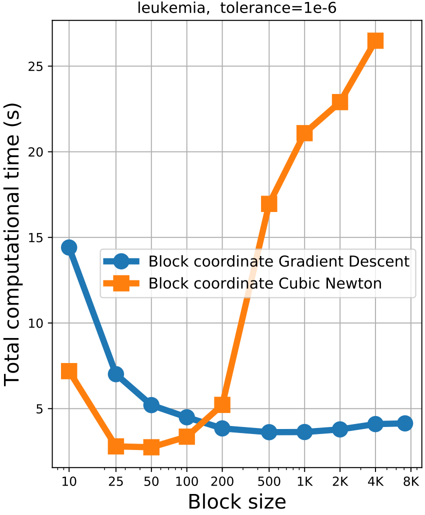
- Let $1 \le \tau \le n$ be fixed

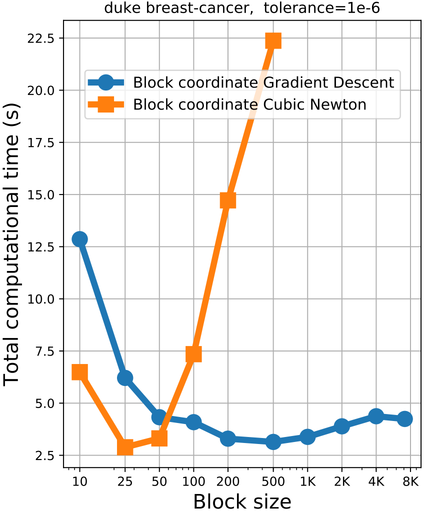
Theorem. To reach $\mathbb{E}\big[\|\nabla f(x_k)\|\big] \le \varepsilon$ it is enough to do

$$k \;=\; \mathcal{O}\Big(\big[\tfrac{n}{\tau}\big]^{3/2} \frac{\sqrt{L_2}(f(x_0)-f^\star)}{\varepsilon^{3/2}} + n^{1/2}\big(1 - \tfrac{\tau(\tau-1)}{n(n-1)}\big)^{1/2} \big[\tfrac{n}{\tau}\big]^{2} \frac{L_1(f(x_0)-f^\star)}{\varepsilon^{2}} \Big)$$

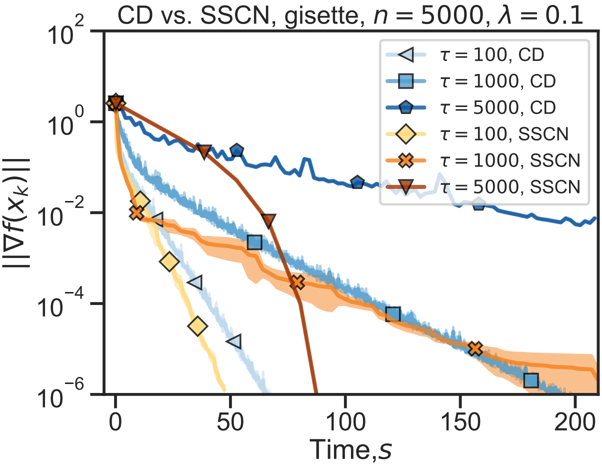
- $\tau = n$: Full Cubic Newton
- $\tau = 1$: Coordinate Descent
- Arithmetic complexity of each iteration is $\mathcal{O}(\tau^3)$

Experiment: Total Computational Time





- **In practice:** use block size of a moderate size



- **the best:** Stochastic Subspace Cubic Newton with $\tau = 100$

Outline

I. Introduction: regularization of Newton's method

II. Stochastic subspaces

III. Stochastic oracles and momentum

IV. Experiments and conclusions

Applications in Machine Learning and Statistics:

no access to exact $\nabla f(x)$ and $\nabla^2 f(x)$

- Let for every x we have an access to stochastic $\nabla f_\xi(x)$ and $\nabla^2 f_\xi(x)$, where ξ is a random variable

Assume:

- unbiased estimates:

$$\mathbb{E}\left[\nabla f_\xi(x)\right] = \nabla f(x), \qquad \mathbb{E}\left[\nabla^2 f_\xi(x)\right] = \nabla^2 f(x),$$

- bounded variance:

$$\mathbb{E}\left[\|\nabla f_\xi(x) - \nabla f(x)\|^2\right] \leq \sigma_g^2,$$

$$\mathbb{E}\left[\|\nabla^2 f_\xi(x) - \nabla^2 f(x)\|^2\right] \leq \sigma_h^2$$

- a.s bound (technical):

$$\|\nabla^2 f_\xi(x) - \nabla^2 f(x)\| \leq \delta_h$$

- **First attempt:** substitute stochastic oracles $\nabla f_\xi(x), \nabla^2 f_\xi(x)$ instead of $\nabla f(x), \nabla^2 f(x)$ in the cubic model

Stochastic Cubic Newton. Iterate $k \geq 0$:

1. Sample $\xi_k \sim \mathcal{D}$
2. Set $g_k = \nabla f_{\xi_k}(x_k)$, $A_k = \nabla^2 f_{\xi_k}(x_k)$
3. Form stochastic cubic model:

$$m_k(y) \stackrel{\text{def}}{=} \langle g_k, y - x_k \rangle + \frac{1}{2}\langle A_k(y - x_k), y - x_k \rangle + \frac{H}{6}\|y - x_k\|^3$$

3. Compute step: $x_{k+1} = \operatorname*{argmin}_y m_k(y)$

Lemma. Let $H \geq L_2$. Then, for some numerical constant $c > 0$, we have

$$c \cdot \big[f(x_k) - f(x_{k+1}) \big]$$

$$\geq \quad \frac{1}{\sqrt{H}} \|\nabla f(x_{k+1})\|^{3/2} - \frac{\|g_k - \nabla f(x_k)\|^{3/2}}{\sqrt{H}} - \frac{\|A_k - \nabla^2 f(x_k)\|^3}{H^2}$$

- No progress at every step due to approximation errors
- Need to bound different moments

$\Rightarrow$ telescoping this inequality, we prove the convergence rate.

Theorem. For every $k \geq 1$, we have

$$\mathbb{E}\big[\|\nabla f(\bar{x}_k)\|^{3/2}\big] \;\; \leq \;\; \mathcal{O}\Big(\frac{\sqrt{H}(f(x_0) - f^\star)}{k} + \frac{\sigma_h^3}{H^{3/2}} + \sigma_g^{3/2} \Big)$$

- A freedom to choose $H \geq L_2 \Rightarrow$ decrease the σ_h^3 term
- No control of $\sigma_g^{3/2} \Rightarrow$ convergence to a ball! Only for $\boxed{\varepsilon \geq \sigma_g^{3/2}}$

Mini-Batching

Let at each iteration $k \geq 0$ we sample b_g gradients and b_h Hessians:

$$g_k := \frac{1}{b_g} \sum_{i \in [b_g]} \nabla f_{\xi_i}(x_k)$$

$$A_k := \frac{1}{b_h} \sum_{i \in [b_h]} \nabla^2 f_{\xi_i}(x_k)$$

Then,

$$\mathbb{E}\|g_k - \nabla f(x_k)\|^{3/2} \leq \frac{\sigma_g^{3/2}}{b_g^{3/4}},$$

and by Matrix Concentration, e.g. [Chen-Gittens-Tropp, 2012],

$$\mathbb{E}\|A_k - \nabla^2 f(x_k)\|^3 \leq \mathcal{O}\Bigl(\log(n)^{3/2} \frac{\sigma_h^3}{b_h^{3/2}} + \log(n)^3 \frac{\delta_h^3}{b_h^3}\Bigr) \underset{b_h \gg 1}{\approx} \tilde{\mathcal{O}}\Bigl(\frac{\sigma_h^3}{b_h^{3/2}}\Bigr)$$

Thus, we can converge to any accuracy:

$$\mathbb{E}\bigl[\|\nabla f(\bar{x}_k)\|^{3/2}\bigr] \leq \tilde{\mathcal{O}}\Bigl(\frac{\sqrt{H}(f(x_0) - f^\star)}{k} + \frac{\sigma_h^3}{H^{3/2} b_h^{3/2}} + \frac{\sigma_g^{3/2}}{b_g^{3/4}}\Bigr)$$

[Kohler-Lucchi, 2017; Chayti-Jaggi-D, 2023]

Momentum

- **Idea:** instead of forming new batch each iteration,

reuse old gradients and Hessians

Momentum in optimization:

- Heavy-Ball Method [Polyak, 1964]

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) + \beta_k (x_k - x_{k-1})$$

- Fast Gradient Method [Nesterov, 1984]

$$y_k = x_k + \beta_k (x_k - x_{k-1})$$
$$x_{k+1} = y_k - \alpha_k \nabla f(y_k)$$

- Deep Learning
- Stochastic Optimization [Cutkosky-Orabona, 2020; Cutkosky-Mehta, 2020; Arnold-Manzagol-Babanezhad-Mitliagkas-Roux, 2019; Gao-Rodomanov-Stich, 2024]

Second-Order Momentum

Let $0 < \alpha, \beta \leq 1$. **Define**, for $k \geq 1$:

$$g_k = (1-\alpha)g_{k-1} + \alpha \nabla f_{\xi_k}\left(x_k + \tfrac{1-\alpha}{\alpha}(x_k - x_{k-1})\right)$$

and

$$A_k = (1-\beta)A_{k-1} + \beta \nabla^2 f_{\xi_k}(x_k)$$

- **NB:** $\alpha = \beta = 1$ implies $g_k = \nabla f_{\xi_k}(x_k)$ and $A_k = \nabla^2 f_{\xi_k}(x_k)$
- By choosing $\alpha, \beta < 1$ we aim to decrease the variance of our estimators

Lemma.

$$\frac{1}{k}\sum_{i=1}^{k} \mathbb{E}\|g_i - \nabla f(x_i)\|^{3/2}$$
$$\leq \mathcal{O}\Bigg(\underbrace{\alpha^{3/4}\sigma_g^{3/2}}_{\text{variance}} + \underbrace{\frac{\sigma_g^{3/2}}{\alpha k} + \frac{(1-\alpha)^{3/2}L^{3/2}}{\alpha^3}\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\|x_i - x_{i-1}\|^3}_{\text{bias}}\Bigg)$$

- Similar bounds for the Hessians

Stochastic Cubic Newton with Momentum

Init: $x_0 \in \mathbb{R}^n$, $g_0 = \nabla f_{\xi_0}(x_0)$, $A_0 = \nabla^2 f_{\xi_0}(x_0)$, $0 < \alpha, \beta \leq 1$

Iteration, $k \geq 0$

1. Sample $\xi_k \sim \mathcal{D}$
2. Set

$$\begin{aligned} g_k &= (1-\alpha)g_{k-1} + \alpha \nabla f_{\xi_k}\big(x_k + \tfrac{1-\alpha}{\alpha}(x_k - x_{k-1})\big) \\ A_k &= (1-\beta)A_{k-1} + \beta \nabla^2 f_{\xi_k}(x_k) \end{aligned}$$

3. Form stochastic cubic model:

$$m_k(y) \stackrel{\text{def}}{=} \langle g_k, y - x_k \rangle + \tfrac{1}{2}\langle A_k(y - x_k), y - x_k \rangle + \tfrac{H}{6}\|y - x_k\|^3$$

4. Compute step: $x_{k+1} = \operatorname{argmin}_y m_k(y)$

Theorem. For any $H \geq L_2$, we have

$$\mathbb{E}\big[\|\nabla f(\bar{x}_k)\|^{3/2}\big] \leq \mathcal{O}\Big(\frac{\sqrt{H}(f(x_0) - f^\star)}{k} + \frac{L^{3/2}\sigma_h^3}{H^3} + \frac{L^{3/8}\sigma_g^{3/2}}{H^{3/8}}\Big)$$

- Convergence with arbitrary noise level!

How many stochastic samples to find $\mathbb{E}\big[\|\nabla f(\bar{x})\|\big] \leq \varepsilon$?

- **Stochastic Gradient Descent (SGD)** [Lan, 2020]

$$\mathcal{O}\left(\frac{1}{\varepsilon^2} + \frac{\sigma_g}{\varepsilon^4}\right)$$

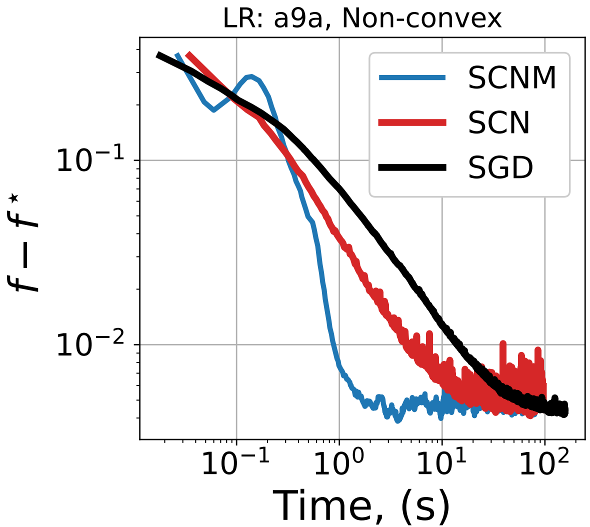
- **Normalized SGD with momentum** [Cutkosky-Mehta, 2020a]

$$\mathcal{O}\left(\frac{1}{\varepsilon^2} + \frac{\sigma_g^2}{\varepsilon^{7/2}}\right)$$

- **Stochastic Cubic Newton with momentum** (ours)

$$\mathcal{O}\left(\frac{1}{\varepsilon^{3/2}} + \frac{\sigma_h^{1/2}}{\varepsilon^{7/4}} + \frac{\sigma_g^2}{\varepsilon^{7/2}}\right)$$

Improvements due to second-order smoothness


LR: a9a, Non-convex
SCNM
SCN
SGD
10^{-1}
10^{-2}
$f - f^\star$
10^{-1}
10^{0}
10^{1}
10^{2}
Time, (s)

Outline

I. Introduction: regularization of Newton's method

II. Stochastic subspaces

III. Stochastic oracles and momentum

IV. Conclusions

Conclusions

- Global convergence of Newton's Method ⇒ regularization
  - **Cubic Regularization**
  - **Gradient Regularization**
- Large dimension n of the problem ⇒ restrict the model to stochastic subspaces of size τ
  - **Preserves the global convergence**
  - **Cheap subproblem if τ is small**
- Stochastic oracles
  - **Momentum reduces the variance of stochastic estimates**

References:

1. Zhao, J., Lucchi, A., and Doikov, N., 2024. Cubic regularized subspace Newton for non-convex optimization. **arXiv:2406.16666**
2. Chayti, E.M., Doikov, N., and Jaggi M., 2024. Improving Stochastic Cubic Newton with Momentum. **arXiv:2410.19644**

Open Questions

- **Better concentration inequalities**; analysis for heavy tails

[Gorbunov-Sadiev-Danilova-Horváth-Gidel-Dvurechensky-Gasnikov-Richtárik, 2024]

- **Lower complexity bounds**; using smoothness breaks the classical lower bound $\Omega\left(\frac{1}{\varepsilon^4}\right) \mapsto \mathcal{O}\left(\frac{1}{\varepsilon^{7/2}}\right)$

- **Accelerated** second-order methods for convex optimization

[Agafonov-Kamzolov-Gasnikov-Kavis-Antonakopoulos-Cevher-Takáč, 2023]

Thank you for your attention!