Cubic regularized subspace Newton for non-convex optimization

Jim Zhao (University of Basel)

Aurelien Lucchi (University of Basel)

Nikita Doikov (EPFL)

AISTATS, Thailand

May 4, 2025

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

f is differentiable, can be non-convex

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

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

Let the gradient be Lipschitz: $\|\nabla f(y) - \nabla f(x)\| \leq L_1 \|y - x\|$.
Set $\boxed{\alpha := 1/L_1}$ Then

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

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

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

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

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 $H \geq L_2$:

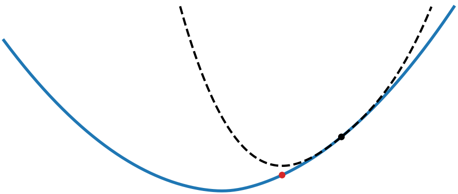
$$f(y) \quad \leq \quad \Omega_2(x; y) + \frac{H}{6}\|y - x\|^3, \qquad \forall x, y \in \mathbb{R}^n$$

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

Cubic Newton: Analysis

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

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

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

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

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

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

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

- Price: more expensive steps. $\mathcal{O}(n^3)$ arithmetic operations to solve the subproblem

- 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 $h \in \mathbb{R}^n$:

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

- 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 compute next step as:

$$\boxed{\begin{aligned} x_{k+1} \quad &= \quad x_k + \operatorname*{argmin}_h m_{x_k,S}(h) \\ &= \quad x_k - \Big(\nabla^2 f(x_k)_{[S]} + \beta_k I \Big)^{-1} \nabla f(x_k)_{[S]} \quad \text{for} \quad \beta_k > 0 \end{aligned}}$$

Stochastic Subspace Cubic Newton

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

Iteration, $k \ge 0$:

1. Sample $S_k \subset \{1, \ldots, n\}$ of size $|S_k| = \boldsymbol{\tau}$
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 $\mathcal{O}(\boldsymbol{\tau}^3)$
- Very efficient for small $\boldsymbol{\tau} \ll n$

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

New Result: Global Convergence

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

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

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

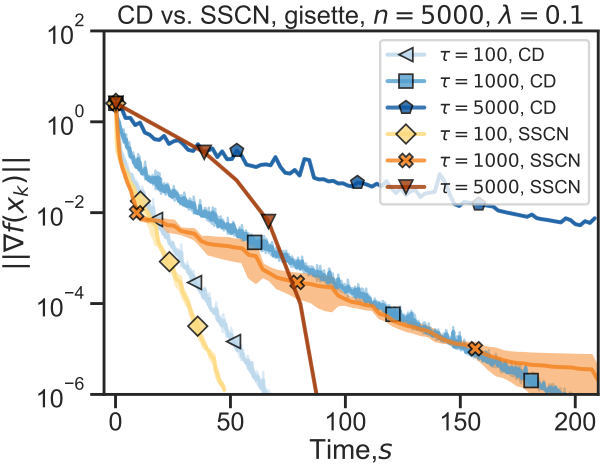
- The error $\to 0$ with $\tau \to n$

Theorem. To reach $\mathbb{E}\big[\|\nabla f(x_K)\|\big] \leq \varepsilon$ it is enough to perform

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

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

Experiment: Logistic Regression



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

- Idea: at iteration $k \geq 0$ sample different number of coordinates $\tau(S_k)$

Theorem. Set

$$\tau(S_k) \quad = \quad n \cdot \max\left\{1 - \frac{\delta^{-2}\|x_k - x_{k-1}\|^2}{\|\nabla f(x_k)\|^2},\ \sqrt{1 - \frac{\delta^{-2}\|x_k - x_{k-1}\|}{\|\nabla^2 f(x_k)\|_F^2}}\right\}$$

Then, with probability at least $1 - \delta$ we have

$$\max\left\{\|\nabla f(\bar{x}_K)\|^{3/2}, [-\lambda_{\min}(\nabla^2 f(x))]^3\right\} \;\leq\; \mathcal{O}\left(\frac{1}{K}\right)$$

- Convergence to a second-order stationary point
- It can be $\tau(S_k) \ll n$

Conclusions

- Cubic regularization $\Rightarrow$ global convergence for Newton's method
- Stochastic subspaces significantly reduce the arithmetic cost

$$\mathcal{O}(n^3) \quad \mapsto \quad \mathcal{O}(\boldsymbol{\tau}^3)$$

- We show that Stochastic Subspace Cubic Newton possesses
  - fast global rates for non-convex optimization
  - convergence to a second-order stationary point under adaptive sampling
  - good practical performance

Thank you for your attention!