Second-Order Optimization with Lazy Hessians

Nikita Doikov **El Mahdi Chayti** **Martin Jaggi**

EPFL

Problem and Motivation

We want to solve **unconstrained minimization** problem:

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

- f is differentiable and can be **non-convex**
- First-order gradient methods: **cheap** to implement, but **slow** rates
- Second-order methods (Newton's Method): **fast** rates but **expensive**

This work: we propose to use a previously seen Hessian for several iterations (*lazy Hessian updates*):

- **Provable improvement** of the total arithmetic complexity

Newton's Method with Cubic Regularization

Assume that the **Hessian is Lipschitz Continuous**:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \;\leq\; L\|x-y\|, \qquad \forall x, y \in \mathbb{R}^d$$

$\Rightarrow$ **global upper model** of the objective, for $H \geq L$:

$$f(y) \leq \Omega(x; y) + \tfrac{H}{6}\|y-x\|^3, \qquad \forall x, y \in \mathbb{R}^d,$$

where

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

Cubic Newton Method [1]. Iterate, $k \geq 0$:

$$x_{k+1} = \mathop{\mathrm{argmin}}_{y \in \mathbb{R}^d} \Big\{ M_H(x_k; y) \;\equiv\; \Omega(x_k; y) + \tfrac{H}{6}\|y - x_k\|^3 \Big\}$$

Theorem. Let $H := L$. Then, to find $\|\nabla f(\bar{x}_k)\| \leq \varepsilon$, the Cubic Newton needs

$$K = \mathcal{O}\Big(\tfrac{\sqrt{L}(f(x_0) - f^\star)}{\varepsilon^{3/2}}\Big)$$

- For the Gradient Method, we need $\mathcal{O}(1/\varepsilon^2)$ iterations
- We also can prove convergence to a **second-order stationary point** for the Cubic Newton: $\nabla^2 f(\bar{x}_k) \succeq -\sqrt{L\varepsilon} I$
- **Adaptive strategy** for H: ensure $f(x_{k+1}) \leq M_H(x_k; x_{k+1})$ [1, 2] — the method becomes **universal**, adapting automatically to the most appropriate problem class [3, 4]

Solving the Cubic Subproblem

How to compute one step? $h^+ = \mathop{\mathrm{argmin}}_{h \in \mathbb{R}^d} \Big\{ \langle g, h \rangle + \tfrac{1}{2}\langle Ah, h\rangle + \tfrac{H}{6}\|h\|^3 \Big\}$

- **Step 1**: compute **factorization** of $A = U \Lambda U^\top$, where U is orthonormal basis $UU^\top = I$, and Λ is **diagonal** or **tridiagonal** — $\mathcal{O}(d^3)$ arithmetic operations (the most expensive part)
- **Step 2:** solve the **dual problem** (concave univariate maximization):
$$\max_{\tau \in \mathbb{R} \,:\, \tau > [-\lambda_{\min}]_+} \Big\{ -\tfrac{1}{2}\langle (\Lambda + \tau I)^{-1} \bar{g}, \bar{g} \rangle - \tfrac{2^4}{3H^2}\tau^3 \Big\} \;\Rightarrow\; \boxed{h^+ = -(A + \tau^\star I)^{-1} g}$$

Lazy Hessian Updates

Main Idea: use the same Hessian for $m \geq 1$ iterations

Lazy Hessian Updates: compute new Hessian once per m iterations.

Hessians:	$\nabla^2 f(x_0)$	reuse Hessian $\longrightarrow$			$\nabla^2 f(x_m)$	reuse Hessian $\longrightarrow$	
Gradients:	$\nabla f(x_0)$	$\nabla f(x_1)$	$\dots$	$\nabla f(x_{m-1})$	$\nabla f(x_m)$	$\nabla f(x_{m+1})$	$\dots$

- Appeared first in [5]

Algorithm: Cubic Newton with Lazy Hessians

Define step of the method with Hessian at some previous point z:

$$T_H(x, z) \overset{\text{def}}{=} \mathop{\mathrm{argmin}}_{y \in \mathbb{R}^d} \Big\{ \langle \nabla f(x), y - x\rangle + \tfrac{1}{2}\langle \nabla^2 f(z)(y - x), y - x\rangle + \tfrac{H}{6}\|y - x\|^3 \Big\}$$

Define $\pi(k) \overset{\text{def}}{=} k - k \bmod m$

Cubic Newton with Lazy Hessians

Iterate, $k \geq 0$:

1. Set last snapshot point $z_k = x_{\pi(k)}$
2. Compute lazy cubic step $x_{k+1} = T_H(x_k, z_k)$

Theory: Convergence Rate

Theorem. Let $H := 6mL$. Then, to find $\|\nabla f(\bar{x})\| \leq \varepsilon$, the method needs

$$K = \mathcal{O}\Big(\tfrac{\sqrt{mL}(f(x_0) - f^\star)}{\varepsilon^{3/2}}\Big)$$

- Worse than the full Cubic Newton by the factor $\sqrt{m}$

Note: the total number of **Hessian updates** during these steps is

$$\tfrac{K}{m} = \mathcal{O}\Big(\tfrac{\sqrt{L}(f(x_0) - f^\star)}{\sqrt{m}\varepsilon^{3/2}}\Big)$$

Choice of m?

$\gg$ **Optimize the total cost:**

$$\texttt{Arithmetic complexity} = K \times \texttt{GradCost} + \tfrac{K}{m} \times \texttt{HessCost}$$

In many problems: $\boxed{\texttt{HessCost} = d \times \texttt{GradCost}} \;\Rightarrow\; m := d$

- Generalized Linear Models (**Logistic Regression**)
- Log-sum-exp (**Soft Max**)
- **Neural Networks**: computing $\nabla^2 f(x) h$ is the same cost as $\nabla f(x)$, for any x, h, by using backpropagation. Then
$$\nabla^2 f(x) = \big[\nabla^2 f(x) e_1 \,\big|\, \dots \,\big|\, \nabla^2 f(x) e_d\big]$$

Total Arithmetic Complexity

- Gradient Method:
$$\mathcal{O}\Big(\tfrac{L_1(f(x_0) - f^\star)}{\varepsilon^2}\Big) \times \texttt{GradCost}$$
- Full Cubic Newton:
$$\mathcal{O}\Big(\tfrac{\sqrt{L_2}(f(x_0) - f^\star)}{\varepsilon^{3/2}}\Big) \times \texttt{GradCost} \times d$$
- Lazy Cubic Newton ($m := d$):
$$\mathcal{O}\Big(\tfrac{\sqrt{L_2}(f(x_0) - f^\star)}{\varepsilon^{3/2}}\Big) \times \texttt{GradCost} \times \sqrt{d}$$

This provably improves the total arithmetic complexity of the Cubic Newton by factor $\sqrt{d}$

Locally, we also have superlinear convergence

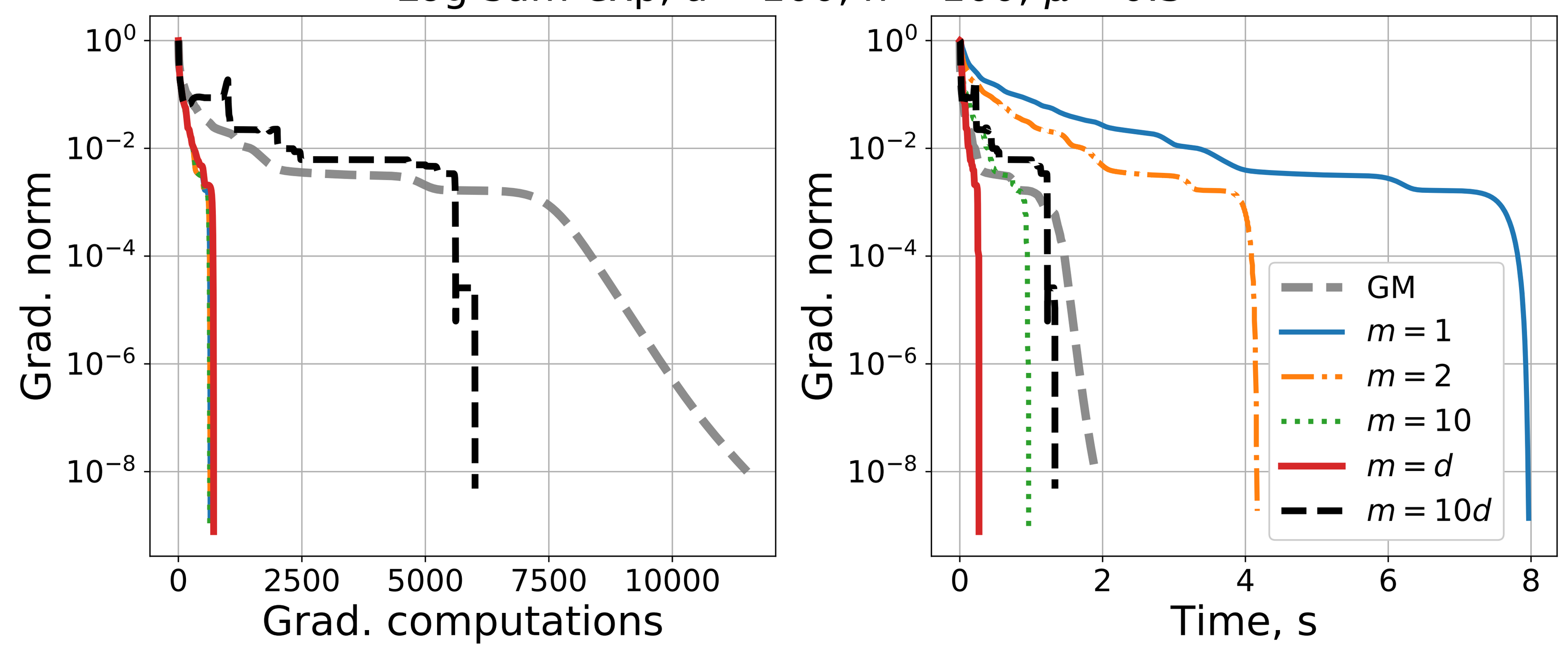
For **convex problems**, we can use the Gradient Regularization technique with lazy Hessian updates, achieving the same global rates:

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

Experiment: Soft Max

$$\min_{x \in \mathbb{R}^d} f(x) \;=\; \mu \ln\Big(\sum_{i=1}^n \exp\big(\tfrac{\langle a_i, x\rangle - b_i}{\mu}\big)\Big) \;\approx\; \max_{1 \leq i \leq n} \big[\langle a_i, x\rangle - b_i\big]$$

Log-sum-exp, $d = 100$, $n = 100$, $\mu = 0.5$



References

[1] Yurii Nesterov and Boris Polyak. "Cubic regularization of Newton's method and its global performance". In: *Mathematical Programming* 108.1 (2006), pp. 177–205

[2] Coralia Cartis, Nicholas IM Gould, and Philippe L Toint. "Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results". In: *Mathematical Programming* 127.2 (2011), pp. 245–295

[3] Geovani N Grapiglia and Yurii Nesterov. "Regularized Newton Methods for Minimizing Functions with Hölder Continuous Hessians". In: *SIAM Journal on Optimization* 27.1 (2017), pp. 478–506

[4] Nikita Doikov and Yurii Nesterov. "Minimizing uniformly convex functions by cubic regularization of Newton method". In: *Journal of Optimization Theory and Applications* (2021), pp. 1–23

[5] VE Shamanskii. "A modification of Newton's method". In: *Ukrainian Mathematical Journal* 19.1 (1967), pp. 118–122