Gradient-Normalized Smoothness for Optimization with Approximate Hessians

Andrei Semenov † Martin Jaggi † Nikita Doikov ‡

†EPFL ‡Cornell University

Second-Order Methods with Inexact Hessians Can Be Provably Fast & Efficient!

Problem: While complexity guarantees for the Newton method are well-established, theory for *inexact Hessians* is limited, often suggesting convergence as slow as Gradient Descent

A Key Innovation: Gradient-Normalized Smoothness

... very basic method, but it's all about the right (second-order) step-size γ_k:

$$\boldsymbol{x}_{k+1} \;\leftarrow\; \boldsymbol{x}_k - \Big(\mathbf{H}(\boldsymbol{x}_k) + \tfrac{\|\nabla f(\boldsymbol{x}_k)\|_*}{\gamma_k}\mathbf{B}\Big)^{-1}\nabla f(\boldsymbol{x}_k), \qquad \gamma_k \approx \gamma(x_k).$$

Goal: $\nabla f(\boldsymbol{x}+\boldsymbol{h}) \approx \nabla f(\boldsymbol{x}) + \mathbf{H}(\boldsymbol{x})\boldsymbol{h}$, $\mathbf{H}(\boldsymbol{x}) \approx \nabla^2 f(\boldsymbol{x}) \succeq 0$ is symmetric positive-definite.

Our Main Definition:

$$\gamma(\boldsymbol{x}) := \max\Big\{\gamma \geq 0 : \big\|\nabla f(\boldsymbol{x}+\boldsymbol{h}) - \nabla f(\boldsymbol{x}) - \mathbf{H}(\boldsymbol{x})\boldsymbol{h}\big\|_* \leq \tfrac{\|\boldsymbol{g}\|_*\|\boldsymbol{h}\|}{\gamma},\ \forall \boldsymbol{h}\ :\ \|\boldsymbol{h}\| \leq \gamma \ \text{ and } \ \|\boldsymbol{h}\|_{\boldsymbol{x}}^2 + \langle \boldsymbol{g}, \boldsymbol{h}\rangle \leq 0\Big\}$$

Theory Automatically Covers Inexact Hessians

$$\big\|\nabla^2 f(\boldsymbol{x}) - \mathbf{H}(\boldsymbol{x})\big\|_* \;\leq\; \mathrm{C}_1 + \mathrm{C}_2\|\nabla f(\boldsymbol{x})\|_*^{1-\beta}, \quad 0 \leq \beta \leq 1 \ (\bigstar)$$

When $\alpha \geq \beta$ and $\mathrm{C}_1 \approx 0$: the **smoothness class dominates** the Hessian inexactness, and the method achieves the **same global rate as exact Newton**. This holds for Fisher, Gauss-Newton, and other approximations in logistic regression, LogSumExp, and nonlinear equations

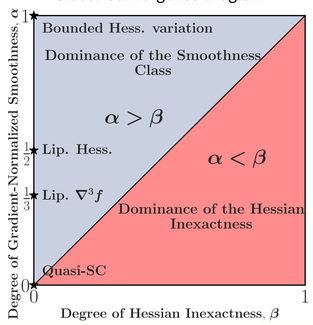
If $\gamma_1(\boldsymbol{x})$ is the Gradient-Normalized Smoothness for the exact Hessian, then for the Hessian approximation we have:

$$\gamma(\boldsymbol{x}) \geq \left(\gamma_1(\boldsymbol{x})^{-1} + \frac{\mathrm{C}_1}{\|\nabla f(\boldsymbol{x})\|_*} + \frac{\mathrm{C}_2}{\|\nabla f(\boldsymbol{x})\|_*^{\beta}}\right)^{-1}$$

Inexact Hessian: Convex Functions

$$K = \widetilde{O}\Big(\mathcal{C}_{\text{NEWTON}}(\varepsilon) + \frac{\mathrm{C}_1 D^2}{\varepsilon} + \frac{\mathrm{C}_2 D^{1+\beta}}{\varepsilon^{\beta}}\Big)$$

Global Convergence Diagram



Takeaways

- **Universal algorithm:** automatically adapts to the local degree of smoothness
- Convergence for both convex and **non-convex** objectives
- **State-of-the-art rates** recovered for both true and inexact Hessians
- **Aligning with practice:** all theoretical conditions are satisfied by popular Hessian approximations
- When $\alpha \geq \beta$ and $\mathrm{C}_1 \approx 0$: the inexact-Hessian method **converges as fast as the exact Newton**

[1] Nesterov: *Lectures on Convex Optimization*, 2018. [2] Nesterov & Polyak: *Cubic regularization of Newton method*, 2006. [3] Doikov, Mishchenko & Nesterov: *Super-universal regularized Newton method* 2024. [4] Doikov, *Minimizing quasi-self-concordant functions by gradient regularization of Newton method*, 2025

Problem Classes We Cover

Problem Class	Exact Case $(\mathrm{C}_1{=}\mathrm{C}_2{=}0)$		Inexact Hess. (ours)
Bounded Hess. variation	$O\Big(\frac{M_0 D^2}{\varepsilon}\Big)$	[1]	$O\Big(\frac{(M_0+\mathrm{C}_2)D^2}{\varepsilon} + \frac{\mathrm{C}_1 D^2}{\varepsilon}\Big)$
Lip. Hess.	$O\Big(\frac{M_{1/2}\, D^{3/2}}{\varepsilon^{1/2}}\Big)$	[2]	$O\Big(\frac{(M_{1/2}+\mathrm{C}_2)D^{3/2}}{\varepsilon^{1/2}} + \frac{\mathrm{C}_1 D^2}{\varepsilon}\Big)$
Lip. $\nabla^3 f$	$O\Big(\frac{M_{1/3}\, D^{4/3}}{\varepsilon^{1/3}}\Big)$	[3]	$O\Big(\frac{(M_{1/3}+\mathrm{C}_2)D^{4/3}}{\varepsilon^{1/3}} + \frac{\mathrm{C}_1 D^2}{\varepsilon}\Big)$
Gen.-SC, $0 < \alpha \leq 1/2$	$O\Big(\frac{M_{1-\alpha}\, D^{1+\alpha}}{\varepsilon^{\alpha}}\Big)$	(ours)	$O\Big(\frac{(M_{1-\alpha}+\mathrm{C}_2)D^{1+\alpha}}{\varepsilon^{\alpha}} + \frac{\mathrm{C}_1 D^2}{\varepsilon}\Big)$
Quasi-SC, $\alpha = 0$	$\widetilde{O}\big(M_1 D\big)$	[4]	$\widetilde{O}\Big((M_1+\mathrm{C}_2)D + \frac{\mathrm{C}_1 D^2}{\varepsilon}\Big)$

Table 1. **Global complexities for our Algorithm on different problem classes with convex objectives and using inexact Hessian.** Notice that with $\gamma(\boldsymbol{x}) \geq \|\nabla f(\boldsymbol{x})\|_*^{\alpha} M_{1-\alpha}^{-1}$, our framework unifies and recovers state-of-the-art rates for standard problem classes (e.g., Hölder and Quasi-SC) **in the exact case**. For example, $\alpha = 0$ corresponds to Quasi-SC functions, resulting in $\gamma(\boldsymbol{x}) \geq 1/M_1$.

The Holy Grail of Our Theory

LogReg $f(\boldsymbol{x}) = \sum_{i=1}^n \log\big(1 + \exp\big(\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle - b_i\big)\big)$ and **Fisher approximation** $\mathbf{H}(\boldsymbol{x}) = \sum_{i=1}^n \nabla f_i(\boldsymbol{x}) \nabla f_i(\boldsymbol{x})^\top \succeq \mathbf{0}$.

We have $(\bigstar)$ satisfied with $\mathrm{C}_1 = f^\star$, $\mathrm{C}_2 = D$, and $\beta = 0$. Then $\gamma(\boldsymbol{x}) \geq \Big(M_1 + D + \frac{f^\star}{\|\nabla f(\boldsymbol{x})\|_*}\Big)^{-1}$, and we get a complexity $K = O\Big(\Big[M_1 D + D^2 + \frac{f^\star D^2}{\varepsilon}\Big] \cdot \log \frac{1}{\varepsilon}\Big)$. When $f^\star \approx 0$, we result in the **global linear rate of converge**

Paper, Code & Additional Contributions

Paper Code

More Results in the Paper:

- **Non-convex** convergence guarantees: $K \leq \frac{8F_0}{\gamma_* \varepsilon} + \log \frac{\|\nabla f(\boldsymbol{x}_0)\|_*}{\varepsilon}$
- Composite optimization
- Generalized self-concordant functions (**new global rates**)
- **Global linear rate** for logistic regression with Fisher approximation
- **Adaptive step-size** — no smoothness-class knowledge needed
- **Properties** of Gradient-Normalized Smoothness & **bounds** on Hessian approximations
- More experiments...

Practical Approximations & Experiments

LogSumExp: $f(\boldsymbol{x}) = \mu \log \sum_{i=1}^d \exp\Big(\frac{\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle - b_i}{\mu}\Big)$ (1)

Nonlinear Equations: $f(\boldsymbol{x}) = \frac{1}{p}\|\boldsymbol{u}(\boldsymbol{x})\|^p$ (2)

Chebyshev Polynomials: $f(\boldsymbol{x}) = \frac{1}{p}\|\boldsymbol{u}(\boldsymbol{x})\|^p$ with $u_1(\boldsymbol{x}) = \frac{1}{2}(1 - x_1)$, $u_i(\boldsymbol{x}) = x_i - \mathrm{p}_2(x_{i-1})$, $\mathrm{p}_2(\tau) = 2\tau^2 - 1$ (3)

Table 2. **Hessian approximations aligned with our theory, evaluated on convex and non-convex problems.**

Problem	Naming	Approximation $\mathbf{H}(\boldsymbol{x})$
LogSumExp (1)	Weighted Gauss-Newton	$\frac{1}{\mu}\mathbf{A}^\top \mathrm{Diag}\left(\mathrm{softmax}\left(\mathbf{A}, \boldsymbol{x}\right)\right)\mathbf{A}$
Equations with linear operator $\boldsymbol{u}(\boldsymbol{x})$	Fisher Term of $\mathbf{H}$	$\frac{p-2}{\|\boldsymbol{u}(\boldsymbol{x})\|^p}\nabla f(\boldsymbol{x})\nabla f(\boldsymbol{x})^\top$
Nonlinear Equations & Chebyshev (3)	Inexact Hessian	$\|\boldsymbol{u}(\boldsymbol{x})\|^{p-2}\nabla \boldsymbol{u}(\boldsymbol{x})^\top \mathbf{G} \nabla \boldsymbol{u}(\boldsymbol{x}) + \frac{p-2}{\|\boldsymbol{u}(\boldsymbol{x})\|^p}\nabla f(\boldsymbol{x})\nabla f(\boldsymbol{x})^\top$