Gradient-Normalized Smoothness for Optimization with Approximate Hessians

Nikita Doikov

Cornell University, ORIE

VOCAL Optimization Conference: Advanced Algorithms

June 11, 2026, Mosonmagyaróvár

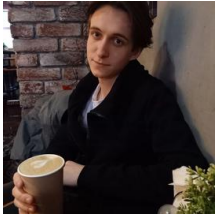
Optimization Problem

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

where $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable and "smooth"; can be non-convex

- **The goal:** to analyze **second-order methods**

1. To have global convergence (from an arbitrary initialization) with fast rates
2. To be able approximate the Hessians: $H_k \approx \nabla^2 f(x_k)$



"*Gradient-Normalized Smoothness for Optimization with Approximate Hessians.*"
Andrii Semenov, Martin Jaggi, Nikita Doikov. **ICLR 2026**

The Result

Algorithm. Choose $x_0 \in \mathbb{R}^n$. Iterate, for $k \geq 0$:

$$x_{k+1} \quad = \quad x_k - \left(H_k + \frac{\|\nabla f(x_k)\|}{\gamma_k} I \right)^{-1} \nabla f(x_k).$$

- $H_k = H_k^\top \succeq 0$ is an approximation of the Hessian: $H_k \approx \nabla^2 f(x_k)$
- NB: for non-convex functions, we might have $\nabla^2 f(x_k) \not\succeq 0$

Examples:

- $H_k := \nabla^2 f(x_k) \quad \Rightarrow \quad$ Newton's method with gradient regularization
- $H_k := 0$ (also a valid choice!) $\quad \Rightarrow \quad x_{k+1} = x_k - \gamma_k \frac{\nabla f(x_k)}{\|\nabla f(x_k)\|}$
- Fisher and Gauss-Newton approximations

$$H_k \quad := \quad \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_k) \nabla f_i(x_k)^\top \quad \overset{?}{\approx} \quad \nabla^2 f(x_k)$$

Our work: we show a universal (problem-class-free) choice of $\gamma_k := \gamma_f(x_k)$

- We call $\gamma_f(\cdot)$ the *Gradient-Normalized Smoothness* of the objective

Outline

I. Newton's Method: Global Convergence

II. Gradient-Normalized Smoothness

III. Examples and Applications

Cubic Regularization of Newton's Method

- Assume that the Hessian is Lipschitz: $\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\|$
- Global second-order upper model of the objective, for $H \geq L$:

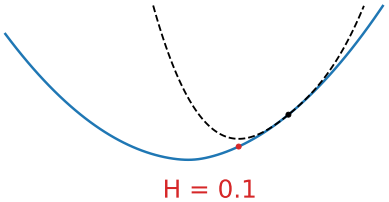
$$f(y) \quad \leq \quad \Omega_2(x; y) + \frac{H}{6}\|y - x\|^3$$

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 $k \geq 0$:

$$x_{k+1} \quad := \quad \arg\min_{y \in \mathbb{R}^n} \left[\Omega_2(x_k; y) + \frac{H}{6}\|y - x_k\|^3 \right]$$

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



Cubic Newton: Theory

$$x_{k+1} \quad = \quad \arg\min_{y \in \mathbb{R}^n} \left[\Omega_2(x_k; y) + \frac{H}{6} \|y - x_k\|^3 \right]$$

Theorem. Let $H := L$ (the Lipschitz constant of the Hessian). Then to find $\|\nabla f(\bar{x}_k)\| \leq \varepsilon$ the method needs

$$k \quad = \quad O\left(\frac{1}{\varepsilon^{3/2}} \right)$$

iterations (second-order oracle calls) [Nesterov-Polyak, 2006]

- For the gradient method, we assume that the gradient is Lipschitz
- The complexity is: $O\left(\frac{1}{\varepsilon^2} \right)$

Gradient Regularization

- **Cubic Newton step:**

$$\begin{aligned} x^+ &= \arg\min_{y \in \mathbb{R}^d} \left\{ \Omega_2(x; y) + \tfrac{H}{6}\|y - x\|^3 \right\} \\ &= x - \left(\nabla^2 f(x) + \tfrac{H}{2} r I\right)^{-1} \nabla f(x), \end{aligned}$$

where r is the solution of a dual problem. We have $r = \|x^+ - x\|$.

- Let f be convex. Then,

$$r := \|x^+ - x\| \leq \tfrac{2}{Hr}\|\nabla f(x)\| \quad \Rightarrow \quad r \leq \sqrt{\tfrac{H\|\nabla f(x)\|}{2}}$$

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

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

+ one linear system

+ fast global rates as for the Cubic Newton

− requires $\nabla^2 f(x) \succeq 0$ (see [Gratton-Jerad-Toint, 2023] for employing negative curvature)

Quasi-Self-Concordant Functions

- **Global norm:** $\|u\| := \langle u, u \rangle^{1/2}$

Functions with Lipschitz Hessian:

$$\nabla^3 f(x)[u, u, u] \quad \leq \quad L\|u\|^3, \qquad \forall x, u$$

- **Local norm:** $\|u\|_x := \langle \nabla^2 f(x) u, u \rangle^{1/2}$

Assume that f is quasi-self-concordant with constant $M \geq 0$:

$$\boxed{\nabla^3 f(x)[u, u, v] \quad \leq \quad M\|u\|_x^2 \|v\|, \qquad \forall u, v}$$

- Combination of the Lipschitzness and classic Self-Concordance

[Nesterov-Nemirovski, 1994]

[Bach, 2010]

[Sun–Tran-Dinh, 2019; Karimireddy–Stich–Jaggi, 2018]

Examples

$$\boxed{\nabla^3 f(x)[u,u,v] \quad \leq \quad M\|u\|_x^2\|v\|}$$

Example 0: f is quadratic. Then $\boxed{M=0}$

Example 1: $f(x)=e^x$. Then $\boxed{M=1}$

Example 2: $f(x)=\ln(1+e^x)$. Then $\boxed{M=1}$

Example 3: (Generalized Linear Models):

$$f(x) \quad = \quad \frac{1}{m}\sum_{i=1}^{m}\ell(\langle a_i,x\rangle),$$

and $\ell:\mathbb{R}\to\mathbb{R}$ is quasi-SC loss function $\Rightarrow$ $f(x)$ is quasi-SC.

...

Global Linear Convergence

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

Theorem. Set

$$\boxed{\beta_k := M\|\nabla f(x_k)\|}$$

Then, we have the global linear rate:

$$f(x_k) - f^* \;\leq\; \exp\left(-\tfrac{k}{8MD}\right)\left(f(x_0) - f^*\right) + \exp\left(-\tfrac{k}{4}\right) g_0 D,$$

where $D := \max\{\|x - x^*\| \;:\; f(x) \leq f(x_0)\}$.

$\Rightarrow$ **the global complexity:** $\mathcal{O}\left(MD \ln \frac{1}{\varepsilon}\right)$ to find $f(x_k) - f^* \leq \varepsilon$ [D, 2025]

NB: compare with

- The gradient method: $\mathcal{O}(\frac{L_1 D^2}{\varepsilon})$
- The cubic Newton: $\mathcal{O}(\sqrt{\frac{L_2 D^3}{\varepsilon}})$

Inexact Hessians

We want to use $H_k \approx \nabla^2 f(x)$

- **Inexactness condition:**

$$\|\nabla^2 f(x_k) - H_k\| \quad \leq \quad C_1 + C_2 \|\nabla f(x_k)\|^{1-\beta}$$

where $C_1 > 0$, $C_2 > 0$ and $0 \leq \beta \leq 1$

- For $H_k \equiv 0$, this condition was called *(L_0, L_1)-smoothness* in machine learning

[Zhang-He-Sra-Jadbabaie, 2020]

[Gorbunov-Tupitsa-Choudhury-Aliev-Richtárik-Horváth-Takáč, 2025]

[Vankov-Rodomanov-Nedich-Sankar-Stich, 2025]

1. How to analyze inexact Hessians?
2. Which problem class to choose? (Lipschitz Hessian, QSC, (L_0, L_1)-functions, ...)

Our work: a universal problem-class-free analysis of inexact second-order methods

$$\Rightarrow \quad \textit{Gradient-Normalized Smoothness}$$

Outline

I. Newton's Method: Global Convergence

II. Gradient-Normalized Smoothness

III. Examples and Applications

Definition: Gradient-Normalized Smoothness

The goal:

$$\nabla f(x+h) \quad \approx \quad \nabla f(x) + H(x)h,$$

where $H(x) \approx \nabla^2 f(x)$ is symmetric positive-definite
Fix $x \in \mathbb{R}^n$ and $g \in \mathbb{R}^n$ (e.g. $g = \nabla f(x)$). Denote:

- Euclidean ball: $B_\gamma := \{h \ : \ \|h\| \leq \gamma\}$
- Local region: $\mathcal{O}_x \ := \ \{\|h\|_x^2 + \langle g, h\rangle \ \leq \ 0\}$
- This is an ellipsoid centered around Newton's direction:

$$\mathcal{O}_x \ = \ \left\{h \ : \ \|h + \tfrac{1}{2}\nabla^2 f(x)^{-1} g\|_x^2 \ \leq \ \tfrac{1}{4}\|g\|_{x,*}^2\right\}$$

Main Condition: for all $h \in B_\gamma \cap \mathcal{O}_x$

$$\|\nabla f(x+h) - \nabla f(x) - H(x)h\| \ \leq \ \tfrac{\|g\| \cdot \|h\|}{\gamma} \qquad (*)$$

Definition. The Gradient-Normalized Smoothness is the maximal $\gamma > 0$

$$\gamma \ \equiv \ \gamma_f(x, g)$$

such that (*) holds. Denote $\gamma_f(x) := \gamma_f(x, \nabla f(x))$

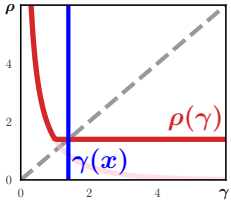
Well-definedness

$$\|\nabla f(x+h) - \nabla f(x) - H(x)h\| \quad \leq \quad \frac{\|\nabla f(x)\| \cdot \|h\|}{\gamma} \qquad (*)$$

- Denote:

$$\rho(\gamma) \quad := \quad \min_{h \in B_\gamma \cap \mathcal{O}_x} \Big\{ \|\nabla f(x+h) - \nabla f(x) - H(x)h\|^{-1} \cdot \|\nabla f(x)\| \cdot \|h\| \Big\}, \quad \gamma \geq 0$$

- Condition (*) means that $\gamma \leq \rho(\gamma)$
- $\rho(\cdot)$ is monotonically decreasing in γ



- $\gamma_f(x)$ is the solution of the non-linear equation: $\gamma^\star = \rho(\gamma^\star)$

Basic Properties

1. Scale-Invariance. Let $\varphi(x) := c \cdot f(x)$, for some $c > 0$. Then,

$$\gamma_\varphi(x) \equiv \gamma_f(x)$$

2. Affine-Substitution. Let $\varphi(x) = f(Ax + b)$. Then,

$$\gamma_\varphi(x) \geq \gamma_f(x) \cdot \|A\|^{-1}$$

3. Sum of Functions. Let $f(x) = \sum_{i=1}^{N} f_i(x)$. Then,

$$\gamma_f(x, g) \geq \left(\sum_{i=1}^{N} \gamma_{f_i}(x, g)^{-1} \right)^{-1}$$

Examples (Exact Hessian)

1. Lipschitz gradient: $\|\nabla^2 f(x)\| \leq L$. Then

$$\gamma_f(x) \quad \geq \quad \frac{\|\nabla f(x)\|}{2L}$$

2. Lipschitz Hessian: $\|\nabla^2 f(x)\| \leq L$. Then

$$\gamma_f(x) \quad \geq \quad \sqrt{\frac{2\|\nabla f(x)\|}{L}}$$

3. Quasi-Self-Concordant Functions, $M > 0$. Then

$$\gamma_f(x) \quad \geq \quad \frac{1}{M}.$$

- Now, we can sum up functions! $f(x) = \sum_{i=1}^{N} f_i(x)$

Global Convergence

Algorithm. Iterate, $k \geq 0$:

$$x_{k+1} = x_k - \left(H_k + \frac{\|\nabla f(x_k)\|}{\gamma_k} I \right)^{-1} \nabla f(x_k)$$

Main Lemma. Let $0 < \gamma_k \leq \gamma_f(x_k)$. Then,

$$f(x_k) - f(x_{k+1}) \geq \frac{\gamma_k}{8} \cdot \frac{\|\nabla f(x_{k+1})\|^2}{\|\nabla f(x_k)\|} \qquad (*)$$

- Convergence result (non-convex functions):

Theorem. Assume $(*)$ holds. Denote $\gamma_* := \min\limits_{0 \leq i \leq N} \gamma_i > 0$. Then, to find a solution: $\|\nabla f(x_k)\| \leq \varepsilon$, it is enough to perform

$$k = \frac{8(f(x_0) - f^*)}{\gamma_* \varepsilon} + \log \frac{\|\nabla f(x_0\|)}{\varepsilon} = O\left(\frac{1}{\gamma_* \varepsilon} \right)$$

Examples

General complexity result:

$$K \quad = \quad O\left(\frac{1}{\gamma_* \varepsilon}\right)$$

iterations to get $\varepsilon > 0$ accuracy. $\boxed{\gamma_* = ?}$

- Lipschitz gradient:

$$\gamma_f(x) \quad \geq \quad \frac{\|\nabla f(x)\|}{2L} \quad \geq \quad \frac{\varepsilon}{2L} \quad \Rightarrow \quad \gamma_* := \frac{\varepsilon}{2L}$$

and we get: $K = O\left(\frac{1}{\varepsilon^2}\right)$ (the same as for the gradient method)

- Lipschitz Hessian:

$$\gamma_f(x) \quad \geq \quad \sqrt{\frac{2\|\nabla f(x)\|}{L}} \quad \geq \quad \sqrt{\frac{2\varepsilon}{L}} \quad \Rightarrow \quad \gamma_* := \sqrt{\frac{2\varepsilon}{L}}$$

  and we get $K = O\left(\frac{1}{\varepsilon^{3/2}}\right)$ (the same as for the Cubic Newton)

- ...
- Even better complexities for convex functions

The Choices of Gamma

- The Gradient Normalized Smoothness $\gamma_f(x_k)$ is the best local step size:

$$x_{k+1} \;=\; x_k - \left(H_k + \frac{\|\nabla f(x_k)\|}{\gamma_k}\right)^{-1} \nabla f(x_k), \qquad \gamma_k \approx \gamma_f(x_k)$$

- How to choose it?

1. Theory choice: $\gamma_k := \gamma_f(x_k)$ — infeasible in practice
2. Constant choice: $\gamma_k := \gamma_*$ for all $k \geq 0$
   where

$$\gamma_* \;:=\; \inf\left\{\gamma_f(x) \;:\; \|\nabla f(x)\| \geq \varepsilon\right\}$$

   - $\gamma_* > 0$ is one constant for all iterates
3. Adaptive search: all we need is the following progress

$$f(x_k) - f(x_{k+1}) \;\geq\; \frac{\gamma_k}{8} \frac{\|\nabla f(x_{k+1})\|^2}{\|\nabla f(x_k)\|}$$

   - This is the condition on γ_k!

Adaptive Method with Approximate Hessians

Init: Choose $x_0 \in \mathbb{R}^n$ and $\gamma_0 > 0$.

Iteration, $k \geq 0$**:**

1. Choose $H_k = H_k^\top \succ 0$
2. Find the smallest $t_k \geq 0$ s.t. for $\gamma := \frac{1}{2^{t_k}}\gamma_k$ and for
$$x^+ = x_k - \left[H_k + \frac{\|\nabla f(x_k)\|}{\gamma_k} I\right]^{-1} \nabla f(x_k)$$
it holds
$$f(x_k) - f(x^+) \geq \frac{\gamma}{8} \frac{\|\nabla f(x^+)\|^2}{\|\nabla f(x_k)\|}$$
3. Set $x_{k+1} = x^+$, and $\gamma_{k+1} = \frac{1}{2^{t_k - 1}}\gamma_k$.

- Automatically achieves the best complexity [D-Mishchenko-Nesterov, 2024]
- Works for non-convex functions
- $\gamma_k \approx \gamma_f(x_k)$

Outline

I. Newton's Method: Global Convergence

II. Gradient-Normalized Smoothness

III. Inexact Hessian and Applications

Inexact Hessian

So far, we saw bounds for $\gamma_f(x)$ for the exact Hessian. Typically

$$\gamma_f(x) \quad \geq \quad \frac{\|\nabla f(x)\|^{\alpha}}{M}, \qquad 0 \leq \boldsymbol{\alpha} \leq 1$$

or combinations of those.

- $\alpha = 1$ — Lipschitz gradient (first-order methods; sublinear rate)
- $\alpha = 0$ — Quasi-Self-Concordance (global linear rate)

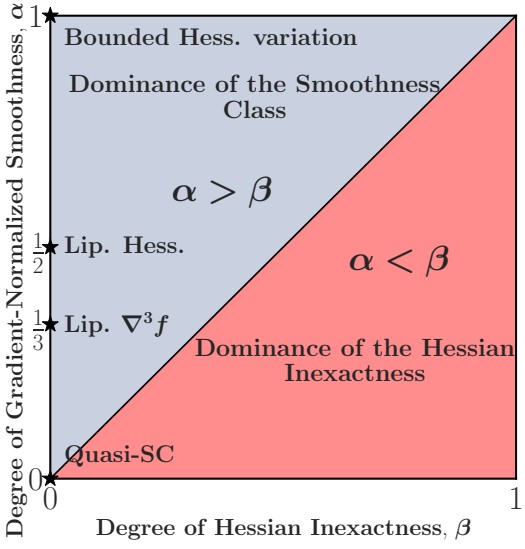
Assume that $\|\nabla^2 f(x_k) - H_k\| \leq C_1 + C_2 \|\nabla f(x)\|^{1-\boldsymbol{\beta}}$, for $\quad 0 \leq \boldsymbol{\beta} \leq 1$

Then, the Gradient Normalized Smoothness $\bar{\gamma}_f(\cdot)$ for inexact Hessian is bounded as:

$$\bar{\gamma}_f(x) \quad \geq \quad \left(\frac{M}{\|\nabla f(x)\|^{\alpha}} + \frac{C_1}{\|\nabla f(x)\|} + \frac{C_2}{\|\nabla f(x)\|^{\boldsymbol{\beta}}} \right)^{-1}$$

- We want $C_1 \to 0$ and $\boxed{\boldsymbol{\beta} \leq \boldsymbol{\alpha}}$

Global Convergence Diagram



Logistic Regression

The Problem

$$\min_{x \in \mathbb{R}^n} \left[f(x) \quad := \quad \frac{1}{N} \sum_{i=1}^{N} \ell(\langle a_i, x \rangle - b_i) \right]$$

where $\ell(t) = \ln(1 + e^t)$ — Quasi-Self-Concordant loss.

- Exact Newton with gradient regularization: $H_k = \nabla^2 f(x_k)$, global linear rate.

$$O(MD \ln \tfrac{1}{\varepsilon}) \qquad \text{iterations to solve the problem}$$

- Fisher Approximation: $H_k = \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_k) \nabla f_i(x_k)^\top$

Lemma. $\|\nabla^2 f(x_k) - H_k\| \leq f^\star + D\|\nabla f(x_k)\|$

$$\Rightarrow \qquad \text{The global complexity:} \qquad O\left(\left[MD + D^2 + \tfrac{f^\star D^2}{\varepsilon} \right] \ln \tfrac{1}{\varepsilon} \right)$$

- Global linear rate when $f^\star \approx 0$ (overparametrized data)

Nonlinear Equations

Non-convex Problem:

$$\min_{x \in \mathbb{R}^n} \left[f(x) \;:=\; \tfrac{1}{p}\|u(x)\|^p \right], \qquad p \geq 2$$

where $u : \mathbb{R}^n \to \mathbb{R}^d$ be a nonlinear operator.

- Consider the matrix:

$$H_k \;=\; \|u(x_k)\|^{p-2} \nabla u(x_k)^\top \nabla u(x_k) + \tfrac{p-2}{\|u(x_k)\|^p} \nabla f(x_k) \nabla f(x_k)^\top \;\succeq\; 0$$

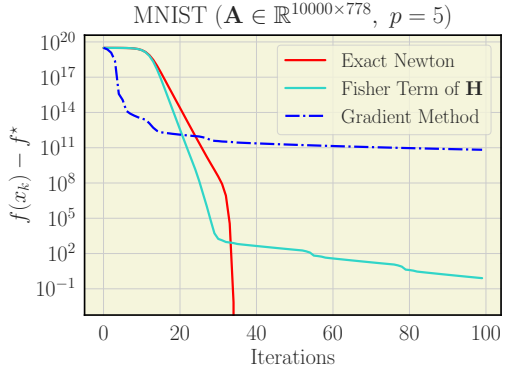
- $p = 2$: this is the standard Gauss-Newton approximation

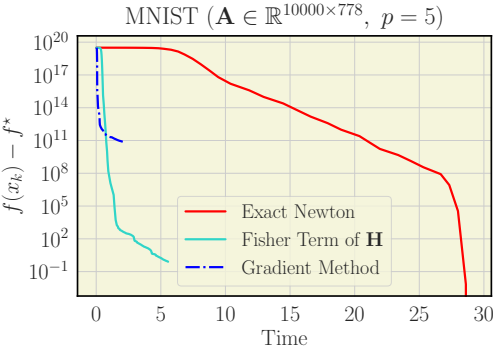
Lemma. Let $\nabla u(x_k) \nabla u(x_k)^\top \succeq \mu I$ (i.e. $n \gg d$) and $\|\nabla^2 u(x)\| \leq \xi$. Then

$$\|\nabla^2 f(x_k) - H_k\| \;\leq\; \tfrac{\xi}{\sqrt{\mu}} \|\nabla f(x_k)\|$$

- Gauss-Newton method (with the gradient regularization) has the same rate as Cubic Newton!

Experiment: Powered Norm and Fisher Approximation





Conclusions

- Modern applications: non-standard smoothness conditions
- The same function can belong to several problem classes simultaneously
- Universal algorithms: automatically adapt to the local "degree of smoothness"

Gradient Normalized Smoothness $\gamma_f(x)$**:** an attempt to formalize this

- Gradient Regularization of the Newton method:

$$x_{k+1} = x_k - \left(H_k + \frac{\|\nabla f(x)\|}{\gamma_k} I\right)^{-1} \nabla f(x_k), \qquad \gamma_k \approx \gamma_f(x_k)$$

- We do not need very exact Hessians $H_k \approx \nabla^2 f(x_k)$

Open Questions

- Instead of the full Fisher approximation:

$$H_k = \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_k) \nabla f_i(x_k)^\top$$

to analyze the dynamic low-rank version:

$$H_k = \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(x_{k-i}) \nabla f_i(x_{k-i})^\top$$

Natural Gradient Descent [Frantar-Kurtic-Alistarh, 2021]
[Martens, 2020]
[Kunstner-Hennig-Balles, 2019]

- Stochastic gradients
- Accelerated second-order methods
- Application domains

Thank you for your attention!