Regularized Newton Method with Global $O(1/k^2)$ Convergence

Konstantin Mishchenko

CNRS, École Normale Supérieure, Inria Sierra

One World Optimization Seminar

[Submitted on 3 Dec 2021]

Regularized Newton Method with Global $O(1/k^2)$ Convergence

Konstantin Mishchenko

We present a Newton-type method that converges fast from any initialization and for arbitrary convex objectives with Lipschitz Hessians. We achieve this by merging the ideas of cubic regularization with a certain adaptive Levenberg--Marquardt penalty. In particular, we show that the iterates given by $x^{k+1} = x^k - \left(\nabla^2 f(x^k) + \sqrt{H\|\nabla f(x^k)\|}\mathbf{I}\right)^{-1}\nabla f(x^k)$, where $H > 0$ is a constant, converge globally with a $\mathcal{O}(\frac{1}{k^2})$ rate. Our method is the first variant of Newton's method that has both cheap iterations and provably fast global convergence. Moreover, we prove that locally our method converges superlinearly when the objective is strongly convex. To boost the method's performance, we present a line search procedure that does not need hyperparameters and is provably efficient.

Comments:	19 pages, 1 figure
Subjects:	**Optimization and Control (math.OC)**; Machine Learning (cs.LG)
Cite as:	arXiv:2112.02089 **[math.OC]**
	(or arXiv:2112.02089v1 **[math.OC]** for this version)

Submission history

From: Konstantin Mishchenko [view email]

[v1] Fri, 3 Dec 2021 18:55:50 UTC (2,244 KB)

Talk structure

1. Problem
2. Historical remarks
3. Key idea
4. Theory overview
5. Experiments
6. Conclusion

Problem

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

Convex and has Lipschitz Hessian

Problem

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

$$\nabla^2 f(x) \succcurlyeq 0$$

Problem

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

$$\nabla^2 f(x) \succcurlyeq 0$$

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq 2H\|x - y\|$$

Some constant

Newton's method

$$x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

Newton's method

$$x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

In practice, solve $\mathbf{A}\delta = b$

with $\mathbf{A} = \nabla^2 f(x^k), \quad b = -\nabla f(x^k)$

$$x^{k+1} = x^k + \delta$$

Newton's method

$$x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

In practice, solve $\mathbf{A}\delta = b$

with $\mathbf{A} = \nabla^2 f(x^k), \quad b = -\nabla f(x^k)$

$$x^{k+1} = x^k + \delta$$

Linear systems are easy

Newton's method

$$x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

Extremely fast locally

Newton's method

$$x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

Extremely fast locally

Does not converge globally

Line search and trust-region

$$x^{k+1} = x^k - t_* (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

Line search and trust-region

$$x^{k+1} = x^k - t_*(\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

$$t_* = \arg\min_{t \geq 0} f(x^k - t(\nabla^2 f(x^k))^{-1} \nabla f(x^k))$$

Line search and trust-region

$$x^{k+1} = x^k - t_*(\nabla^2 f(x^k))^{-1}\nabla f(x^k)$$

$$t_* = \arg\min_{t \geq 0} f(x^k - t(\nabla^2 f(x^k))^{-1}\nabla f(x^k))$$

Full Length Paper | Published: 24 May 2015

Simple examples for the failure of Newton's method with line search for strictly convex minimization

Florian Jarre & Philippe L. Toint

Mathematical Programming **158**, 23–34 (2016) | Cite this article

Talk structure

1. Problem
2. Historical remarks
3. Key idea
4. Theory overview
5. Experiments
6. Conclusion

Historical remarks

- **Ancient Greek, Babylonian and Arab mathematicians, solving equations by improving estimates**

HISTORICAL DEVELOPMENT OF THE NEWTON–RAPHSON METHOD*

TJALLING J. YPMA†

Abstract. This expository paper traces the development of the Newton–Raphson method for solving nonlinear algebraic equations through the extant notes, letters, and publications of Isaac Newton, Joseph Raphson, and Thomas Simpson. It is shown how Newton's formulation differed from the iterative process of Raphson, and that Simpson was the first to give a general formulation, in terms of fluxional calculus, applicable to nonpolynomial equations. Simpson's extension of the method to systems of equations is exhibited.

Historical remarks

- Ancient Greek, Babylonian and Arab mathematicians, solving equations by improving estimates
- Viète, 1600, solving polynomial equation p(x)=0
- Newton, 1669, improved method of Viète (polynomials)
- Raphson, 1690, simpler iterative approach (not just polynomials)
- Simpson, 1740, solving systems of 2 equations using derivatives

Early development

Historical remarks

- Ancient Greek, Babylonian and Arab mathematicians, solving equations by improving estimates
- Viète, 1600, solving polynomial equation p(x)=0
- Newton, 1669, improved method of Viète (polynomials)
- Raphson, 1690, simpler iterative approach (not just polynomials)
- Simpson, 1740, solving systems of 2 equations using derivatives
- Kantorovich, 1939, linear convergence rate for general F(x)=0
- Mysovskikh, 1949, modern proof with quadratic rate
- Davidon, 1959, DFP
- Broyden, Fletcher, Goldfarb, and Shanno, 1970, BFGS

Modern methods and proofs

Historical remarks

- Ancient Greek, Babylonian and Arab mathematicians, solving equations by improving estimates
- Viète, 1600, solving polynomial equation p(x)=0
- Newton, 1669, improved method of Viète (polynomials)
- Raphson, 1690, simpler iterative approach (not just polynomials)
- Simpson, 1740, solving systems of 2 equations using derivatives
- Kantorovich, 1939, linear convergence rate for general F(x)=0
- Mysovskikh, 1949, modern proof with quadratic rate
- Davidon, 1959, DFP
- Broyden, Fletcher, Goldfarb, and Shanno, 1970, BFGS
- Griewank, 1981, numerics for cubic Newton
- Nesterov & Polyak, 2006, theory for cubic Newton

Most relevant works: Cubic Newton

Talk structure

1. Problem
2. Historical remarks
3. Key idea
4. Theory overview
5. Experiments
6. Conclusion

Newton and cubic Newton methods

$$x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

Newton and cubic Newton methods

$$x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

$$f(x) \approx f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2} \langle \nabla^2 f(x^k)(x - x^k), x - x^k \rangle$$

2nd-order Taylor approximation

Newton and cubic Newton methods

$$x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

$$f(x) \approx f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2} \langle \nabla^2 f(x^k)(x - x^k), x - x^k \rangle$$

?

Newton and cubic Newton methods

$$x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

$$f(x) \approx f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2} \langle \nabla^2 f(x^k)(x - x^k), x - x^k \rangle$$

?

$$\|x - x^k\| \approx 0$$

Local by design

Newton and cubic Newton methods

$$x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

$$f(x) \leq f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2} \langle \nabla^2 f(x^k)(x - x^k), x - x^k \rangle + \frac{H}{3} \|x - x^k\|^3$$

Global bound

Newton and cubic Newton methods

$$x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

$$f(x) \leq f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2} \langle \nabla^2 f(x^k)(x - x^k), x - x^k \rangle + \frac{H}{3} \|x - x^k\|^3$$

Global bound

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq 2H \|x - y\|$$

Newton and cubic Newton methods

~~$$x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$~~

$$x^{k+1} = \arg\min_x \left\{ \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2} \langle \nabla^2 f(x^k)(x - x^k), x - x^k \rangle + \frac{H}{3} \|x - x^k\|^3 \right\}$$

Nesterov & Polyak, 2006

Griewank, 1981

Newton and cubic Newton methods

$$~~x^{k+1} = x^k - \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)~~$$

$$x^{k+1} = \arg\min_x \left\{ \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2} \langle \nabla^2 f(x^k)(x - x^k), x - x^k \rangle + \frac{H}{3} \|x - x^k\|^3 \right\}$$

$$\nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k) + H \|x^{k+1} - x^k\| (x^{k+1} - x^k) = 0$$

Newton and cubic Newton methods

$$x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

$$x^{k+1} = \arg\min_x \left\{ \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2} \langle \nabla^2 f(x^k)(x - x^k), x - x^k \rangle + \frac{H}{3} \|x - x^k\|^3 \right\}$$

$$\nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k) + H \|x^{k+1} - x^k\| (x^{k+1} - x^k) = 0$$

Newton and cubic Newton methods

$$\sout{x^{k+1} = x^k - \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)}$$

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

Newton and cubic Newton methods

$$\sout{x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)}$$

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

Converges globally

Newton and cubic Newton methods

$$\sout{x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)}$$

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

Converges globally

But no closed-form expression!

Newton and cubic Newton methods

~~$$x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$~~

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

Lemma. It holds $H\|x^{k+1} - x^k\| \approx \sqrt{H\|\nabla f(x^{k+1})\|}$.

Newton and cubic Newton methods

$$\sout{x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)}$$

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

Lemma. It holds $H\|x^{k+1} - x^k\| \approx \sqrt{H\|\nabla f(x^{k+1})\|}$.

Newton and cubic Newton methods

$$x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

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

Lemma. It holds $H\|x^{k+1} - x^k\| \approx \sqrt{H\|\nabla f(x^{k+1})\|}$.

Idea. Maybe $H\|x^{k+1} - x^k\| \approx \sqrt{H\|\nabla f(x^k)\|}$?

Newton and cubic Newton methods

$$~~x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)~~$$

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

Lemma. It holds $H\|x^{k+1} - x^k\| \approx \sqrt{H\|\nabla f(x^{k+1})\|}$.

Idea. Maybe $H\|x^{k+1} - x^k\| \approx \sqrt{H\|\nabla f(x^k)\|}$?

Almost!

Newton and cubic Newton methods

$$~~x^{k+1} = x^k - \left(\nabla^2 f(x^k)\right)^{-1} \nabla f(x^k)~~$$

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

Lemma. It holds $H\|x^{k+1} - x^k\| \approx \sqrt{H\|\nabla f(x^{k+1})\|}$.

Idea. Maybe $H\|x^{k+1} - x^k\| \approx \sqrt{H\|\nabla f(x^k)\|}$?

Almost!

We can prove one side

$$H\|x^{k+1} - x^k\| \leq \sqrt{H\|\nabla f(x^k)\|}$$

Talk structure

1. Problem
2. Historical remarks
3. Key idea
4. Theory overview
5. Experiments
6. Conclusion

Proposed method

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

Proposed method

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

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq 2H\|x - y\|$$

Global convergence

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

Theorem. For any initialization $x^0 \in \mathbb{R}^d$

$$f(x^k) - f(x^*) = \mathcal{O}\left(\frac{1}{k^2}\right)$$

Global convergence

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

Theorem. For any initialization $x^0 \in \mathbb{R}^d$

$$f(x^k) - f(x^*) = \mathcal{O}\left(\frac{1}{k^2}\right)$$

Convex and has Lipschitz Hessian

Global convergence

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

Theorem. For any initialization $x^0 \in \mathbb{R}^d$

$$f(x^k) - f(x^*) = \mathcal{O}\left(\frac{1}{k^2}\right)$$

Any optimum

Global convergence

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

Theorem. For any initialization $x^0 \in \mathbb{R}^d$

$$f(x^k) - f(x^*) = \mathcal{O}\left(\frac{1}{k^2}\right)$$

No line search or subproblems!

Global convergence

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

Theorem. For any initialization $x^0 \in \mathbb{R}^d$

$$f(x^k) - f(x^*) = \mathcal{O}\left(\frac{1}{k^2}\right)$$

No line search or subproblems!

Matches rate of cubic Newton!

Proof idea: mimic cubic Newton

Define $r_k = \|x^{k+1} - x^k\|$

Proof idea: mimic cubic Newton

Define $r_k = \|x^{k+1} - x^k\|$

Cubic Newton:

$$x^{k+1} = x^k - \left(\nabla^2 f(x^k) + H r_k \mathbf{I}\right)^{-1} \nabla f(x^k)$$

Proof idea: mimic cubic Newton

Define $r_k = \|x^{k+1} - x^k\|$

Cubic Newton:

$$x^{k+1} = x^k - \left(\nabla^2 f(x^k) + H r_k \mathbf{I}\right)^{-1} \nabla f(x^k)$$

Our Newton:

$$x^{k+1} = x^k - \left(\nabla^2 f(x^k) + \lambda_k \mathbf{I}\right)^{-1} \nabla f(x^k)$$

Proof idea: mimic cubic Newton

Define $r_k = \|x^{k+1} - x^k\|$ and $\lambda_k = \sqrt{H\|\nabla f(x^k)\|}$

Cubic Newton:

$$x^{k+1} = x^k - \left(\nabla^2 f(x^k) + H r_k \mathbf{I}\right)^{-1} \nabla f(x^k)$$

Our Newton:

$$x^{k+1} = x^k - \left(\nabla^2 f(x^k) + \lambda_k \mathbf{I}\right)^{-1} \nabla f(x^k)$$

Proof idea: mimic cubic Newton

Define $r_k = \|x^{k+1} - x^k\|$ and $\lambda_k = \sqrt{H\|\nabla f(x^k)\|}$

Cubic Newton:

$$x^{k+1} = x^k - \left(\nabla^2 f(x^k) + Hr_k \mathbf{I}\right)^{-1} \nabla f(x^k)$$

Our Newton:

$$x^{k+1} = x^k - \left(\nabla^2 f(x^k) + \lambda_k \mathbf{I}\right)^{-1} \nabla f(x^k)$$

Idea. Everything is fine if $Hr_k \approx \lambda_k$

Proof idea: mimic cubic Newton

Define $r_k = \|x^{k+1} - x^k\|$ and $\lambda_k = \sqrt{H\|\nabla f(x^k)\|}$

Cubic Newton:

$$x^{k+1} = x^k - \left(\nabla^2 f(x^k) + H r_k \mathbf{I}\right)^{-1} \nabla f(x^k)$$

Our Newton:

$$x^{k+1} = x^k - \left(\nabla^2 f(x^k) + \lambda_k \mathbf{I}\right)^{-1} \nabla f(x^k)$$

Idea. Everything is fine if $H r_k \approx \lambda_k$

Easy part: $H r_k \leq \lambda_k$

Proof sketch

$$r_k = \|x^{k+1} - x^k\| \qquad \lambda_k = \sqrt{H\|\nabla f(x^k)\|}$$

Proof sketch

$$r_k = \|x^{k+1} - x^k\| \qquad \lambda_k = \sqrt{H\|\nabla f(x^k)\|}$$

Lemma 1. $Hr_k \le \lambda_k$ **(Regularization is big enough)**

Proof sketch

$$r_k = \|x^{k+1} - x^k\| \qquad\qquad \lambda_k = \sqrt{H\|\nabla f(x^k)\|}$$

Lemma 1. $Hr_k \leq \lambda_k$ **(Regularization is big enough)**

$$r_k = \|(\nabla^2 f(x^k) + \lambda_k \mathbf{I})^{-1} \nabla f(x^k)\| \leq \frac{1}{\lambda_k} \|\nabla f(x^k)\| = \frac{\lambda_k}{H}$$

Proof sketch

$r_k = \|x^{k+1} - x^k\|$ $\quad\quad$ $\lambda_k = \sqrt{H\|\nabla f(x^k)\|}$

Lemma 1. $Hr_k \leq \lambda_k$

Lemma 2. $\|\nabla f(x^{k+1})\| \leq 2\|\nabla f(x^k)\|$

(No blow-up in gradients)

Proof sketch

$r_k = \|x^{k+1} - x^k\|$ $\qquad$ $\lambda_k = \sqrt{H\|\nabla f(x^k)\|}$

Lemma 1. $Hr_k \leq \lambda_k$

Lemma 2. $\|\nabla f(x^{k+1})\| \leq 2\|\nabla f(x^k)\|$

Lemma 3. $f(x^{k+1}) \leq f(x^k) - \frac{2}{3}\lambda_k r_k^2$ **(Descent)**

Follows from Lemma 1

Proof sketch

$$r_k = \|x^{k+1} - x^k\| \qquad \lambda_k = \sqrt{H\|\nabla f(x^k)\|}$$

Lemma 1. $Hr_k \leq \lambda_k$

Lemma 2. $\|\nabla f(x^{k+1})\| \leq 2\|\nabla f(x^k)\|$

Lemma 3. $f(x^{k+1}) \leq f(x^k) - \frac{2}{3}\lambda_k r_k^2$

Not sufficient to show a good rate.
It's time for a new trick!

Proof sketch

$r_k = \|x^{k+1} - x^k\|$ $\lambda_k = \sqrt{H\|\nabla f(x^k)\|}$

Lemma 1. $Hr_k \leq \lambda_k$

Lemma 2. $\|\nabla f(x^{k+1})\| \leq 2\|\nabla f(x^k)\|$

Lemma 3. $f(x^{k+1}) \leq f(x^k) - \frac{2}{3}\lambda_k r_k^2$

$$\mathcal{I}_\infty = \left\{ k : \|\nabla f(x^{k+1})\| \geq \frac{1}{4}\|\nabla f(x^k)\| \right\}$$

Proof sketch

$$r_k = \|x^{k+1} - x^k\| \qquad \lambda_k = \sqrt{H\|\nabla f(x^k)\|}$$

Lemma 1. $Hr_k \leq \lambda_k$

Lemma 2. $\|\nabla f(x^{k+1})\| \leq 2\|\nabla f(x^k)\|$

Lemma 3. $f(x^{k+1}) \leq f(x^k) - \frac{2}{3}\lambda_k r_k^2$

$$\mathcal{I}_\infty = \left\{k : \|\nabla f(x^{k+1})\| \geq \frac{1}{4}\|\nabla f(x^k)\|\right\}$$

Proof sketch

$$r_k = \|x^{k+1} - x^k\| \qquad \lambda_k = \sqrt{H\|\nabla f(x^k)\|}$$

Lemma 1. $Hr_k \le \lambda_k$

Lemma 2. $\|\nabla f(x^{k+1})\| \le 2\|\nabla f(x^k)\|$

Lemma 3. $f(x^{k+1}) \le f(x^k) - \frac{2}{3}\lambda_k r_k^2$

$$\mathcal{I}_\infty = \left\{k : \|\nabla f(x^{k+1})\| \ge \frac{1}{4}\|\nabla f(x^k)\|\right\}$$

Lemma 4. If $k \in \mathcal{I}_\infty$, $f(x^{k+1}) \le f(x^k) - c(f(x^k) - f^*)^{\frac{3}{2}}$

if $k \notin \mathcal{I}_\infty$, $\|\nabla f(x^{k+1})\| \le \frac{1}{4}\|\nabla f(x^k)\|$

Proof sketch

$$r_k = \|x^{k+1} - x^k\| \qquad \lambda_k = \sqrt{H\|\nabla f(x^k)\|}$$

Lemma 1. $Hr_k \le \lambda_k$

Lemma 2. $\|\nabla f(x^{k+1})\| \le 2\|\nabla f(x^k)\|$

Lemma 3. $f(x^{k+1}) \le f(x^k) - \frac{2}{3}\lambda_k r_k^2$

$$\mathcal{I}_\infty = \left\{k \colon \|\nabla f(x^{k+1})\| \ge \frac{1}{4}\|\nabla f(x^k)\|\right\}$$

Lemma 4. If $k \in \mathcal{I}_\infty$, $f(x^{k+1}) \le f(x^k) - c\left(f(x^k) - f^*\right)^{\frac{3}{2}}$

if $k \notin \mathcal{I}_\infty$, $\|\nabla f(x^{k+1})\| \le \frac{1}{4}\|\nabla f(x^k)\|$

Proof sketch

$$r_k = \|x^{k+1} - x^k\| \qquad \lambda_k = \sqrt{H\|\nabla f(x^k)\|}$$

Lemma 1. $Hr_k \le \lambda_k$

Lemma 2. $\|\nabla f(x^{k+1})\| \le 2\|\nabla f(x^k)\|$

Lemma 3. $f(x^{k+1}) \le f(x^k) - \frac{2}{3}\lambda_k r_k^2$

$$\mathcal{I}_\infty = \left\{k \colon \|\nabla f(x^{k+1})\| \ge \frac{1}{4}\|\nabla f(x^k)\|\right\}$$

Lemma 4. If $k \in \mathcal{I}_\infty$, $f(x^{k+1}) \le f(x^k) - c(f(x^k) - f^*)^{\frac{3}{2}}$

if $k \notin \mathcal{I}_\infty$, $\|\nabla f(x^{k+1})\| \le \frac{1}{4}\|\nabla f(x^k)\|$

Bonus: superlinear rate

$$x^{k+1} = x^k - \left(\nabla^2 f(x^k) + \lambda_k \mathbf{I}\right)^{-1} \nabla f(x^k)$$

Theorem. If $\nabla^2 f(x) \succcurlyeq \mu \mathbf{I}$ and $\|\nabla f(x^0)\| \leq \frac{\mu^2}{4H}$

$$\|\nabla f(x^{k+1})\| \leq \frac{2\sqrt{H}}{\mu} \|\nabla f(x^k)\|^{\frac{3}{2}}$$

Bonus: superlinear rate

$$x^{k+1} = x^k - \left(\nabla^2 f(x^k) + \lambda_k \mathbf{I}\right)^{-1} \nabla f(x^k)$$

Theorem. If $\nabla^2 f(x) \succcurlyeq \mu \mathbf{I}$ and $\|\nabla f(x^0)\| \leq \frac{\mu^2}{4H}$

$$\|\nabla f(x^{k+1})\| \leq \frac{2\sqrt{H}}{\mu} \|\nabla f(x^k)\|^{\frac{3}{2}}$$

Bonus: superlinear rate

$$x^{k+1} = x^k - \left(\nabla^2 f(x^k) + \lambda_k \mathbf{I}\right)^{-1} \nabla f(x^k)$$

Theorem. If $\nabla^2 f(x) \succcurlyeq \mu \mathbf{I}$ and $\|\nabla f(x^0)\| \leq \frac{\mu^2}{4H}$

$$\|\nabla f(x^{k+1})\| \leq \frac{2\sqrt{H}}{\mu} \|\nabla f(x^k)\|^{\frac{3}{2}}$$

Bonus: superlinear rate

$$x^{k+1} = x^k - \left(\nabla^2 f(x^k) + \lambda_k \mathbf{I}\right)^{-1} \nabla f(x^k)$$

$\textbf{Theorem.}$ If $\nabla^2 f(x) \succcurlyeq \mu \mathbf{I}$ and $\|\nabla f(x^0)\| \le \frac{\mu^2}{4H}$

$$\|\nabla f(x^{k+1})\| \le \frac{2\sqrt{H}}{\mu} \|\nabla f(x^k)\|^{\frac{3}{2}}$$

$\|\nabla f(x^k)\| \le \varepsilon$ **after** $\mathcal{O}\left(\log\log\frac{1}{\varepsilon}\right)$ **iterations**

Summary

$$x^{k+1} = x^k - \left(\nabla^2 f(x^k) + \lambda_k \mathbf{I}\right)^{-1} \nabla f(x^k)$$

$$f(x^k) - f(x^*) = \mathcal{O}\left(\frac{1}{k^2}\right)$$

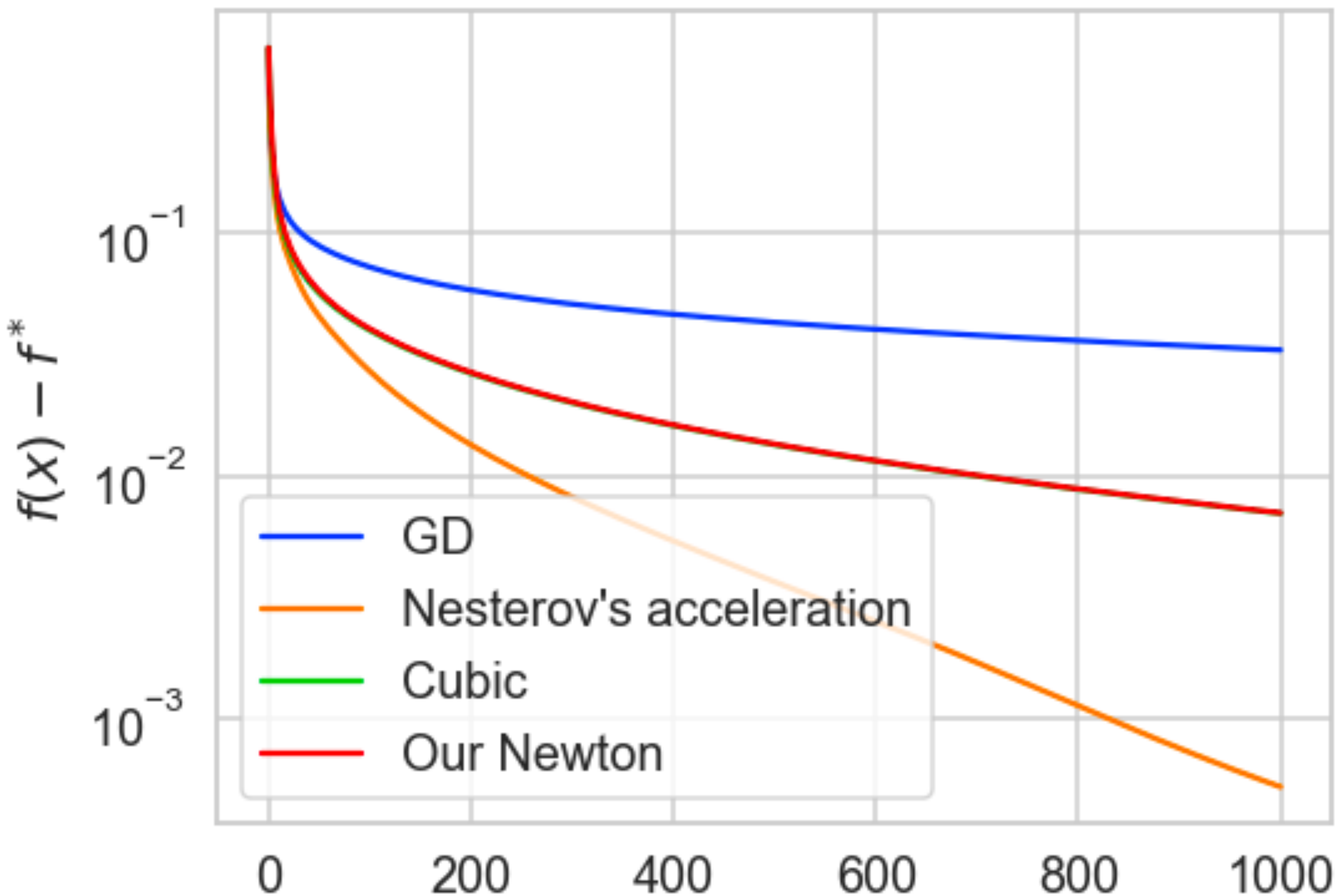
$$\mathcal{O}\left(\log\log\frac{1}{\varepsilon}\right)$$ **locally**

Intuition: it's almost like cubic Newton

Talk structure

1. Problem
2. Historical remarks
3. Key idea
4. Theory overview
5. Experiments
6. Conclusion

Experiments



What if we don't know H

Input: $x^0 \neq x^1 \in \mathbb{R}^d$

Initialize $H_0 = \frac{\|\nabla f(x^1) - \nabla f(x^0) - \nabla^2 f(x^0)(x^1 - x^0)\|}{\|x^1 - x^0\|^2}$

for $k = 1, 2, \ldots$ **do**

$\quad M_k = \frac{\|\nabla f(x^k) - \nabla f(x^{k-1}) - \nabla^2 f(x^{k-1})(x^k - x^{k-1})\|}{\|x^k - x^{k-1}\|^2}$

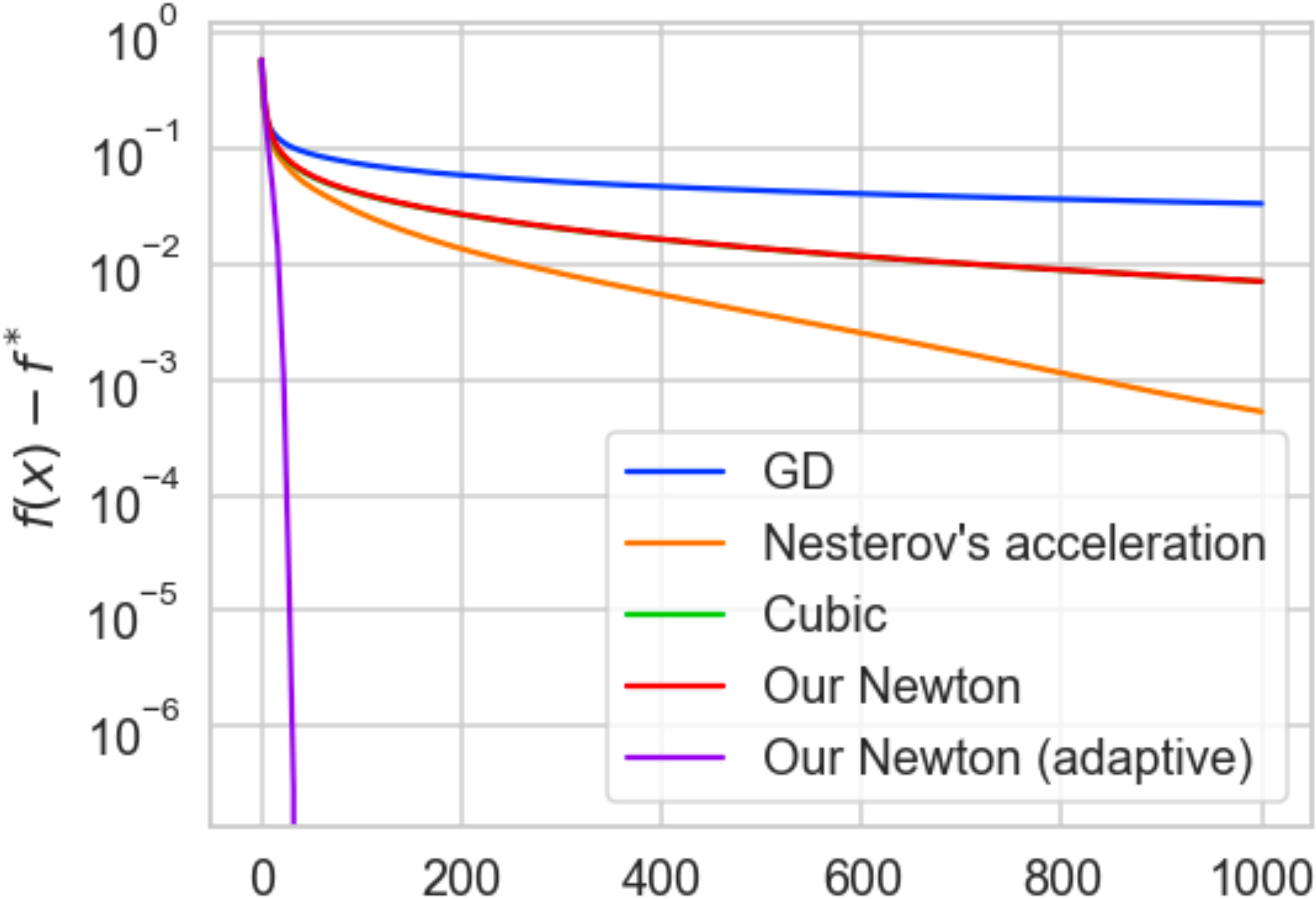
$\quad H_k = \max\left\{M_k, \frac{H_{k-1}}{2}\right\}$

$\quad \lambda_k = \sqrt{H_k \|\nabla f(x^k)\|}$

$\quad$ Compute $x^{k+1} = x^k - (\nabla^2 f(x^k) + \lambda_k \mathbf{I})^{-1} \nabla f(x^k)$

end for

Experiments



Talk structure

1. Problem
2. Historical remarks
3. Key idea
4. Theory overview
5. Experiments
6. Conclusion

What’s next?

1. Nonsmooth problems
2. Inexact/subspace updates
3. Acceleration
4. Minmax optimization
5. Stochastic Newton (very hard!)
6. Practical quasi-Newton variants

90% is due to Nesterov & Polyak

Polyak

Nesterov

90% is due to Nesterov & Polyak

"If I have seen further it is by standing on ye sholders of Giants", Newton, 1676

Polyak

Nesterov