Recent Advances in Global Convergence of Newton Methods

Konstantin Mishchenko

CNRS, École Normale Supérieure, Inria Sierra

Talk structure

1. Overview of methods
2. Newton and Cubic Newton
3. Regularized Newton
4. Experiments
5. Crazy result on 3rd derivatives

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

Overview of methods

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

GD: $f(x) \approx f(x^k) + \langle \nabla f(x^k), x - x^k \rangle$

Overview of methods

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

GD: $f(x) \approx f(x^k) + \langle \nabla f(x^k), x - x^k \rangle$

$\mathcal{O}\left(\frac{1}{k}\right)$ rate (without acceleration)

Overview of methods

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

GD: $f(x) \approx f(x^k) + \langle \nabla f(x^k), x - x^k \rangle$

$\mathcal{O}\left(\frac{1}{k}\right)$ rate (without acceleration)

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

Overview of methods

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

GD: $f(x) \approx f(x^k) + \langle \nabla f(x^k), x - x^k \rangle$

$\mathcal{O}\left(\frac{1}{k}\right)$ rate (without acceleration)

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

$\mathcal{O}\left(\frac{1}{k^2}\right)$ rate (global)

Overview of methods

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

GD: $f(x) \approx f(x^k) + \langle \nabla f(x^k), x - x^k \rangle$

$\mathcal{O}\left(\frac{1}{k}\right)$ rate (without acceleration)

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

$\mathcal{O}\left(\frac{1}{k^2}\right)$ rate (global)

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

$+ \frac{1}{6}\nabla^3 f(x^k)[x - x^k]^3$

Overview of methods

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

GD: $f(x) \approx f(x^k) + \langle \nabla f(x^k), x - x^k \rangle$

$\mathcal{O}\left(\frac{1}{k}\right)$ rate (without acceleration)

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

$\mathcal{O}\left(\frac{1}{k^2}\right)$ rate (global)

Tensor: $f(x) \approx f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2}\nabla^2 f(x^k)[x - x^k]^2 + \frac{1}{6}\nabla^3 f(x^k)[x - x^k]^3$

$\mathcal{O}\left(\frac{1}{k^3}\right)$ rate

Talk structure

1. Overview of methods
2. Newton and Cubic Newton
3. Regularized Newton
4. Experiments
5. Crazy result on 3rd derivatives

Newton's method

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

Newton's method

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

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

2nd-order Taylor approximation

Newton'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

Does not converge globally

Why not line search?

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

Why not line search?

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

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

Why not line search?

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

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

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

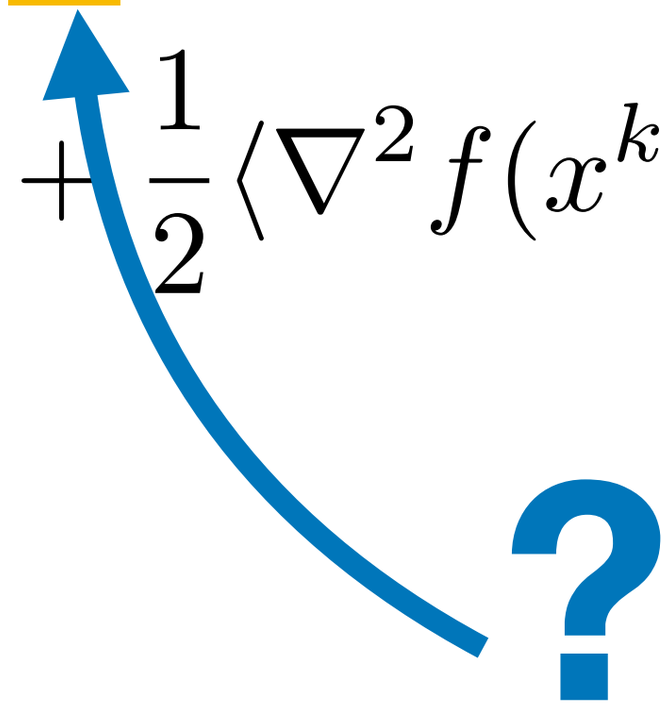
Mathematical Programming **158**, 23–34 (2016) | Cite this article

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

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

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

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

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

Newton and cubic Newton methods

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

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

Newton and cubic Newton methods

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

Converges globally

But no closed-form expression!

Newton and cubic Newton methods

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

Converges globally

But no closed-form expression!

Curious property:

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

Talk structure

1. Overview of methods
2. Newton and Cubic Newton
3. Regularized Newton
4. Experiments
5. Crazy result on 3rd derivatives

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 $x^0 \in \mathbb{R}^d$

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

No line search or subproblems!

Matches the rate of cubic Newton!

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

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

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

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

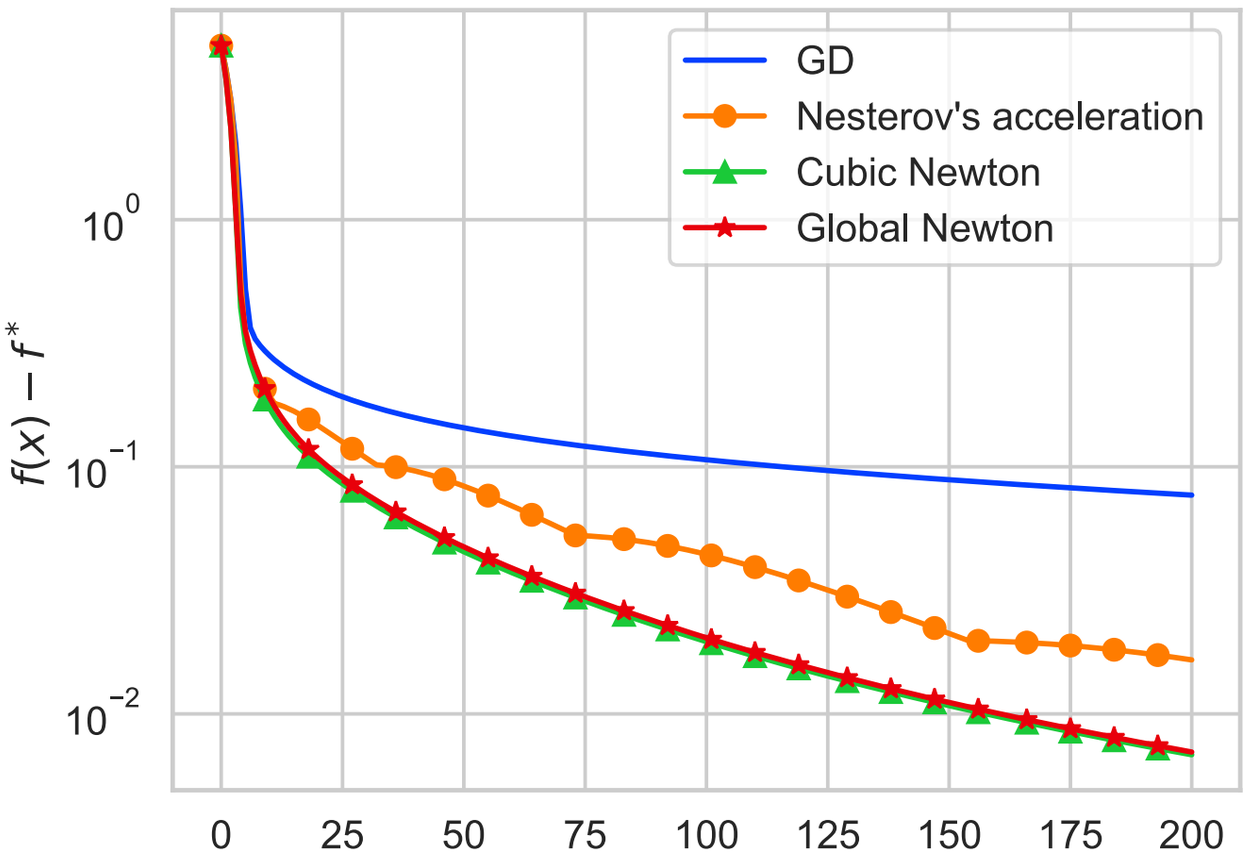
$\|\nabla f(x^k)\| \leq \varepsilon$ **after** $\mathcal{O}\left(\log\log \frac{1}{\varepsilon}\right)$ **iterations**

Talk structure

1. Overview of methods
2. Newton and Cubic Newton
3. Regularized Newton
4. Experiments
5. Crazy result on 3rd derivatives

Experiments

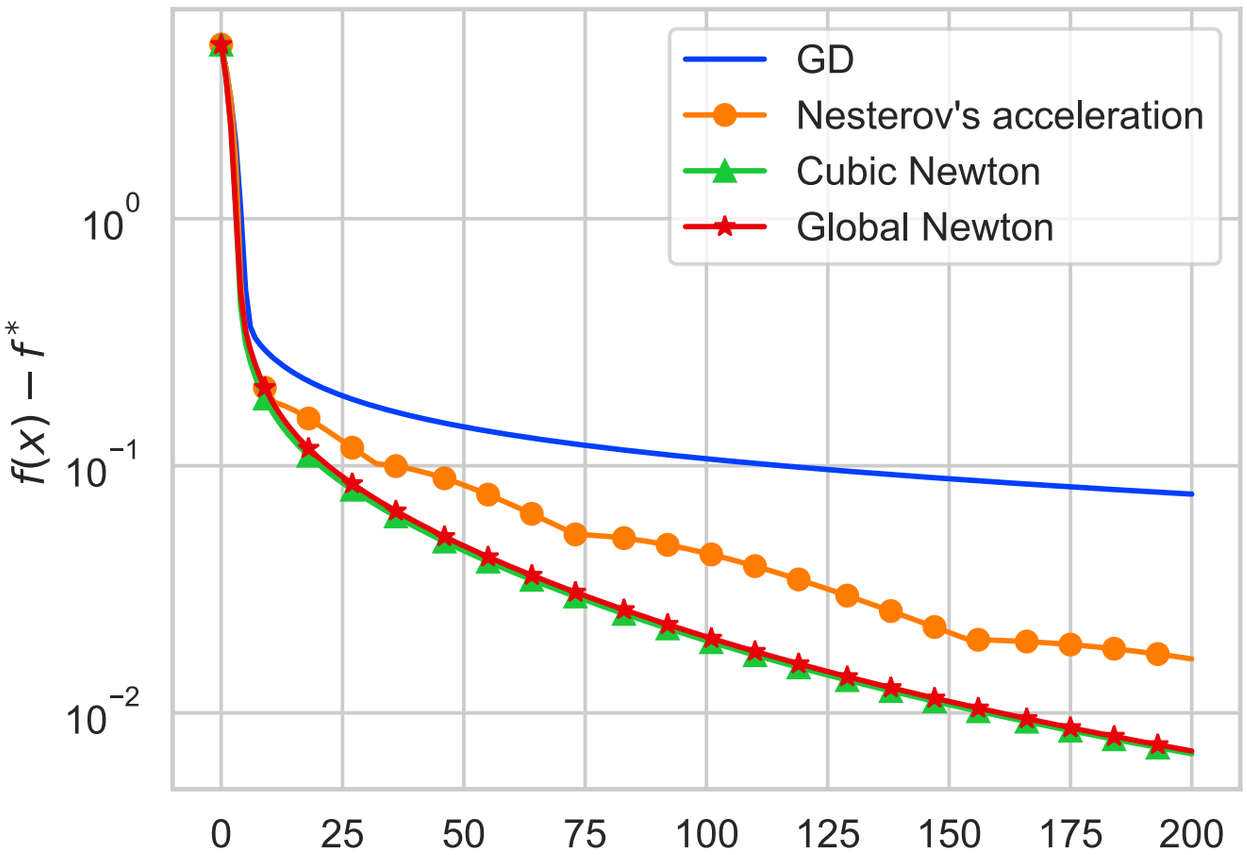
Logistic regression

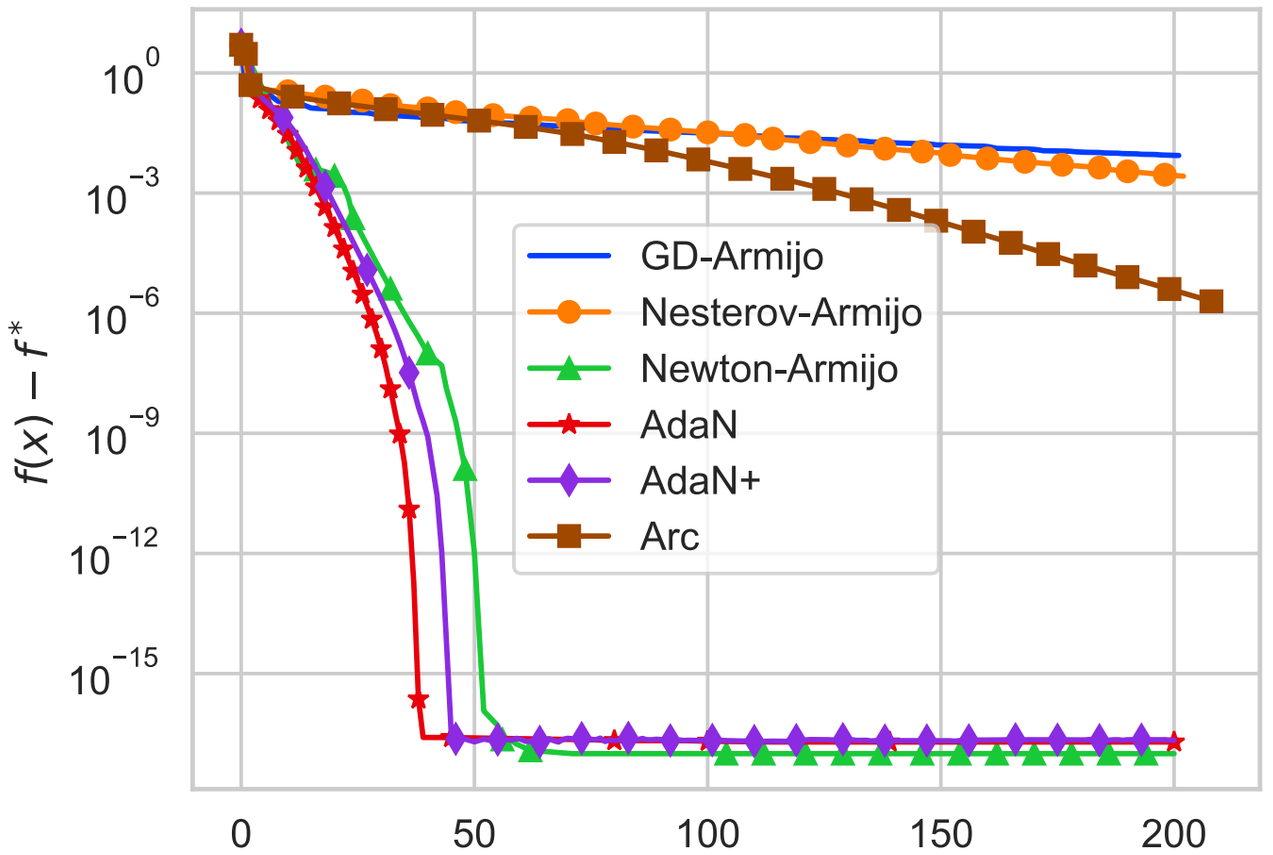


Non-adaptive methods

Experiments

Logistic regression





Non-adaptive methods

Adaptive methods

Experiments

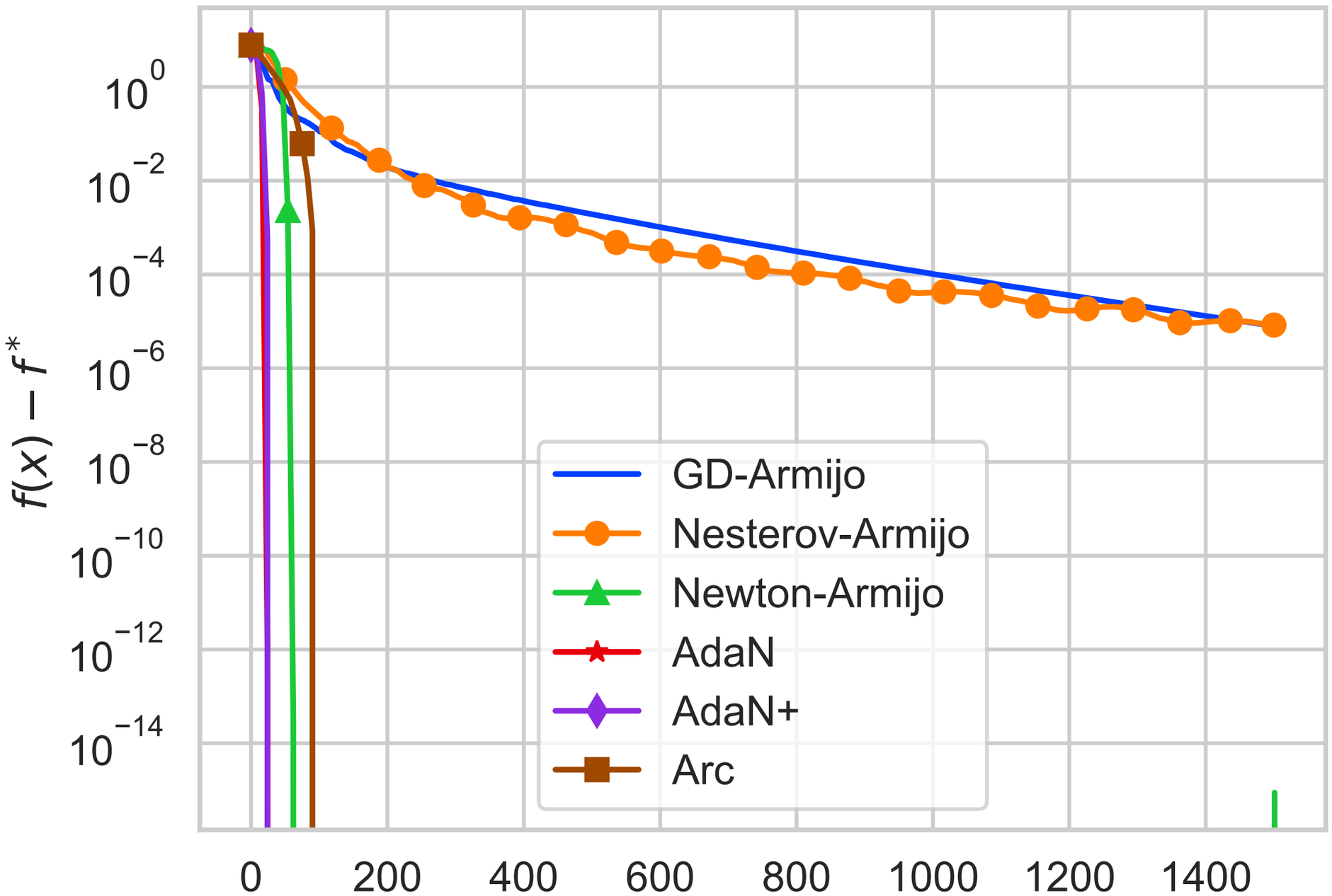
Log-sum-exp

$$\min_{x \in \mathbb{R}^d} \rho \log \left(\sum_{i=1}^{n} \exp \left(\frac{a_i^\top x - b_i}{\rho} \right) \right)$$

Small ρ implies ill-conditioning

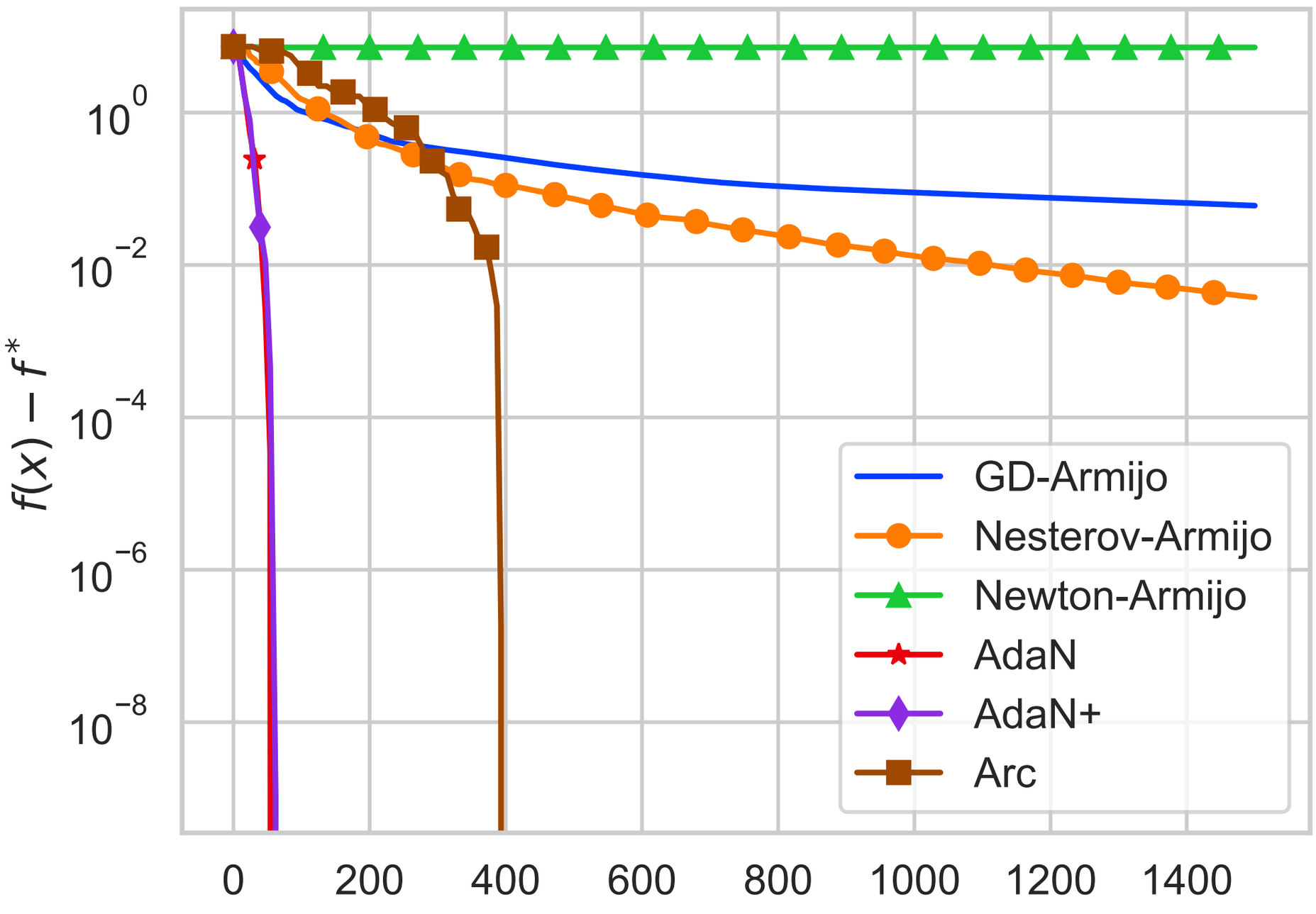
Experiments

Log-sum-exp ($\rho = 0.5$, well-conditioned)



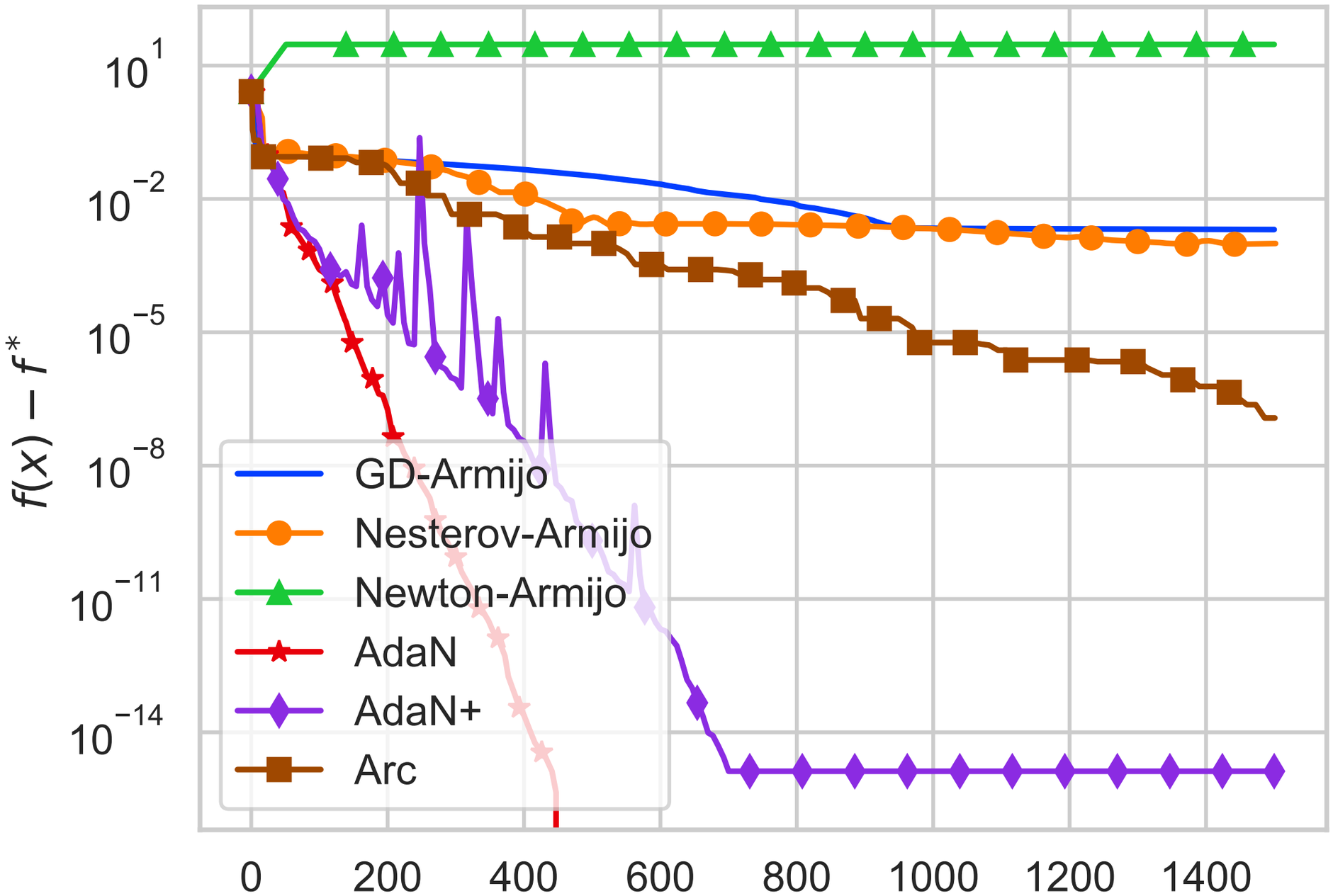
Experiments

Log-sum-exp ($\rho = 0.25$, medium)



Experiments

Log-sum-exp ($\rho = 0.05$, ill-conditioned)



Talk structure

1. Overview of methods
2. Newton and Cubic Newton
3. Regularized Newton
4. Experiments
5. Crazy result on 3rd derivatives

Reference

[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.

3 December 2021
(first appearance in July 2021)

Surprise paper

[Submitted on 6 Dec 2021]

Gradient Regularization of Newton Method with Bregman Distances

Nikita Doikov, Yurii Nesterov

In this paper, we propose a first second-order scheme based on arbitrary non-Euclidean norms, incorporated by Bregman distances. They are introduced directly in the Newton iterate with regularization parameter proportional to the square root of the norm of the current gradient. For the basic scheme, as applied to the composite optimization problem, we establish the global convergence rate of the order $O(k^{-2})$ both in terms of the functional residual and in the norm of subgradients. Our main assumption on the smooth part of the objective is Lipschitz continuity of its Hessian. For uniformly convex functions of degree three, we justify global linear rate, and for strongly convex function we prove the local superlinear rate of convergence. Our approach can be seen as a relaxation of the Cubic Regularization of the Newton method, which preserves its convergence properties, while the auxiliary subproblem at each iteration is simpler. We equip our method with adaptive line search procedure for choosing the regularization parameter. We propose also an accelerated scheme with convergence rate $O(k^{-3})$, where k is the iteration counter.

6 December 2021
(cites the version from July)

March 2022

Me

Nikita

Yurii

Result of my visit

[Submitted on 11 Aug 2022]

Super-Universal Regularized Newton Method

Nikita Doikov, Konstantin Mishchenko, Yurii Nesterov

We analyze the performance of a variant of Newton method with quadratic regularization for solving composite convex minimization problems. At each step of our method, we choose regularization parameter proportional to a certain power of the gradient norm at the current point. We introduce a family of problem classes characterized by Hölder continuity of either the second or third derivative. Then we present the method with a simple adaptive search procedure allowing an automatic adjustment to the problem class with the best global complexity bounds, without knowing specific parameters of the problem. In particular, for the class of functions with Lipschitz continuous third derivative, we get the global $O(1/k^3)$ rate, which was previously attributed to third-order tensor methods. When the objective function is uniformly convex, we justify an automatic acceleration of our scheme, resulting in a faster global rate and local superlinear convergence. The switching between the different rates (sublinear, linear, and superlinear) is automatic. Again, for that, no a priori knowledge of parameters is needed.

Extra smoothness

$$\|\nabla^3 f(x) - \nabla^3 f(y)\| \leq L\|x - y\|$$

Extra smoothness

$$\|\nabla^3 f(x) - \nabla^3 f(y)\| \leq L\|x - y\|$$

Prior work (Nesterov): for any τ, it holds

$$\nabla^3 f(x)[h]^3 \leq \frac{1}{\tau}\nabla^2 f(x)[h]^2 + \frac{\tau}{2}L\|h\|^4$$

Extra smoothness

$$\|\nabla^3 f(x) - \nabla^3 f(y)\| \leq L\|x - y\|$$

Prior work (Nesterov): for any τ, it holds

$$\nabla^3 f(x)[h]^3 \leq \frac{1}{\tau}\nabla^2 f(x)[h]^2 + \frac{\tau}{2}L\|h\|^4$$

$$f(y) \leq \underbrace{f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}\nabla^2 f(x)[y - x]^2 + \frac{1}{6}\nabla^3 f(x)[y - x]^3}_{\text{Taylor approximation of } f(y)} + \frac{L}{24}\|y - x\|^4$$

Extra smoothness

$$\|\nabla^3 f(x) - \nabla^3 f(y)\| \leq L\|x - y\|$$

Prior work (Nesterov): for any τ, it holds

$$\nabla^3 f(x)[h]^3 \leq \frac{1}{\tau}\nabla^2 f(x)[h]^2 + \frac{\tau}{2}L\|h\|^4$$

$$f(y) \leq \underbrace{f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}\nabla^2 f(x)[y - x]^2 + \frac{1}{6}\nabla^3 f(x)[y - x]^3}_{\text{Taylor approximation of } f(y)} + \frac{L}{24}\|y - x\|^4$$

Extra smoothness

$$f(y) \leq \underbrace{f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}\nabla^2 f(x)[y-x]^2 + \frac{1}{6}\nabla^3 f(x)[y-x]^3}_{\text{Taylor approximation of } f(y)} + \frac{L}{24}\|y-x\|^4$$

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \left(\frac{1}{2} + \frac{1}{6\tau}\right) \nabla^2 f(x)[y-x]^2 + \frac{L + 2\tau}{24} L\|y-x\|^4$$

Extra smoothness

$$f(y) \leq \underbrace{f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}\nabla^2 f(x)[y-x]^2 + \frac{1}{6}\nabla^3 f(x)[y-x]^3}_{\text{Taylor approximation of } f(y)} + \frac{L}{24}\|y-x\|^4$$

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \left(\frac{1}{2} + \frac{1}{6\tau}\right) \nabla^2 f(x)[y-x]^2 + \frac{L+2\tau}{24} L\|y-x\|^4$$

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

Extra smoothness

$$f(y) \leq \underbrace{f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}\nabla^2 f(x)[y-x]^2 + \frac{1}{6}\nabla^3 f(x)[y-x]^3}_{\text{Taylor approximation of } f(y)} + \frac{L}{24}\|y-x\|^4$$

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \left(\frac{1}{2} + \frac{1}{6\tau}\right) \nabla^2 f(x)[y-x]^2 + \frac{L + 2\tau}{24} L\|y-x\|^4$$

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

$$\lambda_k = \sqrt[3]{6L\|\nabla f(x^k)\|^2}$$

Extra smoothness

$$f(y) \leq \underbrace{f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}\nabla^2 f(x)[y-x]^2 + \frac{1}{6}\nabla^3 f(x)[y-x]^3}_{\text{Taylor approximation of } f(y)} + \frac{L}{24}\|y-x\|^4$$

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \left(\frac{1}{2} + \frac{1}{6\tau}\right) \nabla^2 f(x)[y-x]^2 + \frac{L + 2\tau}{24} L\|y-x\|^4$$

$$x^{k+1} = x^k - (\nabla^2 f(x^k) + \lambda_k \mathbf{I})^{-1} \nabla f(x^k) \qquad \lambda_k = \sqrt[3]{6L\|\nabla f(x^k)\|^2}$$

Theorem. If f is convex and has Lipschitz third derivatives, then

$$f(x^k) - f^* = \mathcal{O}\left(\frac{1}{k^3}\right).$$

Composite objective

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

Constraints, ℓ_1 penalty, etc.

Composite objective

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

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

Composite objective

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

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

$$\lambda_k \propto \|\nabla f(x^k) + \psi'(x^k)\|^\alpha$$

Full method

Algorithm 2 Super-Universal Newton Method

Input: $x_0 \in \operatorname{dom}\psi$, $\psi'(x_0) \in \partial\psi(x_0)$. Choose arbitrary $\alpha \in \left[\frac{2}{3}, 1\right]$, $H_0 > 0$

for $k = 0, 1, \ldots$ **do**

  $g_k = \|\nabla f(x_k) + \psi'(x_k)\|_*$

  for $j_k = 0, 1, \ldots$ **do**

    $\lambda_k = 4^{j_k} H_k g_k^{\alpha}$

    $x_+ = \operatorname{argmin}_x \left\{ \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2}\nabla^2 f(x_k)[x - x_k]^2 + \frac{\lambda_k}{2}\|x - x_k\|^2 + \psi(x) \right\}$

    $\psi'(x_+) \overset{\text{def}}{=} -\nabla f(x_k) - \nabla^2 f(x_k)(x_+ - x_k) - \lambda_k B(x_+ - x_k)$

    $F'(x_+) \overset{\text{def}}{=} \nabla f(x_+) + \psi'(x_+)$

  until $\langle F'(x_+), x_k - x_+ \rangle \geq \frac{\|F'(x_+)\|_*^2}{4\lambda_k}$

  $x_{k+1} = x_+$

  $H_{k+1} = \frac{4^{j_k} H_k}{4}$

Full method

Algorithm 2 Super-Universal Newton Method

Input: $x_0 \in \operatorname{dom}\psi$, $\psi'(x_0) \in \partial\psi(x_0)$. Choose arbitrary $\alpha \in \left[\frac{2}{3}, 1\right]$, $H_0 > 0$

for $k = 0, 1, \ldots$ **do**

  $g_k = \|\nabla f(x_k) + \psi'(x_k)\|_*$

  for $j_k = 0, 1, \ldots$ **do**

    $\lambda_k = 4^{j_k} H_k g_k^{\alpha}$

    $x_+ = \operatorname{argmin}_x \left\{ \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2}\nabla^2 f(x_k)[x - x_k]^2 + \frac{\lambda_k}{2}\|x - x_k\|^2 + \psi(x) \right\}$

    $\psi'(x_+) \stackrel{\text{def}}{=} -\nabla f(x_k) - \nabla^2 f(x_k)(x_+ - x_k) - \lambda_k B(x_+ - x_k)$

    $F'(x_+) \stackrel{\text{def}}{=} \nabla f(x_+) + \psi'(x_+)$

  until $\langle F'(x_+), x_k - x_+ \rangle \geq \frac{\|F'(x_+)\|_*^2}{4\lambda_k}$

  $x_{k+1} = x_+$

  $H_{k+1} = \frac{4^{j_k} H_k}{4}$

Full method

Algorithm 2 Super-Universal Newton Method

Input: $x_0 \in \operatorname{dom}\psi$, $\psi'(x_0) \in \partial\psi(x_0)$. Choose arbitrary $\alpha \in \left[\frac{2}{3}, 1\right]$, $H_0 > 0$

for $k = 0, 1, \ldots$ **do**

  $g_k = \|\nabla f(x_k) + \psi'(x_k)\|_*$

  for $j_k = 0, 1, \ldots$ **do**

    $\lambda_k = 4^{j_k} H_k g_k^{\alpha}$

    $x_+ = \operatorname{argmin}_x \left\{ \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2}\nabla^2 f(x_k)[x - x_k]^2 + \frac{\lambda_k}{2}\|x - x_k\|^2 + \psi(x) \right\}$

    $\psi'(x_+) \overset{\text{def}}{=} -\nabla f(x_k) - \nabla^2 f(x_k)(x_+ - x_k) - \lambda_k B(x_+ - x_k)$

    $F'(x_+) \overset{\text{def}}{=} \nabla f(x_+) + \psi'(x_+)$

  until $\langle F'(x_+), x_k - x_+ \rangle \geq \frac{\|F'(x_+)\|_*^2}{4\lambda_k}$

  $x_{k+1} = x_+$

  $H_{k+1} = \frac{4^{j_k} H_k}{4}$

Full method

Algorithm 2 Super-Universal Newton Method

Input: $x_0 \in \operatorname{dom}\psi$, $\psi'(x_0) \in \partial\psi(x_0)$. Choose arbitrary $\alpha \in \left[\frac{2}{3}, 1\right]$, $H_0 > 0$

for $k = 0, 1, \ldots$ **do**
  $g_k = \|\nabla f(x_k) + \psi'(x_k)\|_*$
  for $j_k = 0, 1, \ldots$ **do**
    $\lambda_k = 4^{j_k} H_k g_k^{\alpha}$
    $x_+ = \operatorname{argmin}_x \left\{\langle \nabla f(x_k), x - x_k\rangle + \frac{1}{2}\nabla^2 f(x_k)[x - x_k]^2 + \frac{\lambda_k}{2}\|x - x_k\|^2 + \psi(x)\right\}$
    $\psi'(x_+) \overset{\text{def}}{=} -\nabla f(x_k) - \nabla^2 f(x_k)(x_+ - x_k) - \lambda_k B(x_+ - x_k)$
    $F'(x_+) \overset{\text{def}}{=} \nabla f(x_+) + \psi'(x_+)$
  until $\langle F'(x_+), x_k - x_+\rangle \geq \frac{\|F'(x_+)\|_*^2}{4\lambda_k}$
  $x_{k+1} = x_+$
  $H_{k+1} = \frac{4^{j_k} H_k}{4}$

Provably works for any α and any function class

One method for all rates

If $\|\nabla^3 f(x) - \nabla^3 f(y)\| \leq L\|x - y\|$ for all x, y, then

$$f(x^k) - f^* = \mathcal{O}\left(\frac{1}{k^3}\right).$$

One method for all rates

If $\|\nabla^3 f(x) - \nabla^3 f(y)\| \leq L\|x - y\|$ for all x, y, then

$$f(x^k) - f^* = \mathcal{O}\left(\frac{1}{k^3}\right).$$

If $\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq 2H\|x - y\|$ for all x, y, then

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

One method for all rates

If $\|\nabla^3 f(x) - \nabla^3 f(y)\| \leq L\|x - y\|$ for all x, y, then

$$f(x^k) - f^* = \mathcal{O}\left(\frac{1}{k^3}\right).$$

If $\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq 2H\|x - y\|$ for all x, y, then

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

If $\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq M = \text{const}$ for all x, y, then

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