Complexity of Cubically Regularized Newton Method for Minimizing Uniformly Convex Functions

Nikita Doikov Yurii Nesterov

Catholic University of Louvain, Belgium

FGS-19, Nice
September 18, 2019

Plan of the talk

1. Introduction: first-order nondegeneracy
2. Regularization of Newton method
3. Universal Cubic Newton
4. Numerical examples

Plan of the talk

1. Introduction: first-order nondegeneracy
2. Regularization of Newton method
3. Universal Cubic Newton
4. Numerical examples

Review: Gradient descent for smooth strongly convex functions

Optimization problem

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

- Problem class: $\mu I \;\preceq\; \nabla^2 f(x) \;\preceq\; LI, \quad \forall x \in \mathbb{R}^n$

- Gradient Method: $x_{k+1} \;=\; x_k \;-\; \frac{1}{\alpha_k}\nabla f(x_k)$

$$= \operatorname*{argmin}_{y\in\mathbb{R}^n}\Big\{ f(x_k) + \langle \nabla f(x_k), y - x_k\rangle + \frac{\alpha_k}{2}\|y - x_k\|^2 \Big\}$$

- Linear convergence rate. Set $\alpha_k = L$. Then

$$f(x_k) - f(x_{k+1}) \;\geq\; \tfrac{1}{2L}\|\nabla f(x_k)\|^2 \;\geq\; \tfrac{\mu}{L}(f(x_k) - f^*).$$

- Problem class: $\mu I \preceq \nabla^2 f(x) \preceq LI, \quad \forall x \in \mathbb{R}^n$.

  The **condition number** of f:

$$\kappa \equiv \frac{\mu}{L} \leq 1.$$

- Iteration complexity. $f(x_N) - f^* \leq \varepsilon$

  Gradient Method: $N = O\left(\frac{1}{\kappa} \log \frac{f(x_0)-f^*}{\varepsilon}\right)$.

  Fast Gradient Method: $N = O\left(\sqrt{\frac{1}{\kappa}} \log \frac{f(x_0)-f^*}{\varepsilon}\right)$.

  Second Order Methods: ?

Damped Newton method

$$x_{k+1} \;=\; x_k - \tfrac{1}{\alpha_k}\big(\nabla^2 f(x_k)\big)^{-1}\nabla f(x_k)$$

$$= \operatorname*{argmin}_{y\in\mathbb{R}^n}\Big\{ f(x_k) + \langle \nabla f(x_k), y - x_k\rangle + \tfrac{\alpha_k}{2}\big\langle \nabla^2 f(x_k)(y - x_k), y - x_k\big\rangle \Big\}$$

- Global convergence. Set $\alpha_k = L/\mu$.

$$\begin{aligned} f(x_k) - f(x_{k+1}) \;&\geq\; \tfrac{\mu}{2L}\langle (\nabla^2 f(x_k))^{-1}\nabla f(x_k), \nabla f(x_k)\rangle \\ &\geq\; \tfrac{\mu}{2L^2}\|\nabla f(x_k)\|^2 \;\geq\; \tfrac{\mu^2}{L^2}\big(f(x_k) - f^*\big). \end{aligned}$$

- Complexity: $N = O\Big(\frac{1}{\boldsymbol{\kappa}^2}\log\frac{f(x_0)-f^*}{\varepsilon}\Big)$. It is worse than GM!

Our work: we use **Regularization of Newton method** [Nesterov & Polyak 06; Grapiglia & Nesterov 17] and show global linear rate of convergence which is **faster** than for Gradient Methods.

Plan of the talk

1. Introduction: first-order nondegeneracy

2. Regularization of Newton method

3. Universal Cubic Newton

4. Numerical examples

Functions with Hölder continuous Hessian

For $\nu \in [0, 1]$ Hessian of f is Hölder continuous of degree ν on a convex set $C \subseteq \mathrm{dom}\, f$ iff

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \;\leq\; \mathcal{H}_f(\nu)\|y - x\|^\nu, \quad \forall x, y \in C.$$

- $\nu = 1$. Lipschitz continuous Hessian:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \;\leq\; \mathcal{H}_f(1)\|y - x\|$$

- $\nu = 0$. Hessian with bounded variation:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \;\leq\; \mathcal{H}_f(0)$$

Note: for $\mu I \;\preceq\; \nabla^2 f(x) \;\preceq LI$ we have

$$\mathcal{H}_f(0) \;\leq\; L - \mu.$$

For $\nu \in [0,1]$:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \;\le\; \mathcal{H}_f(\nu)\|y-x\|^{\nu}, \quad \forall x, y \in C.$$

Bound for Quadratic approximation:

- $|f(y) - Q(x;y)| \;\le\; \frac{\mathcal{H}_f(\nu)}{(1+\nu)(2+\nu)}\|y-x\|^{2+\nu}$, where

$$Q(x;y) \;\stackrel{\mathrm{def}}{=}\; f(x) + \langle \nabla f(x), y-x\rangle + \tfrac{1}{2}\langle \nabla^2 f(x)(y-x), y-x\rangle.$$

Regularized Newton step:

$$T_{\nu,H}(x) \;\stackrel{\mathrm{def}}{=}\; \operatorname*{argmin}_{y\in C}\Big\{ Q(x;y) + \tfrac{H}{(1+\nu)(2+\nu)}\|y-x\|^{2+\nu}\Big\}$$

Iterations: $x_{k+1} \;=\; T_{\nu,H_k}(x_k), \; k \ge 0.$

[Nesterov & Polyak 06; Grapiglia & Nesterov 17]

$f : \mathrm{dom}\, f \to \mathbb{R}$ is called **uniformly convex** of degree $p \geq 2$ on a convex set $C \subseteq \mathrm{dom}\, f$ iff

$$f(y) \;\geq\; f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{\sigma_f(p)}{p} \|x - y\|^p, \qquad \forall x, y \in C.$$

$\sigma_f(p)$ – constant of uniform convexity.

- Strongly convex functions: $p = 2$, $\mu = \sigma_f(2)$.
- Example:

  $$f(x) \;=\; \tfrac{1}{p} \|x - x_0\|^p, \quad p \geq 2$$

  is uniformly convex of degree p with constant $\sigma_f(p) = 2^{2-p}$.
- Sum of convex and uniformly convex functions gives uniformly convex.

$\min_x f(x)$

- Hölder Hessian of degree $\nu \in [0,1]$:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \;\le\; \mathcal{H}_f(\nu)\|x-y\|^\nu, \quad \forall x, y \in C.$$

- Uniformly convex of degree $p \ge 2$:

$$f(y) \;\;\ge\;\; f(x) + \langle \nabla f(x), y - x\rangle + \tfrac{\sigma_f(p)}{p}\|x-y\|^p, \quad \forall x, y \in C.$$

Which ν and p to use?

$$O\Big((\mathcal{H}_f(\nu))^\alpha(\sigma_f(p))^\beta \log\frac{f(x_0)-f^*}{\varepsilon}\Big) \quad \#\text{ of iterations}$$

Physical dimensions: $[f]$ and $[x]$. Then

$[\mathcal{H}_f(\nu)] = \frac{[f]}{[x]^{2+\nu}}$, $[\sigma_f(p)] = \frac{[f]}{[x]^p}$, $\Big[(\mathcal{H}_f(\nu))^\alpha(\sigma_f(p))^\beta\Big] = \frac{[f]^{\alpha+\beta}}{[x]^{\alpha(2+\nu)+\beta p}}$.

$\alpha + \beta = 0$ and $\alpha(2+\nu) + \beta p = 0 \quad \Rightarrow \quad \boxed{\boldsymbol{p = 2 + \nu}}$

Fix $\boldsymbol{\nu \in [0,1]}$. Model of the objective:

$$M_{\nu,H}(x;y) \stackrel{\mathrm{def}}{=} Q(x;y) + \frac{H}{(1+\nu)(2+\nu)}\|y-x\|^{2+\nu}$$

Regularized Newton step: $\quad T_{\nu,H}(x) \stackrel{\mathrm{def}}{=} \operatorname*{argmin}_{y \in C} M_{\nu,H}(x;y).$

Iterations: $\quad \boxed{x_{k+1} = T_{\nu,H_k}(x_k), \quad k \geq 0}$

New result: Theorem 1

- Let $0 < \mathcal{H}_f(\nu) < +\infty$.
- Let $\sigma_f(2+\nu) > 0$.
- Let $0 \leq H_k \leq \beta \mathcal{H}_f(\nu)$ and $f(x_{k+1}) \leq M_{\nu,H_k}(x_k; x_{k+1})$, $\forall k$.

Then: $f(x_K) - f^* \leq \varepsilon$ for

$$K = O\left(\max\{1, \left(\frac{\mathcal{H}_f(\nu)}{\sigma_f(2+\nu)}\right)^{\frac{1}{1+\nu}}\} \cdot \log \frac{f(x_0)-f^*}{\varepsilon}\right).$$

Condition number of degree ν.

Denote

$$\gamma_f(\nu) \stackrel{\mathrm{def}}{=} \frac{\sigma_f(2+\nu)}{\mathcal{H}_f(\nu)}$$

second-order condition number of degree $\nu \in [0, 1]$.

Properties (unbounded case, $\mathrm{diam}\, C = +\infty$):

1. $\gamma_f(\nu) \leq \frac{1}{1+\nu}, \quad \nu \in (0, 1]$.
2. Let for some $\nu > 0$ we have $\gamma_f(\nu) > 0$. Then

$$\gamma_f(\nu') = 0 \qquad \forall \nu' \in [0, 1] \setminus \{\nu\}.$$

<u>In practice we may not know exact value for ν.</u> Universal methods?

Plan of the talk

1. Introduction: first-order nondegeneracy

2. Regularization of Newton method

3. Universal Cubic Newton

4. Numerical examples

Composite optimization problem

$$F^* = \min_{x \in \mathrm{dom}\, F} F(x) \stackrel{\mathrm{def}}{=} f(x) + h(x).$$

- f is a smooth part, $\gamma_f(\nu) > 0$ for <u>some</u> $\nu \in [0,1]$.
- $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a composite part (possibly nonsmooth but *simple*).

Cubically regularized (composite) Newton:

$$\boxed{x_{k+1} = T_{H_k}(x_k), \quad k \geq 0}$$

for

$$T_H(x) \stackrel{\mathrm{def}}{=} \operatorname*{argmin}_{y \in \mathrm{dom}\, F} M_H(x; y)$$

where $M_H(x; y) \stackrel{\mathrm{def}}{=} Q(x; y) + \frac{H}{6}\|y - x\|^3 + h(y)$.

$$Q(x; y) \stackrel{\mathrm{def}}{=} f(x) + \langle \nabla f(x), y - x \rangle + \tfrac{1}{2}\langle \nabla^2 f(x)(y - x), y - x \rangle$$

$$M_H^*(x) \stackrel{\mathrm{def}}{=} M_H(x; T_H(x))$$

Algorithm 1 Adaptive cubically regularized Newton method

Initialization: Choose $x_0 \in \mathrm{dom}\, F$, $H_0 > 0$.

Iterations: $k \geq 0$.

1: Find minimum integer $i_k \geq 0$ s.t. it holds

$$F(T_{H_k 2^{i_k}}(x_k)) \leq M^*_{H_k 2^{i_k}}(x_k).$$

2: Make the Cubic step $x_{k+1} := T_{H_k 2^{i_k}}(x_k)$.

3: Set $H_{k+1} := H_k 2^{i_k - 1}$.

[Nesterov-Polyak 06; Cartis-Gould-Toint 11; Grapiglia-Nesterov 17]

Global convergence

New result: Theorem 2

- Let $\gamma_f(\nu) > 0$ **for some $\nu \in [0,1]$**.

Then $F(x_K) - F^* \leq \varepsilon$ for

$$K \;=\; O\left(\max\{1, \left(\gamma_f(\nu)\right)^{\frac{-1}{1+\nu}}\} \cdot \log \frac{F(x_0)-F^*}{\varepsilon}\right).$$

Total number of oracle calls N_K during K iterations is bounded as:

$$N_K \;\leq\; 2K + \log_2 \frac{\kappa_f(\nu)}{\varepsilon^{(1-\nu)/(2+\nu)} H_0}$$

The algorithm **does not use** real $\nu \in [0,1]$ and adapt to **the best possible bound:**

$$O\left(\max\{1, \inf_{\nu\in[0,1]} \left(\gamma_f(\nu)\right)^{\frac{-1}{1+\nu}}\} \cdot \frac{F(x_0)-F^*}{\varepsilon}\right).$$

Classic Newton method achieves local quadratic convergence under the conditions

1. Strong convexity: $\sigma_f(2) > 0$.
2. Hessian is Lipschitz continuous: $\mathcal{H}_f(1) < +\infty$.
3. Initial point x_0 *close* to the optimum x^*.

For **Adaptive cubically regularized Newton** we have:

$$F(x_{k+1}) - F^* \quad \leq \quad \frac{9\mathcal{H}_f^2(1)}{2\sigma_f^3(2)}\big(F(x_k) - F^*\big)^2, \quad \forall k.$$

And the region of quadratic convergence is

$$\mathcal{Q} \quad = \quad \left\{ x \in \mathrm{dom}\, F \;:\; F(x) - F^* \leq \frac{2\sigma_f^3(2)}{9\mathcal{H}_f^2(1)} \right\}.$$

Number of iterations to find ε-solution:

$$O\left(\log\log \frac{\sigma_f^3(2)}{\mathcal{H}_f^2(1)\varepsilon}\right).$$

The same order as for Classic Newton.

Cubic Newton vs. Gradient Method

Composite minimization problem: $\min_{x \in \mathbb{R}^n} F(x) \overset{\text{def}}{=} f(x) + h(x)$.

- Problem class: $\mu I \preceq \nabla^2 f(x) \preceq LI$. **Gradient Method** needs $O\Big(\frac{L}{\mu} \log \frac{F(x_0) - F^*}{\varepsilon}\Big)$ iterations to find ε-solution.

- Problem class: for some $\nu \in [0, 1]$ it holds $\gamma_f(\nu) > 0$.
  Adaptive Cubic Newton needs
  $O\Big(\max\{1, \inf\limits_{\nu \in [0,1]} \big(\gamma_f(\nu)\big)^{\frac{-1}{1+\nu}}\} \log \frac{F(x_0) - F^*}{\varepsilon}\Big)$ iterations to find ε-solution, without knowledge of ν. (Until entering $\mathcal{Q}$).

Note:

$$\big(\gamma_f(0)\big)^{-1} \quad \leq \quad \frac{L-\mu}{\mu} \; < \; \frac{L}{\mu}.$$

Consider $\tilde{f}(x) = f(x) + \frac{1}{2}\langle Ax, x\rangle$. Then

$$\tilde{L} = L + \lambda_{\max}(A) \quad \text{but} \quad \big(\gamma_{\tilde{f}}(\nu)\big)^{-1} \leq \big(\gamma_f(\nu)\big)^{-1}.$$

$\Rightarrow$ Cubic Newton is not affected by any quadratic parts of F.

Composite minimization problem: $\min_{x \in \mathbb{R}^n} F(x) \stackrel{\mathrm{def}}{=} f(x) + h(x)$.

- Arithmetical complexity of every iteration: $O(n^3)$.
- Strongly convex functions with Lipschitz Hessian:

  local quadratic convergence.
- Strongly convex functions with bounded Hessian,

$$\mu I \preceq \nabla^2 f(x) \preceq LI :$$

$O\left(\left(\frac{L}{\mu} \right)^2 \log \frac{F(x_0) - F^*}{\varepsilon} \right)$ for **Damped Newton method**,

$O\left(\frac{L - \mu}{\mu} \log \frac{F(x_0) - F^*}{\varepsilon} \right)$ for **Adaptive Cubic Newton**.

Plan of the talk

1. Introduction: first-order nondegeneracy

2. Regularization of Newton method

3. Universal Cubic Newton

4. Numerical examples

Log-sum-exp

$$\min_{x \in \mathbb{R}^n} f(x) \;=\; \log \left(\sum_{i=1}^{m} e^{\langle a_i, x \rangle} \right) + \tfrac{\mu}{2} \langle x, x \rangle .$$

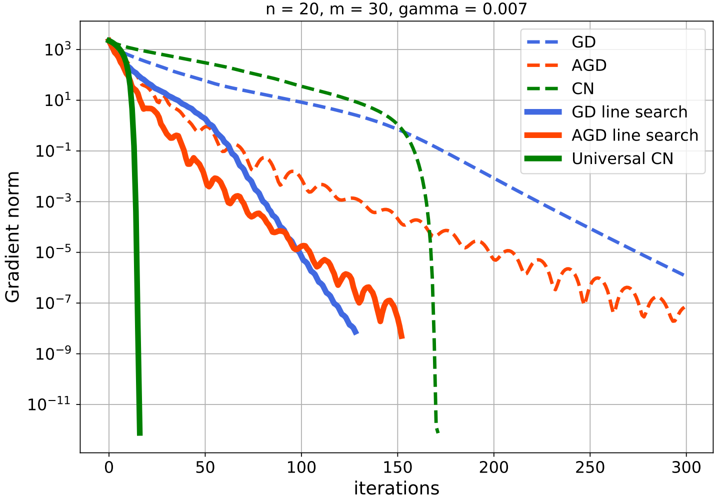
- $a_1, \dots, a_m \in \mathbb{R}^n$ — given data.
- Denote $B \equiv \sum_{i=1}^{m} a_i a_i^T \succeq 0$, and use $\|x\| \equiv \langle Bx, x \rangle^{1/2}$.
- We have

$$\mathcal{H}_f(0) \;\leq\; 1, \qquad \mathcal{H}_f(1) \;\leq\; 2.$$

- For any $\nu \in [0, 1]$:

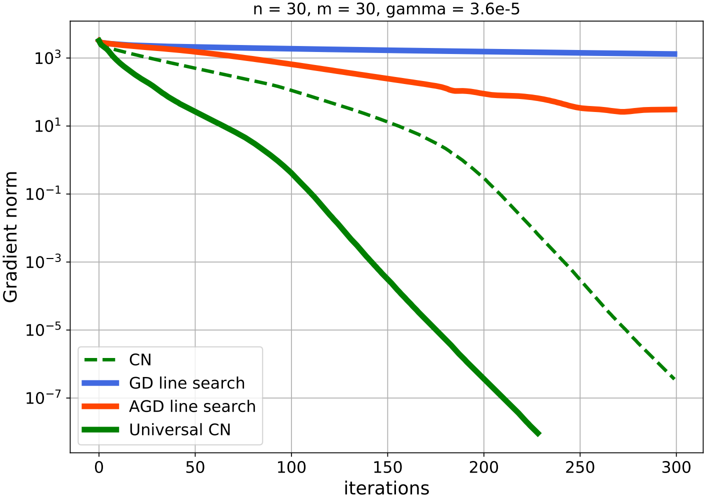
$$\mathcal{H}_f(\nu) \;\;\leq\;\; \bigl(\mathcal{H}_f(0)\bigr)^{1-\nu} \cdot \bigl(\mathcal{H}_f(1)\bigr)^{\nu} \;\leq\; 2^{\nu}.$$

$$\min_{x \in \mathbb{R}^n} f(x) \;=\; \log\left(\sum_{i=1}^{m} e^{\langle a_i, x\rangle}\right) + \frac{\mu}{2}\langle x, x\rangle.$$



Log-sum-exp, ill-conditioned problem

$$\min_{x \in \mathbb{R}^n} f(x) \;=\; \log\left(\sum_{i=1}^{m} e^{\langle a_i, x\rangle}\right) + \frac{\mu}{2}\langle x, x\rangle.$$



- Second-order condition number $\gamma_f(\nu)$ of degree $\nu \in [0, 1]$.
- Adaptive cubically regularized Newton achieves linear rate of convergence when $\gamma_f(\nu) > 0$ for some $\nu \in [0, 1]$. The main complexity factor is $\big(\gamma_f(\nu)\big)^{\frac{-1}{1+\nu}}$.
- The algorithm is universal: it does not need to know right $\nu \in [0, 1]$ and any other parameters of the class.
- For the class $f \in \mathcal{S}^{2,1}_{\mu,L}$: $\mu I \preceq \nabla^2 f(x) \preceq LI$ provided complexity bounds for Cubic Newton are always better than for Gradient Method.

Thank you for your attention!