Super-Universal Regularized Newton Method

Nikita Doikov (EPFL, Switzerland)

Joint work with
Konstantin Mishchenko (Samsung, UK) and
Yurii Nesterov (UCLouvain, Belgium)

EUROPT Workshop on Continuous Optimization, Budapest
August 25, 2023

Convex Optimization Problem

$$\min_x f(x), \qquad x \in \mathbb{R}^n$$

f is convex and differentiable.

The Goal: efficient **second-order** optimization methods with global convergence guarantees.

- **This work:** a very simple variant of the Newton Method that automatically achieves *fast global rates* for wide classes of convex problems.

Notation

Fix matrix $B = B^T \succ 0$ and denote the Euclidean norm

$$\|x\| \;=\; \langle Bx, x\rangle^{1/2}, \qquad x \in \mathbb{R}^n.$$

$\Rightarrow$ induced norm for multilinear forms.

- Gradients:

$$\|\nabla f(x)\|_* \;=\; \max_{\|h\|\leq 1} \langle \nabla f(x), h\rangle \;=\; \langle \nabla f(x), B^{-1}\nabla f(x)\rangle^{1/2}$$

- High-order Tensors: $\|D^p f(x)\| = \max\limits_{\|h\|\leq 1} |D^p f(x)[h]^p|$

Assume that the pth derivative is Lipschitz continuous ($p \geq 1$):

$$\|D^p f(x) - D^p f(y)\| \;\leq\; L_p\|x - y\|, \qquad \forall x, y \in \mathbb{R}^n$$

The Plan

1. Tensor Methods in Convex Optimization
2. Super-Universal Newton Method
3. Uniformly Convex Functions
4. Experiments and Conclusions

Basic Tensor Method

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

Global upper model of our function, for $H \geq L_p$:

$$f(y) \;\; \leq \;\; \Omega_p(x;y) + \tfrac{H}{(p+1)!}\|y-x\|^{p+1}, \qquad \forall x, y \in \mathbb{R}^n,$$

where Ω_p is Taylor polynomial:

$$\Omega_p(x;y) \;\; \stackrel{\text{def}}{=} \;\; f(x) + \langle \nabla f(x), y-x \rangle + \sum_{i=2}^{p} \tfrac{1}{i!} D^i f(x)[y-x]^i.$$

Basic Tensor Method of order $p \geq 1$. Iterate, $k \geq 0$:

$$\boxed{x_{k+1} \;\; = \;\; \operatorname*{argmin}_{y \in \mathbb{R}^n} \Big\{ \, \Omega_p(x_k;y) + \tfrac{H}{(p+1)!}\|y-x_k\|^{p+1} \, \Big\}}$$

[Birgin et al., 2017; Nesterov, 2019; Cartis-Gould-Toint, 2020; Grapiglia-Nesterov, 2020]

Global Convergence

Theorem. Let $H := L_p \Rightarrow$ global rate of the Tensor Method:

$$f(x_k) - f^\star \;\; \leq \;\; \mathcal{O}\big(1/k^p\big)$$

$p = 1$: **the Gradient Method**. $x_{k+1} = x_k - \frac{1}{H} B^{-1} \nabla f(x_k)$

$p = 2$: **the Cubic Newton** [Nesterov-Polyak, 2006]

$$x_{k+1} \;\; = \;\; x_k - \Big(\nabla^2 f(x_k) + \tfrac{H r_k}{2} B\Big)^{-1} \nabla f(x_k),$$

where r_k is the solution to a univariate dual problem.

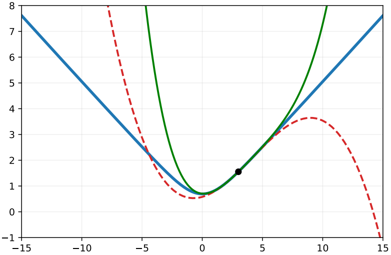
$p = 3$: **the Third-Order Tensor Method**

$$x_{k+1} \;\; = \;\; \operatorname*{argmin}_y \Big\{ \, \Omega_3(x_k; y) + \tfrac{H}{24} \|y - x_k\|^4 \, \Big\}$$

$\ldots$

Convexity of the High-Order Model

Note: $\Omega_p(x; y)$ is **nonconvex** for $p \geq 3$.



Theorem [Nesterov, 2018]: Let $f(\cdot)$ be convex and $H \geq pL_p$. Then

$$M(y) \quad := \quad \Omega_p(x; y) + \tfrac{H}{(p+1)!}\|y - x\|^{p+1}$$

is convex in y.

- We can use efficient tools of Convex Optimization to solve the subproblem

Gradient Regularization Technique

Step of the Cubic Newton:

$$x^+ \;=\; x - \big(\nabla^2 f(x) + \lambda B\big)^{-1} \nabla f(x_k),$$

where $\lambda = \frac{H}{2}\|x^+ - x\|$. Note that

$$\begin{aligned} \|x^+ - x\| \;&=\; \|\big(\nabla^2 f(x) + \lambda B\big)^{-1} \nabla f(x)\| \\ &\leq\; \tfrac{1}{\lambda}\|\nabla f(x)\|_* \;=\; \tfrac{2}{H\|x^+ - x\|}\|\nabla f(x)\|_*. \end{aligned}$$

Hence, we have an upper bound:

$$\boxed{\lambda \;=\; \frac{H}{2}\|x^+ - x\| \leq \sqrt{\frac{H\|\nabla f(x)\|_*}{2}}}$$

Gradient Regularization.

$$x^+ \;=\; x - \left(\nabla^2 f(x) + \sqrt{\frac{H\|\nabla f(x)\|_*}{2}} B\right)^{-1} \nabla f(x).$$

- One matrix inversion; fast global rates

[Mishchenko, 2021; D-Nesterov, 2021]

Weakness of the Third Derivative

Third derivative is Lipschitz continuous:

$$\|D^3 f(x) - D^3 f(y)\| \quad \leq \quad L_3 \|x - y\|, \qquad x, y \in \mathbb{R}^n$$

and convexity: $\nabla^2 f(x) \succeq 0$. Then,

$$|D^3 f(x)[h]^3| \quad \leq \quad \tfrac{1}{\tau} \nabla^2 f(x)[h]^2 + \tfrac{\tau}{2} L_3 \|h\|^4, \qquad \forall x \in \mathbb{R}^n, \tau > 0.$$

$\Rightarrow$ we can upper bound the third derivative in Taylor's approximation:

$$\begin{aligned} f(y) \quad &\leq \Omega_3(x; y) + \tfrac{H}{24} \|y - x\|^4 \\ &\leq f(x) + \langle \nabla f(x), y - x \rangle + \left(\tfrac{1}{2} + \tfrac{1}{6\tau} \right) \nabla^2 f(x)[y - x]^2 + \left(\tau + \tfrac{1}{2} \right) \tfrac{L_3}{12} \|y - x\|^4 \end{aligned}$$

- Purely second-order method (the same fast global rates)

The Plan

1. Tensor Methods in Convex Optimization
2. Super-Universal Newton Method
3. Uniformly Convex Functions
4. Experiments and Conclusions

Let $p \in \{2, 3\}$. Fix $\nu \in [0, 1]$ and define

$$L_{p,\nu} \quad \stackrel{\mathrm{def}}{=} \quad \sup_{x \neq y} \frac{\|D^p f(x) - D^p f(y)\|}{\|x - y\|^\nu}$$

$L_{p,\nu}$ is log-convex function of ν: for any $0 \leq \nu_1 \leq \nu_2 \leq 1$ we have

$$L_{p,\nu} \quad \leq \quad \left[L_{p,\nu_1}\right]^{\frac{\nu_2 - \nu}{\nu_2 - \nu_1}} \left[L_{p,\nu_2}\right]^{\frac{\nu - \nu_1}{\nu_2 - \nu_1}} \qquad \forall \nu \in [\nu_1, \nu_2].$$

Define M_q, for $2 \leq q \leq 4$:

$$M_{2+\nu} \quad \stackrel{\mathrm{def}}{=} \quad L_{2,\nu}, \qquad \nu \in [0, 1),$$

$$M_{3+\nu} \quad \stackrel{\mathrm{def}}{=} \quad L_{3,\nu}, \qquad \nu \in [0, 1].$$

Main Assumption: $\boxed{\inf\limits_{2 \leq q \leq 4} M_q \; < \; +\infty}$.

Fix $q \in [2, 4]$. Choose $M_q > 0$.

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

$$x_{k+1} \;\; = \;\; x_k - \Big(\nabla^2 f(x_k) + \lambda_k B\Big)^{-1} \nabla f(x_k),$$

with $\lambda_k := (6M_q \|\nabla f(x_k)\|_*^{q-2})^{\frac{1}{q-1}}$.

Theorem. Global convergence rate:

$$f(x_k) - f^\star \;\; \leq \;\; 6M_q D^q \Big(\tfrac{32(q-1)}{k}\Big)^{q-1} \; + \; \|\nabla f(x_0)\| D \exp\big(-\tfrac{k}{4}\big)$$

where D is diameter of the initial sublevel set.

Note: $\|\nabla f(x_0)\| D \exp\big(-\frac{k}{4}\big) \leq \varepsilon$ for $k \geq 4 \ln \frac{\|\nabla f(x_0)\| D}{\varepsilon}$.

Which problem class to choose?

Global rate: $f(x_k) - f^* \leq O\left(\frac{M_q D^q}{k^{q-1}}\right), \quad 2 \leq q \leq 4.$

$q = 2$: **Bounded variation of the Hessian**

$$\Rightarrow \qquad f(y) \;\leq\; \Omega_2(x; y) + \tfrac{M_2}{2}\|y - x\|^2, \quad \forall x, y$$

$q = 3$: **Lipschitz continuity of the Hessian**

$$\Rightarrow \qquad f(y) \;\leq\; \Omega_2(x; y) + \tfrac{M_3}{6}\|y - x\|^3, \quad \forall x, y$$

$q = 4$: **Lipschitz continuity of the third derivative**

$$\Rightarrow \qquad f(y) \;\leq\; \Omega_3(x; y) + \tfrac{M_4}{24}\|y - x\|^4, \quad \forall x, y$$

Our objective can belong to several problem classes simultaneously!

Main Lemma

Consider the step $x^+ = x - \left(\nabla^2 f(x) + \lambda B\right)^{-1} \nabla f(x)$
with

$$\lambda \quad := \quad H \|\nabla f(x)\|_*^{\alpha}, \quad 0 \leq \alpha \leq 1$$

Lemma. Let $\frac{q-2}{q-1} \leq \alpha \leq 1$, and $H \geq \left(6M_q\right)^{\frac{1}{q-1}} \left(\frac{1}{\|\nabla f(x)\|_*}\right)^{\alpha - \frac{q-2}{q-1}}$.

Then

$$\langle \nabla f(x^+), x - x^+ \rangle \quad \geq \quad \tfrac{1}{4\lambda} \|\nabla f(x^+)\|_*^2.$$

Note: by convexity, we have

$$f(x) - f(x^+) \quad \geq \quad \langle \nabla f(x^+), x - x^+ \rangle \quad \geq \quad \tfrac{1}{4\lambda} \|\nabla f(x^+)\|_*^2.$$

Initialization. Choose $x_0 \in \mathbb{R}^n$. Fix arbitrary $\alpha \in \left[\frac{2}{3}, 1\right]$, $H_0 > 0$.

Iteration $k \geq 0$:

Find smallest $j_k \geq 0$ s.t. for $\lambda_k := 4^{j_k} H_k \|\nabla f(x_k)\|_*^\alpha$ and

$$x^+ \;\; = \;\; x_k - \Big(\nabla^2 f(x_k) + \lambda_k B\Big)^{-1} \nabla f(x_k)$$

it holds

$$\langle \nabla f(x^+), x_k - x^+ \rangle \;\; \geq \;\; \tfrac{1}{4\lambda_k} \|\nabla f(x^+)\|_*^2.$$

Set $x_{k+1} = x^+$ and $H_{k+1} = \frac{4^{j_k} H_k}{4}$.

Theorem. The method is well defined. We have

$$f(x_k) - f^\star \;\; \leq \;\; 6 M_q D^q \Big(\tfrac{32(q-1)}{k}\Big)^{q-1} \; + \; \|\nabla f(x_0)\| D \exp\big(-\tfrac{k}{4}\big)$$

- The average number of adaptive steps per iterations is two.

The method does not know q. To reach $f(x_k) - f^\star \leq \varepsilon$, we need

$$k \;\; = \;\; \mathcal{O}\Big(\inf_{q \in [2,4]} \big[\tfrac{M_q D^q}{\varepsilon}\big]^{1/q} \; + \; \ln \tfrac{1}{\varepsilon}\Big)$$

second-order oracle calls.

The Plan

1. Tensor Methods in Convex Optimization
2. Super-Universal Newton Method
3. Uniformly Convex Functions
4. Experiments and Conclusions

Strictly Convex Functions

Initial sublevel set $\mathcal{F}_0 \stackrel{\mathrm{def}}{=} \left\{ x \,:\, f(x) \le f(x_0) \right\}$ and its diameter:

$$D \quad \stackrel{\mathrm{def}}{=} \quad \sup_{x,y \in \mathcal{F}_0} \|x - y\|.$$

Symmetrized Bregman Divergence:

$$\beta_f(x,y) \quad \stackrel{\mathrm{def}}{=} \quad \langle \nabla f(x) - \nabla f(y), x - y \rangle \quad > \quad 0.$$

and normalization:

$$\boxed{\xi_f(x,y) \quad \stackrel{\mathrm{def}}{=} \quad \frac{1}{V_F} \beta_f(x,y)}$$

where $V_f \stackrel{\mathrm{def}}{=} \sup\limits_{x,y \in \mathcal{F}_0} \beta_f(x,y)$.

Relative s-size:

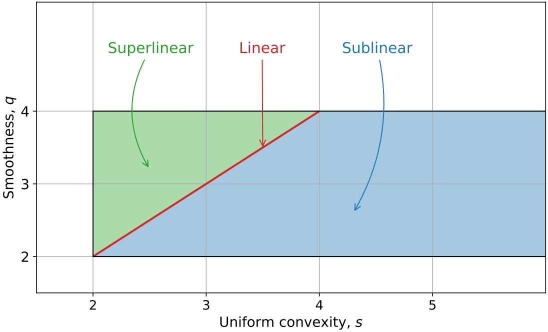
$$D_s \quad \stackrel{\mathrm{def}}{=} \quad \sup_{x \neq y} \|x - y\| \cdot \xi_f(x,y)^{-1/s}, \qquad s \ge 2.$$

Assumption: $D_s < +\infty$ for some $s \in [2, +\infty]$.

Summary of Complexities

- Level of smoothness $2 \le q \le 4$ is fixed.

$2 \le s < q$	$s = q$	$q < s < \infty$	$s = \infty$
$\left(M_q \frac{D_s^s D^{q-s}}{V_F}\right)^{\frac{1}{q-1}} + \ln\ln\frac{1}{\varepsilon}$	$\left(M_q \frac{D_q^q}{V_F}\right)^{\frac{1}{q-1}} \ln\frac{1}{\varepsilon}$	$\left(M_q \frac{D_s^q}{(V_F^q \varepsilon^{s-q})^{1/s}}\right)^{\frac{1}{q-1}}$	$\left(M_q \frac{D^q}{\varepsilon}\right)^{\frac{1}{q-1}}$



The Plan

1. Tensor Methods in Convex Optimization

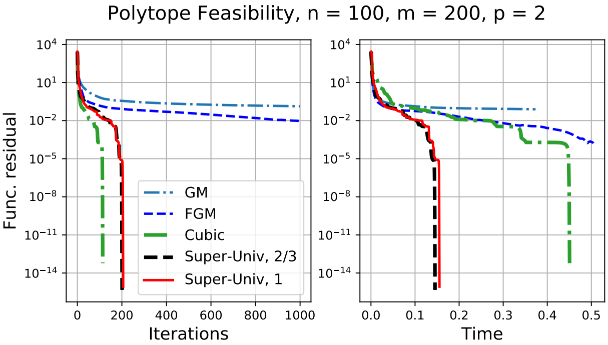
2. Super-Universal Newton Method

3. Uniformly Convex Functions

4. Experiments and Conclusions

$$\min_{x \in \mathbb{R}^n} \Big[\ f(x) \quad := \quad \sum_{i=1}^{m} \big(\langle a_i, x\rangle - b_i\big)_+^p\ \Big],$$

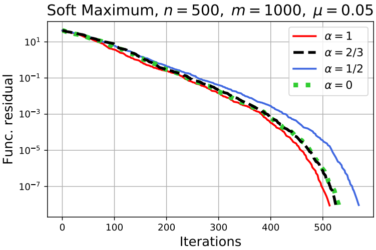
where $(t)_+ \stackrel{\mathrm{def}}{=} \max\{0, t\}$

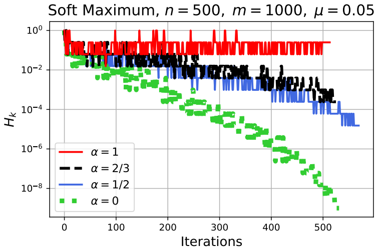


$$\min_x f(x) \quad := \quad \mu \ln\Big(\sum_{i=1}^{m} \exp\big(\tfrac{\langle a_i, x\rangle - b_i}{\mu}\big) \Big) \quad \approx \quad \max_{1 \le i \le m} \big[\langle a_i, x\rangle - b_i \big].$$

Set $B := \sum_{i=1}^{m} a_i a_i^T \succeq 0$ (the primal norm $\|x\| = \langle Bx, x\rangle^{1/2}$)

- $M_q \le \frac{12}{\mu^{q-1}}, \qquad \forall q \in [2, 4]$





Conclusions

1. To globalize the Newton's method we need to do regularization
   - Cubic Newton — explicit regularizer, $\| \cdot \|^3$
   - We can reduce the power to $\| \cdot \|^2$ by **Gradient Regularization**

2. Method $\leftrightarrow$ Problem class

3. **Super-universal** methods: adjust automatically to the best problem class
   - Achieved by using an adaptive search

4. We can solve Composite Problems

$$\min_x \left\{ F(x) \quad := \quad f(x) + \psi(x) \right\}$$

where ψ is a nonsmooth part (e.g. ℓ_1-regularizer; indicator of a convex set)

Open Questions

1. **Accelerated** (optimal) Super-Universal second-order methods?

   [Grapiglia-Nesterov, 2019; Carmon-Hausler-Jambulapati-Jin-Sidford, 2022]

2. **Quasi-Newton** methods?

   - Nonasymptotic complexity bounds: local superlinear rates

     [Rodomanov-Nesterov, 2021]

   - Lazy Hessian updates: update Hessian once per n steps

     [D-Chayti-Jaggi, 2022]

3. Consequences for **nonconvex** and **stochastic** optimization?

Thank you very much for your attention!