Minimizing quasi-self-concordant functions by gradient regularization of Newton method

Nikita Doikov
EPFL, Switzerland

33rd European Conference on Operational Research
EURO, Copenhagen
July 1, 2024

Outline

I. Introduction: Newton's method

II. Quasi-self-concordant functions

III. Acceleration

IV. Experiments and conclusions

Optimization problem

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

f is convex and differentiable

- Fix a symmetric matrix $B = B^\top \succ 0$. Global Euclidean norms:

$$\|u\| \;:=\; \langle Bu, u\rangle^{1/2}, \qquad \|s\|_* \;:=\; \langle s, B^{-1}s\rangle^{1/2}$$

For example, $B := I$

Newton's method. Iterate, with some $\beta_k \geq 0$:

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

[Newton, 1669; Raphson, 1690; Fine-Bennett, 1916; Kantorovich, 1948]

- Local quadratic convergence when $\beta_k \to 0$
- Globalization $\beta_k > 0$ [Levenberg, 1944; Marquardt, 1963]

Global complexity bounds? $\Leftrightarrow$ **a suitable problem class?**

I. Functions with **Hölder Hessian**, for $\nu \in [0, 1]$:

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \quad \leq \quad L_{2,\nu}\|x - y\|^{\nu}, \qquad \forall x, y \in \mathbb{R}^n$$

[Nesterov-Polyak, 2006; Cartis-Gould-Toint, 2011; Grapiglia-Nesterov, 2017; D-Nesterov, 2021]

- $\nu = 1$: Lipschitz Hessian (**Cubic Regularization**)
- $\nu = 0$: Functions with bounded variation of the Hessian

NB: $L_{2,0} \leq 2L_1$, where L_1 is the Lipschitz constant of the gradient

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

Theorem [D-Mishchenko-Nesterov, 2022]. Set

$$\boxed{\beta_k := \left(6L_{2,\nu}\|\nabla f(x_k)\|_*^{\nu}\right)^{\frac{1}{\nu+1}}}$$

Then, we have the global rate:

$$f(x_k) - f^\star \quad \leq \quad 6L_{2,\nu}D^{2+\nu}\left(\frac{32(1+\nu)}{k}\right)^{1+\nu} \quad + \quad \|\nabla f(x_0)\|D\exp\left(-\frac{k}{4}\right)$$

II. Functions with **Hölder Third Derivative**, for $\nu \in [0, 1]$:

$$\|\nabla^3 f(x) - \nabla^3 f(y)\| \quad \leq \quad L_{3,\nu}\|x - y\|^\nu, \qquad \forall x, y \in \mathbb{R}^n$$

First attempt: High-order **Tensor Methods**:

$$x_{k+1} \quad = \quad x_k + \operatorname*{argmin}_h \Big[\sum_{i=1}^{3} \tfrac{1}{i!}\nabla^i f(x_k)[h]^i + \tfrac{L_{3,\nu}}{(1+\nu)(3+\nu)}\|h\|^{3+\nu} \Big]$$

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

- Global rate: $f(x_k) - f^\star \leq \mathcal{O}\big(\frac{L_{3,\nu} D^{3+\nu}}{k^{2+\nu}}\big) \Rightarrow \mathcal{O}(1/k^3)$ for $\nu = 1$
- The inner subproblem is convex and efficiently solvable [Nesterov, 2019]

Recent advancements: **no need in third-order information $\nabla^3 f$**

$$\boxed{x_{k+1} \quad = \quad x_k - \Big(\nabla^2 f(x_k) + (6L_{3,\nu}\|\nabla f(x_k)\|_*^{1+\nu})^{\frac{1}{2+\nu}} B\Big)^{-1}\nabla f(x_k)}$$

- The same global rates! [D-Mishchenko-Nesterov, 2022]

- Instead of choosing ν, we can use a simple **adaptive search**:

Init: Choose $x_0 \in \mathbb{R}^n$, $g_0 = \|\nabla f(x_0)\|_*$, and $\sigma_0 > 0$.
Iteration, $k \geq 0$:

1. Find smallest $j_k \geq 0$ s.t. for $\beta_k := 4^{j_k} \sigma_k g_k$ and for
$$x^+ \;=\; x_k - \left[\nabla^2 f(x_k) + \beta_k B\right]^{-1} \nabla f(x_k)$$
it holds
$$\langle \nabla f(x^+), x_k - x^+ \rangle \;\geq\; \frac{1}{2\beta_k} \|\nabla f(x^+)\|_*^2.$$
2. Set $x_{k+1} = x^+$, $g_{k+1} = \|\nabla f(x^+)\|_*$, and $\sigma_{k+1} = \frac{4^{j_k}\sigma_k}{4}$.

[D-Mishchenko-Nesterov, 2022]

- The method does not need to know any parameters
- Automatic adjustment to the right problem class
- In average: one extra oracle call per iteration

Global complexities: Summary

Classical Newton's method

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

with gradient regularization $\beta_k \propto \|\nabla f(x_k)\|_*^{\alpha}$

- fix α according to the problem class
- use adaptive search

Global Complexity: $f(x_k) - f^\star \le \varepsilon$?

1. Bounded variation of the Hessian: $k = \mathcal{O}\left(\frac{L_{2,0} D^2}{\varepsilon}\right)$
2. Lipschitz Hessian: $k = \mathcal{O}\left(\left[\frac{L_{2,1} D^3}{\varepsilon}\right]^{1/2}\right)$
3. Lipschitz Third Derivative: $k = \mathcal{O}\left(\left[\frac{L_{3,1} D^4}{\varepsilon}\right]^{1/3}\right)$
4. . . . **Can we do better? Yes!**

Outline

Functions with Lipschitz Hessian:

$$\nabla^3 f(x)[u,u,u] \quad \leq \quad L_{2,1}\|u\|^3, \qquad \forall x, u$$

- Fixed global norm (no affine-invariance) $\|u\| := \langle Bu, u\rangle^{1/2}$
- Main example: $f(x) = \frac{1}{3}|x|^3$

Self-Concordant functions [Nesterov-Nemirovski, 1994]:

$$\nabla^3 f(x)[u,u,u] \quad \leq \quad M_{\mathfrak{sc}}\langle \nabla^2 f(x)u, u\rangle^{3/2} \quad \equiv \quad M_{\mathfrak{sc}}\|u\|_x^3, \qquad \forall x, u$$

- Affine-invariant
- Efficiency of the damped Newton method for logarithmic barriers, e.g. $f(x) = -\ln x$

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

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

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

- Combination of the Lipschitzness and classic Self-Concordance

[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 $f'''(x)=f''(x)=e^x \Rightarrow \boxed{M=1}$.

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

$$f'(x) \;=\; \tfrac{1}{1+e^{-x}}, \qquad f''(x) \;=\; f'(x)\cdot(1-f'(x)),$$

$$f'''(x) \;=\; f''(x)\cdot(1-2f'(x)).$$

Thus

$$|f'''(x)| \;=\; f''(x)\cdot\left|1-\tfrac{2}{1+e^{-x}}\right| \;\leq\; f''(x) \;\Rightarrow\; \boxed{M=1}.$$

Example 3: (Generalized Linear Models):

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

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

Example 4: (Soft Maximum):

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

$f(x)$ is quasi-SC with $\boxed{M = \frac{2}{\mu}}$ for $B := \sum_{i=1}^{m} a_i a_i^\top$.

Example 5: (Matrix Scaling, $A \in \mathbb{R}_+^{n \times n}$):

$$f(x, y) \quad = \quad \sum_{1 \le i,j \le n} A_{ij} e^{x_i - x_j}, \qquad x, y \in \mathbb{R}^n$$

is quasi-SC with $\boxed{M = \sqrt{2}}$ for $B := I$.

1. $f(\cdot) = f_1(\cdot) + f_2(\cdot)$ is quasi-SC with $M = \max\{M_1, M_2\}$

2. Adding to f an arbitrary convex quadratic function does not change M

3. **Scale-invariance:** $f(\cdot) \mapsto cf(\cdot)$, $c > 0$, does not change M

4. For an **affine substitution**, $f(x) = g(Ax + b)$, we need to update the global norm:

$$B_f \;=\; A^\top B_g A$$

   (no affine invariance)

Main Bounds

Lemma. for quasi-SC f we have, for any x, y:

$$\nabla^2 f(x) e^{-M\|x-y\|} \preceq \nabla^2 f(y) \preceq \nabla^2 f(x) e^{M\|x-y\|}$$

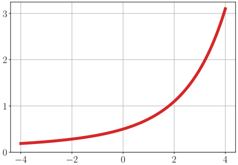
$\Rightarrow$ the Hessian is stable: For any x, y s.t. $\|x-y\| \leq r := \frac{1}{M}$ it holds

$$\tfrac{1}{e}\nabla^2 f(x) \preceq f(y) \preceq e\nabla^2 f(x).$$

[Cohen-Madry-Tsipras-Vladu, 2017; Karimireddy-Stich-Jaggi, 2018]

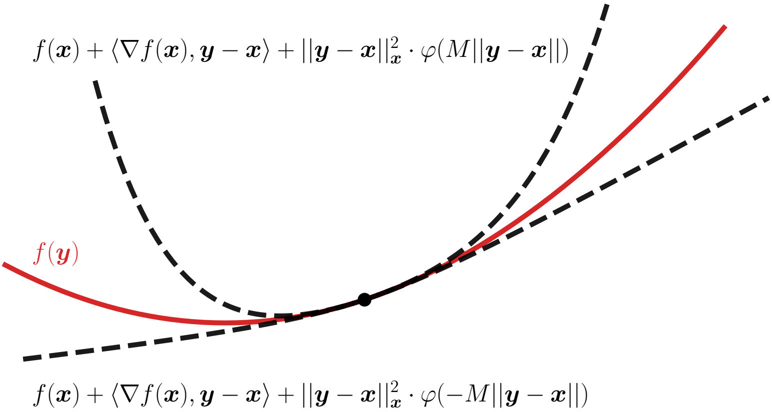
Define $\varphi(t) := \frac{e^t - t - 1}{t^2} > 0$

- convex
- monotone



Bounds on the Function

Using $\varphi(t) := \frac{e^t - t - 1}{t^2} > 0$, we have **global upper and lower second-order models:**



Gradient Regularization

Problem: $\min_x f(x)$, where f is quasi-SC

Consider one regularized Newton step, for $\beta \geq 0$:

$$x^+ = \operatorname*{argmin}_y \Big[\langle \nabla f(x), y - x \rangle + \tfrac{1}{2}\|y - x\|_x^2 + \tfrac{\beta}{2}\|y - x\|^2 \Big]$$

$$\Leftrightarrow \qquad x^+ = x - [\nabla^2 f(x) + \beta B]^{-1} \nabla f(x)$$

Lemma. Set $\beta := \sigma \|\nabla f(x)\|_*$ and $\sigma \geq M$. Then,

1. $\|x^+ - x\| \leq \frac{1}{M}$
2. $\|x^+ - x\|_x^2 \leq \frac{\|\nabla f(x)\|_*}{M}$
3. $\langle \nabla f(x^+), x - x^+ \rangle \geq \frac{1}{2\beta} \|\nabla f(x^+)\|_*^2$

NB: by convexity, $f(x) - f(x^+) \geq \frac{1}{2\beta}\|\nabla f(x^+)\|_*^2$

Main result

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

Theorem. Set

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

Then, we have the global linear rate:

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

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

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

Outline

I. Introduction: Newton's method

II. Quasi-self-concordant functions

III. Acceleration

IV. Experiments and conclusions

Proximal-Point Method:

$$x_{k+1} \approx \operatorname*{argmin}_y \Big[h_k(y) = f(y) + \tfrac{1}{2a_{k+1}} \|y - x_k\|^2 \Big]$$

[Moreau, 1965; Rockafellar, 1976; Martinet, 1978; Solodov-Svaiter, 2002]

Note: the subproblem $h_k(\cdot)$ is **strongly convex** with constant $\mu = \frac{1}{a_{k+1}}$. We have

$$\nabla h_k(y) = \nabla f(y) + \tfrac{1}{a_{k+1}} B(y - x_k).$$

The neighborhood of local quadratic convergence:

$$\|\nabla h_k(x_k)\|_* = \|\nabla f(x_k)\|_* \overset{(?)}{\leq} \tfrac{\mu}{2M} = \tfrac{1}{2a_{k+1}M}.$$

Set: $\boxed{a_{k+1} := \frac{1}{2M\|\nabla f(x_k)\|_*}}$ $\Rightarrow$ we can minimize $h_k(\cdot)$ up to any accuracy by Newton's method!

Dual Newton Scheme

Init: $x_0 \in \mathbb{R}^n$, $g_0 = \|\nabla f(x_0)\|_*$, and $\delta > 0$

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

1. Set $z_0 = x_k$
2. **For** $t \geq 0$ **iterate:**
   - Perform Newton's step
     $$z_{t+1} \quad = \quad z_t - \left[\nabla^2 f(z_t) + Mg_k B\right]^{-1} \nabla f(z_t)$$
   - **Until** $\|\nabla f(z_{t+1}) - \nabla f(z_t) - \nabla^2 f(z_t)(z_{t+1} - z_t)\|_* \leq \frac{2Mg_k\delta}{(k+1)^2}$,
3. Set $x_{k+1} = z_{t+1}$ and $g_{k+1} = \|\nabla f(x_{k+1})\|_*$
4. If $g_{k+1} \leq \delta$ then **return** x_{k+1}

Theorem. We have the global linear rate for the gradient norm:

$$\|\nabla f(x_k)\|_* \quad \leq \quad \exp\Big(2M^2(\|x_0 - x^\star\|^2 + 2\delta)^2 - \tfrac{k}{2}\Big)\|\nabla f(x_0)\|_*$$

The total number N_k of second-order oracle calls is bounded as

$$N_k \quad \leq \quad k \cdot \Big(1 + \tfrac{1}{\ln 2}\ln\ln\tfrac{(k+1)^2}{2M\delta}\Big).$$

$\Rightarrow$ the method stops after $\mathcal{O}(M^2\|x_0 - x^\star\|^2)$ iterations.

- \+ Possibility of restarts
- \+ Convergence in terms of the gradient norm
- − The condition number is worse: $(MD)^2$ vs. MD

Idea. Contraction $+$ regularization, for $\gamma \in (0, 1)$, set $A_k := A_0(1-\gamma)^{-k}$. Solve:

$$\min_y \Big[h_k(y) \;\; = \;\; A_{k+1} f\big(\gamma y + (1-\gamma)x_k\big) + \tfrac{1}{2}\|y - v_k\|^2 \Big]$$

Contracting Proximal Method. Iteration, $k \geq 0$:

$$\begin{aligned} v_{k+1} &\approx \operatorname*{argmin}_y h_k(y) \\ x_{k+1} &= \gamma v_{k+1} + (1-\gamma)x_k \end{aligned}$$

[Nesterov, 1983; Güler, 1991; Lin-Mairal-Harchaoui, 2018; D-Nesterov, 2020]

Theorem. $A_k(f(x_k) - f^\star) + \frac{1}{2}\sum_{i=1}^{k} \|v_i - v_{i-1}\|^2 \; \leq \; \mathcal{O}\Big(\|x_0 - x^\star\|^2\Big)$

- Global linear rate by design: $f(x_k) - f^\star \leq \mathcal{O}\Big(\frac{\|x_0 - x^\star\|^2}{\exp(\gamma k)}\Big)$
- Control over $\|v_i - v_{i-1}\|$

Choice of γ

How to minimize $v_{k+1} \approx \underset{y}{\operatorname{argmin}}\, h_k(y)$**?**

Consider $\varphi(y) = f(\gamma y + (1-\gamma)x_k), \quad \gamma \in (0,1)$

- $\gamma = 0$, we have $\varphi(y) \equiv f(x_k)$
- $\gamma = 1$, we have $\varphi(y) \equiv f(y)$

The parameter of quasi-SC is $\boxed{M_\varphi \;=\; \gamma M_f.}$

Hence, the **Dual Newton Method** needs the following number of iterations at step $k \geq 0$, to approximate $v_k^\star = \underset{y}{\operatorname{argmin}}\, h_k(y)$:

$$I_k \;\leq\; \mathcal{O}\Bigl(M_\varphi^2 \|v_k - v_k^\star\|^2\Bigr) \;=\; \mathcal{O}\Bigl(\gamma^2 M_f^2 \|v_k - v_{k+1}\|^2\Bigr)$$

Totally, after k steps:

$$\sum_{i=1}^k I_i \;\leq\; \mathcal{O}\Bigl(\gamma^2 M_f^2 \sum_{i=1}^k \|v_i - v_{i-1}\|^2\Bigr) \leq \mathcal{O}\Bigl(\gamma^2 M_f^2 \|x_0 - x^\star\|^2\Bigr) \overset{(?)}{=} \frac{1}{\gamma}$$

$\Rightarrow$ **optimal choice:** $\boxed{\gamma = \bigl[M_f \|x_0 - x^\star\|\bigr]^{-2/3}}$

Summary

Problem: $\min_x f(x)$, where f is quasi-SC with parameter $M > 0$

1. **Primal Newton with Gradient Regularization:**

$$\mathcal{O}\left(\boldsymbol{MD} \ln \tfrac{1}{\varepsilon}\right) \quad \text{second-order oracle calls for } f$$

2. **Dual Newton:**

$$\mathcal{O}\left(\left[\boldsymbol{M}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|\right]^2 \ln \tfrac{1}{\varepsilon} \ln \ln \tfrac{1}{\varepsilon^2}\right)$$

3. **Accelerated Newton:**

$$\tilde{\mathcal{O}}\left(\left[\boldsymbol{M}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|\right]^{2/3}\right)$$

Optimal? Most probably yes!

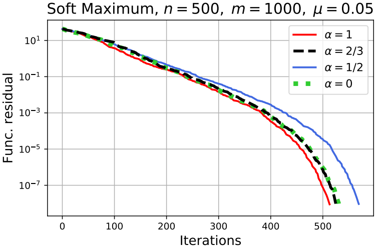
- Matches the lower bound for the *ball minimization oracle* [Carmon-Jambulapati-Jiang-Jin-Lee-Sidford-Tian, 2020]

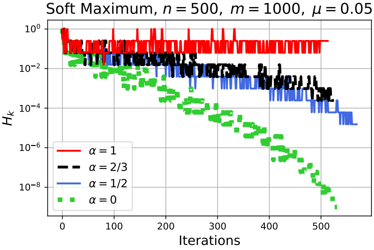
Outline

$$\boxed{\min_x f_\mu(x)}$$

Iterate $k \geq 0$:

$$x_{k+1} \;=\; x_k - \left(\nabla^2 f_\mu(x_k) + \left(\sigma\|\nabla f_\mu(x_k)\|\right)^\alpha B\right)^{-1} \nabla f(x_k)$$





Conclusions

- Quasi-SC functions $\approx$ loss functions with exponential tails
- The Newton method is very efficient in this case (fast global linear rate): $\mathcal{O}\Big(MD \ln \frac{1}{\varepsilon}\Big)$
- We can accelerate: $MD \mapsto (MD)^{2/3}$
- Solving
$$\min_x \Big[F(x) = f(x) + \psi(x) \Big]$$
is as difficult as
$$\min_x \Big[\langle Ax, x \rangle - \langle b, x \rangle + \psi(x) \Big]$$

References:

1. Doikov, N., Mishchenko, K. and Nesterov, Y., 2022. Super-universal regularized Newton method. *SIAM Journal on Optimization.*
2. Doikov, N., 2023. Minimizing quasi-self-concordant functions by gradient regularization of Newton method. *arXiv:2308.14742.* (under review)

- Lower complexity bounds
- Practical accelerated schemes (currently, **no local superlinear convergence**)
- Comparison with polynomial-time **Interior-Point schemes**
- Consequences for non-convex optimization

Thank you for your attention!