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

Nikita Doikov
EPFL, Switzerland

Workshop on Nonsmooth Optimization and Applications in Honor of the 75th Birthday of Boris Mordukhovich

University of Antwerp, Belgium
April 9, 2024

Outline

I. Introduction: Quasi-Self-Concordance

II. Gradient Regularization of Newton Method

III. Dual and Accelerated Newton

IV. Applications and Conclusions

Two black-box convex functions:

$$\min_x \Big[F(x) \;=\; f(x) + \psi(x) \Big] \qquad (*)$$

- f is differentiable (sufficiently smooth)
- $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a proper closed convex function

In this talk: we show that the main cost of solving $(*)$ is in ψ.

- Assume we can solve problems with a quadratic smooth part:

$$\min_x \Big[\tfrac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle + \psi(x) \Big]$$

Main Theorem. For solving $(*)$, it is enough to solve the quadratic problem

$$\mathcal{O}\Big(MD \cdot \ln \tfrac{1}{\varepsilon} \Big) \quad \text{times.}$$

NB: No strong/uniform convexity needed!

Complexity of the gradient methods:

- Bounds on the second derivative

For example, Lipschitz gradient:

$$0 \preceq \nabla^2 f(x) \preceq L_1 I, \qquad \forall x$$

or, Relative smoothness:

$$\alpha \nabla^2 d(x) \preceq \nabla^2 f(x) \preceq \beta \nabla^2 d(x), \qquad \forall x$$

[Bauschke-Bolte-Teboulle, 2016; Van Nguyen, 2017; Lu-Freud-Nesterov, 2018]

Complexity of the second-order methods:

- Bounds on the third derivative!

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

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

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

Functions with Lipschitz Hessian:

$$\nabla^3 f(x)[u,u,u] \;\; \leq \;\; L_2 \|u\|^3, \qquad \forall x, u$$

- Fixed global norm (no affine-invariance)
- Efficiency of the Cubic regularization of Newton's method

[Nesterov-Polyak, 2006; Cartis-Gould-Toint 11; Grapiglia-Nesterov, 2017]

Functions with Lipschitz third derivative:

$$\nabla^3 f(x)[u,u,v] \;\; \leq \;\; \sqrt{2L_3} \langle \nabla^2 f(x) u, u \rangle^{1/2} \|u\| \|v\|, \qquad \forall x, u, v$$

- Faster rates for second-order schemes

[Nesterov, 2018; 2021; Kamzolov-Gasnikov-Dvurechensky, 2021;
D-Mishchenko-Nesterov, 2022]

The standard Euclidean norm for some fixed operator $B = B^\top \succ 0$:

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

and denote the local norm:

$$\|u\|_x \;:=\; \langle \nabla^2 f(x)u, u\rangle^{1/2}.$$

In this talk, we assume that f is quasi-self-concordant with constant $M \geq 0$:

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

- 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):

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

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

Example 4: (Soft Maximum):

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

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

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

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

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

Basic Operations

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

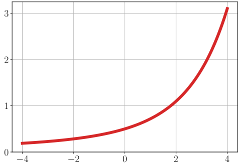
$$\tfrac{1}{e}\nabla^2 f(x) \preceq f(y) \preceq e\nabla^2 f(x).$$

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

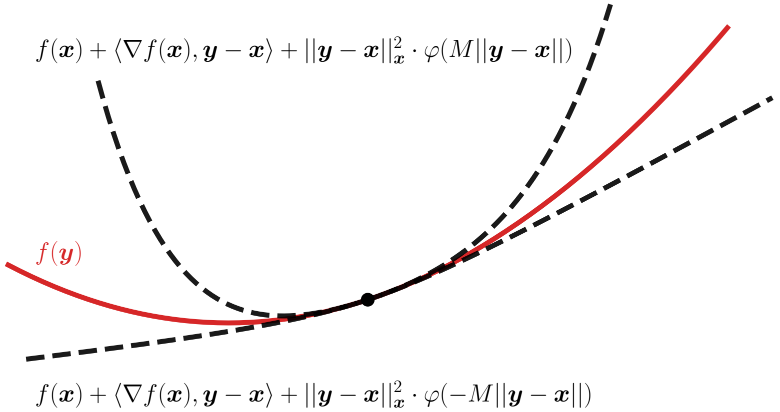
The gradient approximation:

- $\|\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y-x)\|_* \leq M\|y-x\|_x^2 \cdot \varphi(M\|y-x\|)$,

where $\varphi(t) := \frac{e^t - t - 1}{t^2} \geq 0$ is a convex and monotone function:



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



Outline

I. Introduction: Quasi-Self-Concordance

II. Gradient Regularization of Newton Method

III. Dual and Accelerated Newton

IV. Applications and Conclusions

Problem: $\min_x \Big[F(x) = f(x) + \psi(x) \Big]$, where f is quasi-SC.

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

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

$$\Leftrightarrow \qquad \nabla f(x) + [\nabla^2 f(x) + \beta B](x^+ - x) \;\in\; -\partial\psi(x^+).$$

Lemma. Set $\beta := \sigma\|\nabla f(x) + s\|_*$ for any $s \in \partial\psi(x)$ and $\sigma \geq M$. Then,

1. $\|x^+ - x\| \leq \frac{1}{M}$
2. $\|x^+ - x\|_x^2 \leq \frac{\|\nabla f(x) + s\|_*}{M}$

[Polyak, 2009; Ueda-Yamashita, 2009; Mishchenko, 2021; D-Nesterov, 2021]

Main Lemma. Let $\beta := \sigma \|F'(x)\|_*$ for $F'(x) \in \partial F(x)$ and $\sigma \geq M$.

Then, for the specific subgradient

$$F'(x^+) \; := \; \nabla f(x^+) - \nabla f(x) - [\nabla^2 f(x) + \beta B](x^+ - x) \; \in \; \partial F(x^+),$$

we have

$$\langle F'(x^+), x - x^+ \rangle \;\; \geq \;\; \tfrac{1}{2\beta} \|F'(x^+)\|_*^2.$$

Note: by convexity, we conclude

$$F(x) - F(x^+) \;\; \geq \;\; \langle F'(x^+), x - x^+ \rangle \;\; \geq \;\; \tfrac{1}{2\beta} \|F'(x^+)\|_*^2.$$

Init: $x_0 \in \mathrm{dom}\,\psi$ and $g_0 = \|F'(x_0)\|_*$ for any $F'(x_0) \in \partial F(x_0)$.

Iteration, $k \geq 0$:

- For some $\sigma_k \geq 0$, compute x_{k+1} s.t.

$$\nabla f(x_k) + \left[\nabla^2 f(x_k) + \sigma_k g_k B\right](x_{k+1} - x_k) \quad \in \quad -\partial\psi(x_{k+1})$$

- Update $g_{k+1} = \|\nabla f(x_{k+1}) - \nabla f(x_k) - [\nabla^2 f(x_k) + \sigma_k g_k B](x_{k+1} - x_k)\|_*$

Theorem. Set $\sigma_k := M$. Then, we have the global linear rate:

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

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 $F(x_k) - F^\star \leq \varepsilon$

Super-Universal Newton Method

In our method, we set $\sigma_k := M$

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

Init: Choose $x_0 \in \mathrm{dom}\,\psi$, $g_0 = \|F'(x_0)\|_*$, 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 x^+:
$$\nabla f(x_k) + \left[\nabla^2 f(x_k) + \beta_k B\right](x^+ - x_k) \quad \in \quad -\partial\psi(x^+)$$
it holds
$$\langle F'(x^+), x_k - x^+\rangle \quad \geq \quad \frac{1}{2\beta_k}\|F'(x^+)\|_*^2.$$
2. Set $x_{k+1} = x^+$, $g_{k+1} = \|F'(x^+)\|_*$, 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

The classic Newton's method has a local quadratic convergence, when close to the solution [Fine, 1916; Bennett, 1916; Kantorovich, 1948]

- We have the same local rate for our method!

Theorem. Let $\nabla^2 f(x) \succeq \mu I, \forall x$. Let

$$\|F'(x_0)\|_* \quad \leq \quad \frac{\mu}{2M} \qquad \text{(a neighborhood of the solution)}$$

Then,

$$\|F'(x_k)\|_* \quad \leq \quad \frac{\mu}{2M} \cdot \left(\frac{1}{e}\right)^{2^k}.$$

- Quadratic convergence: to find $\|F'(x_k)\|_* \leq \varepsilon$ it is enough to perform $k = \mathcal{O}(\ln \ln \frac{\mu}{M\varepsilon})$ steps.

Outline

I. Introduction: Quasi-Self-Concordance

II. Gradient Regularization of Newton Method

III. Dual and Accelerated Newton

IV. Applications and Conclusions

Proximal Viewpoint

Proximal-Point Method:

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

[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

$$h_k'(y) \quad = \quad F'(y) + \tfrac{1}{a_{k+1}} B(y - x_k).$$

The neighborhood of local quadratic convergence:

$$\|h_k'(x_k)\|_* \quad = \quad \|F'(x_k)\|_* \quad \stackrel{(?)}{\leq} \quad \tfrac{\mu}{2M} \quad = \quad \tfrac{1}{2a_{k+1}M}.$$

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

Init: $x_0 \in \mathrm{dom}\,\psi$ and $g_0 = \|F'(x_0)\|_*$ for $F'(x_0) \in \partial F(x_0)$, $\delta > 0$.

Iteration, $k \geq 0$:

1. Set $z_0 = x_k$
2. **For** $t \geq 0$ **iterate:**
   - Compute z_{t+1} s.t.

   $$\nabla f(z_t) + \left[\nabla^2 f(z_t) + M g_k B\right](z_{t+1} - z_t) \quad \in \quad -\partial\psi(z_{t+1})$$

   - **Until** $\|s_{t+1}\|_* \leq \frac{2Mg_k\delta}{(k+1)^2}$, where

   $$s_{t+1} \quad := \quad \nabla f(z_{t+1}) - \nabla f(z_t) - \nabla^2 f(z_t)(z_{t+1} - z_t).$$
3. Set $x_{k+1} = z_{t+1}$ and $g_{k+1} = \|s_{t+1} - 2Mg_k B(x_{k+1} - x_k)\|_*$
4. If $g_{k+1} \leq \delta$ then **return** x_{k+1}

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

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

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

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

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

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

Idea. Contraction + regularization, for $\gamma \in (0,1)$:

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

where $A_k := A_0(1-\gamma)^{-k}$, $a_k := A_k - A_{k-1}$.

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 $v_{k+1} \approx \operatorname*{argmin}_y h_k(y)$?

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 $\varphi(y) = f(y)$

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

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

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

Totally, after k steps:

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

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

Convergence Rates Summary

Problem: $\min_x \Big[F(x) = f(x) + \psi(x) \Big]$

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

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

2. **Dual Newton:**

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

3. **Accelerated Newton:**

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

Optimal? Most probably yes!

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

Outline

I. Introduction: Quasi-Self-Concordance

II. Gradient Regularization of Newton Method

III. Dual and Accelerated Newton

IV. Applications and Conclusions

Non-smooth problem:

$$\min_{x \in \mathbb{R}^n} \Big[f(x) \;=\; \max_{1 \le i \le m} \big[\langle a_i, x \rangle - b_i \big] \Big]$$

Find x_k s.t. $f(x_k) - f^\star \le \varepsilon$.

1. **Subgradient method:** $\mathcal{O}(1/\varepsilon^2)$ [Shor, 1962]
2. Smoothing technique: $f_\mu(x) \le f(x) \le f_\mu(x) + \mu D^2$ where

$$f_\mu(x) \;:=\; \mu \ln\Big(\sum_{i=1}^m \exp\big(\tfrac{\langle a_i, x\rangle - b_i}{\mu} \big) \Big)$$

   - Need to choose $\mu = \mathcal{O}(\varepsilon)$
   - Lipschitz gradient $L(f_\mu) = 1/\mu = \mathcal{O}(1/\varepsilon)$
   - **Fast Gradient Method**

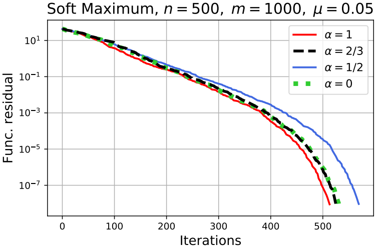
$$\mathcal{O}\Big(\sqrt{L(f_\mu) D^2 \varepsilon^{-1}} \Big) \;=\; \mathcal{O}(1/\varepsilon) \quad \text{[Nesterov, 2003]}$$

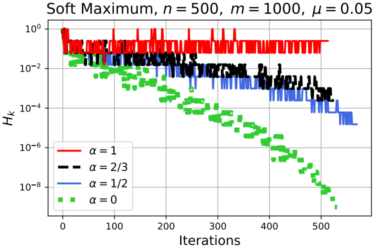
3. Newton's Method. f_μ is quasi-SC with $M = 2/\mu = \mathcal{O}(1/\varepsilon)$.
   - **Primal Newton Method:** $\tilde{\mathcal{O}}(MD) = \tilde{\mathcal{O}}(1/\varepsilon)$
   - **Accelerated Newton:** $\tilde{\mathcal{O}}((MD)^{2/3}) = \tilde{\mathcal{O}}(1/\varepsilon^{2/3})$ **!**

$$\min_x f_\mu(x)$$

Iterate $k \geq 0$:

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





Conclusions

- Quasi-SC functions $\approx$ loss functions with exponential tails
- The Newton method is very efficient in this case (fast global linear rate): $\mathcal{O}\left(MD \ln \frac{1}{\varepsilon}\right)$
- We can accelerate: $\tilde{\mathcal{O}}\left((MD)^{2/3}\right)$
- 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.*

Open Questions

- Lower complexity bounds
- Practical accelerated schemes (currently, no local superlinear convergence)
- Comparison with polynomial-time Interior-Point schemes
- Other problem classes? Minimizing an arbitrary convex analytic function
- Consequences for non-convex optimization

Thank you very much for your attention!

Happy Birthday to Prof. Boris Mordukhovich!