Local convergence of tensor methods

Nikita Doikov

Joint work with Yurii Nesterov

UCLouvain, Belgium

Workshop on Advances in Continuous Optimization, EUROPT
July 7, 2021

Optimization Problem:

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

- f is a convex differentiable function

The Newton Method [Newton, 1669; Raphson, 1690; Fine, 1916; Bennett, 1916; Kantorovich, 1948]**:**

$$\begin{aligned} x_{k+1} &= \operatorname*{argmin}_y \Big\{ \langle \nabla f(x_k), y - x_k \rangle + \tfrac{1}{2} \langle \nabla^2 f(x_k)(y - x_k), y - x_k \rangle \Big\} \\ &= x_k - \big(\nabla^2 f(x_k)\big)^{-1} \nabla f(x_k), \qquad k \geq 0. \end{aligned}$$

Local quadratic convergence: $\mathcal{O}(\log_2 \log_2 \frac{1}{\varepsilon})$ iterations to find an ε-solution. **Assumptions:**

1. Strong convexity. $\forall x: \ \nabla^2 f(x) \succeq \mu I$
2. Lipschitz Hessian. $\forall x, y: \ \|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L_2 \|x - y\|$
3. x_0 is close to x^*

Cubic Newton Method [Nesterov-Polyak, 2006]**:**

$$
\begin{aligned}
x_{k+1} &= \operatorname*{argmin}_y \Big\{ \langle \nabla f(x_k), y - x_k \rangle + \tfrac{1}{2} \langle \nabla^2 f(x_k)(y - x_k), y - x_k \rangle \\
&\qquad\qquad + \tfrac{H}{6} \|y - x\|^3 \Big\} \\
&= x_k - \Big(\nabla^2 f(x_k) + \tfrac{H \|x_{k+1} - x_k\|}{2} I \Big)^{-1} \nabla f(x_k), \qquad k \geq 0.
\end{aligned}
$$

- $H := 0 \quad \Rightarrow \quad$ The Classical Newton (no global convergence).
- $H := L_2 \quad \Rightarrow \quad$ Global convergence: $f(x_k) - f^* \leq \mathcal{O}(1/k^2)$.

For strongly convex functions: local quadratic rate as well.

Taylor's polynomial of degree p at point x:

$$f(y) \approx \Omega_p(x;y) \stackrel{\mathrm{def}}{=} f(x) + \sum_{i=1}^{p} \frac{1}{i!} D^i f(x)[y-x]^i.$$

Tensor Method of order $p \geq 1$:

$$x_{k+1} = \operatorname*{argmin}_y \Big\{ \Omega_p(x_k;y) + \frac{H}{(p+1)!}\|y-x_k\|^{p+1} \Big\}, \qquad k \geq 0.$$

- $p = 1$: The Gradient Method. $p = 2$: The Cubic Newton.
- Let pth derivative be Lipschitz continuous:

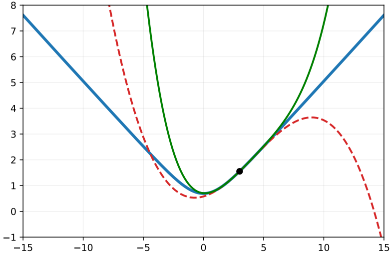
$$\|D^p f(x) - D^p f(y)\| \leq L_p \|x-y\|, \qquad \forall x, y.$$

Set $H := L_p \Rightarrow$ Global rate: $f(x_k) - f^* \leq \mathcal{O}(1/k^p)$ [Baes, 2009].

How to solve the subproblem?

Convex Tensor Model

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



- **Theorem** [Nesterov, 2018]:

Let $f(\cdot)$ be a convex function and $H \geq pL_p$. Then $\forall x$ the model

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

is convex in y

- **For $p = 3$:** efficient implementation using only second-order oracle is available [Nesterov, 2019]. The cost is $\mathcal{O}(n^3) + \tilde{O}(n)$.

- **Accelerated** Tensor Methods: $F(x_k) - F^* \leq O(1/k^{p+1})$ [Baes, 2009; Nesterov, 2018]

- **Optimal** Tensor Methods: $F(x_k) - F^* \leq O(1/k^{\frac{3p+1}{2}})$ [Gasnikov et al., 2019; Kamzolov-Gasnikov-Dvurechensky, 2020]

  The oracle complexity matches the lower bound (up to logarithmic factor) from [Arjevani-Shamir-Shiff, 2017]

- **Universal** Tensor Methods: [Grapiglia-Nesterov, 2019], [Cartis-Gould-Toint, 2020]

- **Stochastic** Tensor Methods: [Lucchi-Kohler, 2019]

- ...

Uniformly Convex Functions

f is called **uniformly convex** of degree $q \geq 2$ iff $\forall x, y$

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \quad \geq \quad \sigma_q \|x - y\|^q.$$

$\sigma_q > 0$ is a parameter.

- Strongly convex functions: $q = 2$
- Example: $f(x) = \frac{1}{q}\|x - x_0\|^q$ is uniformly convex of degree q with constant $\sigma_q = 2^{2-q}$
- Sum of convex and uniformly convex functions gives uniformly convex

Tensor Method of order $p \geq 2$:

$$x_{k+1} \;=\; \operatorname*{argmin}_{y} \Big\{ \Omega_p(x_k; y) + \tfrac{H}{(p+1)!}\|y - x_k\|^{p+1} \Big\}, \qquad k \geq 0.$$

- Set $H := pL_p$.

New result: **Theorem.** Assume the objective is uniformly convex of degree

$$q \;\in\; [2, p+1)$$

with parameter $\sigma_q > 0$. Let

$$f(x_0) - f^* \;\leq\; \mathcal{O}\bigg(\Big[\tfrac{\sigma_q^{p+1}}{L_p^q} \Big]^{\frac{1}{p-q+1}} \bigg) \quad \text{(the local region)}.$$

Then, the Tensor Method needs $K = \mathcal{O}\big(\log_{\frac{p}{q-1}} \log_2 \frac{1}{\varepsilon}\big)$ iterations to find an ε-solution.

Composite Optimization Problem:

$$\min_{x\in\mathbb{R}^n}\left\{F(x) \stackrel{\text{def}}{=} f(x)+\psi(x)\right\}$$

- ψ is a *simple* convex function taking values in $\mathbb{R}\cup\{+\infty\}$
- f is convex and differentiable (the *difficult part*)

Examples:

1. Let Q be a simple convex set, $\psi(x) = \begin{cases} 0, & x \in Q \\ +\infty, & \text{otherwise.} \end{cases}$
2. $\psi(x) = \lambda\|x\|_1$ (adding ℓ_1-Regularizer to the problem).

Composite Tensor Method, $p \geq 2$:

$$x_{k+1} = \operatorname*{argmin}_y \Big\{ \Omega_p(x_k; y) + \frac{H}{(p+1)!} \|y - x_k\|^{p+1} + \psi(y) \Big\}, \; k \geq 0.$$

Let the full objective be uniformly convex of degree $q \in [2, p+1)$:

$$\langle G_x - G_y, x - y \rangle \geq \sigma_q \|x - y\|^q, \quad \forall G_x \in \partial F(x), \; G_y \in \partial F(y),$$

and the smooth part have Lipschitz continuous pth derivative:

$$\|D^p f(x) - D^p f(y)\| \leq L_p \|x - y\|.$$

Theorem. Let $F(x_0) - F^* \leq \mathcal{O}\Big(\Big[\frac{\sigma_q^{p+1}}{L_p^q} \Big]^{\frac{1}{p-q+1}} \Big)$ (the local region). Then the Composite Tensor Method needs $K = \mathcal{O}\big(\log_{\frac{p}{q-1}} \log_2 \frac{1}{\varepsilon} \big)$ iterations to find an ε-solution.

- We also established the convergence in terms of the minimal subgradient $\eta(x) \overset{\text{def}}{=} \min\limits_{g \in \partial \psi(x)} \|\nabla f(x) + g\|_*$.

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

Proximal-Point Algorithm [Rockafellar, 1976]**:**

$$x_{k+1} \;=\; \operatorname*{argmin}_y \Big\{ f(y) + \tfrac{1}{2a_{k+1}} \|y - x_k\|^2 \Big\}, \qquad k \geq 0.$$

- If f is convex, the objective of the subproblem $h_{k+1}(y) = f(y) + \frac{1}{2a_{k+1}}\|y - x_k\|^2$ is strongly convex.
- The Gradient Method needs $\tilde{\mathcal{O}}(a_{k+1}L_1)$ iterations to minimize h_{k+1}.
- It is enough to use for x_{k+1} an inexact minimizer of h_{k+1}.

[Solodov-Svaiter, 2001; Schmidt-Roux-Bach, 2011; Salzo-Villa, 2012]

Set $a_{k+1} = \frac{1}{L_1}$. Then $f(\bar{x}_k) - f^* \leq \frac{L_1 \|x_0 - x^*\|^2}{2k}$.

What about High-Order methods?

Globalizing the Local Convergence

$$h_{k+1}(y) \quad = \quad f(y) + \frac{1}{2a_{k+1}}\|y - x_k\|^2 \quad \to \quad \min_y$$

Idea: Choose $a_{k+1} > 0$ to ensure that x_k in the *region of local convergence* of the Tensor Method, $p \geq 2$.

$$\boxed{a_{k+1} \quad \approx \quad \left(\frac{1}{\|\nabla f(x_k)\|_*}\right)^{\frac{p-1}{p}} \cdot \left(\frac{1}{L_p}\right)^{\frac{1}{p}}} \qquad (*)$$

Then we can solve the subproblem very efficiently.

Theorem. For the inexact Proximal-Point algorithm with $(*)$, we have:

$$f(\bar{x}_k) - f^* \quad \leq \quad \mathcal{O}\left(\frac{L_p\|x_0 - x^*\|^{p+1}}{k^{\frac{p+1}{2}}}\right).$$

- For the Gradient Method we had $\mathcal{O}(1/k)$.
- $\mathcal{O}(1/k^{\frac{p+1}{2}})$ is worse than the rate of the direct TM: $\mathcal{O}(1/k^p)$.

Conclusions

1. We need to use regularization for high-order ($p \geq 3$) Taylor's approximation of the objective
   - Ensures convexity of the model
   - Efficient implementation for $p = 3$ (no need to store tensors)

2. Local superlinear convergence of the Composite Tensor Method, $p \geq 2$:
   - Rate: $\mathcal{O}\left(\log_{\frac{p}{q-1}} \log \frac{1}{\varepsilon}\right)$ — bigger base in the logarithm
   - Degree of uniform convexity $q \in [2, p+1)$ — wider class of functions

3. Globalize the local method by doing Proximal-Point iterations
   - Accelerated Methods – ?

Thank you for your attention!