Inexact Tensor Methods with Dynamic Accuracies

Nikita Doikov **Yurii Nesterov**

UCLouvain, Belgium

ICML 2020

Plan of the talk

1. Introduction: Tensor Methods in Convex Optimization
2. Inexact Tensor Methods
3. Acceleration
4. Numerical Example

Plan of the talk

1. Introduction: Tensor Methods in Convex Optimization

2. Inexact Tensor Methods

3. Acceleration

4. Numerical Example

Composite optimization problem

$$\min_{x \in \mathrm{dom}\, F} F(x) := f(x) + \psi(x),$$

- f is convex and smooth;
- $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is convex (possibly nonsmooth, but *simple*).

The Gradient Method:

$$x_{k+1} \;\; = \;\; \mathop{\mathrm{argmin}}_{y} \Big\{ \langle \nabla f(x_k), y - x_k \rangle + \frac{H}{2}\|y - x_k\|^2 + \psi(y) \Big\}, \;\; k \geq 0.$$

- Gradient of f is Lipschitz continuous:

  $\|\nabla f(y) - \nabla f(x)\| \;\leq\; L_1 \|y - x\| \quad \Rightarrow \quad H := L_1$
- Global sublinear convergence: $F(x_k) - F^* \;\leq\; O(1/k)$.

Newton Method with Cubic Regularization

- Hessian of f is Lipschitz continuous:

$$\|\nabla^2 f(y) - \nabla^2 f(x)\| \leq L_2 \|y - x\|.$$

Cubic Newton:

$$x_{k+1} = \operatorname*{argmin}_y \Big\{ \langle \nabla f(x_k), y - x_k \rangle + \tfrac{1}{2} \langle \nabla^2 f(x_k)(y - x_k), y - x_k \rangle + \tfrac{H}{6} \|y - x_k\|^3 + \psi(y) \Big\}, \quad k \geq 0.$$

- $H := 0 \quad \Rightarrow \quad$ Classical Newton.
- $H := L_2 \quad \Rightarrow \quad$ Global convergence: $F(x_k) - F^* \leq O(1/k^2)$.

[Nesterov-Polyak, 2006]

Let $x \in \mathbb{R}^n$ be fixed, consider arbitrary $h \in \mathbb{R}^n$ and one-dimensional

$$\phi(t) \quad := \quad f(x + th), \quad t \in \mathbb{R}.$$

Then $\phi(0) \; = \; f(x)$, $\phi'(0) \; = \; \langle \nabla f(x), h \rangle$, $\phi''(0) \; = \; \langle \nabla^2 f(x) h, h \rangle$.
Denote:

$$D^p f(x)[h]^p \quad := \quad \phi^{(p)}(0).$$

The model:

$$\Omega_H(x; y) \quad := \quad \sum_{k=1}^{p} \tfrac{1}{k!} D^k f(x)[y-x]^k + \tfrac{H}{(p+1)!} \|y - x\|^{p+1} + \psi(y).$$

Tensor Method of order $p \geq 1$:

$$x_{k+1} \quad = \quad \operatorname*{argmin}_y \Omega_H(x_k; y), \quad k \geq 0.$$

- p-th derivative is Lipschitz continuous:

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

- Global convergence: $F(x_k) - F^* \leq O(1/k^p)$. [Baes, 2009]

At each iteration $k \geq 0$, the subproblem is

$$\min_y \Omega_H(x_k; y) \quad := \quad \sum_{k=1}^{p} \frac{1}{k!} D^k f(x)[y-x]^k + \frac{H}{(p+1)!}\|y-x\|^{p+1} + \psi(y).$$

- $H \geq pL_p \quad \Rightarrow \quad \Omega_H(x_k; y)$ is **convex** in y. [Nesterov, 2018]
- For $p = 3$: efficient implementation, using Gradient Method with relative smoothness condition [Van Nguyen, 2017; Bauschke-Bolte-Teboulle, 2016; Lu-Freund-Nesterov, 2018].

  The cost of minimizing $\Omega_H(x_k; \cdot)$ is: $O(n^3) + \tilde{O}(n)$.

- **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].

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

- ...

Plan of the talk

1. Introduction: Tensor Methods in Convex Optimization

2. Inexact Tensor Methods

3. Acceleration

4. Numerical Example

Definition of Inexactness

Use a point $T = T_{H,\delta}(x_k)$ with <u>small residual in function value</u>:

$$\Omega_H(x_k; T) - \min_y \Omega_H(x_k; y) \leq \delta.$$

- Easier to achieve by inner method.
- Can be controlled in practice using the duality gap.

Set $H := pL_p$. We have

$$F(T) \leq F(x_k) + \delta.$$

- Inexact step can be nonmonotone.

Initialization: choose $x_0 \in \mathrm{dom}\, F$, set $H := pL_p$.

Iterations: $k \geq 0$.

1: Pick up $\delta_{k+1} \geq 0$.

2: Compute inexact monotone tensor step T, such that

$$\begin{aligned} &\Omega_H(x_k; T) - \min_y \Omega_H(x_k; y) \;\leq\; \delta_{k+1}, \\ &\text{and} \quad F(T) \;<\; F(x_k). \end{aligned}$$

3: $x^{k+1} \;:=\; T$.

Theorem 1. Set $\boxed{\delta_k := \frac{c}{k^{p+1}},}$ for $c \geq 0$. Then

$$F(x_k) - F^* \quad \leq \quad O\left(\tfrac{1}{k^p}\right).$$

Adaptive Strategy for Inner Accuracy

Let us set $\boxed{\delta_k := c(F(x_{k-2}) - F(x_{k-1})).}$

Theorem 2. (General convex case)

$$F(x_k) - F^* \quad \leq \quad O\left(\tfrac{1}{k^p}\right).$$

Theorem 3. (Uniformly convex objective) Let

$$F(y) \quad \geq \quad F(x) + \langle F'(x), y - x \rangle + \tfrac{\sigma_{p+1}}{p+1} \|y - x\|^{p+1}.$$

Denote $\omega_p := \max\{ \frac{(p+1)^2 L_p}{p! \sigma_{p+1}}, 1 \}$. Then we have linear rate

$$F(x_{k+1}) - F^* \quad \leq \quad \left(1 - \tfrac{p \omega_p^{-1/p}}{2(p+1)}\right)(F(x_k) - F^*).$$

- This works for methods, starting from $p \geq 1$.

Theorem 4. For $p \geq 2$ and strongly convex objective, we have local superlinear rate.

Plan of the talk

1. Introduction: Tensor Methods in Convex Optimization
2. Inexact Tensor Methods
3. Acceleration
4. Numerical Example

- Fix prox-function $d(x)$.

Bregman divergence: $\beta_d(x; y) := d(y) - d(x) - \langle \nabla d(x), y - x \rangle$.

- Two sequences of points $\{x_k\}_{k \geq 0}$, $\{v_k\}_{k \geq 0}$, $v_0 = x_0$.
- Sequence of positive coefficients $\{a_k\}_{k \geq 0}$, $A_k \stackrel{\mathrm{def}}{=} \sum_{i=1}^{k} a_i$.

Iterations, $k \geq 0$:

1. Compute

$$v_{k+1} \;=\; \operatorname*{argmin}_y \Big\{ A_{k+1} f\big(\tfrac{a_{k+1} y + A_k x_k}{A_{k+1}}\big) + a_{k+1}\psi(y) + \beta_d(v_k; y) \Big\}.$$

2. Put $x_{k+1} = \frac{a_{k+1} v_{k+1} + A_k x_k}{A_{k+1}}$.

The rate of convergence: $F(x_k) - F^* \;\leq\; \frac{\beta_d(x_0; x^*)}{A_k}$.

[Doikov-Nesterov, 2019]

For Tensor Method of order $p \geq 1$:

- Set $d(x) := \frac{1}{p+1}\|x - x_0\|^{p+1}$.
- $A_{k+1} := \frac{(k+1)^{p+1}}{L_p}$.

For contracted objective with regularization

$$h_{k+1}(y) \quad := \quad A_{k+1} f\left(\frac{a_{k+1} y + A_k x_k}{A_{k+1}}\right) + a_{k+1}\psi(y) + \beta_d(v_k; y),$$

we compute inexact minimizer v_{k+1}:

$$h_{k+1}(v_{k+1}) - h^*_{k+1} \quad \leq \quad \frac{c}{(k+1)^{p+2}}.$$

- It requires $\tilde{O}(1)$ inexact Tensor Steps.

Theorem. For outer iterations, we obtain accelerated rate:

$$F(x_k) - F^* \quad \leq \quad O\left(\frac{1}{k^{p+1}}\right).$$

Plan of the talk

1. Introduction: Tensor Methods in Convex Optimization
2. Inexact Tensor Methods
3. Acceleration
4. Numerical Example

$$\min_{x \in \mathbb{R}^n} f(x) := \mu \log\left(\sum_{i=1}^{m} \exp\left(\frac{\langle a_i, x\rangle - b_i}{\mu} \right) \right) \qquad \text{(SoftMax)}.$$

- $a_1, \ldots, a_m, b$ — given data.
- $\mu > 0$ — smoothing parameter.
- Denote $B \equiv \sum_{i=1}^{m} a_i a_i^T \succeq 0$, and use $\|x\| \equiv \langle Bx, x\rangle^{1/2}$.

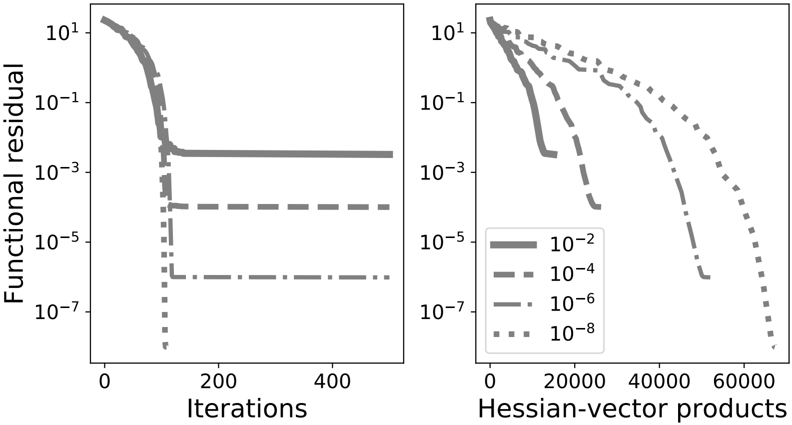
We have

$$L_1 \le \frac{1}{\mu}, \quad L_2 \le \frac{2}{\mu^2}, \quad L_3 \le \frac{4}{\mu^3}.$$

- Cubic Newton ($p = 2$).
- Compute each step (inexactly) by Fast Gradient Method.

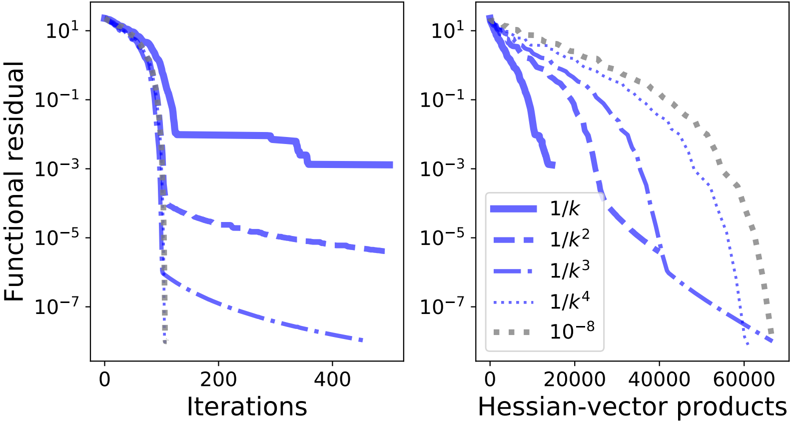
- $\delta_k :=$ const.

Log-sum-exp, $\mu = 0.05$: constant strategies

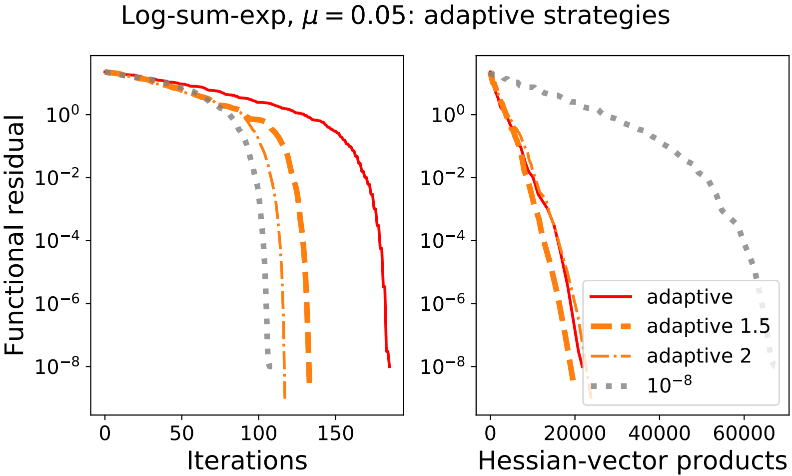


- $\delta_k := 1/k^\alpha$.

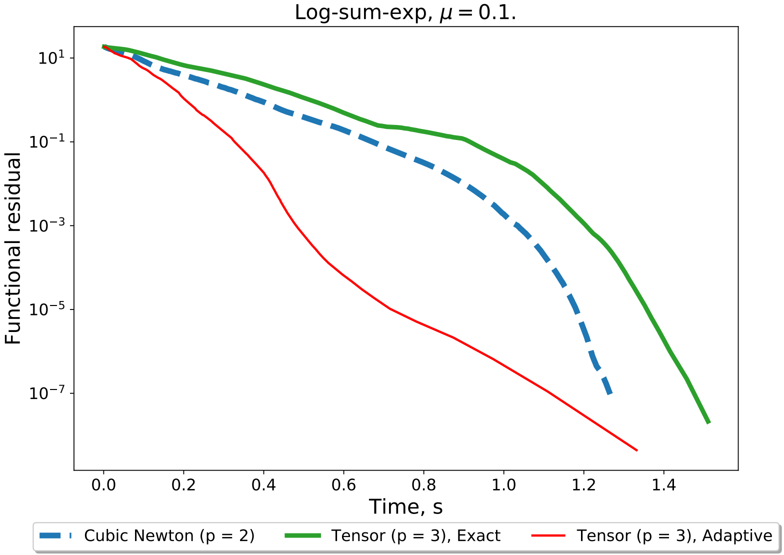
Log-sum-exp, $\mu = 0.05$: dynamic strategies



- $\delta_k := (F(x_{k-1}) - F(x_k))^{\alpha}$.



Log-sum-exp: Cubic Newton vs. Tensor Method



- H is fixed.

Inexact Tensor Methods of degree $p \geq 1$:

$p = 1$: Gradient Method.

$p = 2$: Newton method with Cubic regularization.

$p = 3$: Third order Tensor method.

We admit to solve the subproblem inexactly, δ_k — accuracy in functional residual for the subproblem.

- Dynamic strategy $\delta_k := \frac{c}{k^{p+1}}$.
- Adaptive strategy $\delta_k := c(F(x_k) - F(x_{k-1}))$.

Global rate of convergence: $F(x_k) - F^* \leq O(\frac{1}{k^p})$.

- Using contracting proximal iterations we obtain accelerated $O(\frac{1}{k^{p+1}})$ rate.

Thank you for your attention!