Inexact Tensor Methods with Dynamic Accuracies

Nikita Doikov nikita.doikov@uclouvain.be

Yurii Nesterov yurii.nesterov@uclouvain.be

UCLouvain University

Problem and Motivation

Composite optimization problem

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

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

High-order methods for solving this problem can tackle the ill-conditioning, and improve the rate of convergence. However, at each step they require to solve some nontrivial subproblem (for example, minimization of quadratic function with a regularizer, and some additional nondifferentiable parts).

In this work we study: *which level of exactness we need to ensure at each step for not loosing the fast convergence of the initial method*. We propose two simple strategies for choosing the inner accuracy. The required precision changes dynamically with iterations, improving the practical performance.

Tensor Methods of Degree $p \geq 1$

For $p \geq 1$, consider the following model of F:

$$F(y) \;\approx\; \Omega_H(x; y) \;:=\; f(x) + \sum_{i=1}^{p} \frac{1}{i!} D^i f(x)[y-x]^i + \frac{H\|y-x\|^{p+1}}{(p+1)!} + \psi(y),$$

this is Taylor approximation of f around x, with regularization.

Tensor Method: $\boxed{x_{k+1} \;\in\; \mathop{\mathrm{Argmin}}_y \Omega_H(x_k; y), \quad k \geq 0.}$

- $p = 1$: Gradient Method (for composite objective).
- $p = 2$: Newton Method with Cubic regularization.

Rate of convergence [1]:

$$F(x_k) - F^* \;\leq\; O(1/k^p),$$

for functions with Lipschitz continuous p-th derivative:

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

Solving the Subproblem in Tensor Methods

- $H \geq pL_p \quad \Rightarrow \quad \Omega_H(x; y)$ is **convex** in y [3].
- For $p = 1$: inexact computation of proximal operator (see [5]).
- For $p = 2$: we can apply linear algebra techniques (usually require $O(n^3)$ arithmetical operations), or first-order optimization schemes for solving the subproblem (Hessian-free methods).
- For $p = 3$: efficient implementation, using Gradient Method with relative smoothness condition [3].
  The cost of minimizing $\Omega_H(x_k; \cdot)$ is: $O(n^3) + \tilde{O}(n)$.

Definition of Inexactness

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

$$\Omega_H(x_k; T) - \min_y \Omega_H(x_k; y) \;\leq\; \delta_{k+1}. \qquad (1)$$

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

Algorithm 1: Monotone Inexact Tensor Method

Initialization: choose $x^0 \in dom F$, set $H := pL_p$.

Iterations: $k \geq 0$.

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

2: Compute point T which satisfies (1) **and**

$$F(T) < F(x_k).$$

3: Make the step: $x_{k+1} \;:=\; T$.

Dynamic Strategy

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

$$F(x_k) - F^* \leq \frac{(p+1)^{p+1} L_p D^{p+1}}{p!\, k^p} + \frac{c}{k^p},$$

where D is the radius of the initial level set of the objective.

Adaptive Strategy

Let us set $\boxed{\delta_{k+1} := c \cdot \bigl(F(x_{k-1}) - F(x_k)\bigr),}$ for some $c \geq 0$.

Theorem 2. (General convex case)

$$F(x_k) - F^* \leq O\bigl(\tfrac{1}{k^p}\bigr).$$

Theorem 3. (Uniformly convex objective) Let

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

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

$$F(x_{k+1}) - F^* \leq \Bigl(1 - \frac{p\,\omega_p^{-1/p}}{2(p+1)}\Bigr)(F(x_k) - F^*).$$

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

Local convergence

Let $\boxed{\delta_{k+1} := c \cdot \bigl(F(x_{k-1}) - F(x_k)\bigr)^{\frac{p+1}{2}},}$ for some $c \geq 0$.

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

Contracting Proximal Scheme

- 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 increasing coefficients A_k, $a_{k+1} := A_{k+1} - A_k$.

Iterations, $k \geq 0$:

1. Compute

$$v_{k+1} = \mathop{\mathrm{argmin}}_y \Bigl\{ A_{k+1} f\bigl(\tfrac{a_{k+1} y + A_k x_k}{A_{k+1}}\bigr) + a_{k+1}\psi(y) + \beta_d(v_k; y) \Bigr\}. \qquad (2)$$

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

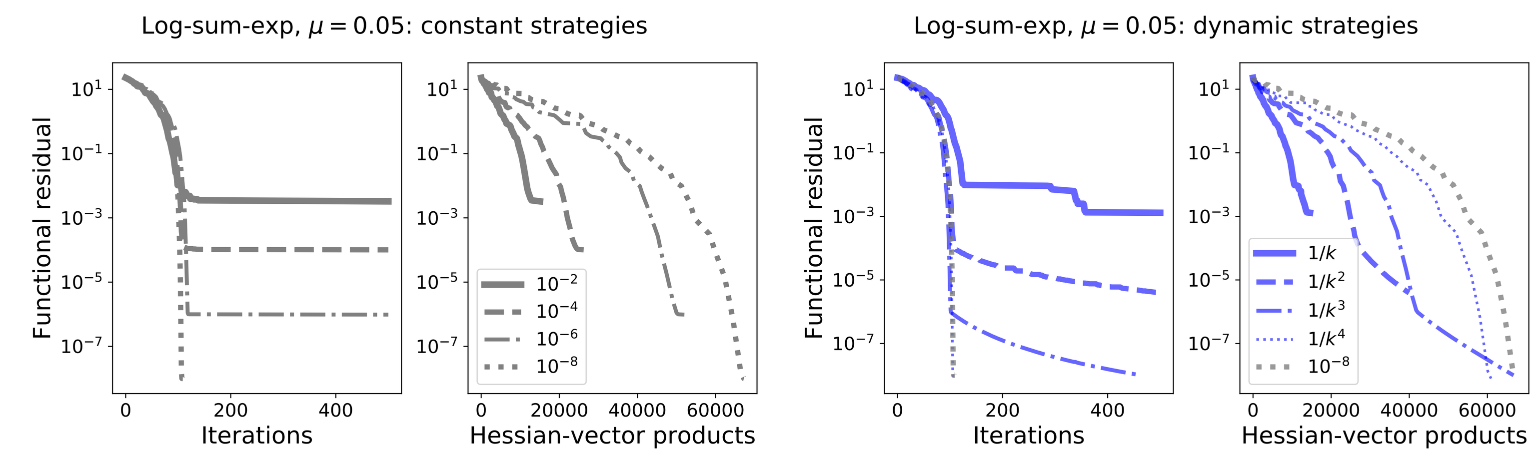
The rate of convergence [2]: $F(x_k) - F^* \;\leq\; \frac{\beta_d(x_0; x^*)}{A_k}$.

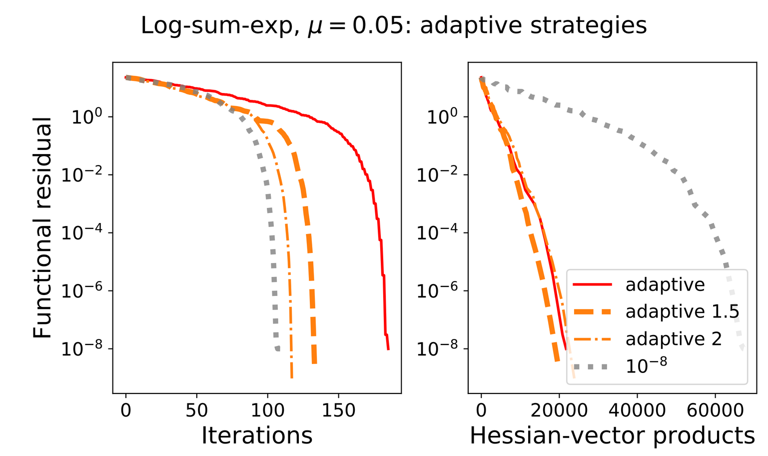
- Set $d(x) := \frac{1}{p}\|x - x_0\|^{p+1}$, $A_k := \frac{k^{p+1}}{L_p}$. We obtain **accelerated** rate

$$F(x_k) - F^* \leq O\bigl(\tfrac{L_p \|x^* - x_0\|^{p+1}}{k^{p+1}}\bigr).$$

- To compute (2), it is enough $\tilde{O}(1)$ inexact Tensor Steps (1).

Experiments





Minimization of Log-sum-exp objective by **Cubic Newton** ($p = 2$);

Each step is computed inexactly by **Fast Gradient Method**.

References

[1] Michel Baes. "Estimate sequence methods: extensions and approximations". In: *Institute for Operations Research, ETH, Zürich, Switzerland* (2009)

[2] Nikita Doikov and Yurii Nesterov. "Contracting proximal methods for smooth convex optimization". In: *CORE Discussion Papers 2019/27* (2019)

[3] Yurii Nesterov. "Implementable tensor methods in unconstrained convex optimization". In: *Mathematical Programming* (2019), pp. 1–27

[4] Yurii Nesterov. "Inexact basic tensor methods". In: *CORE Discussion Papers 2019/23* (2019)

[5] Mark Schmidt, Nicolas L Roux, and Francis R Bach. "Convergence rates of inexact proximal-gradient methods for convex optimization". In: *Advances in neural information processing systems*. 2011, pp. 1458–1466