Inexact Tensor Methods and Their Application to Stochastic Convex Optimization

Artem Agafonov, Dmitry Kamzolov, Pavel Dvurechensky, Alexander Gasnikov

Problem Statement: online setting

$$\min_{x \in \mathbb{E}} f(x) := \mathrm{E}_{\xi \sim \mathcal{D}}[f(x;\xi)],$$

functions $f, \nabla f, \nabla^2 f, \nabla^3 f, \ldots, \nabla^p f$ are Lipschitz continuous for all $i \in \{0,1,\ldots,p\}$, $x, y \in \mathbb{E}$:

$$\|\nabla^i f(x) - \nabla^i f(y)\| \leq L_i \|x - y\|.$$

And

$$\|\nabla^i f(x,\xi) - \nabla^i f(x)\| \leq M_i.$$

Problem Statement: offline setting

$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x).$$

functions $f, \nabla f, \nabla^2 f, \nabla^3 f, \dots, \nabla^p f$ are Lipschitz continuous for all $i \in \{0,1,\dots,p\},\ x, y \in \mathbb{E}$:

$$\|\nabla^i f(x) - \nabla^i f(y)\| \leq L_i \|x - y\|.$$

Some Definitions

Power prox function:

$$d_p(x) = \frac{1}{p}\|x\|^p$$

Taylor approximation of function f:

$$\Phi_{\mathrm{x},p}(\mathrm{y}) \stackrel{\mathrm{def}}{=} f(\mathrm{x}) + \sum_{i=1}^{p} \frac{1}{i!} D^i f(\mathrm{x})[\mathrm{y}-\mathrm{x}]^i, \quad y \in \mathbb{E}$$

Nesterov's model:

$$\Omega_{\mathrm{x},p}(\mathrm{y}) = \Phi_{\mathrm{x},p}(\mathrm{y}) + \frac{M}{(p-1)!} d_{p+1}(\mathrm{y}-\mathrm{x})$$

Some Definitions

Inexact Taylor approximation

$$\phi_{x_k,p}(y) = f(x_k) + g_k^\top (y - x_k) + \frac{1}{2}(y - x_k)^\top B_k (y - x_k) + \frac{1}{6} T_k[(y - x_k)]^3,$$

where g_k, B_k, T_k are approximate derivatives $\nabla f(x_k), \nabla^2 f(x_k), \nabla^3 f(x_k)$ through sampling.

Sampling Conditions

For a given ε accuracy, one can choose the size of the sample sets $\mathcal{S}_i$ for sufficiently small $\kappa_i > 0$ such that $\forall y \in \mathbb{R}^d$

$$\|(G_{\mathrm{x}_k,i} - \nabla^i f(x_k))[\mathrm{y} - \mathrm{x}_k]^{i-1}\| \leq \kappa_i \varepsilon^{(p-i+1)/p} \|\mathrm{y} - \mathrm{x}_k\|^{i-1}.$$

For sampled gradient, Hessian and tensor of third-order partial derivatives

$$\|g_k - \nabla f(x_k)\| \leq \kappa_g \varepsilon,$$
$$\|(B_k - \nabla^2 f(x_k))[y - x_k]\| \leq \kappa_b \varepsilon^{2/3} \|y - x_k\|,$$
$$\|T_k[y - x_k]^2 - \nabla^3 f(x_k)[y - x_k]^2\| \leq \kappa_t \varepsilon^{1/3} \|y - x_k\|^2.$$

Sampling Conditions

Lemma. Let Assumptions be satisfied. Then for any fixed small constants $\kappa_i > 0$ we can choose sample set $\mathcal{S}_i$ sizes of approximate derivatives G_i

$$n_i = \tilde{\mathcal{O}}\left(\frac{(L_{i-1} + M_i)^2}{\kappa_i^2} \cdot \varepsilon^{-\frac{2(p-i+1)}{p}}\right)$$

so that with probability $1 - \delta$ sampling conditions hold.

Inexact Model

$$\omega_{x,p}(y) = \phi_{x,p}(y) + \delta_1 \|y - x\| + \sum_{i=2}^{p} \frac{\delta_i}{(i-2)!} d_i(y - x) + \frac{\sigma}{(p-1)!} d_{p+1}(y - x)$$

Theorem 1.
Model $\omega_{x,p}(y)$ majorizes the function f:

$$f(x) \leq \omega_{x,p}(y).$$

Theorem 2.
Model $\omega_{x,p}(y)$ is convex for all $y \in \mathbb{E}$.

Algorithm

$$x_{t+1} = \arg\min_{y \in \mathbb{R}^n} \omega_{x_t,p}(y)$$

Theorem 3. If Condition 1 is satisfied and $\sigma \geq L_p$ then

$$f(x_{T+1}) - f(x_*) \leq$$

$$O\left(\kappa_1 \varepsilon D + \sum_{i=2}^{p} \frac{\kappa_i \varepsilon^{\frac{p+1-i}{p}} D^i}{T^{i-1}} + \frac{\sigma D^{p+1}}{T^p}\right).$$

Complexity $O\left(\left(L_p D^{p+1}/\varepsilon\right)^{1/p}\right)$.

Accelerated Stochastic Tensor Method

Sampling Conditions

For a given ε accuracy, one can choose the size of the sample sets $\mathcal{S}_i$ for sufficiently small $\kappa_i > 0$ such that $\forall y \in \mathbb{R}^d$

$$\frac{1}{2}\kappa_i \varepsilon^{\frac{p+1-i}{p+1}} \|\mathrm{y}-\mathrm{x}\|^{i-2} I \preccurlyeq (G_{\mathrm{x},i} - \nabla^i f(\mathrm{x}))[\mathrm{y}-\mathrm{x}]^{i-2}$$
$$\preccurlyeq \kappa_i \varepsilon^{\frac{p+1-i}{p+1}} \|\mathrm{y}-\mathrm{x}\|^{i-2} I, i = 2, \ldots, p.$$

Corollary.

$$\|(G_{\mathrm{x},i} - \nabla^i f(x))[\mathrm{y}-\mathrm{x}]^{i-1}\| \leq \kappa_i \varepsilon^{(p-i+1)/p} \|\mathrm{y}-\mathrm{x}\|.$$

Algorithm 2 Accelerated Inexact Tensor Method

1: **Input:** convex function f such that $\nabla^p f$ is L_p-Lipschitz; ε is target objective residual; x_0 is starting point; constants $\sigma \geq 3L_p$, $\beta > 0$; nonnegative nondecreasing sequences $\{\bar{\kappa}_i^t\}_{t\geq 0}$ for $i = 2, \ldots, p$, and

$$\alpha_t = \frac{p+1}{t+p+1}, \quad A_t = \prod_{i=1}^{t}(1-\alpha_t). \tag{41}$$

2: **Precomputation:** Call the inexact oracle to compute $G_{x_0,i}$ for $i = 1, \ldots, p$ such that Condition 2 is satisfied, compute

$$x_1 = \arg\min_{x\in\mathbb{R}^n}\{\phi_{x_0,p}(x) + \frac{\sigma}{(p-1)!}d_{p+1}(x-x_0)\} \tag{42}$$

$$y_1 = \arg\min_{x\in\mathbb{R}^n}\left\{\psi_1(x) := f(x_1) + \sum_{i=2}^{p}\frac{\bar{\kappa}^0}{(i-1)!}d_i(x-x_0) + \frac{\beta}{(p-1)!}d_{p+1}(x-x_0)\right\}. \tag{43}$$

3: **for** $t \geq 0$ **do**

4: Call the inexact oracle to compute $G_{x_t,i}$ for $i = 1, \ldots, p$ such that Condition is satisfied.

5: Set

$$u_t = (1-\alpha_t)x_t + \alpha_t y_t, \tag{44}$$

$$x_{t+1} = \arg\min_{x\in\mathbb{R}^n}\{\phi_{u_t,p}(x) + \frac{\sigma}{(p-1)!}d_{p+1}(x-u_t)\}. \tag{45}$$

6: Compute

$$y_{t+1} = \arg\min_{x\in\mathbb{R}^n}\left\{\psi_{t+1}(x) := \psi_t(x) + \sum_{i=2}^{p}\frac{\bar{\kappa}_i^t - \bar{\kappa}_i^{t-1}}{(i-1)!}d_i(x-x_0) + \frac{\alpha_t}{A_t}\Phi_{x_{t+1},1}(x)\right\}. \tag{46}$$

7: **end for**

Rate of Convergence

After T iterations of Accelerated Stochastic Tensor Method function f will satisfy:

$$f(x_T) - f(x_*) \leq$$
$$\leq O\left(\sum_{i=2}^{p} \frac{\kappa_i \varepsilon^{\frac{p+1-i}{p+1}} R^i}{T^i} + \frac{(L_p + p\beta)R^{p+1}}{T^{p+1}}\right).$$

Complexity $O(\varepsilon^{-\frac{1}{p+1}})$.

Implementation Details

Solution of the Auxiliary Problems

Smooth $\omega_{x,p}(s)$ using the following inequality

$$\|x\| \leq \frac{\|x\|^2}{2\alpha} + \frac{\alpha}{2}.$$

On each step we need to solve the following problem:

$$\zeta_{x,3}(y) = \phi_{x,3}(y) + \left(\frac{\sigma}{2} + \frac{2\kappa_t}{3}\right) d_4(y - x) +$$

$$+ \left(\frac{\kappa_g \varepsilon^{\frac{2}{3}}}{2} + \kappa_b \varepsilon^{\frac{2}{3}} + \frac{\kappa_t \varepsilon^{\frac{2}{3}}}{2}\right) d_2(y - x) + \frac{\kappa_g \varepsilon^{\frac{4}{3}}}{2} \to \min_{y \in \mathbb{E}}.$$

Solution of the Auxiliary Problem

Lemma 1.
Function $\zeta_k(\mathsf{s})$ satisfies the strong relative convexity and relative smoothness conditions

$$\nabla^2 \rho_{\mathsf{x}}(\mathsf{s}) \preccurlyeq \nabla^2 \zeta(\mathsf{s}) \preccurlyeq \left(\frac{\tau+2}{\tau-2}\right) \nabla^2 \rho_{\mathsf{x}}(\mathsf{s}).$$

with

$$\rho_x = \frac{1}{2}\left(1-\frac{2}{\tau}\right)\langle \mathrm{B}h, h\rangle + \frac{\sigma - L_3\tau}{2} d_4(\mathsf{s}) + \left(1-\frac{2}{\tau}\right) C_2 d_2(\mathsf{s}) + \left(1-\frac{2}{\tau}\right)\frac{2\kappa_t}{3} d_4(\mathsf{s}).$$

This condition allows us to solve the auxiliary problem very efficiently.

Solution of the Auxiliary Problem

We solve the auxiliary problem with the following algorithm:

$$\mathsf{h}_{k+1} = \arg\min_{\mathsf{h}\in\mathbb{E}} \left\{ \langle \nabla\zeta\left(\mathsf{h}_k\right), \mathsf{h} - \mathsf{h}_k \rangle + \kappa(\tau)\beta_{\rho_x}\left(\mathsf{h}_k, \mathsf{h}\right) \right),$$

where $\beta_{\rho_x}(\mathsf{u},\mathsf{v})$ is the Bregman divergence of function $\rho_x(\cdot)$:
$\beta_{\rho_x}(\mathsf{u},\mathsf{v}) = \rho_x(\mathsf{v}) - \rho_x(\mathsf{u}) - \langle \nabla\rho_x(\mathsf{u}), \mathsf{v}-\mathsf{u}\rangle.$

This method has linear rate of convergence.

Implementation Details

Accelerated Stochastic Tensor Method

Sampling Conditions

For a given ε accuracy, one can choose the size of the sample sets $\mathcal{S}_i$ for sufficiently small $\kappa_i > 0$ such that $\forall \mathbf{y} \in \mathbb{R}^d$

$$\frac{1}{2}\kappa_i \varepsilon^{\frac{p+1-i}{p+1}} \|\mathrm{y} - \mathrm{x}\|^{i-2} I \preccurlyeq (G_{\mathrm{x},i} - \nabla^i f(\mathrm{x}))[\mathrm{y} - \mathrm{x}]^{i-2}$$
$$\preccurlyeq \kappa_i \varepsilon^{\frac{p+1-i}{p+1}} \|\mathrm{y} - \mathrm{x}\|^{i-2} I, i = 2, \ldots, p.$$

Sampling Conditions

To satisfy sampling conditions for accelerated method we can do the following:

- Sample $G_{x,i}$ (as in non-accelerated method) s.t. $\|(\mathbf{G}_{x,i} - \nabla^i f(\mathbf{x}))[y - x]^{i-1}\| \leq \kappa_i \varepsilon^{(p-i+1)/p} \|y - x\|$.
- Add regularization $3\kappa_i \varepsilon^{\frac{p+i-1}{p+1}} E_i[y - x]^{i-2}$.

We obtain

$$2\kappa_i \varepsilon^{\frac{p+1-i}{p+1}} \|y - x\|^{i-2} I \preccurlyeq (G_{x,i} + 3\kappa_i \varepsilon^{\frac{p+1-i}{p+1}} E_i - \nabla^i f(x))[y -$$

$$\preccurlyeq 4\kappa_i \varepsilon^{\frac{p+1-i}{p+1}} \|y - x\|^{i-2} I, i = 2, \dots, p.$$

Auxiliary Problem

On each step of accelerated method we need to solve the following subproblem

$$x_{t+1} = \arg\min_{x \in \mathbb{R}^n} \{ \tilde{\phi}_{u_t,p}(x) + \frac{\sigma}{(p-1)!} d_{p+1}(x - u_t) \}.$$

Using regularization from the previous slide the problem becomes:

$$x_{t+1} = \arg\min_{x \in \mathbb{E}} \{ \phi_{u_t,p}(x) + \sum_{i=2}^{p} 2\kappa_i \varepsilon^{\frac{p+1-2}{p+1}} d_i(x - u_t) + \frac{\sigma}{(p-1)!} d_{p+1}(x - u_t) \}.$$

That subproblem can be solved for $p = 3$ like in non-accelerated method.