Derivative-Free Method For Decentralized Distributed Non-Smooth Optimization

A. Beznosikov, E. Gorbunov, A. Gasnikov

Moscow Institute of Physics and Technology

2 December 2019

Content

Original problem

- Composite optimization problem

$$\Psi_0(x) = f(x) + g(x) \rightarrow \min_{x \in X}$$

- $X \subseteq \mathbb{R}^n$ is a compact and convex set with diameter D_X.
- Function g is convex and L-smooth on X,
- Function f is convex differentiable function on X with bounded gradient ($x \in X$ we have $\|\nabla f(x)\|_* \leq M$).

Oracles

- Gradient $\nabla g(x)$ is available.
- For f we have only stochastic zeroth-order oracle

$$\tilde{f}(x) = f(x) + \Delta(x) + \xi(x)$$

where $\Delta(x)$ is the bounded noise of unknown nature

$$|\Delta(x)| \leq \Delta$$

and $\xi(x)$ is a stochastic noise which satisfies

$$\mathbb{E}[\xi(x) \mid x] = 0, \quad \mathbb{E}[\xi^4(x) \mid x] \leq B^4.$$

- Stochastic approximation of $\nabla f(x)$:

$$\tilde{f}'_r(x) = \frac{n}{2r}(\tilde{f}(x + re) - \tilde{f}(x - re))e$$

where u is a random vector uniformly distributed on the Euclidean sphere and r is smoothing parameter.

Smoothed problem

- Smoothed version of $f(x)$

$$F(x) = \mathbb{E}_e[f(x + re)]$$

- Smoothed problem

$$\Psi(x) = F(x) + g(x) \to \min_{x \in X}$$

Algorithm

Algorithm 1 Zeroth-Order Sliding Algorithm (`zoSA`)

Input: Initial point $x_0 \in X$ and iteration limit N.
Let $\beta_k \in \mathbb{R}_{++}, \gamma_k \in \mathbb{R}_+$, and $T_k \in \mathbb{N}$, $k = 1, 2, \ldots$, be given and set $\bar{x}_0 = x_0$.
for $k = 1, 2, \ldots, N$ **do**

1. Set $\underline{x}_k = (1 - \gamma_k)\bar{x}_{k-1} + \gamma_k x_{k-1}$,
   and let $h_k(\cdot) \equiv l_g(\underline{x}_{k-1}, \cdot) = g(x) + \langle \nabla g(x), y - x \rangle$.
2. Set
$$(x_k, \tilde{x}_k) = \text{PS}(h_k, x_{k-1}, \beta_k, T_k);$$
3. Set $\bar{x}_k = (1 - \gamma_k)\bar{x}_{k-1} + \gamma_k \tilde{x}_k$.

end for
Output: $\bar{x}_N$.

Algorithm

Algorithm 2 The PS (prox-sliding) procedure

procedure $(x^+, \tilde{x}^+) = \mathrm{PS}(h, x, \beta, T)$

Let the parameters $p_t \in \mathbb{R}_{++}$ and $\theta_t \in [0, 1]$, $t = 1, \ldots,$ be given. Set $u_0 = \tilde{u}_0 = x$.

for $t = 1, 2, \ldots, T$ **do**

$$u_t = \arg\min_{u \in X} \left\{ h(u) + \langle \tilde{f}_r'(x), u \rangle + \beta V(x, u) + \beta p_t V(u_{t-1}, u) \right\},$$
$$\tilde{u}_t = (1 - \theta_t)\tilde{u}_{t-1} + \theta_t u_t.$$

end for

Set $x^+ = u_T$ and $\tilde{x}^+ = \tilde{u}_T$.

end procedure

Convergence

Theorem Suppose $\{p_t\}$, $\{\theta_t\}$, $\{\beta_k\}$, $\{\gamma_k\}$, $\{T_k\}$ satisfy some conditions. Then

$$\mathbb{E}[\Psi(\overline{x}_N) - \Psi(x^*)] \leq \frac{12LD_X^2}{N(N+1)} + \frac{n\Delta D_X p_*}{r} \quad \forall N \geq 1.$$

Convergence

Corollary For all $N \geq 1$:

$$\mathbb{E}[\Psi_0(\overline{x}_N) - \Psi_0(x^*)] \leq 2rM + \frac{12LD_X^2}{N(N+1)} + \frac{n\Delta D_X p_*}{r}$$

If

$$r = \Theta\left(\frac{\varepsilon}{M}\right), \Delta = O\left(\frac{\varepsilon^2}{nMD_X}\right), B = O\left(\frac{\varepsilon}{\sqrt{n}}\right)$$

then the number of evaluations for ∇g and $\tilde{f}_r'$ to find a ε-solution can be bounded by

$$O\left(\sqrt{\frac{LD_X^2}{\varepsilon}}\right)$$

$$O\left(\sqrt{\frac{LD_X^2}{\varepsilon}} + \frac{D_X^2 p_*^2 n M^2 (C_1^2 + 1)}{\varepsilon^2}\right).$$

Convergence: special cases

- Euclidean case, i.e. $\|\cdot\| = \|\cdot\|_2$. $p_* = C_1 = C_2 = 1$ and the number of $\tilde{f}'_r$ oracle calls reduces to

$$O\left(\sqrt{\frac{LD_X^2}{\varepsilon}} + \frac{D_X^2 n M^2}{\varepsilon^2}\right)$$

- Case when $\|\cdot\| = \|\cdot\|_1$. $p_* = O\left(\ln(n)/n\right)$ and $C_1 = 1$, $C_2 = \sqrt{n}$. The number of $\tilde{f}'_r(x)$ computations:

$$O\left(\sqrt{\frac{LD_X^2}{\varepsilon}} + \frac{D_X^2 M^2 \ln n}{\varepsilon^2}\right).$$

When X is a probability simplex we have $D_X = 2$.

Convex Optimization with Affine Constraints

- $$f(x) \to \min_{Ax=0, x\in X},$$
  where $A \succeq 0$ and $\mathrm{Ker} A \neq \{0\}$ and X is convex compact in $\mathbb{R}^n$ with diameter D_X.
- Penalized problem
  $$F(x) = f(x) + \frac{R_y^2}{\varepsilon}\|Ax\|_2^2 \to \min_{x\in X},$$
  where $\varepsilon > 0$ is some positive number.

Convergence

zoSA Algorithm requires

$$O\left(\sqrt{\frac{\lambda_{\max}(A^\top A)R_y^2D_X^2}{\varepsilon^2}}\right) \text{ calculations of } A^\top Ax$$

and

$$O\left(\sqrt{\frac{\lambda_{\max}(A^\top A)R_y^2D_X^2}{\varepsilon^2}} + \frac{nD_X^2M^2}{\varepsilon^2}\right) \text{ calculations of } \tilde{f}(x)$$

Decentralized Distributed Optimization

- $$f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x_i) \to \min_{\substack{x_1 = \ldots = x_m \\ x_1,\ldots,x_m \in X}},$$
  where $x^\top = (x_1^\top, \ldots, x_m^\top)^\top \in \mathbb{R}^{nm}$
- Equivalent problem
  $$f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x_i) \to \min_{\substack{\sqrt{W}x = 0, \\ x_1,\ldots,x_m \in X}}.$$

$$W = \overline{W} \otimes I_n \qquad \overline{W}_{ij} = \begin{cases} -1, & \text{if } (i,j) \in E, \\ \deg(i), & \text{if } i = j, \\ 0 & \text{otherwise}, \end{cases}$$

Convergence

zoSA Algorithm requires

$$O\left(\sqrt{\frac{\chi(W)M^2D_X^2}{\varepsilon^2}}\right) \text{ communication rounds}$$

and

$$O\left(\sqrt{\frac{\chi(W)M^2D_X^2}{\varepsilon^2}} + \frac{nD_X^2M^2}{\varepsilon^2}\right) \text{calculations of } \tilde{f}(x)$$