Derivative-Free Method For Decentralized Distributed Non-Smooth Optimization

A. Beznosikov, E. Gorbunov, A. Gasnikov

Moscow Institute of Physics and Technology (National Research University),
Sirius University of Science and Technology

12 July 2020

Original problem

- Composite optimization problem

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

- $X \subseteq \mathbb{R}^n$ is a compact and convex set with diameter D_X.
- Function g is convex and L-smooth on X.
- Function f is convex, differentiable on X with a bounded gradient.

Oracles

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

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

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

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

and ξ is responsible for a stochastic noise

$$\mathbb{E}[f(x, \xi)] = f(x), \quad \|\nabla f(x, \xi)\|_2 \leq M(\xi), \quad \mathbb{E}[M^2(\xi)] = M^2.$$

- Approximation of $\nabla f(x)$:

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

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

Smoothed problem

- Smoothed version of $f(x)$

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

- Smoothed problem

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

Prox first-order method

Algorithm 1 Accelerated proximal first-order method

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

1. $\underline{x}_k = \bar{x}_{k-1} + \gamma_k x_{k-1}$.
2. $$x_k = \arg\min_{u \in X} \{g(\underline{x}_k) + \langle \nabla g(\underline{x}_k), u - \underline{x}_k \rangle + f(u) + \beta_k V(x_{k-1}, u)\}$$
3. $\bar{x}_k = \bar{x}_{k-1} + \gamma_k x_k$.

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

Sliding Algorithm

Algorithm 2 Sliding Algorithm

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 $h_k(\cdot) \equiv g(\underline{x}_{k-1}) + \langle \nabla g(\underline{x}_{k-1}), \cdot - \underline{x}_{k-1} \rangle$.
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: $\bar{x}_N$.

Sliding Algorithm

Algorithm 3 The PS (prox-sliding) procedure

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

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**

$$u_t = \arg\min_{u \in X} \left\{ h(u) + \langle \nabla f(u_{t-1}), u \rangle + \beta V(x,u) + \beta p_t V(u_{t-1}, u) \right\},$$

$$\tilde{u}_t = (1-\theta_t)\tilde{u}_{t-1} + \theta_t u_t.$$

end for

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

end procedure

Convergence

- Accelerated proximal first-order method. The number of ∇g and ∇f oracles calls

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

- Sliding Algorithm. The number of ∇g and ∇f computations:

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

New Algorithm

Algorithm 4 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 $h_k(\cdot) \equiv g(\underline{x}_{k-1}) + \langle \nabla g(\underline{x}_{k-1}), \cdot - \underline{x}_{k-1} \rangle$.
2. Set
$$(x_k, \tilde{x}_k) = \mathrm{PS}(h_k, x_{k-1}, \beta_k, T_k);$$
3. Set $\bar{x}_k = (1 - \gamma_k)\bar{x}_{k-1} + \gamma_k \tilde{x}_k$.

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

New Algorithm

Algorithm 5 The PS (prox-sliding) procedure

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

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**

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

$$\tilde{u}_t = (1-\theta_t)\tilde{u}_{t-1} + \theta_t u_t.$$

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

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

where N – number of iterations.

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

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

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. $\|\cdot\| = \|\cdot\|_2$. 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 $\|\cdot\| = \|\cdot\|_1$. The number of $\tilde{f}'_r(x)$ computations:

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

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 R_y, ε are some positive numbers.

Convergence

zoSA Algorithm requires

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

and

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