Gradient Free Methods for Non-Smooth Convex Stochastic Optimization with Heavy Tails on Convex Compact

Nikita Kornilov, Alexander Gasnikov, Pavel Dvurechensky, Darina Dvinskikh

Moscow Institute of Physics and Technology

20 May, 2023

Plan

Problem

Consider stochastic non-smooth convex minimization problem over compact convex set $\mathcal{S} \subset \mathbb{R}^d$ with function $f : \mathbb{R}^d \to \mathbb{R}$

$$\min_{x \in \mathcal{S}} f(x) \triangleq \mathbb{E}_\xi[f(x, \xi)],$$

where the values of the objective are available only through a zeroth-order noisy corrupted oracle , i.e.

$$\phi(x, \xi) = f(x, \xi) + \delta(x).$$

We consider two-point zeroth-order oracle, i.e. for two query points $x, y \in \mathcal{S}$ we are given two outputs $\phi(x, \xi)$ and $\phi(y, \xi)$ with the same ξ.

Assumptions

1. Function $f(x, \xi)$ is convex w.r.t. x for any ξ on $\mathcal{S}$.
2. Function $f(x, \xi)$ is $M_2(\xi)$-Lipschitz continuous w.r.t. x in the l_2-norm, i.e., for all $x_1, x_2 \in \mathcal{S}$

   $$|f(x_1, \xi) - f(x_2, \xi)| \leq M_2(\xi)\|x_1 - x_2\|_2.$$

   Moreover, there exist $\kappa \in (0, 1]$ and M_2 such that $\mathbb{E}_\xi[M_2^{1+\kappa}(\xi)] \leq M_2^{1+\kappa}$.
3. For all $x \in \mathcal{S} : |\delta(x)| \leq \Delta < \infty$

Approximation and Sampling

In order to make approximation of objective function gradient we sample vector $\mathbf{e}$ from uniform distribution on Euclidean sphere $\{\mathbf{e} : ||\mathbf{e}||_2 = 1\}$.
Smoothed function

$$\hat{f}_\tau(x) = \mathbb{E}_{\mathbf{e}}[f(x + \tau\mathbf{e})]$$

Its gradient

$$\nabla \hat{f}_\tau(x) = \mathbb{E}_{\mathbf{e}}\left[\frac{d}{\tau} f(x + \tau\mathbf{e})\mathbf{e}\right]$$

Gradient approximation

$$g(x, \xi, \mathbf{e}) = \frac{d}{2\tau}(\phi(x + \tau\mathbf{e}, \xi) - \phi(x - \tau\mathbf{e}, \xi))\mathbf{e}$$

for $\tau > 0$.

Smoothing Example

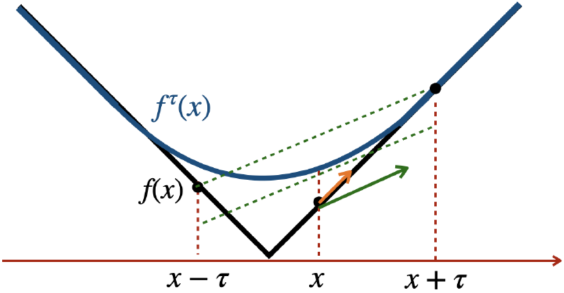

Figure: Smoothed function

Approximation Quality

Approximation quality

$$\sup_{x \in \mathcal{S}} |\hat{f}_\tau(x) - f(x)| \leq \tau M_2.$$

Gradient $(1+\kappa)$-th moment is bounded

$$\mathbb{E}_{\xi,\mathbf{e}}[\|g(x,\xi,\mathbf{e})\|_q^{1+\kappa}] \leq 2^\kappa \left(\frac{\sqrt{d}}{2^{1/4}} a_q M_2\right)^{1+\kappa} + 2^\kappa \left(\frac{d a_q \Delta}{\tau}\right)^{1+\kappa} = \sigma_q^{1+\kappa},$$

where $a_q = d^{\frac{1}{q} - \frac{1}{2}} \min\{\sqrt{32 \ln d - 8}, \sqrt{2q-1}\}$.

Stochastic Mirror Descent

Let function $\Psi : \mathbb{R}^d \to \mathbb{R}$ be 1 strongly-convex w.r.t. the ℓ_p-norm and continuously differentiable. We denote its Fenchel conjugate and its Bregman divergence respectively as

$$\Psi^*(y) = \sup_{x \in \mathbb{R}^d} \{\langle x, y \rangle - \Psi(x)\}$$

$$D_\Psi(y, x) = \Psi(y) - \Psi(x) - \langle \nabla\Psi(x), y - x \rangle.$$

The Stochastic Mirror Descent updates with stepsize ν and update vector g_{k+1} are as follows:

$$y_{k+1} = \nabla(\Psi^*)(\nabla\Psi(x_k) - \nu g_{k+1}), \quad x_{k+1} = \arg\min_{x \in \mathcal{S}} D_\Psi(x, y_{k+1}). \tag{1}$$

Convexity Generalization

From article "Mirror Descent Strikes Again: Optimal Stochastic Convex Optimization under Infinite Noise Variance" by Nuri Mert Vural.

Definition

Uniform convex. Consider a differentiable convex function $\psi : \mathbb{R}^d \to \mathbb{R}$, an exponent $r \geq 2$, and a constant $K > 0$. Then, ψ is (K, r)-uniformly convex w.r.t. p-norm if for any $x, y \in \mathbb{R}^d$

$$\psi(y) - \psi(x) - \langle \psi(x), y - x \rangle \geq \frac{K}{r} ||x - y||_p^r.$$

Theorem

Consider some $\kappa \in (0,1], p \in [1,\infty]$, *$q$ defined by the equality* $\frac{1}{q} + \frac{1}{p} = 1$, *and function* Ψ_p *which is* $\left(1, \frac{1+\kappa}{\kappa}\right)$*-uniformly convex w.r.t. p norm. Then, for the SMD Algorithm outlined in* (1) *with the corresponding map function* $\nabla\Psi_p$, *after T iterations with any* $g_k \in \mathbb{R}^d, k \in \overline{1, T}$ *and starting point* $x_0 = \arg\min\limits_{x \in \mathcal{S}} \Psi(x)$ *we have*

$$\frac{1}{T}\sum_{k=0}^{T-1}\langle g_{k+1}, x_k - x^*\rangle \le \frac{\kappa}{\kappa+1}\frac{R_0^{\frac{1+\kappa}{\kappa}}}{\nu T} + \frac{\nu^\kappa}{1+\kappa}\frac{1}{T}\sum_{k=0}^{T-1}\|g_{k+1}\|_q^{1+\kappa}, \tag{2}$$

where $R_0^{\frac{1+\kappa}{\kappa}} = \frac{1+\kappa}{\kappa}(\Psi_p(x^*) - \Psi_p(x_0))$.

Zeroth-Order Robust SMD Algorithm

procedure $\textsc{Zero Robust SMD}$(Number of iterations T, stepsize ν, prox-function Ψ_p, smoothing constant τ)

  $x_0 \leftarrow \arg\min_{x \in \mathcal{S}} \Psi_p(x)$

  for $k = 0, 1, \ldots, T-1$ **do**

    Sample $\mathbf{e}_k \sim \mathrm{Uniform}(\{\mathbf{e} : \|\mathbf{e}\|_2 = 1\})$ independently

    Sample ξ_k independently

    $g_{k+1} = \frac{d}{2\tau}(\phi(x_k + \tau\mathbf{e}_k, \xi_k) - \phi(x_k - \tau\mathbf{e}_k, \xi_k))\mathbf{e}_k$

    Calculate $y_{k+1} \leftarrow \nabla(\Psi_p^*)(\nabla\Psi_p(x_k) - \nu g_{k+1})$

    Calculate $x_{k+1} \leftarrow \arg\min_{x \in \mathcal{S}} D_{\Psi_p}(x, y_{k+1})$

  end for

  return $\overline{x}_T \leftarrow \frac{1}{T}\sum_{k=0}^{T-1} x_k$

end procedure

Robust SMD Algorithm Convergence

Theorem

Let $q \in [2, \infty]$, arbitrary number of iterations T, smoothing constant $\tau > 0$ be given. Choose $\left(1, \frac{1+\kappa}{\kappa}\right)$-uniformly convex w.r.t. the p-norm function $\Psi_p(x)$. Set the stepsize $\nu = \frac{R_\Psi^{1/\kappa}}{\sigma_q} T^{-\frac{1}{1+\kappa}}$, $R_0^{\frac{1+\kappa}{\kappa}} = \frac{1+\kappa}{\kappa}(\Psi_p(x^*) - \Psi_p(x_0))$ and $\mathcal{D}_\Psi^{\frac{1+\kappa}{\kappa}} = \frac{1+\kappa}{\kappa} \sup\limits_{x,y \in \mathcal{S}} D_{\Psi_p}(x, y)$. Let $\overline{x}_T$ be the output of Algorithm with the above parameters

$$\mathbb{E}[f(\overline{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi + \frac{R_0\sigma_q}{T^{\frac{\kappa}{1+\kappa}}}, \tag{3}$$

where $\sigma_q^{1+\kappa} = 2^\kappa \left(\frac{\sqrt{d}}{2^{1/4}} a_q M_2\right)^{1+\kappa} + 2^\kappa \left(\frac{d a_q \Delta}{\tau}\right)^{1+\kappa}$.

Clipping technique

Given a constant $c > 0$, the clipping operator applied to a vector g is given by

$$\hat{g} = \frac{g}{\|g\|} \min(\|g\|, c).$$

If g is an unbiased stochastic gradient, then, on the one hand, $\hat{g}$ is bounded, and, on the other hand, is a biased stochastic gradient. Thus, the constant c allows playing with the trade-off between the faster convergence and bias $\|\mathbb{E}[\hat{g} - g]\|$ when $c \to 0$.

Clipping Algorithm

procedure $\textsc{Zero CLIP}$(Number of iterations T, stepsize ν, clipping constant c, prox-function Ψ_p, smoothing constant τ)

$\quad x_0 \leftarrow \arg\min_{x \in \mathcal{S}} \Psi_p(x)$

$\quad$ **for** $k = 0, 1, \ldots, T-1$ **do**

$\quad\quad$ Sample $\mathbf{e}_k \sim \mathrm{Uniform}(\{\mathbf{e} : \|\mathbf{e}\|_2 = 1\})$ independently

$\quad\quad$ Sample ξ_k independently

$\quad\quad$ $g_{k+1} = \frac{d}{2\tau}(\phi(x_k + \tau \mathbf{e}_k, \xi_k) - \phi(x_k - \tau \mathbf{e}_k, \xi_k))\mathbf{e}_k$

$\quad\quad$ Calculate $\hat{g}_{k+1} = \frac{g_{k+1}}{\|g_{k+1}\|_q} \min(\|g_{k+1}\|_q, c)$

$\quad\quad$ Calculate $y_{k+1} \leftarrow \nabla(\Psi_p^*)(\nabla \Psi_p(x_k) - \nu \hat{g}_{k+1})$

$\quad\quad$ Calculate $x_{k+1} \leftarrow \arg\min_{x \in \mathcal{S}} D_{\Psi_p}(x, y_{k+1})$

$\quad$ **end for**

$\quad$ **return** $\overline{x}_T \leftarrow \frac{1}{T} \sum_{k=0}^{T-1} x_k$

end procedure

Clipping Algorithm Convergence

Theorem

Let $q \in [2, \infty]$, arbitrary number of iterations T, smoothing constant $\tau > 0$ be given. Choose 1-strongly convex w.r.t. the p-norm prox-function $\Psi_p(x)$. Set the clipping constant $c = T^{\frac{1}{(1+\kappa)}} \sigma_q$. After set the stepsize $\nu = \frac{\mathcal{D}_\Psi}{c}$ with diameter$\mathcal{D}^2_\Psi = 2 \sup\limits_{x,y \in \mathcal{S}} D_{\Psi_p}(x, y)$.

Let $\overline{x}_T$ be a point obtained by Clipping Algorithm with the above parameters

$$\mathbb{E}_{\xi,\mathbf{e}}[f(\overline{x}_T)] - f(x^*) \leq 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D}_\Psi + \frac{\mathcal{D}_\Psi \sigma_q}{T^{\frac{\kappa}{1+\kappa}}}, \tag{4}$$

where $\sigma_q^{1+\kappa} = 2^\kappa \left(\frac{\sqrt{d}}{2^{1/4}} a_q M_2\right)^{1+\kappa} + 2^\kappa \left(\frac{d a_q \Delta}{\tau}\right)^{1+\kappa}$.
Or with probability at least $1 - \delta$ right part of (4) correct is up to $\log \frac{1}{\delta}$ factor denoted by $\tilde{O}$.

Maximum level of adversarial noise

Let ε be desired function value accuracy, i.e. with probability at least $1-\delta : f(\overline{x}_T) - f(x^*) \leq \varepsilon$.
If there is no adversarial noise, i.e., $\Delta = 0$, then the number of iterations is $T^{\frac{\kappa}{1+\kappa}} = \tilde{O}\left(\frac{\mathcal{D}_\Psi \sqrt{d} a_q M_2}{\varepsilon}\right)$ when $\tau \to 0$.
Also,

$$\text{when} \quad \tau = \frac{\varepsilon}{M_2} \text{ and } \Delta \leq \frac{\varepsilon^2}{M_2 \sqrt{d} \mathcal{D}_\Psi} \Rightarrow \text{rate is the same.}$$

Otherwise, when $\Delta > \frac{\varepsilon^2}{M_2 \sqrt{d} \mathcal{D}_\Psi}$, the convergence rate is twice as bad and we can't achieve accuracy less than $\sqrt{M_2 \sqrt{d} \Delta \mathcal{D}_\Psi}$.

d dependency

In the smooth case to approximate gradient it suffices to use $d+1$ function values. For the first-order stochastic methods, the optimal oracle complexity is proportional to $\varepsilon^{-\frac{1+\kappa}{\kappa}}$, thus for zeroth-order oracle we may expect the bound $d\varepsilon^{-\frac{1+\kappa}{\kappa}}$. In this paper we obtain the bound $\left(\sqrt{d}/\varepsilon\right)^{\frac{1+\kappa}{\kappa}}$ matching the expected bound only for $\kappa = 1$.

Is this bound optimal?

For smooth stochastic convex optimization problems with $(d+1)$-points stochastic zeroth-order oracle the answer is negative and the optimal bound is $\sim d\varepsilon^{-\frac{1+\kappa}{\kappa}}$.

Recommendations for choosing Ψ

The two main setups are given by

1. Ball setup:
$$p = 2, \Psi(x) = \frac{1}{2}\|x\|_2^2, \tag{5}$$
2. Entropy setup:
$$p = 1, \Psi(x) = (1+\gamma)\sum_{i=1}^{d}(x_i + \gamma/d)\log(x_i + \gamma/d), \gamma > 0. \tag{6}$$

Introduce standard sets $B^p = \{x \in \mathbb{R}^d : \|x\|_p \leq 1\}$ and $\Delta_d^+ = \{x \in \mathbb{R}^d : x \geq 0, \sum_i x_i = 1\}$.

Recommendations for choosing Ψ

Table: $T^{\frac{\kappa}{1+\kappa}}$ for Clipping Algorithm

Δ_d^+	B^1	B^2	B^∞
Entropy	Entropy	Ball	Ball
$\ln dM_2/\varepsilon$	$\ln dM_2/\varepsilon$	$\sqrt{d}M_2/\varepsilon$	dM_2/ε

Table: Maximum feasible noise level Δ up to $O(1)$ factor for Clipping Algorithm

Δ_d^+	B^1	B^2	B^∞
Entropy	Entropy	Ball	Ball
$\varepsilon^2/(\sqrt{d \ln d}M_2)$	$\varepsilon^2/(\sqrt{d \ln d}M_2)$	$\varepsilon^2/(\sqrt{d}M_2)$	$\varepsilon^2/(dM_2)$

Restart technique

Definition

Function f is r-growth function if there is $r \geq 1$ and $\mu_r \geq 0$ such that for all x
$$\frac{\mu_r}{2}\|x - x^*\|_p^r \leq f(x) - f(x^*),$$
where x^* is problem solution.

In particular, μ-strong convex w.r.t. the p-norm functions are 2-growth.
For functions with r-growth condition there is restart technique for algorithms acceleration.

Restart Algorithm

procedure ZEROTH-ORDER RESTART(Algorithm type $\mathcal{A}$, number of restarts N, sequence of number of steps $\{T_k\}_{k=1}^N$, sequence of smoothing constants $\{\tau_k\}_{k=1}^N$, sequence of stepsizes $\{\nu_k\}_{k=1}^N$, sequence of clipping constants $\{c_k\}_{k=1}^N$(if necessary), prox-function Ψ_p)

  $x_0 \leftarrow \arg\min\limits_{x \in \mathcal{S}} \Psi_p(x)$ or randomly

  for $k = 0, 1, \dots, N$ **do**

    Set parameters $\nu_k, (c_k), \Psi_p, \tau_k$ of the Algorithm $\mathcal{A}$

    Compute T_k iterations of the Algorithm $\mathcal{A}$ with starting point x_0 and get x_{final}

    $x_0 \leftarrow x_{\text{final}}$

  end for

  return x_{final}

end procedure

Restart Algorithm Convergence

Theorem

For the Clipping Algorithm let ε be fixed accuracy with probability at least $1-\delta$ and r-growth Assumption is held for $r \geq 1$. Then with certain parameters Restart Algorithm achieves desired accuracy after total number of steps and total number of restarts

$$T = \tilde{O}\left(\left[\frac{a_q M_2 \sqrt{d}}{\mu_r^{1/r}} \cdot \frac{1}{\varepsilon^{\frac{(r-1)}{r}}}\right]^{\frac{1+\kappa}{\kappa}}\right), N = \tilde{O}\left(\frac{1}{r}\log_2\left(\frac{\mu_r R_0^r}{2\varepsilon}\right)\right).$$

Moreover, Adversarial Noise Assumption must be held with

$$\Delta_k = \tilde{O}\left(\frac{\mu_r^2 R_0^{(2r-1)}}{M_2\sqrt{d}} \frac{1}{2^{k(2r-1)}}\right) \geq \tilde{O}\left(\frac{\mu_r^{1/r}}{M_2\sqrt{d}}\varepsilon^{(2-1/r)}\right), \quad 1 \leq k \leq N.$$

Online Optimization and Nonlinear Bandits

One needs to find sequence $\{x_t\} \in \mathcal{S}$ to minimize pseudo-regret

$$\mathcal{R}_T(\{f_t(\cdot)\}, \{x_t\}) = \sum_{t=1}^{T} f_t(x_t) - \min_{x \in \mathcal{S}} \sum_{t=1}^{T} f_t(x).$$

After each choice of x_t we get loss $\phi_t(x_t, \xi_t) = f_t(x_t, \xi_t) + \delta_t(x_t)$ given by one point oracle. Choice of x_t can be based only on available information

$$\{\phi_1(x_1, \xi_1), \ldots, \phi_{t-1}(x_{t-1}, \xi_{t-1})\}.$$

Theorem

Let ε be desired average pseudo regret accuracy. Under almost similar Assumptions on loss functions and adversarial noise with modified Clipping Algorithm number of iterations to achieve accuracy is

$$T = O\left(\left(\frac{M_2 d a_q \mathcal{D}_\psi}{\varepsilon^2}\right)^{\frac{\kappa+1}{\kappa}}\right).$$

Questions?

Thank You For Your Attention!