Gradient Free Optimization with Infinite Variance

Nikita Kornilov

Moscow Institute of Physics and Technology

14 December, 2022

Plan

1. Problem Statement
2. Function Smoothing
3. Mirror Descent with Infinite Noise
4. Main Algorithm and Theorem
5. Future Plans

Problem

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

$$\min_{\mathcal{S}} f(x)$$

where $f(x) = \mathbb{E}_\xi[f(x,\xi)]$ and $f : \mathcal{S} \to \mathbb{R}$ is convex and Lipschitz continuous function.
We are given zeroth order oracle $\phi(x,\xi) = f(x,\xi) + \delta(x)$ with adversarial noise $\delta(x)$.

Assumptions

1. Function $f(x, \xi)$ is convex and $M_2(\xi)$ Lipschitz continuous w.r.t. l_2 norm. 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, $\exists \kappa \in (0, 1]$ such that $\mathbb{E}_\xi[M_2^{1+\kappa}(\xi)] \leq M_2^{1+\kappa}$
2. 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

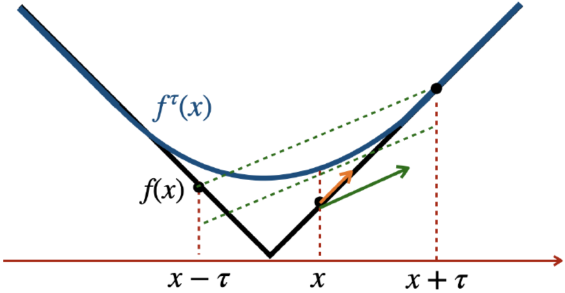

Figure: Smoothed function

Approximation Quality

Let $q \geq 2$. By definition $\mathbb{E}_{\mathbf{e}}\left[||\mathbf{e}||_q^{2(1+\kappa)}\right] \leq a_{q,\kappa}^{2(1+\kappa)}$. Then

$$a_{q,\kappa} = d^{\frac{1}{q}-\frac{1}{2}} \min\{\sqrt{32 \ln d - 8}, \sqrt{2q-1}\}$$

Smoothed Approximation

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

Gradient Norm

$$\mathbb{E}_{\xi,\mathbf{e}}[||g(x,\xi,\mathbf{e})||_q^{1+\kappa}] \leq 32\left(\frac{\sqrt{cd}}{2\tau} a_{q,\kappa} M_2\right)^{1+\kappa} + 4\left(\frac{d a_{q,\kappa}\Delta}{\tau}\right)^{1+\kappa} = \sigma_{q,\kappa}^{1+\kappa}$$

where numerical constant $c = \frac{1}{\sqrt{2}}$

SMD

For function $\Psi : \mathbb{R}^d \to \mathbb{R}$ that is strictly convex w.r.t l_p norm, continuously differentiable, we denote its Fenchel conjugate and Bregman divergence

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

$$D_\Psi(x, y) = \Psi(x) - \Psi(y) - \langle \nabla \Psi(y), y - x \rangle$$
$$\mathcal{D} = \max_{u, v \in \mathcal{S}} \sqrt{2 D_\Psi(u, v)}$$

The stochastic mirror descent updates

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

where g_{k+1} is unbiased estimation of $\nabla f(x_k)$.
With conditions for Ψ it can be proofed that updates are well defined and $(\nabla \Psi)^{-1} = \nabla \Psi^*$. Map $\nabla \Psi$ is transformation map.

Convexity and Smoothness Generalization

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 $x, y \in \mathbb{R}^d$

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

Uniform smoothness. Consider a (K_0, r_0) uniform convex and differentiable function $\psi : \mathbb{R}^d \to \mathbb{R}$, an exponent $r \in (1, 2]$, and a constant $K > 0$. Then, ψ is (K, r)-uniformly convex w.r.t. p-norm if for any $x, y \in \mathbb{R}^d$

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

Uniform Convex Example

For $\kappa \in (0,1], p \in [1+\kappa, \infty)$ and $p^* : \frac{1}{p} + \frac{1}{p^*} = 1$. We define

$$K_p = 10 \max\left\{1, (p-1)^{\frac{1+\kappa}{2}}\right\}, \phi(x) = \frac{1}{1+\kappa}||x||_p^{1+\kappa}$$

The the following statements are true

1. $\phi^*(y) = \frac{\kappa}{1+\kappa}||y||_{p^*}^{\frac{1+\kappa}{\kappa}}$
2. ϕ is $(K_p, 1+\kappa)$-uniformly smooth w.r.t. p-norm
3. ϕ^* is $\left(K_p^{-\frac{1}{\kappa}}, \frac{1+\kappa}{\kappa}\right)$-uniformly convex w.r.t. p^*-norm

Convergence

Consider some $\kappa \in (0,1], q \in [1,\infty]$, q^* defined from $\frac{1}{q}+\frac{1}{q^*}=1$ and function Ψ which is $\left(1, \frac{1+\kappa}{\kappa}\right)$-uniformly convex w.r.t. q^* norm. With any $g_k \in \mathbb{R}^d, k \in \overline{1, T}$ and starting point $x_0 = \arg\min\limits_{x \in \mathcal{S}} \Psi(x)$

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

where $R_0^{\frac{1+\kappa}{\kappa}} = \frac{1+\kappa}{\kappa}\sup\limits_{x \in \mathcal{S}}\{\Psi(x) - \Psi(x_0)\}$

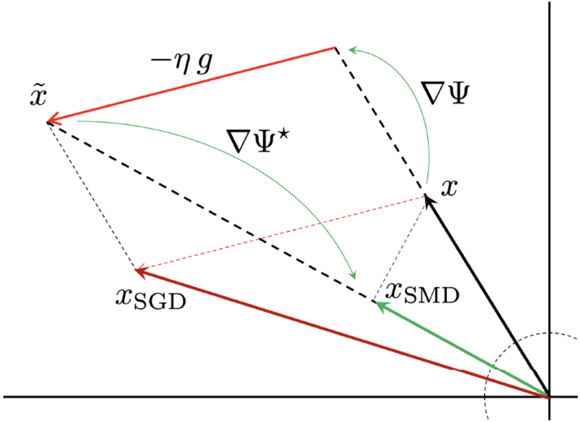

Figure: SMD Stability

Main Algorithm

We select $q \in [1, \infty]$, $\frac{1}{p} + \frac{1}{q} = 1$, norm $||\cdot||_q$ and

$\Psi_p(x) = K_p^{1/\kappa} \phi^*(x)$

Then we choose number of iterations T and step sizes ν and $\tau > 0$

$\sigma_{q,\kappa} = \frac{A}{\tau} \quad \nu = \frac{R_0^{1/\kappa}}{\sigma_{q,\kappa}} T^{-\frac{1}{1+\kappa}}$

procedure $\mathrm{IZ\ SMD}$(Number of iterations T)
    $x_0 \leftarrow \arg\min_{x \in \mathcal{S}} \Psi_p(x)$
    for $k = 0, 1, \ldots, T-1$ **do**
        Sample ξ_k and $\mathbf{e}^k$ from uniform distribution on Euclidean sphere
        Calculate $g_{k+1} = g(x_k, \xi_k, \mathbf{e}_k)$
        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

Main Theorem

Let $\overline{x}_T$ point obtained with T iterations, $x^* \in \arg\min_{x \in \mathcal{S}} f(x)$, then

$$f(\overline{x}_T) - f(x^*) \leq 2M_2\tau + \frac{\sqrt{d}\Delta}{\tau}\mathcal{D} + R_0 \sigma_{q,\kappa} T^{-\frac{\kappa}{1+\kappa}},$$

where $\sigma_{q,\kappa}^{1+\kappa} = 32\left(\frac{\sqrt{cd}}{2\tau} a_{q,\kappa} M_2\right)^{1+\kappa} + 4\left(\frac{d a_{q,\kappa}\Delta}{\tau}\right)^{1+\kappa} = \left(\frac{A}{\tau}\right)^{1+\kappa}$

If $\tau = \sqrt{\frac{\mathcal{D}\Delta\sqrt{d} + AR_0 T^{-\frac{\kappa}{1+\kappa}}}{2M_2}}$, then bound is optimal

$$f(\overline{x}_T) - f(x^*) \leq \sqrt{8M_2\mathcal{D}\Delta\sqrt{d}} + \sqrt{8M_2 A R_0}\frac{1}{T^{\frac{\kappa}{2(1+\kappa)}}}$$

What to do next?

1. Adaptive algorithm
2. Strongly convex f
3. Another sphere norm for sampling

Questions?

Thank You For Attention!