The Power of First-Order Smooth Optimization for Black-Box Non-Smooth Problems


Alexander Gasnikov
MIPT, ISP RAS, MBZUAI


Anton Novitskii
MIPT, ISP RAS


Vasilii Novitskii
MIPT

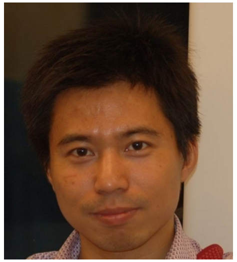

Farshed Abdukhakimov
MBZUAI


Dmitry Kamzolov
MBZUAI


Aleksandr Beznosikov
MIPT, Yandex, HSE, MBZUAI


Martin Takac
MBZUAI


Pavel Dvurechensky
WIAS


Bin Gu
MBZUAI

ICML 2022

Optimization problem

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x)$$

Optimization problem

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x) \longrightarrow \quad f \text{ is convex} \qquad Q \text{ is convex}$$

Optimization problem

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x) \longrightarrow$$

$$f \text{ is convex} \qquad Q \text{ is convex}$$

$$|f(y) - f(x)| \leq M_2 \|y - x\|_2$$

Optimization problem

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x) \longrightarrow$$

f is convex $\quad Q$ is convex

$$|f(y) - f(x)| \leq M_2 \|y - x\|_2$$

$$\downarrow$$

$$\nabla f$$

Optimization problem

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x) \longrightarrow$$

f is convex $\quad Q$ is convex

$$|f(y) - f(x)| \leq M_2 \|y - x\|_2$$

$$\nabla f$$

Optimization problem

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x)$$

f is convex $\quad Q$ is convex

$$|f(y) - f(x)| \leq M_2 \|y - x\|_2$$

$\cancel{\nabla f}$

Only zeroth order information!

Optimization problem

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x) \longrightarrow$$

f is convex $\quad$ Q is convex

$$|f(y) - f(x)| \leq M_2 \|y - x\|_2$$

∇f

Only zeroth order information!

$$d \frac{f(x + \gamma e) - f(x - \gamma e)}{2\gamma} e$$

e uniformly distributed on Euclidean unit sphere

Optimization problem

$$\min_{x \in Q \subseteq \mathbb{R}^d} f(x)$$

f is convex $\quad$ Q is convex

$$|f(y) - f(x)| \leq M_2 \|y - x\|_2$$

∇f

Only zeroth order information!

?

$$d\frac{f(x+\gamma e) - f(x-\gamma e)}{2\gamma} e$$

e uniformly distributed on Euclidean unit sphere

Gradient approximation

$$\nabla f_\gamma(x, e) = d\frac{f(x+\gamma e) - f(x-\gamma e)}{2\gamma}e$$

Gradient approximation

$$\nabla f_\gamma(x, e) = d\frac{f(x+\gamma e) - f(x-\gamma e)}{2\gamma}e$$

$$\mathbb{E}_e\left[\nabla f_\gamma(x, e)\right] = \textbf{?}$$

Gradient approximation

$$\nabla f_\gamma(x,e) = d\frac{f(x+\gamma e) - f(x-\gamma e)}{2\gamma}e$$

$$\downarrow$$

$$\mathbb{E}_e\left[\nabla f_\gamma(x,e)\right] = \,?$$

$$\swarrow$$

$$\nabla f(x)$$

Gradient approximation

$$\nabla f_\gamma(x,e) = d\,\frac{f(x+\gamma e) - f(x-\gamma e)}{2\gamma}\,e$$

$$\mathbb{E}_e\left[\nabla f_\gamma(x,e)\right] = ?$$

$$\nabla f(x)$$

Gradient approximation

$$\nabla f_\gamma(x, e) = d\frac{f(x+\gamma e) - f(x-\gamma e)}{2\gamma}e$$

$$\mathbb{E}_e\left[\nabla f_\gamma(x, e)\right] = \textbf{?}$$

$$\nabla f(x)$$

$$\nabla f_\gamma(x)$$

$$f_\gamma(x) = \mathbb{E}_e f(x+\gamma e)$$

Smooth approximation

$$f_\gamma(x) = \mathbb{E}_e f(x + \gamma e)$$

Smooth approximation

$$f_\gamma(x) = \mathbb{E}_e f(x + \gamma e)$$

$f(x) \longleftrightarrow f_\gamma(x)$

Smooth approximation

$$f_\gamma(x) = \mathbb{E}_e f(x + \gamma e)$$

$$f(x) \longleftrightarrow f_\gamma(x)$$

$$f(x) \leq f_\gamma(x) \leq f(x) + \gamma M_2$$

Smooth approximation

$$f_\gamma(x) = \mathbb{E}_e f(x + \gamma e)$$

$$f(x) \longleftrightarrow f_\gamma(x)$$

$$f(x) \le f_\gamma(x) \le f(x) + \gamma M_2$$

$$\gamma = \varepsilon / (2M_2)$$

Smooth approximation

$$f_\gamma(x) = \mathbb{E}_e f(x + \gamma e)$$

$$f(x) \longleftrightarrow f_\gamma(x)$$

$$f(x) \le f_\gamma(x) \le f(x) + \gamma M_2$$

$$\gamma = \varepsilon/(2M_2)$$

$f_\gamma(x)$ has $L = \dfrac{\sqrt{d}M}{\gamma}$-Lipschitz gradient:

$$\|\nabla f_\gamma(y) - \nabla f_\gamma(x)\|_q \le L\|y - x\|_p$$

Smooth approximation

$$f_\gamma(x) = \mathbb{E}_e f(x + \gamma e)$$

$f(x) \longleftrightarrow f_\gamma(x)$

$$f(x) \leq f_\gamma(x) \leq f(x) + \gamma M_2$$

$$\gamma = \varepsilon/(2M_2)$$

$f_\gamma(x)$ has $L = \dfrac{\sqrt{d}M}{\gamma}$-Lipschitz gradient:

$$\|\nabla f_\gamma(y) - \nabla f_\gamma(x)\|_q \leq L\|y - x\|_p$$

smoothness!

Smooth approximation

$$f_\gamma(x) = \mathbb{E}_e f(x + \gamma e)$$

$$f(x) \longleftrightarrow f_\gamma(x)$$

$$f(x) \leq f_\gamma(x) \leq f(x) + \gamma M_2$$

$$\gamma = \varepsilon/(2M_2)$$

$f_\gamma(x)$ has $L = \dfrac{\sqrt{d}M}{\gamma}$-Lipschitz gradient:

$$\|\nabla f_\gamma(y) - \nabla f_\gamma(x)\|_q \leq L\|y - x\|_p$$

smoothness!

By solving a non-smooth problem, we run the first order SGD for a smooth problem that is very close to the original problem.

SGD in the smooth case

$$x^{k+1} = x^k - \gamma \nabla_x f(x^k, \eta^k)$$

SGD in the smooth case

$$x^{k+1} = x^k - \gamma \nabla_x f(x^k, \eta^k)$$

$$\|\nabla f(y) - \nabla f(x)\|_q \leq L \|y - x\|_p$$

SGD in the smooth case

$$x^{k+1} = x^k - \gamma \nabla_x f(x^k, \eta^k)$$

$$\|\nabla f(y) - \nabla f(x)\|_q \leq L\|y - x\|_p$$

In our case

$$L \leq \frac{2\sqrt{d} M M_2}{\varepsilon}$$

SGD in the smooth case

$$x^{k+1} = x^k - \gamma \nabla_x f(x^k, \eta^k)$$

$$\|\nabla f(y) - \nabla f(x)\|_q \leq L \|y - x\|_p \qquad \mathbb{E}_\eta \left[\|\nabla_x f(x, \eta) - \nabla f(x)\|_q^2 \right] \leq \sigma^2$$

In our case

$$L \leq \frac{2\sqrt{d} M M_2}{\varepsilon}$$

SGD in the smooth case

$$x^{k+1} = x^k - \gamma \nabla_x f(x^k, \eta^k)$$

$$\|\nabla f(y) - \nabla f(x)\|_q \leq L\|y - x\|_p \qquad \mathbb{E}_\eta\left[\|\nabla_x f(x,\eta) - \nabla f(x)\|_q^2\right] \leq \sigma^2$$

In our case

$$L \leq \frac{2\sqrt{d}MM_2}{\varepsilon}$$

$$\sigma^2 \leq 2\kappa(p,d)dM_2^2$$

$$\kappa(p,d) = O\left(\min\{q, \ln d\}\, d^{2/q-1}\right)$$

Richness of the smooth case

- ✔ Stochastic optimization

Richness of the smooth case

- ✔ Stochastic optimization
- ✔ Finite sum-optimization

Richness of the smooth case

- ✔ Stochastic optimization
- ✔ Finite sum-optimization
- ✔ Distributed optimization

Richness of the smooth case

- ✔ Stochastic optimization
- ✔ Finite sum-optimization
- ✔ Distributed optimization
- ✔ Federated Learning

Richness of the smooth case

- ✔ Stochastic optimization
- ✔ Finite sum-optimization
- ✔ Distributed optimization
- ✔ Federated Learning
- ✔ and many other examples

Richness of the smooth case

- ✔ Stochastic optimization
- ✔ Finite sum-optimization
- ✔ Distributed optimization
- ✔ Federated Learning
- ✔ and many other examples

Gradient-free methods for non-smooth problems can be analysed in a unified way if first order smooth stochastic analysis is already available.