Derivative-Free Optimization under Adversarial Noise

Darina Dvinskikh

based on paper *'Gradient-Free Optimization for Non-Smooth Saddle Point Problems under Adversarial Noise'* of D.Dvinskikh, V.Tominin, Ya.Tominin, and A.Gasnikov

Summer school 'Learning, understanding and optimization in artificial intelligence models'

June 24, 2022

Outline

Convex optimization problem

Problem:

$$\min_{x \in \mathcal{X} \subseteq \mathbb{R}^d} \{F(x) := \mathbb{E} f(x, \xi)\}.$$

Let

- function $f(x, \xi)$ is available via a black-box
- the objective function is noisy
- derivative information is unavailable or too expensive

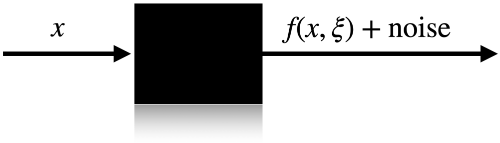
Goal: solve problem with ϵ-precision

$$\mathbb{E}[F(\hat{x}^N)] - \min_{x \in \mathcal{X}} F(x) \le \epsilon,$$

where $\hat{x}^N = \frac{1}{N}\sum_{k=1}^N x^k$ is the output of an algorithm

Black-box zero-order oracle model

Available: only noisy zero-order black-box oracle



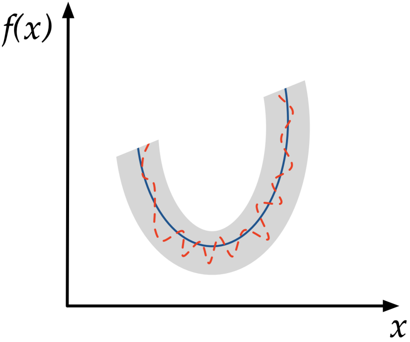
Input: x.
Output: $\varphi(x,\xi) = f(x,\xi) + \delta(x)$, where

$$\delta(x) = \varphi(x,\xi) - f(x,\xi)$$

is the noise (or accuracy).

Application to non-convex problem

Red function means the target function (non-convex): it can be seen as convex blue function with some noise



Contribution and related works

Goal: design an optimal algorithm in terms of its total complexity. Thus, minimize

1. number of oracle calls
2. maximum value of the noise (accuracy): $\max_{x \in \mathcal{X}} \delta(x)$

Paper	Problem	Oracle Calls	Maximum Noise
[Bayandina et al., 2018]	convex	d/ϵ^2	$\epsilon^2/d^{3/2}$
[Beznosikov et al., 2020]	saddle point	d/ϵ^2	ϵ^2/d
[Vasin et al., 2021]	convex	$\text{Poly}\left(d, 1/\epsilon\right)$	$\epsilon^2/\sqrt{d}$
[Risteski and Li, 2016]	convex	$\text{Poly}\left(d, 1/\epsilon\right)$	$\max\left\{\epsilon^2/\sqrt{d}, \epsilon/d\right\}^{(1)}$

(1) $\epsilon/d \lesssim \epsilon^2/\sqrt{d}$, in the large-dimension regime as $\epsilon^{-2} \lesssim d$

Color 'green' means optimal due to lower bounds [Risteski and Li, 2016]

1. number of oracle calls: d/ϵ^2
2. maximum value of the noise: $\epsilon^2/\sqrt{d}$

Outline

Setup for Mirror Descent Algorithm

The setup

- the l_p-norm;
- prox-function $\omega(x)$, that is 1-strongly convex w.r.t. the l_p-norm;
- Bregman divergence associated with $\omega(x)$:

$$V_x(y) = \omega(x) - \omega(y) - \langle \nabla \omega(y), x - y \rangle;$$

- ω-diameter of $\mathcal{X}$:

$$\mathcal{D} = \max_{x,y \in \mathcal{X}} \sqrt{2V_x(y)};$$

- prox-mapping

$$\mathrm{Prox}_x(\beta) = \arg\min_{y \in \mathcal{X}} \left(V_x(y) + \langle \beta, y \rangle \right).$$

Examples

Example (Euclidean setup)

- the ℓ_2-norm in prox-setup
- prox-function $w(x) = \frac{1}{2}\|x\|_2^2$
- Bregman divergence $V_x(y) = \frac{1}{2}\|x - y\|_2^2$
- $\mathcal{D}^2 = \max\limits_{x,y \in \mathcal{X}} \|x - y\|_2^2$
- $\mathrm{Prox}_{x^k}(\gamma g(x^k, \xi^k)) = \pi_{\mathcal{X}}\left(x^k - \gamma g(x^k, \xi^k)\right) \leftarrow$ subgradient descent

Examples

Example (Euclidean setup)

- the ℓ_2-norm in prox-setup
- prox-function $w(x) = \frac{1}{2}\|x\|_2^2$
- Bregman divergence $V_x(y) = \frac{1}{2}\|x - y\|_2^2$
- $\mathcal{D}^2 = \max\limits_{x,y\in\mathcal{X}} \|x - y\|_2^2$
- $\mathrm{Prox}_{x^k}(\gamma g(x^k, \xi^k)) = \pi_{\mathcal{X}}\left(x^k - \gamma g(x^k, \xi^k)\right) \leftarrow$ subgradient descent

Example (Probability simplex)

- $\mathcal{X} = \{x \in \mathbb{R}^d_+ : \|x\|_1 = 1\}$, the ℓ_1-norm in prox-setup
- prox-function $w(x) = \langle x, \log x\rangle$
- Bregman divergence $V_x(y) = \mathrm{KL}(x, y) = \langle x, \log(x/y)\rangle$
- $\mathcal{D}^2 = \mathcal{O}\left(\log d \max\limits_{x,y\in\mathcal{X}} \|y - x\|_1^2\right)$

Algorithm: Gradient-free stochastic mirror descent

Stochastic mirror descent (SMD) [Nemirovski et al., 2009]:

Input: iteration number N, starting point x^1, step size γ

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

1. Sample ξ^k
2. Calculate $g(x^k, \xi^k)$
3. Calculate $x^{k+1} = \mathrm{Prox}_{x^k}(\gamma g(x^k, \xi^k))$

end

Output: $\hat{x}^N = \frac{1}{N}\sum_{k=1}^{N} x^k$

Algorithm: Gradient-free stochastic mirror descent

Stochastic mirror descent (SMD) [Nemirovski et al., 2009]:

Input: iteration number N, starting point x^1, step size γ

for $k = 1, \ldots, N$ **do**

1. Sample ξ^k
2. Calculate $g(x^k, \xi^k)$
3. Calculate $x^{k+1} = \mathrm{Prox}_{x^k}(\gamma g(x^k, \xi^k))$

end

Output: $\hat{x}^N = \frac{1}{N}\sum_{k=1}^{N} x^k$

Goal: estimate $g(x^k, \xi^k)$ by zero-order gradient approximation.

Stochastic mirror descent: convergence rates

Theorem ([Nemirovski et al., 2009])

Let $\mathbb{E}[\|g(\cdot)\|_p^2] \leq M^2$. *Let* N *be the number of iterations of* SMD *and step size be*

$$\gamma = \frac{\mathcal{D}}{M\sqrt{N}}.$$

Then it holds

$$\mathbb{E}\left[F(\hat{x}^N)\right] - \min_{x\in\mathcal{X}} F(x) = \mathcal{O}\left(\frac{M\mathcal{D}}{\sqrt{N}}\right).$$

Corollary

To fulfill $\mathbb{E}\left[F(\hat{x}^N)\right] - \min_{x\in\mathcal{X}} F(x) \leq \epsilon$, *the number of oracle calls is*

$$N = \mathcal{O}\left(\frac{M^2\mathcal{D}^2}{\epsilon^2}\right).$$

Outline

Randomized smoothing of non-smooth function $f(x)$. Euclidean case.

Let us consider deterministic convex problem (for simplicity)

$$\min_{x \in X} f(x),$$

where $f(x)$ is M-Lipschitz continuous w.r.t. the ℓ_2-norm.

Def.

Function $f(x, \xi)$ is M-Lipschitz continuous w.r.t. the ℓ_2-norm, i .e., for all $x_1, x_2 \in \mathcal{X}$:

$$|f(x_1) - f(x_2)| \le M\|x_1 - x_2\|_2.$$

Randomized smoothing

Let $B_2^d = \{u \in \mathbb{R}^d : \|u\|_2 \le 1\}$ be the ℓ_2 unit ball and $u \in B_2^d$ be a random vector. Then a smooth approximation of a non-smooth function $f(x)$ is

$$f^\tau(x) = \mathbb{E}\left[f(x + \tau u) \mid x\right],$$

where $\tau > 0$, $u \in B^d$

Properties of the smoothed approximation

Lemma (properties of $f^\tau(x)$).

function $f^\tau(x)$ is differentiable with

$$\nabla f^\tau(x) = \mathbb{E}\left[\frac{d}{\tau} f(x+\tau e)e \mid x\right],$$

where $e \in S_2^d$ and $S_2^d = \{e \in \mathbb{R}^d : \|e\|_2 = 1\}$ is the ℓ_2 unit sphere.

Intuition behind the Lemma: Divergence (Ostrogradsky–Gauss) theorem

$$\int_{B_2^d} \nabla f(x) dV(x) = \int_{S_2^d} f(x) n(x) dS(x),$$

where $n(x)$ is the normal vector to S_2^d.

Proof of Lemma.

Let $e \in S_2^d$ and $u \in B_2^d$, and $\tau > 0$. Due to Ostrogradsky–Gauss theorem and $f(x)$ is convex

$$\nabla \int_{B_2^d} f(x+\tau u) dV(u) = \frac{1}{\tau} \int_{S_2^d} f(x+\tau e) e dS(e),$$

Then we rewrite it as

$$\nabla \mathbb{E}\left[f(x+\tau u)\right] = \frac{1}{\tau} \frac{\mathsf{Vol}(S_2^d)}{\mathsf{Vol}(B_2^d)} \mathbb{E}\left[f(x+\tau e) e\right],$$

As $\mathsf{Vol}(B_2^d) = d\mathsf{Vol}(S_2^d)$

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

□

Approximation

Lemma

Let function $f(x)$ be M-Lipschitz continuous, then for all $x \in \mathcal{X}$ the following holds

$$|f^\tau(x) - f(x)| \le \tau M.$$

Proof. By the definition of $f^\tau(z)$ it holds

$$\begin{aligned}|f^\tau(x) - f(x)| &= |\mathbb{E}\left[f(x+\tau u) \mid x\right] - f(x)| = \mathbb{E}\left[|f(x+\tau u) - f(x)| \mid x\right] \\ &\le \mathbb{E}\left[M\|\tau u\|_2\right] \le M\tau \quad \text{as } u \in B_2^d.\end{aligned}$$

□

Relation to initial problem

Let the smooth problem

$$\min_{x\in\mathcal{X}} f^\tau(x).$$

be solved with $\epsilon/2$-precision:

$$\mathbb{E}\left[f^\tau(\hat{x}^N)\right] - \min_{x\in\mathcal{X}} f^\tau(x) \le \frac{\epsilon}{2}.$$

Then the initial problem

$$\min_{x\in\mathcal{X}} f(x).$$

will be solved with ϵ-precision if $\tau = \frac{\epsilon}{2M}$:

$$\mathbb{E}\left[f(\hat{x}^N)\right] - \min_{x\in\mathcal{X}} f(x) \le \frac{\epsilon}{2} + \tau M = \epsilon.$$

Outline

Zero-order gradient estimate.

Zero-order gradient estimator with two-point feedback:

$$g(x,\xi,e) = \frac{\mathsf{Vol}(S_q^d)}{\mathsf{Vol}(B_q^d)}\left(\varphi(x+\tau e,\xi) - \varphi(x-\tau e,\xi)\right) n(e),$$

where

e is a vector picked uniformly at random from S_q^d,

$n(e)$ is the normal vector to S_q^d,

$\tau > 0$.

Intuition behind the gradient estimate:

Let $u \in B_q^d$ and $e \in S_q^d$. Due to Ostrogradsky–Gauss theorem

$$\nabla \int_{B_q^d} f(x+\tau u) dV(u) = \frac{1}{\tau}\int_{S_q^d} f(x+\tau e) e dS(e),$$

Then we rewrite it as

$$\nabla \mathbb{E}\left[f(x+\tau u)\right] = \frac{1}{\tau}\frac{\mathsf{Vol}(S_q^d)}{\mathsf{Vol}(B_q^d)}\mathbb{E}\left[f(x+\tau e) n(e)\right].$$

Examples

Gradient estimator (ℓ_2-randmization) [Shamir, 2017]

$$g(x, \xi, e) = \frac{d}{2\tau} \left(\varphi(x + \tau e, \xi) - \varphi(x - \tau e, \xi)\right) e,$$

where $e \in S_2^d$, $\tau > 0$.

Gradient estimator (ℓ_1-randmization) [Gasnikov et al., 2016]

$$g(x, \xi, \zeta) = \frac{d}{2\tau} \left(\varphi(x + \tau \zeta, \xi) - \varphi(x - \tau \zeta, \xi)\right) \operatorname{sign}(\zeta),$$

where $\zeta \in S_1^d$, $\tau > 0$.

Why did we smooth?

Example with ℓ_2-randmization:

Let $d = 1$ and $f(x) = |x|$. Then for $x \in [-\tau, \tau]$ and e is uniform in $\{-1, 1\}$

$$g(x, e) = \frac{1}{2\tau}(f(x+\tau) - f(x-\tau))e = \pm\frac{x}{2\tau}.$$

However,

- $\nabla f(x) = 1$, for all $x > 0$,
- $\nabla f(x) = -1$ for all $x < 0$.

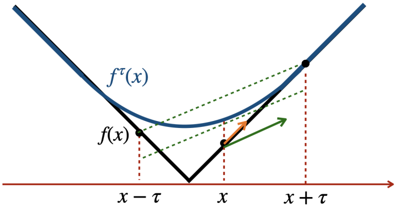


Figure: Smooth approximation of a non-smooth function

Outline

Unbiased estimator

Canceling noise or noise-free setup

For all $x', x'' \in \mathcal{X}$, it holds $\delta(x') = \delta(x'') = \delta$ almost surely.

- $g(x, \xi, e)$ is an unbiased estimation for $\nabla f^\tau(x)$:

$$\mathbb{E}\left[g(x, \xi, e) \mid x\right] = \mathbb{E}\left[\frac{d}{\tau} f(x + \tau e, \xi) \mid x\right] = \nabla f^\tau(x)$$

- $g(x, \xi, e)$ has bounded second moment

$$\mathbb{E}\left[\|g(x, \xi, e)\|_{p^*}^2 \mid x\right] = \mathcal{O}\left(d^{2-\frac{2}{p}} \min\{p/(p-1), \log d\} M^2\right),$$

where $\frac{1}{p} + \frac{1}{p^*} = 1$ (dual norm)

Adversarial Noise

Assumption (Boundedness of the noise)

For all $x \in \mathcal{X}$, it holds $|\delta(x)| \leq \Delta$.

For all $x \in \mathcal{X}$ and $r \in \{r \in \mathbb{R}^d : \|r\|_2 \leq \mathcal{D}\}$

- 'bias':

$$\mathbb{E}\left[\langle g(x,\xi,\ e) - \nabla f^\tau(x), r\rangle \mid x\right] \leq \frac{\sqrt{d}\Delta\mathcal{D}}{\tau}$$

- 'variance'

$$\mathbb{E}\left[\|g(x,\xi,e)\|_{p^*}^2 \mid x\right] = \mathcal{O}\left(d^{2-\frac{2}{p}}\min\{p/(p-1), \log d\}\left(M^2 + d\frac{\Delta^2}{\tau^2}\right)\right)$$

where $\frac{1}{p} + \frac{1}{p^*} = 1$ (dual norm)

Adversarial Noise

Assumption (Lipschitz continuity of the noise)

Function $\delta(x)$ *is* M_δ*-Lipschitz continuous in* $x \in \mathcal{X}$ *w.r.t. the* ℓ_2*-norm.*

Let us consider $\{r \in \mathbb{R}^d : \|r\|_2 \leq \mathcal{D}\}$, then for all $x \in \mathcal{X}$

- $$\mathbb{E}\left[\langle g(x,\xi,\ e) - \nabla f^\tau(x), r\rangle \mid x\right] \leq \sqrt{d} M_\delta \mathcal{D}$$
- $g(x,\xi,e)$ has bounded the second moment is

$$\mathbb{E}\left[\|g(x,\xi,e)\|_{p^*}^2 \mid x\right] = \mathcal{O}\left(d^{2-\frac{2}{p}} \min\{p/(p-1), \log d\} \left(M^2 + M_\delta^2\right)\right).$$

ℓ_2-randmization in other two points

Let

$$x_1 = x, \; x_2 = x + \tau e,$$

where $\tau > 0$ is some constant and $e \in S_2^d$. Then

$$g(x, \xi, e) = \frac{d}{\tau}\left(\varphi(x + \tau e, \xi) - \varphi(x, \xi)\right) e$$

Issue (the second moment is quadratic in d) [Duchi et al., 2015]

Let $f(x) = \|x\|_2$ (non-differentiable function), let $x_1 = 0$ and $x_2 = \tau e$, then

$$\mathbb{E}\left[g(x, e)\right] = \left\|\frac{d}{\tau}\left(f(\tau e) - f(0)\right) e\right\|_2^2 = d^2 \mathbb{E}\left[\|e\|_2\right] = d^2.$$

Outline

Randomized smoothing

Smooth approximation

Let $B_1^d = \{v \in \mathbb{R}^d : \|v\|_1 \leq 1\}$ be the ℓ_2 unit ball and $v \in B_1^d$ be a random vector. Then a smooth approximation of a non-smooth function $f(x, \xi)$ is

$$f^\tau(x) = \mathbb{E}\left[f(x + \tau v, \xi) \mid x\right],$$

where $\tau > 0$, $v \in B_1^d$.

Lemma (properties of $f^\tau(x)$).
Function $f^\tau(x)$ is differentiable with

$$\nabla f^\tau(x) = \mathbb{E}\left[\frac{d}{\tau} f(x + \tau\zeta, \xi) \text{sign}(\zeta) \mid x\right],$$

where $e \in S_1^d$.

Approximation

Lemma

It holds for all $x \in \mathcal{X}$

$$|f^\tau(x) - f(x)| \le \frac{2}{\sqrt{d}}\tau M$$

Proof. By the definition of $f^\tau(z)$ it holds

$$\begin{aligned}|f^\tau(x) - f(x)| &= |\mathbb{E}\left[f(x+\tau v) \mid x\right] - f(x)| = \mathbb{E}\left[|f(x+\tau v) - f(x)| \mid x\right] \\ &\le \tau M \mathbb{E}\left[\|v\|_2\right].\end{aligned}$$

Then we use the next lemma with $p = 2$

Lemma[Akhavan et al., 2022]

Let $q \in [1, \infty)$ and let v be distributed uniformly on B_1^d. Then

$$\mathbb{E}\left[\|v\|_p\right] \le \frac{pd^{\frac{1}{p}}}{d+1}.$$

Relation to initial problem

Let the smooth problem

$$\min_{x \in \mathcal{X}} f^\tau(x).$$

be solved with $\epsilon/2$-precision:

$$\mathbb{E}\left[f^\tau(\hat{x}^N)\right] - \min_{x \in \mathcal{X}} f^\tau(x) \leq \frac{\epsilon}{2}.$$

Relation to initial problem

Let the smooth problem

$$\min_{x\in\mathcal{X}} f^\tau(x).$$

be solved with $\epsilon/2$-precision:

$$\mathbb{E}\left[f^\tau(\hat{x}^N)\right] - \min_{x\in\mathcal{X}} f^\tau(x) \le \frac{\epsilon}{2}.$$

Then the initial problem

$$\min_{x\in\mathcal{X}} F(x).$$

will be solved with ϵ-precision if $\tau = \frac{\sqrt{d}\epsilon}{4M}$:

$$\mathbb{E}\left[F(\hat{x}^N)\right] - \min_{x\in\mathcal{X}} F(x) \le \frac{\epsilon}{2} + \frac{2}{\sqrt{d}}\tau M = \epsilon.$$

Unbiased estimator

Canceling noise or noise-free setup

For all $x', x'' \in \mathcal{X}$, it holds $\delta(x') = \delta(x'') = \delta$ almost surely.

- $g(x, \xi, \zeta)$ is an unbiased estimation for $\nabla f^\tau(x)$:

$$\mathbb{E}\left[g(x,\xi,\zeta) \mid x\right] = \mathbb{E}\left[\frac{d}{2\tau} f(x+\tau\zeta,\xi)\mathrm{sign}(\zeta) \mid x\right] = \nabla f^\tau(x)$$

- $g(x, \xi, \zeta)$ has bounded second moment

$$\mathbb{E}\left[\|g(x,\xi,\zeta)\|_{p^*}^2 \mid x\right] = \mathcal{O}\left(d^{2-\frac{2}{p}} M^2\right),$$

where $\frac{1}{p} + \frac{1}{p^*} = 1$ (dual norm)

Adversarial Noise

Assumption (Boundedness of the noise)

For all $x \in \mathcal{X}$, it holds $|\delta(x)| \leq \Delta$.

For all $x \in \mathcal{X}$ and $r \in \{r \in \mathbb{R}^d : \|r\|_2 \leq \mathcal{D}\}$

- 'bias'

$$\mathbb{E}\left[\langle g(x, \xi,\ \zeta) - \nabla f^\tau(x), r\rangle \mid x\right] = \mathcal{O}\left(\frac{d\Delta\mathcal{D}}{\tau}\right)$$

- 'variance'

$$\mathbb{E}\left[\|g(x, \xi, \zeta)\|^2_{p^*} \mid x\right] = \mathcal{O}\left(d^{2-\frac{2}{p}} M^2 + d^{4-\frac{2}{p}}\frac{\Delta^2}{\tau^2}\right).$$

where $\frac{1}{p} + \frac{1}{p^*} = 1$ (dual norm)

Adversarial Noise

Assumption (Lipschitz continuity of the noise)

Function $\delta(x)$ is M_δ-Lipschitz continuous in $x \in \mathcal{X}$ w.r.t. the ℓ_2-norm.

For all $x \in \mathcal{X}$ and $r \in \{r \in \mathbb{R}^d : \|r\|_2 \leq \mathcal{D}\}$

- 'bias':

$$\mathbb{E}\left[\langle g(x,\xi,\zeta) - \nabla f^\tau(x), r\rangle \mid x\right] = \mathcal{O}\left(\sqrt{d} M_\delta \mathcal{D}\right)$$

- 'variance':

$$\mathbb{E}\left[\|g(x,\xi,\zeta)\|_{p^*}^2 \mid x\right] = \mathcal{O}\left(d^{2-\frac{2}{p}}(M^2 + M_\delta^2)\right).$$

where $\frac{1}{p} + \frac{1}{p^*} = 1$ (dual norm)

Outline

Convergence rate

Theorem

Let $\mathbb{E}[\|g(\cdot)\|_{p^*}^2 \mid x] \le M_{\text{new}}^2$ *for all* $x \in X$*. Let* N *be the number of zero-order* SMD *and step size be chosen as*

$$\gamma = \frac{\mathcal{D}}{M_{\text{new}}\sqrt{N}}.$$

Then it holds

$$\mathbb{E}\left[F(\hat{x}^N)\right] - \min_{x\in\mathcal{X}} F(x) \le \mathcal{O}\left(\frac{M_{\text{new}}\mathcal{D}}{\sqrt{N}} + \underbrace{\text{'bias'}}_{\le\epsilon} + \underbrace{\text{smooth approx.}}_{\le\epsilon}\right).$$

Corollary

To fulfill $\mathbb{E}\left[F(\hat{x}^N)\right] - \min_{x\in\mathcal{X}} F(x) \le \epsilon$*, the number of oracle calls is*

$$N = \mathcal{O}\left(M_{\text{new}}^2\mathcal{D}^2/\epsilon^2\right).$$

Maximal level of noise

Conditions

1. τ: smooth approx. $\le \epsilon \quad \Longrightarrow \quad |f^\tau(x) - F(x)| \le \epsilon$ for all $x \in X$
2. Δ or M_δ: 'bias' $\le \epsilon$
3. Δ or M_δ: $N(\Delta) = N(0) \quad \Longrightarrow \quad M_{\text{new}}(\Delta) = M_{\text{new}}(0)$ (fulfilled due to 'bias' condition)

randomization	τ	Δ	M_δ
ℓ_1-randomization	$\sqrt{d}\frac{\epsilon}{M}$	$\frac{\epsilon^2}{\sqrt{d}M\mathcal{D}}$	$\frac{\epsilon}{\sqrt{d}\mathcal{D}}$
ℓ_2-randomization	$\frac{\epsilon}{M}$	$\frac{\epsilon^2}{\sqrt{d}M\mathcal{D}}$	$\frac{\epsilon}{\sqrt{d}\mathcal{D}}$

Table: Maximal level of bounded noise and smoothing parameter up to $\mathcal{O}(\cdot)$

Comparison under bounded noise

Randomization	Number of iterations N
ℓ_1	$d^{2-\frac{2}{p}} M^2 \mathcal{D}^2/\epsilon^2$
ℓ_2	$d^{2-\frac{2}{p}} \min\{p/(p-1), \log d\} M^2 \mathcal{D}^2/\epsilon^2$

Table: Number of iterations depending on the type of randomization in the ℓ_p- norm of proximal setup up to $\mathcal{O}(\cdot)$

Norm in prox. setup	$p=1$	$p=2$
N with ℓ_1-randomization	$\frac{M^2}{\epsilon^2} \log(d) \max\limits_{x,y\in\mathcal{X}} \|x-y\|_1^2$	$\frac{dM^2}{\epsilon^2} \max\limits_{x,y\in\mathcal{X}} \|x-y\|_2^2$
N with ℓ_2- randomization	$\frac{\log(d) M^2}{\epsilon^2} \log(d) \max\limits_{x,y\in\mathcal{X}} \|x-y\|_1^2$	$\frac{dM^2}{\epsilon^2} \max\limits_{x,y\in\mathcal{X}} \|x-y\|_2^2$

Table: Examples of N

What is in article but beyond the lecture?

- saddle point problems:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \ \mathbb{E}_\xi \left[f(x,y,\xi)\right],$$

- infinite noise:

  function $f(x,\xi)$ is M-Lipschitz continuous: for all $x \in X$ and $\xi \in \Xi$,

$$|f(x,\xi) - f(y,\xi)| \leq M(\xi)\|x-y\|_2,$$

  and there exists a positive constant $\tilde{M}_2$: $\mathbb{E}\left[M_2(\xi)^{1+\kappa}\right] \leq \tilde{M}_2^{1+\kappa}$, $\kappa \in (0,1]$.

- restarts
- r-growth condition:

  there is $r \geq 1$ and $\mu_r > 0$ such that for all $x \in \mathcal{X}$

$$\frac{\mu_r}{2}\|x - x^\star\|_p^r \leq f(x) - f(x^*),$$

  where x^* is problem solution

- large deviations

For the details

For the details, please, see the paper:

D.Dvinskikh, V.Tominin, Ya.Tominin, and A.Gasnikov *'Gradient-Free Optimization for Non-Smooth Saddle Point Problems under Adversarial Noise'* (https://arxiv.org/pdf/2202.06114.pdf)

Akhavan, A., Chzhen, E., Pontil, M., and Tsybakov, A. B. (2022).
A gradient estimator via l1-randomization for online zero-order optimization with two point feedback.
arXiv preprint arXiv:2205.13910.

Bayandina, A. S., Gasnikov, A. V., and Lagunovskaya, A. A. (2018).
Gradient-free two-point methods for solving stochastic nonsmooth convex optimization problems with small non-random noises.
Automation and Remote Control, 79(8):1399–1408.

Beznosikov, A., Sadiev, A., and Gasnikov, A. (2020).
Gradient-free methods with inexact oracle for convex-concave stochastic saddle-point problem.
In *International Conference on Mathematical Optimization Theory and Operations Research*, pages 105–119. Springer.

Duchi, J. C., Jordan, M. I., Wainwright, M. J., and Wibisono, A. (2015).
Optimal rates for zero-order convex optimization: The power of two function evaluations.
IEEE Trans. Information Theory, 61(5):2788–2806.
arXiv:1312.2139.

Gasnikov, A. V., Lagunovskaya, A. A., Usmanova, I. N., and Fedorenko, F. A. (2016).
Gradient-free proximal methods with inexact oracle for convex stochastic nonsmooth optimization problems on the simplex.
Automation and Remote Control, 77(11):2018–2034.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009).

Robust stochastic approximation approach to stochastic programming.
SIAM Journal on Optimization, 19(4):1574–1609.

Risteski, A. and Li, Y. (2016).
Algorithms and matching lower bounds for approximately-convex optimization.
Advances in Neural Information Processing Systems, 29:4745–4753.

Shamir, O. (2017).
An optimal algorithm for bandit and zero-order convex optimization with two-point feedback.
Journal of Machine Learning Research, 18:52:1–52:11.
First appeared in arXiv:1507.08752.

Vasin, A., Gasnikov, A., and Spokoiny, V. (2021).
Stopping rules for accelerated gradient methods with additive noise in gradient.