25 Randomized smoothing

Goal

convolution, duality between smoothness and decay, randomized smoothing, gradient-free, zero-th order optimization

Alert 25.1: Convention

Gray boxes are not required hence can be omitted for unenthusiastic readers.

This note is likely to be updated again soon.

Definition 25.2: Problem

In this lecture we revisit our old problem

$$\min_{\mathbf{w}\in C\subseteq\mathbb{R}^d} f(\mathbf{w}), \tag{25.1}$$

where C is closed convex and f is (non)convex. We impose a new twist: we can only evaluate the function value $f(\mathbf{w})$ at any $\mathbf{w}$ but not its (sub)gradient. Can we still solve (35.1), efficiently? And how?

Definition 25.3: Convolution

The convolution of two functions f and g is defined through integration:

$$(f*g)(\mathbf{w}) = f*g(\mathbf{w}) := \int_{\mathbf{z}} f(\mathbf{w}-\mathbf{z})g(\mathbf{z})\,\mathrm{d}\mathbf{z} = \int_{\mathbf{z}} f(\mathbf{z})g(\mathbf{w}-\mathbf{z})\,\mathrm{d}\mathbf{z} =: (g*f)(\mathbf{w}).$$

Note the similarity to the infimal convolution in Definition 11.8. Recall the Fourier transform and its inverse:

$$(\mathscr{F}f)(\mathbf{w}^*) = \mathscr{F}f(\mathbf{w}^*) = \int_{\mathbf{w}} \exp(-2\pi i\,\langle\mathbf{w},\mathbf{w}^*\rangle)f(\mathbf{w})\,\mathrm{d}\mathbf{w}, \quad (\mathscr{F}^{-1}g)(\mathbf{w}) = \int_{\mathbf{w}^*} \exp(2\pi i\,\langle\mathbf{w},\mathbf{w}^*\rangle)g(\mathbf{w}^*)\,\mathrm{d}\mathbf{w}^*,$$

where $\mathbf{w}$ is usually called the time variable and $\mathbf{w}^*$ the frequency variable. It is well-known that

$$\mathscr{F}(f*g) = \mathscr{F}f\cdot\mathscr{F}g, \quad \mathscr{F}\mathscr{F}^{-1} = \mathscr{F}^{-1}\mathscr{F} = \mathrm{Id}, \quad \mathscr{F}f^{(\mathbf{k})} = (-2\pi i\mathbf{w}^*)^{\mathbf{k}}\mathscr{F}f,$$

where $\mathbf{z}^{\mathbf{k}} := \prod_{j=1}^d z_j^{k_j}$ and $f^{(\mathbf{k})} := \prod_{j=1}^d \partial_{k_j} f$ is the partial derivative. In particular, how fast a function decays (than a polynomial of certain degree) corresponds to how smooth its Fourier transform is, and vice versa. It also follows that

$$\mathscr{F}(f*g)^{(\mathbf{k})} = (-2\pi i\mathbf{w}^*)^{\mathbf{k}}\cdot\mathscr{F}(f*g) = [(-2\pi i\mathbf{w}^*)^{\mathbf{k}}\mathscr{F}f]\mathscr{F}g = \mathscr{F}(f^{(\mathbf{k})}*g) = \mathscr{F}(f*g^{(\mathbf{k})}),$$

and applying the inverse transform we obtain the familiar formula of differentiating under the integral:

$$(f*g)^{(\mathbf{k})} = f^{(\mathbf{k})}*\mathbf{g} = f*g^{(\mathbf{k})},$$

where of course the partial derivative of a function needs proper interpretation.

Alert 25.4: Existence and finiteness of expectation/integral

When one writes the expectation of a random variable, or more generally an integral such as

$$\int_{\mathbf{w}} f(\mathbf{w})g(\mathbf{w})\,\mathrm{d}\mathbf{w},$$

some conditions on f and g are needed to make sure the above integral makes sense. In this lecture we always assume the expectation (integral) exists and is finite, while ignoring to state the standard conditions.

Definition 25.5: Randomized Smoothing

For any (vector-valued) function $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^c$ we define its randomized smoothing as:

$$\mathbf{f}_\gamma(\mathbf{w}) = \mathbb{E}\mathbf{f}(\mathbf{w} + \gamma\boldsymbol{\varepsilon}),$$

where $\boldsymbol{\varepsilon}$ is some random noise with zero mean and identity covariance. W.l.o.g., we may take $\boldsymbol{\varepsilon}$ to be symmetric (i.e., $\pm\boldsymbol{\varepsilon}$ are identically distributed), for otherwise we may replace $\boldsymbol{\varepsilon}$ with $\beta\boldsymbol{\varepsilon}$ where β is an independent $\{\pm\}$-valued Bernoulli random variable. Let p be the probability density function (pdf) of $\boldsymbol{\varepsilon}$ where $p(\mathbf{w}) = p(-\mathbf{w})$ due to symmetry. We define the dilated density $\mathsf{p}_\gamma = \frac{1}{\gamma^d}\mathsf{p}(\frac{1}{\gamma}\cdot)$. Then,

$$\mathbf{f}_\gamma \overset{\text{symmetry}}{=} \mathbb{E}\mathbf{f}(\mathbf{w} - \gamma\boldsymbol{\varepsilon}) = \mathbf{f} * \mathsf{p}_\gamma, \text{ hence } \mathbf{f}_\gamma \to \mathbf{f} \text{ as } \gamma \to 0,$$

as is intuitively expected. (The convergence is pointwise but can be made uniform on compact sets.)

More generally, we can allow non-additive noise:

$$\mathbf{f}_\gamma(\mathbf{w}) = \mathbb{E}\mathbf{f}(\mathbf{w}, \boldsymbol{\varepsilon}).$$

For instance, if $\mathbf{f}$ represents a deep network, we can add the noise $\boldsymbol{\varepsilon}$ to network input $\mathbf{x}$ which transforms into a highly nonlinear random effect on the network weights $\mathbf{w}$.

Exercise 25.6: Moment inequalities

Define the k-th moment of a standard normal random variable X as

$$M_k = \mathbb{E}|X|^k.$$

It is easy to see that $(M_k)^{1/k}$ is an increasing function of k. Prove that

$$\forall k \geq 2, \quad M_k^{1/k} \leq \sqrt{k + M_2^2}.$$

Exercise 25.7: Calculus for randomized smoothing

Prove the following:

- The map $\mathbf{f} \mapsto \mathbf{f}_\gamma$ is linear.
- If f is convex/concave, so is f_γ.
- If f is convex, then $f_\gamma \geq f$.

- If $\mathbf{f}$ is L_0-Lipschitz continuous (w.r.t. $\|\cdot\|_2$ say), so is $\mathbf{f}_\gamma$. Moreover,

$$\|\mathbf{f}_\gamma - \mathbf{f}\|_2 \leq \gamma \mathsf{L}_0 \mathbb{E}\|\boldsymbol{\varepsilon}\|_2 \leq \gamma \mathsf{L}_0 \sqrt{\mathbb{E}\|\boldsymbol{\varepsilon}\|_2^2} = \gamma \mathsf{L}_0 \sqrt{d}.$$

- If f is L_1-smooth (w.r.t. $\|\cdot\|_2$ say), so is f_γ. Moreover,

$$f_\gamma - f \leq \tfrac{\gamma^2 \mathsf{L}_1}{2} \mathbb{E}\|\boldsymbol{\varepsilon}\|_2^2 = \tfrac{\gamma^2 \mathsf{L}_1 d}{2},$$

whereas a two-sided bound holds if both $\pm f$ are L_1-smooth.

- If f is L_2-smooth (w.r.t. $\|\cdot\|_2$ say), i.e.

$$f(\mathbf{z}) \leq f(\mathbf{w}) + \langle \mathbf{z} - \mathbf{w}; \nabla f(\mathbf{w}) \rangle + \tfrac{1}{2} \left\langle \mathbf{z} - \mathbf{w}; \nabla^2 f(\mathbf{w})(\mathbf{z} - \mathbf{w}) \right\rangle + \tfrac{\mathsf{L}_2}{6} \|\mathbf{z} - \mathbf{w}\|_2^3,$$

so is f_γ. Moreover,

$$f_\gamma - f - \tfrac{\gamma^2}{2} \operatorname{tr} \nabla^2 f \leq \tfrac{\gamma^3 \mathsf{L}_2}{6} \mathbb{E}\|\boldsymbol{\varepsilon}\|_2^3 \leq \tfrac{\gamma^3 \mathsf{L}_2}{6} (3 + d)^{3/2},$$

whereas a two-sided bound holds if both $\pm f$ are L_2-smooth.

This last exercise reveals that the square dependence on γ cannot be further improved even if the function f is smoother than L_1-smooth.

Exercise 25.8: Gradient approximation

Prove the following:

- If $\pm f$ is L_1-smooth, then $\|\nabla f_\gamma - \nabla f\|_\circ \leq \gamma \mathsf{L}_1 \sqrt{d}$. In fact, $\nabla f_\gamma = (\nabla f)_\gamma$, and

$$\|\nabla f\|_\circ \leq \|\nabla f_\gamma\|_\circ + \gamma \mathsf{L}_1 \sqrt{d}.$$

- If $\pm f$ is L_2-smooth, then $\|\nabla f_\gamma - \nabla f\|_\circ \leq \gamma^2 \mathsf{L}_2 d/2$. In fact, $\nabla f_\gamma = (\nabla f)_\gamma$ and $\nabla^2 f_\gamma = (\nabla^2 f)_\gamma$.

Remark 25.9: Justifying the name

Differentiating under the integral we obtain

$$f_\gamma^{(\mathbf{k})} := [f * \mathsf{p}_\gamma]^{(\mathbf{k})} = f^{(\mathbf{k-1})} * \mathsf{p}_\gamma^{(\mathbf{l})}, \quad \text{in particular} \quad \nabla^k f_\gamma(\mathbf{w}) = \int_{\mathbf{z}} \nabla^{k-1} f(\mathbf{w} - \mathbf{z}) \otimes \nabla \mathsf{p}_\gamma(\mathbf{z}) \, \mathrm{d}\mathbf{z}.$$

Therefore, if f is L_{k-1}-smooth, then f_γ is L_k-smooth, where

$$\mathsf{L}_k \leq \mathsf{L}_{k-1} \int_{\mathbf{z}} \|\nabla \mathsf{p}_\gamma(\mathbf{z})\|_2 \, \mathrm{d}\mathbf{z} = \tfrac{\mathsf{L}_{k-1}}{\gamma} \int_{\mathbf{z}} \|\nabla \mathsf{p}(\mathbf{z})\|_2 \, \mathrm{d}\mathbf{z} = \tfrac{s \mathsf{L}_{k-1}}{\gamma} \quad \text{where} \quad s := \mathbb{E}\|\nabla \ln \mathsf{p}(\boldsymbol{\varepsilon})\|_2, \quad \boldsymbol{\varepsilon} \sim \mathsf{p}.$$

In other words, f_γ is (at least) 1 degree more smoother than f, as long as the score function $\nabla \ln \mathsf{p}$ has finite expectation (in norm). The case $k = 1$ is of particular interest to us, so we repeat the formula:

$$\begin{aligned}\nabla f_\gamma(\mathbf{w}) = \int_{\mathbf{z}} f(\mathbf{w} - \mathbf{z}) \nabla \mathsf{p}_\gamma(\mathbf{z}) \, \mathrm{d}\mathbf{z} = \tfrac{1}{\gamma} \mathbb{E}[f(\mathbf{w} - \gamma\boldsymbol{\varepsilon}) \nabla \ln \mathsf{p}(\boldsymbol{\varepsilon})] &= -\tfrac{1}{\gamma} \mathbb{E}[f(\mathbf{w} + \gamma\boldsymbol{\varepsilon}) \nabla \ln \mathsf{p}(\boldsymbol{\varepsilon})] \\ &= -\mathbb{E}\left[\frac{f(\mathbf{w} + \gamma\boldsymbol{\varepsilon}) - f(\mathbf{w})}{\gamma} \nabla \ln \mathsf{p}(\boldsymbol{\varepsilon}) \right] \\ &= -\mathbb{E}\left[\frac{f(\mathbf{w} + \gamma\boldsymbol{\varepsilon}) - f(\mathbf{w} - \gamma\boldsymbol{\varepsilon})}{2\gamma} \nabla \ln \mathsf{p}(\boldsymbol{\varepsilon}) \right].\end{aligned}$$

Interestingly, when f is directionally differentiable (e.g. f is convex or an envelope), we have the limit:

$$\nabla f_0(\mathbf{w}) := -\mathbb{E}[f'(\mathbf{w};\boldsymbol{\varepsilon})\nabla \ln \mathsf{p}(\boldsymbol{\varepsilon})], \quad \text{where} \quad f'(\mathbf{w};\boldsymbol{\varepsilon}) := \lim_{\gamma\downarrow 0}[f(\mathbf{w}+\gamma\boldsymbol{\varepsilon}) - f(\mathbf{w})]/\gamma$$
$$= -\mathbb{E}[\sigma_{\partial f(\mathbf{w})}(\boldsymbol{\varepsilon})\nabla \ln \mathsf{p}(\boldsymbol{\varepsilon})]$$

Needless to say, when f is actually differentiable, we have $\nabla f_0 = \nabla f$.

Example 25.10: Gaussian and uniform

Two choices of the noise distribution p are common:

- $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$, i.e. $\mathsf{p}(\boldsymbol{\varepsilon}) = (2\pi)^{d/2}\exp(-\|\boldsymbol{\varepsilon}\|_2^2/2)$ hence $-\nabla \ln \mathsf{p}(\boldsymbol{\varepsilon}) = \boldsymbol{\varepsilon}$, $s = \mathbb{E}\|\nabla \ln \mathsf{p}(\boldsymbol{\varepsilon})\|_2 \le \sqrt{d}$, and

$$\nabla f_\gamma(\mathbf{w}) = \tfrac{1}{\gamma}\mathbb{E}[f(\mathbf{w}+\gamma\boldsymbol{\varepsilon})\boldsymbol{\varepsilon}] = \mathbb{E}\left[\tfrac{f(\mathbf{w}+\gamma\boldsymbol{\varepsilon})-f(\mathbf{w})}{\gamma}\boldsymbol{\varepsilon}\right] = \mathbb{E}\left[\tfrac{f(\mathbf{w}+\gamma\boldsymbol{\varepsilon})-f(\mathbf{w}-\gamma\boldsymbol{\varepsilon})}{2\gamma}\boldsymbol{\varepsilon}\right].$$

  This setting, considered by Nesterov and Spokoiny (2017), is convenient since f_γ is in fact infinitely many times differentiable although it requires f to be defined on entire $\mathbb{R}^d$.

- $\boldsymbol{\varepsilon} \sim \text{Uniform}(K)$, i.e. $\mathsf{p}(\boldsymbol{\varepsilon}) = 1/v_d$ if $\boldsymbol{\varepsilon} \in K$ and 0 otherwise, where v_d is the volume of the (symmetric, isotropic, i.e. $\mathbb{E}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top = \mathbb{I}$) compact set K (with smooth boundary). It follows from multivariate integration by parts (e.g. Stokes' theorem, see Katz 1979) that $\nabla \mathsf{p}(\boldsymbol{\varepsilon}) = \mathbf{1}_{\partial K}\cdot \mathsf{n}(\boldsymbol{\varepsilon})/v_d$, where $\mathsf{n}(\boldsymbol{\varepsilon})$ is the (positively oriented) normal vector at $\boldsymbol{\varepsilon} \in \partial K$. Thus, $s = u_{d-1}/v_d$ where u_{d-1} is the surface area of ∂K, and

$$\nabla f_\gamma(\mathbf{w}) = -\tfrac{s}{\gamma}\mathbb{E}[f(\mathbf{w}+\gamma\boldsymbol{\delta})\mathsf{n}(\boldsymbol{\delta})] = -s\mathbb{E}\left[\tfrac{f(\mathbf{w}+\gamma\boldsymbol{\delta})-f(\mathbf{w})}{\gamma}\mathsf{n}(\boldsymbol{\delta})\right] = -s\mathbb{E}\left[\tfrac{f(\mathbf{w}+\gamma\boldsymbol{\delta})-f(\mathbf{w}-\gamma\boldsymbol{\delta})}{2\gamma}\mathsf{n}(\boldsymbol{\delta})\right],$$

  where $\boldsymbol{\delta} \sim \text{Uniform}(\partial K)$. This setting only requires f to be defined (and bounded) over $C + \gamma K$ if we are only interested in f over C. In particular, let $K = \mathsf{B}_2(\mathbf{0}, \sqrt{d+2})$ we have $\mathsf{n}(\boldsymbol{\delta}) = -\sqrt{d+2}\boldsymbol{\delta}/\|\boldsymbol{\delta}\|_2$ and $s = d/\sqrt{d+2} \le \sqrt{d}$, which was considered in the seminal book of Nemirovski and Yudin (1983) and later used in Flaxman et al. (2005) for online bandits.

Nesterov, Y. and V. Spokoiny (2017). "Random Gradient-Free Minimization of Convex Functions". *Foundations of Computational Mathematics*, vol. 17, pp. 527–566.
Katz, V. J. (1979). "The History of Stokes' Theorem". *Mathematics Magazine*, vol. 52, no. 3, pp. 146–156.
Nemirovski, A. S. and D. B. Yudin (1983). "Problem complexity and method efficiency in optimization". Wiley.
Flaxman, A. D., A. T. Kalai, and H. B. McMahan (2005). "Online convex optimization in the bandit setting: gradient descent without a gradient". In: *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 385–394.

Algorithm 25.11: Randomized smoothing for gradient-free optimization

We can now put everything together:

- We optimize f_γ as a smoothed approximation of f. The approximation error is bounded in Exercise 25.7 for function values and in Exercise 25.8 for gradients.

- We compute an unbiased, stochastic (sub)gradient of f_γ by

  (I). $\hat{\partial}^1 f_\gamma(\mathbf{w}) = -\frac{1}{\gamma}f(\mathbf{w}+\gamma\boldsymbol{\epsilon})\cdot\nabla \ln \mathsf{p}(\boldsymbol{\epsilon})$;

  (II). $\hat{\partial}^{1,0} f_\gamma(\mathbf{w}) = -\frac{f(\mathbf{w}+\gamma\boldsymbol{\varepsilon})-f(\mathbf{w})}{\gamma}\nabla \ln \mathsf{p}(\boldsymbol{\varepsilon})$;

  (III). $\hat{\partial}^{1,1} f_\gamma(\mathbf{w}) = -\frac{f(\mathbf{w}+\gamma\boldsymbol{\varepsilon})-f(\mathbf{w}-\gamma\boldsymbol{\varepsilon})}{2\gamma}\nabla \ln \mathsf{p}(\boldsymbol{\varepsilon})$;

  (IV). $\hat{\partial} f_0(\mathbf{w}) = -f'(\mathbf{w};\boldsymbol{\varepsilon})\nabla \ln \mathsf{p}(\boldsymbol{\varepsilon})$.

Note that except the last choice, we only require 1 or 2 evaluations of the function itself, and these stochastic (sub)gradients, except the last one, are in general biased for the original function f.

- We bound the second moment of the stochastic (sub)gradient, as shown in Exercise 25.13 below.
- We apply the stochastic GDA algorithm in Lecture 23 and obtain convergence towards f_γ.
- We set γ appropriately so that we obtain convergence towards f, in much the same way as in Lecture 11.

Example 25.12: Concrete rates

We give some concrete examples on how to set γ:

- If f is L_0-Lipschitz continuous and convex, then using $\hat\partial^{1,0} f_\gamma$ we obtain from Remark 23.8 that

$$\mathbb{E}[f(\bar{\mathbf{w}}_t) - f(\mathbf{w})] - \gamma \mathsf{L}_0\sqrt{d} \le \mathbb{E}[f_\gamma(\bar{\mathbf{w}}_t) - f_\gamma(\mathbf{w})] \le \frac{\|\mathbf{w}_0 - \mathbf{w}\|_2^2 + \sum_{k=0}^t \eta_k^2[\mathsf{L}_\gamma^2 + \varsigma^2]}{2H_t}$$
$$\le \frac{\|\mathbf{w}_0 - \mathbf{w}\|_2^2 + \sum_{k=0}^t \eta_k^2 \mathsf{L}_0^2(d+1)^2}{2H_t}.$$

Setting

$$\gamma = \frac{\epsilon}{2\mathsf{L}_0\sqrt{d}}, \quad \eta_t = \frac{\mathrm{diam}(C)}{(d+1)\mathsf{L}_0\sqrt{t+1}}$$

we have

$$\mathbb{E}[f(\bar{\mathbf{w}}_t) - f(\mathbf{w})] \le \epsilon, \quad \text{if } t > \tfrac{4(d+1)^2}{\epsilon^2}[\mathrm{diam}(C)]^2\mathsf{L}_0^2,$$

which is d^2 times slower than running subgradient directly on f.

- If f is L_1-smooth and convex, then using again $\hat\partial^{1,0} f_\gamma$ we obtain similarly

$$\mathbb{E}[f(\bar{\mathbf{w}}_t) - f(\mathbf{w})] - \tfrac{\gamma^2\mathsf{L}_1 d}{2} \le \mathbb{E}[f_\gamma(\bar{\mathbf{w}}_t) - f_\gamma(\mathbf{w})] \le \frac{\|\mathbf{w}_0 - \mathbf{w}\|_2^2 + \sum_{k=0}^t \eta_k^2 \mathbb{E}\|\hat\partial^{1,0} f_\gamma(\mathbf{w}_k)\|_2^2}{2H_t}.$$

With $\gamma = O\left(\frac{1}{d}\sqrt{\frac{\epsilon}{\mathsf{L}_1}}\right)$ and $\eta_t \equiv O\left(\frac{1}{d\mathsf{L}_1}\right)$, an ϵ-approximate minimizer of f can be found in $O\left(\frac{d}{\epsilon}\mathsf{L}_1\,\mathrm{diam}^2(C)\right)$ many steps, which is d times slower than running (projected) gradient directly on f.

Exercise 25.13: Second moment bound

Prove the following for the Gaussian smoothing (so that $\nabla \ln \mathsf{p}(\boldsymbol{\varepsilon}) = -\boldsymbol{\varepsilon}$):

- If f is differentiable, then

$$\mathbb{E}\|\hat\partial f_0(\mathbf{w})\|_2^2 = \mathbb{E}\|\boldsymbol{\varepsilon}\|_2^4 \left\langle \tfrac{\boldsymbol{\varepsilon}}{\|\boldsymbol{\varepsilon}\|_2}, \nabla f(\mathbf{w})\right\rangle^2 = \mathbb{E}\|\boldsymbol{\varepsilon}\|_2^4 \cdot \mathbb{E}\left\langle \tfrac{\boldsymbol{\varepsilon}}{\|\boldsymbol{\varepsilon}\|_2}, \nabla f(\mathbf{w})\right\rangle^2 = (d+2)\|\nabla f(\mathbf{w})\|_2^2,$$

where we use the fact that $\|\boldsymbol{\varepsilon}\|_2^2$ and $\frac{\boldsymbol{\varepsilon}}{\|\boldsymbol{\varepsilon}\|_2}$ are independent while the former follows χ_d^2 and the latter follows uniform on the sphere.

- If f is L_0-Lipschitz continuous, then $\mathbb{E}\|\hat\partial^{1,0} f_\gamma(\mathbf{w})\|_2^2 \le \mathsf{L}_0^2 d(d+2)$.
- If ∇f is L_1-Lipschitz continuous, then $\mathbb{E}\|\hat\partial^{1,0} f_\gamma(\mathbf{w})\|_2^2 \le \frac{\gamma^2\mathsf{L}_1^2}{2} d(d+2)(d+4) + 2(d+2)\|\nabla f(\mathbf{w})\|_2^2$.
- If $\pm f$ is $\mathsf{L}_1^\pm$-smooth, then $\mathbb{E}\|\hat\partial^{1,1} f_\gamma(\mathbf{w})\|_2^2 \le \frac{\gamma^2(\mathsf{L}_1^+ + \mathsf{L}_1^-)^2}{8} d(d+2)(d+4) + 2(d+2)\|\nabla f(\mathbf{w})\|_2^2$.

- If $\nabla^2 f$ is L_2-Lipschitz continuous, then $\mathbb{E}\|\hat{\partial}^{1,1} f_\gamma(\mathbf{w})\|_2^2 \leq \frac{\gamma^4 \mathsf{L}_2^2}{18} d(d+2)(d+4)(d+6) + 2(d+2)\|\nabla f(\mathbf{w})\|_2^2$.

Remark 25.14: Square root dependence on dimension

The dependence on dimension has been further reduced in (Ghadimi and Lan 2013; Duchi et al. 2015). See also the recent work in (Auger and Hansen 2016; Bach and Perchet 2016; Shamir 2017; Balasubramanian and Ghadimi 2018; Bergou et al. 2020).

Ghadimi, S. and G. Lan (2013). "Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming". *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2341–2368.

Duchi, J. C., M. I. Jordan, M. J. Wainwright, and A. Wibisono (2015). "Optimal Rates for Zero-Order Convex Optimization: The Power of Two Function Evaluations". *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2788–2806.

Auger, A. and N. Hansen (2016). "Linear Convergence of Comparison-based Step-size Adaptive Randomized Search via Stability of Markov Chains". *SIAM Journal on Optimization*, vol. 26, no. 3, pp. 1589–1624.

Bach, F. and V. Perchet (2016). "Highly-Smooth Zero-th Order Online Optimization". In: *Proceedings of the 29th Annual Conference on Learning Theory*, pp. 257–283.

Shamir, O. (2017). "An Optimal Algorithm for Bandit and Zero-Order Convex Optimization with Two-Point Feedback". *Journal of Machine Learning Research*, vol. 18, pp. 1–11.

Balasubramanian, K. and S. Ghadimi (2018). "Zeroth-order (Non)-Convex Stochastic Optimization via Conditional Gradient and Gradient Updates". In: *Advances in Neural Information Processing Systems 31*, pp. 3455–3464.

Bergou, E. H., E. Gorbunov, and P. Richtárik (2020). "Stochastic Three Points Method for Unconstrained Smooth Minimization". *SIAM Journal on Optimization*, vol. 30, no. 4, pp. 2726–2749.

Alert 25.15: When to use?

- Same dependence on ϵ!
- Only 1 or 2 evaluation of the function per step!
- Convergence in terms of expectation or high probability.
- Much worse dependence on the dimension!

$\boxed{\text{Use only if you have to!}}$