Zero-order Convex Optimization: An Introduction

Ziniu Li
ziniuli@link.cuhk.edu.cn

The Chinese University of Hong Kong, Shenzhen, Shenzhen, China

Nov. 28, 2020

Mainly based on the paper:
Duchi, John C., et al. "Optimal rates for zero-order convex optimization: The power of two function evaluations." IEEE Transactions on Information Theory 61.5 (2015): 2788-2806.

Outline

Introduction to Zero-order Optimization

- We consider the following optimization:

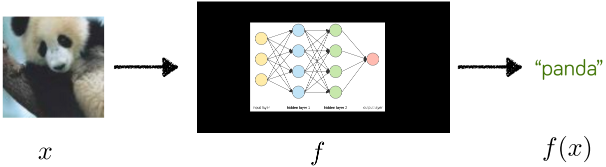
$$\min_{x \in \mathcal{X}} \quad f(x).$$

- When f is convex and importantly differentiable, many first-order (i.e., gradient-based) methods can be applied [Nesterov, 2018].
  - Typically, the convergence rate is dimension-free.
- However, if f is non-differentiable and only zero-order information is available?
  - We have access to $f(x)$ but not $\nabla f(x)$.
  - Even $\nabla f(x)$ could not be properly defined.

ZO Application: Adversarial Attack

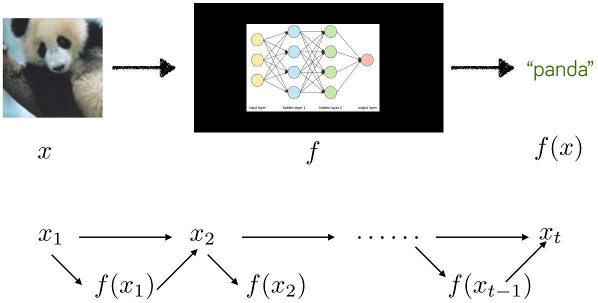
- Imagine there is a hacker who wants to attack the trained neural nets.
  - He can send a query to the "black-box" model and get the feedback.
- The objective to find some adversarial examples that incurs large losses:

$$\min_{\epsilon \in \mathbb{R}^d} \quad -\mathcal{L}(f(x_0+\epsilon), y), \quad \text{s.t. } ||\epsilon|| \le \delta.$$

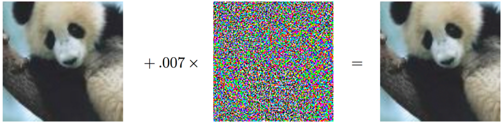


ZO Application: Adversarial Attack

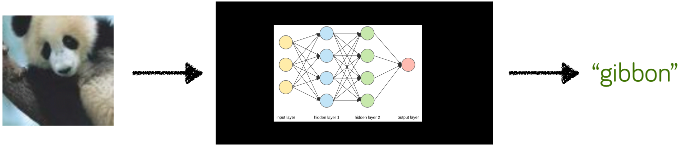
- The hacker can only adopt a ZO algorithm to optimize the adversarial example.



ZO Application: Adversarial Attack



$x_1 \longrightarrow \cdots\cdots \longrightarrow x_t$



x_t f $f(x_t)$

More Applications of Zero-order Optimization

- Bandit Optimization [Flaxman et al., 2005, Bartlett et al., 2008, Agarwal et al., 2010].
- Simulation-based optimization [Spall, 2005].
- Graphical model inference [Wainwright and Jordan, 2008].
- Policy optimization [Wierstra et al., 2014, Salimans et al., 2017].
- Escaping the local minimum in ERM [Jin et al., 2018].

Main Difficulty of Zero-Order Optimization

- (The curse of dimension) Convergence rate of ZO methods scales up with dimension d [Duchi et al., 2012b, Jamieson et al., 2012, Shamir, 2013, Duchi et al., 2015].
- Consider to optimize a Lipschitz continuous function f: $|f(x) - f(y)| \leq L||x - y||$ with only zero-order information.
  - The lower bound of total evaluation numbers suggests an exponential dependence on d.

$$\text{Lower Bound: } \left\lfloor \frac{L}{2\epsilon} \right\rfloor^d$$

  - The simple method of grid search is minimax optimal!

$$\text{Upper Bound: } \left(\left\lfloor \frac{L}{2\epsilon} \right\rfloor + 1 \right)^d$$

Outline

A General Start: Stochastic Optimization

- We need to restrict our attention to not-so-hard class: convex function class.

$$\min_{\theta \in \Theta} f(\theta) := \mathbb{E}_P \left[F(\theta; X) \right] = \int_{\mathcal{X}} F(\theta; x)\, dP(x). \tag{1}$$

  where $\Theta \subseteq \mathbb{R}^d$ is a compact convex set, P is a distribution over $\mathcal{X}$ and for every $x \in \mathcal{X}$ we have $F(\cdot; x)$ is closed and convex.
- Each iteration, we have access to $F(\theta; x)$ by drawing x from P (this process is not controlled by algorithms).
  - In machine learning, x is a training sample, $F_i(\theta; x)$ is the individual loss and $f(\theta)$ is the population/empirical loss.
  - We do not know $\nabla f(\theta)$ or even $\nabla F(\theta; x)$.

Intuition of Zero-order Optimization

- We can utilize multiple function evaluations to approximate the directional derivative:

$$F'(\theta; z, x) = \lim_{u \downarrow 0} \frac{1}{u} \left(F(\theta + uz; x) - F(\theta; x) \right) = \langle \nabla F(\theta; x), z \rangle.$$

- In high-level, zero-order algorithms sample a noisy gradient to optimize.

$$\frac{1}{u} \left(F(\theta + uz; x) - F(\theta; x) \right) z \approx zz^\top \nabla F(\theta; x).$$

  where $u > 0$ is a small perturbation size and z is a random vector.

- Taking the expectation on both sides and with the assumption that $\mathbb{E}\left[zz^\top\right] = \mathbb{I}_d$, we obtain an estimate of $\nabla F(\theta; x)$.

Algorithmic Assumptions

- We consider a mirror descent type algorithm:

$$\theta^{t+1} = \operatorname*{argmin}_{\theta \in \Theta} \left\{ \langle g^t, \theta \rangle + \frac{1}{\alpha(t)} D_\psi \left(\theta, \theta^t \right) \right\}, \tag{2}$$

  - $\{\alpha(t)\}_{t=1}^{\infty}$ is a non-increasing sequence of step sizes.
  - $g^t \in \mathbb{R}^d$ is a (subgradient) vector.
  - D_ψ is a Bregman distance defined by the proximal function ψ: $D_\psi(\theta, \tau) := \psi(\theta) - \psi(\tau) - \langle \nabla \psi(\tau), \theta - \tau \rangle$.

Algorithmic Assumptions

Assumption 1.

The proximal function ψ is 1-strongly convex with respect to the norm $||\cdot||$. The domain Θ is compact and there exists $R < \infty$ such that $D_\psi(\theta^*, \theta) \leq \frac{1}{2}R^2$ for $\theta \in \Theta$.

- If we consider $||\cdot||$ as ℓ_2-norm, $\psi(\theta) = \frac{1}{2}||\theta||_2^2$ and $\Theta = \mathbb{R}^n$, we have $D_\psi(\theta, \tau) = \frac{1}{2}\|\theta - \tau\|_2^2$, and,

$$
\begin{aligned}
\theta^{t+1} &= \underset{\theta \in \Theta}{\operatorname{argmin}} \left\{ \langle g^t, \theta \rangle + \frac{1}{\alpha(t)} D_\psi\left(\theta, \theta^t\right) \right\} \\
&= \theta^t - \alpha(t) g^t.
\end{aligned}
$$

Algorithmic Assumptions

Assumption 2.

There is a constant $G < \infty$ *such that the (sub)gradient* g *satisfies that* $\mathbb{E}\left[||g(\theta; X)||^2\right] \leq G^2$ *for all* $\theta \in \Theta$.

- The variance of (sub)gradient is controlled by G.
- This holds when $F(\cdot; x)$ are G-Lipschitz continuous with respect to the norm $||\cdot||$.

Outline

Main Idea

- The directional gradient estimate can approximate the gradient:

$$\mathtt{G}_{\mathrm{sm}}(\theta; u, z, x) := \frac{F(\theta + uz; x) - F(\theta; x)}{u} z, \quad (3)$$

$$\mathbb{E}\left[\mathtt{G}_{\mathrm{sm}}(\theta; u, z, x)\right] = \nabla f(\theta) + u \cdot \mathtt{bias}, \quad (4)$$

  here we assume $x \sim P(x)$ and $z \sim \mu(z)$ and the $\mathtt{bias}$ term will be shown later.
- We use a noisy gradient estimate g^t and shrink the parameter u to control the bias.

$$g^t = \mathtt{G}_{\mathrm{sm}}\left(\theta^t; u_t, Z^t, X^t\right) = \frac{F\left(\theta^t + u_t Z^t; X^t\right) - F\left(\theta^t; X^t\right)}{u_t} Z^t. \quad (5)$$

More Assumptions about Smooth Optimization

- Different from stochastic mirror descent, zero-order algorithms need to ensure the parameter domain is well-defined.

Assumption 3.

The domain of Functions F and support of μ satisfies

$$\mathbf{dom}\, F(\cdot; x) \supset \Theta + u \operatorname{supp} \mu \qquad \text{for } x \in \mathcal{X}.$$

and,

$$\mathbb{E}_{\mu}\left[ZZ^{\top}\right] = \mathbb{I}_d.$$

More Assumptions about Smooth Optimization

Assumption 4.

For $Z \sim \mu$*, the quantity* $M(\mu) = \sqrt{\mathbb{E}\left[\|Z\|^4 \|Z\|_*^2\right]}$ *is finite. Moreover, there is a function* $s : \mathbb{N} \to \mathbb{R}_+$ *such that*

$$\mathbb{E}\left[\|\langle g, Z\rangle Z\|_*^2\right] \leq s(d)\|g\|_*^2 \quad \text{for any vector } g \in \mathbb{R}^d. \tag{6}$$

- For example, μ is a standard Gaussian distribution $\mathcal{N}(0, \mathbb{I}_d)$ and $||\cdot||$ is the ℓ_2-norm.
- $M(\mu) = \sqrt{\mathbb{E}\left[\|Z\|^6\right]} = \sqrt{15d^6} \lesssim d^3$.
- $\mathbb{E}\left[||\langle g, Z\rangle Z||_2^2\right] = \mathbb{E}\left[g^\top ZZ^\top ZZ^\top g\right] = g^\top \mathbb{E}\left[(2+d)\mathbb{I}\right] g \Longrightarrow s(d) \lesssim d$.

More Assumptions about Smooth Optimization

Assumption 5.

There is a function $L : \mathcal{X} \to \mathbb{R}_+$ *such that for* P*-almost every* $x \in \mathcal{X}$*, the function* $F(\cdot; x)$ *has* $L(x)$*-Lipschitz continuous gradient with respect to the norm* $||\cdot||$ *and moreover the quantity* $L(P) := \sqrt{\mathbb{E}\left[(L(X))^2\right]}$ *is finite.*

Gradient Approximation

Lemma 1.

Under Assumption 4 and 5, the gradient estimate (3) has the expectation:

$$\mathbb{E}\left[G_{\mathrm{sm}}(\theta; u, Z, X)\right] = \nabla f(\theta) + uL(P)v(\theta, u), \tag{7}$$

for a vector $v = v(\theta, u)$ *such that* $||v||_* \leq \frac{1}{2}\mathbb{E}\left[||Z||^2||Z||_*\right]$. *Its expected squared norm has the bound*

$$\mathbb{E}\left[\|G_{\mathrm{sm}}(\theta; u, Z, X)\|_*^2\right] \leq 2s(d)\mathbb{E}\left[\|g(\theta; X)\|_*^2\right] + \frac{1}{2}u^2 L(P)^2 M(\mu)^2. \tag{8}$$

Here $g(\theta; x) \in \partial F(\theta; x)$ *is a subgradient with* $\mathbb{E}\left[g(\theta; x)\right] \in \partial f(\theta)$.

Implication of Lemma 1

- The estimate g^t is unbiased up to a correction term of u.

$$\mathbb{E}\left[\mathtt{G}_{\mathrm{sm}}(\theta; u, Z, X)\right] = \nabla f(\theta) + uL(P)v(\theta, u).$$

- The second moment is also unbiased up to an order u_t^2 correction–within a factor $s(d)$.

$$\mathbb{E}\left[\|\mathtt{G}_{\mathrm{sm}}(\theta; u, Z, X)\|_*^2\right] \leq 2s(d)\mathbb{E}\left[\|g(\theta; X)\|_*^2\right] + \frac{1}{2}u^2 L(P)^2 M(\mu)^2.$$

- $\rightsquigarrow$ As long as we shrink u_t, we can obtain arbitrary accurate estimates of the directional derivative.

Proof of Lemma 1: Preliminary

- We start with a general convex function h with L_h-Lipschitz continuous gradient w.r.t the norm $||\cdot||$.
- For any $u > 0$, we have that

$$h'(\theta, z) = \frac{\langle \nabla h(\theta), uz \rangle}{u} \leq \frac{h(\theta + uz) - h(\theta)}{u} \leq \frac{\langle \nabla h(\theta), uz \rangle + (L_h/2) \|uz\|^2}{u}$$
$$= h'(\theta, z) + \frac{L_h u}{2} \|z\|^2,$$

- Therefore for any $z \in \mathbb{R}^d$, we have that

$$\frac{h(\theta + uz) - h(\theta)}{u} z = h'(\theta, z) z + \frac{L_h u}{2} \|z\|^2 \gamma(u, \theta, z) z, \tag{9}$$

where γ is some function with range contained in $[0, 1]$.

Proof of Lemma 1: Preliminary

- By our assumption that $\mathbb{E}\left[ZZ^\top\right] = \mathbb{I}_d$, (9) implies that

$$\mathbb{E}\left[\frac{h(\theta + uZ) - h(\theta)}{u} Z\right] = \mathbb{E}\left[h'(\theta, Z)Z + \frac{L_h u}{2}\|Z\|^2 \gamma(u, \theta, Z)Z\right] \tag{10}$$

$$= \mathbb{E}\left[\langle \nabla h(\theta), Z\rangle Z\right] + \mathbb{E}\left[\frac{L_h u}{2}\|Z\|^2 \gamma(u, \theta, Z)Z\right] \tag{11}$$

$$= \nabla h(\theta) + u L_h v(\theta, u), \tag{12}$$

where $v(\theta, u) \in \mathbb{R}^d$ is an error vector with $||v(\theta, u)||_* \leq \frac{1}{2}\mathbb{E}\left[\|Z\|^2 \|Z\|_*\right]$.

Proof of Lemma 1: The First Moment

- Recalling the gradient estimate in (7), expression (12) implies that

$$\mathbb{E}\left[\mathrm{G}_{\mathrm{sm}}(\theta; u, Z, x)\right] = \nabla F(\theta; x) + uL(x)v(\theta, u), \tag{13}$$

for some vector $v = v(\theta, u)$ with $2||v||_* \leq \mathbb{E}\left[\|Z\|^2 \|Z\|_*\right]$.

- Now taking the expectation over X, for the first term we have $\mathbb{E}\left[\nabla F(\theta; X)\right] = \nabla f(\theta^t)$. For the second term, by Jensen's inequality we have that

$$\mathbb{E}\left[L(X)\|v(\theta, u)\|_*\right] \leq \sqrt{\mathbb{E}\left[L(X)^2\right]}\, \|v\|_* \leq \frac{1}{2} L(P) \mathbb{E}\left[\|Z\|^2 \|Z\|_*\right],$$

from which the bound (7) follows.

Proof of Lemma 1: The Second Moment

- Applying (9) to $F(\cdot; X)$, we obtain that

$$\mathrm{G}_{\mathrm{sm}}(\theta; u, Z, X) = \langle \mathrm{g}(\theta; X), Z \rangle Z + \frac{L(X)u}{2} \|Z\|^2 \gamma Z,$$

for some function $\gamma \equiv \gamma(u, \theta, Z, X) \in [0, 1]$.

- To upper bound the second moment, we use the relation $(a+b)^2 \leq 2a^2 + 2b^2$:

$$\begin{aligned}
\mathbb{E}\left[\|G_{\mathrm{sm}}(\theta; u, Z, X)\|_*^2\right] &\leq \mathbb{E}\left[\left(\|\langle \mathrm{g}(\theta, X), Z \rangle Z\|_* + \frac{1}{2}\|L(X)u\|Z\|^2 \gamma Z\|_*\right)^2\right] \\
&\leq 2\mathbb{E}\left[\|\langle \mathrm{g}(\theta, X), Z \rangle Z\|_*^2\right] + \frac{u^2}{2}\mathbb{E}\left[L(X)^2 \|Z\|^4 \|Z\|_*^2\right] \\
&\leq 2s(d)\mathbb{E}\left[\|g(\theta; X)\|_*^2\right] + \frac{1}{2}u^2 L(P)^2 M(\mu)^2.
\end{aligned}$$

Key Result For Smooth Optimization

Theorem 1.

Under Assumption 1, 2 3, 4 and 5, consider a sequence $\{\theta^t\}_{t=1}^{\infty}$ generated by the mirror descent update (2) using the gradient estimator (5), with step and perturbation parameter

$$\alpha(t) = \alpha \frac{R}{2G\sqrt{s(d)}\sqrt{t}} \quad \text{and} \quad u_t = u \frac{G\sqrt{s(d)}}{L(P)M(\mu)} \cdot \frac{1}{t} \quad \text{for } t = 1, 2, \ldots$$

Then for all k,

$$\mathbb{E}\left[f(\widehat{\theta}(k)) - f(\theta^*)\right] \leq 2\frac{RG\sqrt{s(d)}}{\sqrt{k}} \max\left\{\alpha, \alpha^{-1}\right\} + \alpha u^2 \frac{RG\sqrt{s(d)}}{k} + u\frac{RG\sqrt{s(d)}\log(2k)}{k}, \tag{14}$$

where $\widehat{\theta}(k) = \frac{1}{k}\sum_{t=1}^{k}\theta^t$ and the expectation is taken w.r.t. samples X and Z.

Implication of Theorem 1

- We first compare the result with stochastic mirror descent with first-order information.

Method	Step Size	Perturbation Size	Optimality Gap
First-order	$\frac{\alpha}{\sqrt{t}}$		$\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$
Zero-order	$\alpha\frac{R}{2G\sqrt{s(d)}\sqrt{t}}$	$u\frac{G\sqrt{s(d)}}{L(P)M(\mu)} \cdot \frac{1}{t}$	$\mathcal{O}\left(\frac{\sqrt{s(d)}}{\sqrt{k}}\right)$

- The convergence rate only slows down by $\sqrt{s(d)}$!
  - If we consider μ a Gaussian distribution over $\mathbb{R}^d$ or a uniform distribution over ℓ_2-ball, $s(d) \lesssim d$.
  - This is partially because we have to use a small step size in ZO algorithms.

Implication of Theorem 1

- We see that a small perturbation size is applied to control the bias.
- Variance-control can also be achieved by multiple independent samples $Z^{t,i}$, $i = 1, \cdots, m$ to construct a more accurate gradient estimate.

$$g^t = \frac{1}{m} \sum_{i=1}^{m} \mathtt{G}_{\mathrm{sm}}(\theta^t; u_t, Z^{t,i}, X^t).$$

- In this way, we may achieve a standard $RG/\sqrt{k}$ convergence rate (see the next page).

Smooth Optimization with Multiple Function Evaluations

Corollary 1.

Let $Z^{t,i}$, $i = 1, \cdots, m$ be sampled independently according to μ and at each iteration of mirror descent use the gradient estimate $g^t = \frac{1}{m}\sum_{i=1}^{m} G_{\rm sm}(\theta^t; u_t, Z^{t,i}, X^t)$ with the step and perturbation sizes

$$\alpha(t) = \alpha \frac{R}{2G\max\{\sqrt{d/m}, 1\}} \cdot \frac{1}{\sqrt{t}} \quad \textit{and} \quad u_t = u\frac{G}{L(P)d^{3/2}} \cdot \frac{1}{t}.$$

There exists a universal constant $C \leq 5$ such that for all k,

$$\mathbb{E}\left[f(\widehat{\theta}(k)) - f(\theta^*)\right] \leq C\frac{RG\sqrt{1 + d/m}}{\sqrt{k}}\left[\max\left\{\alpha, \alpha^{-1}\right\} + \alpha u^2 \frac{1}{\sqrt{k}} + u\frac{\log(2k)}{k}\right].$$

More Cooments on Corollary 1 and Theorem 1

- We could achieve a "dimension-free" result by choose $m \approx d$ in Corollary 1.
- However, if we define the sample complexity as the total number of function evaluations, the sample complexity is clearly not dimension-free.
  - There is a (currently unknown) trade-off about how many evaluations to apply.
- In high dimensional scenarios, we can properly choose the proximal function and the norm $||\cdot||$ to release the true power of mirror descent.
  - Refer to [Duchi et al., 2015, Corollary 3] and [Beck and Teboulle, 2003].
- The term of $\max\{\alpha, \alpha^{-1}\}$ is said to be robust in stochastic optimization.
  - i.e., if we specify α wrongly, the final result is not so bad (ref to [Nemirovski et al., 2009]).

Outline

Difficulty in Non-smooth Optimization

- Recall that by L-Lipschitz continuous gradient of $F(\cdot; x)$, we have

$$\mathbb{E}\left[\|\mathtt{G}_{\mathrm{sm}}(\theta; u, Z, X)\|_*^2\right] \leq 2s(d)\mathbb{E}\left[\|g(\theta; X)\|_*^2\right] + \frac{1}{2}u^2 L(P)^2 M(\mu)^2$$

$$\lesssim \underbrace{s(d)}_{\lesssim d} \underbrace{\mathbb{E}\left[\|g(\theta; X)\|_*^2\right]}_{\leq G^2} \qquad \left(u \propto \frac{\sqrt{s(d)\mathbb{E}\left[\|g(\theta; X)\|_*^2\right]}}{L(P)M(\mu)}\right)$$

- For non-smooth case, we only have G-Lipschitz continuity, we have that

$$\mathbb{E}\left[\|\mathtt{G}_{\mathrm{sm}}(\theta; u, Z, X)\|_*^2\right] \leq \mathbb{E}\left[\left\|\frac{F(\theta + uZ; x) - F(\theta; x)}{u} Z\right\|_2^2\right] \leq G^2 \underbrace{\mathbb{E}\left[\|Z\|_2^4\right]}_{\geq d^2}.$$

- That is, the $\mathcal{O}(d^2)$ term for non-smooth in contrast to the $\mathcal{O}(d)$ term for the smooth case.

Solution: Smoothing the Non-smooth Functions

- For a general function $f(\theta)$, we can define the smoothed objective function,

$$f_u(\theta) := \mathbb{E}[f(\theta + uZ)] = \int f(\theta + uz)\, d\mu(z).$$

- If f is Lipschitz continuous and convex, we can show that $f_u(\theta)$ is differentiable even though f is not [Duchi et al., 2012a, Nesterov and Spokoiny, 2017].
- Implication: if we smooth the non-smooth function $F(\theta; x)$ slightly, we may achieve a convergence rate that is roughly the same as that in smooth case.

Gradient Estimate for Non-smooth Functions

- Based on the above intuition, we can construct the gradient estimate:

$$\mathrm{G}_{\mathrm{ns}}\left(\theta; u_1, u_2, z_1, z_2, x\right) := \frac{F\left(\theta + u_1 z_1 + u_2 z_2; x\right) - F\left(\theta + u_1 z_1; x\right)}{u_2} z_2. \tag{15}$$

  Here z_1, z_2 are independently drawn from distributions μ_1 and μ_2 and $\{u_{1,t}\}_{t=1}^{\infty}$ and $\{u_{2,t}\}_{t=1}^{\infty}$ are two positive non-increasing sequences with $u_{2,t} \leq u_{1,t}$.

- And similarly, we have

$$g^t = \frac{F\left(\theta^t + u_{1,t} Z_1^t + u_{2,t} Z_2^t; X^t\right) - F\left(\theta^t + u_{1,t} Z_1^t; X^t\right)}{u_{2,t}} Z_2^t, \tag{16}$$

  where Z_1^t serves the "smoothing" function and Z_2^t severs the gradient estimate function.

More Assumptions For Non-smooth Functions

Assumption 6.

There is a function $G : \mathcal{X} \to \mathbb{R}_+$ such that for every $x \in \mathcal{X}$, the function $F(\cdot; x)$ is $G(x)$-Lipschitz with respect to the ℓ_2-norm $|| \cdot ||$ and the quantity $G(P) = \sqrt{\mathbb{E}\left[G(X)^2\right]}$ is finite.

More Assumptions For Non-smooth Functions

Assumption 7.

The smoothing distributions are one of the following pairs:

- *both* μ_1 *and* μ_2 *are standard normal distribution in* $\mathbb{R}^d$.
- *both* μ_1 *and* μ_2 *are uniform on the* ℓ_2*-ball of radius* $\sqrt{d+2}$.
- μ_1 *is uniformly on the* ℓ_2*-ball of radius* $\sqrt{d+2}$ *and* μ_2 *is uniform on the* ℓ_2*-sphere of radius* $\sqrt{d}$.

In addition, the domain of $F(\cdot;x)$ *is well defined:*

$$\operatorname{dom} F(\cdot;x) \supset \Theta + u_{1,1} \operatorname{supp} \mu_1 + u_{2,1} \operatorname{supp} \mu_2 \qquad \text{for } x \in \mathcal{X}.$$

Gradient Approximation For Non-smooth Functions

Lemma 2.

Under Assumption 6 and 7, the gradient estimator (15) has the expectation:

$$\mathbb{E}\left[G_{\mathrm{ns}}\left(\theta; u_1, u_2, Z_1, Z_2, X\right)\right] = \nabla f_{u_1}(\theta) + \frac{u_2}{u_1} G(P) v\left(\theta, u_1, u_2\right), \tag{17}$$

where $v = v(\theta, u_1, u_2)$ *has bound* $\|v\|_2 \leq \frac{1}{2}\mathbb{E}\left[\|Z_2\|_2^3\right]$. *There exists a universal constant* c *such that*

$$\mathbb{E}\left[\|G_{\mathrm{ns}}\left(\theta; u_1, u_2, Z_1, Z_2, X\right)\|_2^2\right] \leq cG(P)^2 d\left(\sqrt{\frac{u_2}{u_1}} d + 1 + \log d\right). \tag{18}$$

Comments on Lemma 2

- Compared to Lemma 1, Lemma 2 suggests that the gradient estimate is nearly unbiased.
- If we could choose a small u_2 such that $\sqrt{\frac{u_2}{u_1}}d$ is almost negligible, then we recover the convergence rate of smooth optimization.
- Actually, there still is an additional $\log d$ term but it is expected to remove this in future works.

Convergence Result For Non-smooth Optimization

Theorem 2.

Under Assumption 1, 6 and 7, consider a sequence $\{\theta^t\}_{t=1}^{\infty}$ generated according to mirror descent update 2 using the gradient estimator (16) with step and perturbation sizes

$$\alpha(t) = \alpha \frac{R}{G(P)\sqrt{d\log(2d)}\sqrt{t}}, \quad u_{1,t} = u\frac{R}{t}, \quad \text{and} \quad u_{2,t} = u\frac{R}{d^2t^2}.$$

Then there exists a universal constant c such that for all k,

$$\mathbb{E}\left[f(\widehat{\theta}(k)) - f(\theta^*)\right] \le c\max\left\{\alpha, \alpha^{-1}\right\}\frac{RG(P)\sqrt{d\log(2d)}}{\sqrt{k}} + cuRG(P)\sqrt{d}\frac{\log(2k)}{k}, \tag{19}$$

where $\widehat{\theta}(k) = \frac{1}{k}\sum_{t=1}^{k}\theta^t$ and the expectation is taken w.r.t. samples X and Z.

Comments on Theorem 2

- Compared to Theorem 1, Theorem 2 suggests that two-point zero-order algorithms for non-smooth functions is at worst a factor $\sqrt{\log d}$ worse than the rate for smooth functions.

Outline

Minimax Error and Minimax Optimal

- Let $\mathcal{F}$ be a collection of pairs (F, P), each of which defines a problem instance (1).
- Let $\mathbb{A}_k$ denote the collection of all algorithms that receives a sequences $(Y_1, \cdots, Y_k)$, each of which contains two-point evaluations:

$$Y^t = \left[F(\theta^t, X^t), F(\tau^t, X^t)\right].$$

  Here (θ^t, τ^t) can be determined by the algorithm.
- Given an algorithm $\mathcal{A} \in \mathbb{A}_k$ and a pair $(F, P) \in \mathcal{F}$, the optimality gap is defined as

$$\epsilon_k(\mathcal{A}, F, P, \Theta) := f(\widehat{\theta}(k)) - \inf_{\theta \in \Theta} f(\theta) = \mathbb{E}_P[F(\widehat{\theta}(k); X)] - \inf_{\theta \in \Theta} \mathbb{E}_P[F(\theta; X)],$$

  where $\widehat{\theta}(k)$ is the output of algorithm $\mathcal{A}$ at iteration k.

Minimax Error and Minimax Optimal

- The minimax error is defined as

$$\epsilon_k^*(\mathcal{F}, \Theta) := \inf_{\mathcal{A} \in \mathbb{A}_k} \sup_{(F,P) \in \mathcal{F}} \mathbb{E}\left[\epsilon_k(\mathcal{A}, F, P, \Theta)\right], \tag{20}$$

where expectation is taken over the observations $(Y^1, \cdots, Y^k)$ and any additional randomness in $\mathcal{A}$.

- An algorithm $\mathcal{A}$ is called minimax optimal if its upper bound matches the lower bound up to constant and logarithmic terms.

Lower Bound For Two-point Evaluations

- For a given ℓ_p-norm $||\cdot||_p$, we consider the class of linear functionals:

$$\mathcal{F}_{G,p} := \left\{(F,P) \mid F(\theta;x) = \langle \theta, x\rangle \quad \text{with} \quad \mathbb{E}_P\left[\|X\|_p^2\right] \leq G^2\right\}.$$

- Each of which satisfies Assumption 6 (i.e., Lipschitz continuity).
- Moreover, $\nabla F(\cdot;x)$ has Lipschitz constant 0 for all x.
- We consider the domain is equal to some ℓ_q-ball of radius, i.e.,

$$\Theta = \left\{\theta \in \mathbb{R}^d \middle| \, \|\theta\|_q \leq R\right\}.$$

Lower Bound For Two-point Evaluations

Proposition 1.

For the class $\mathcal{F}_{G,2}$ *and* $\Theta = \left\{\theta \in \mathbb{R}^d \middle| \, \|\theta\|_q \leq R\right\}$*, we have*

$$\epsilon_k^* \left(\mathcal{F}_{G,2}, \Theta\right) \geq \frac{1}{12} \left(1 - \frac{1}{q}\right) \frac{GR}{\sqrt{k}} \min\left\{d^{1-1/q}, k^{1-1/q}\right\}. \tag{21}$$

- For $k \geq d$, this lower bound translates to $\Omega\left(\frac{GR}{\sqrt{k}} d^{1-1/q}\right)$.

Comments on Lower Bound 1

- For $q \geq 2$, the ℓ_2-ball of radius $d^{1/2-1/q}R$ contains the ℓ_q-ball of radius R, so the upper bound in Theorem 1 and 2 be analyzed here.
- In particular, we have the upper bound that

$$\frac{RG\sqrt{d}}{\sqrt{k}} \leq \frac{RG\sqrt{d}d^{1/2-1/q}}{\sqrt{k}} = \frac{RGd^{1-1/q}}{\sqrt{k}}.$$

- This implies that the algorithm for smooth optimization is optimal up to constant factors and the algorithm for non-smooth optimization is also tight to within logarithmic factors.

Lower Bound For Multiple Evaluations

- An inspection of the proof of Proposition 1 yields that

$$\epsilon_k^* \left(\mathcal{F}_{G,2}, \Theta\right) \geq \frac{1}{10}\left(1-\frac{1}{q}\right)\frac{GR}{\sqrt{mk}}\min\left\{d^{1-1/q}, k^{1-1/q}\right\}. \tag{22}$$

- In Corollary 1, we have the upper bound $\mathcal{O}\left(RG\frac{\sqrt{d/m}}{\sqrt{k}}\right)$.
- This indicates that when $m \to d$, the algorithm also achieves minimax optimal.

Outline

Outline

Proof of Theorem 1

- By mirror descent with Assumption 1, we have that [Beck and Teboulle, 2003, Nemirovski et al., 2009, Shalev-Shwartz, 2012]:

$$\sum_{t=1}^{t} f(\theta^t) - f(\theta^*) \leq \sum_{t=1}^{k} \langle g^t, \theta^t - \theta^* \rangle \leq \frac{1}{2\alpha(k)} R^2 + \sum_{t=1}^{k} \frac{\alpha(t)}{2} \left\| g^t \right\|_*^2 . \tag{23}$$

- Now let's introduce the error vector $e^t := \nabla f(\theta^t) - g^t$,

$$\begin{aligned} \sum_{t=1}^{k} \left(f\left(\theta^t\right) - f\left(\theta^*\right) \right) &\leq \sum_{t=1}^{k} \langle g^t, \theta^t - \theta^* \rangle + \sum_{t=1}^{k} \langle e^t, \theta^t - \theta^* \rangle \\ &\leq \frac{1}{2\alpha(k)} R^2 + \sum_{t=1}^{k} \frac{\alpha(t)}{2} \left\| g^t \right\|_*^2 + \sum_{t=1}^{k} \langle e^t, \theta^t - \theta^* \rangle . \end{aligned} \tag{24}$$

Proof of Theorem 1

- For the second moment term, by (8) in Lemma 1, we have that

$$\mathbb{E}\left[\left\|g^t\right\|_*^2\right] \leq 2s(d)G^2 + \frac{1}{2}u_t^2 L(P)^2 M(\mu)^2. \tag{25}$$

  We can properly choose $u_t \propto \frac{\sqrt{s(d)}G}{L(P)M(\mu)}$ to control this term.

- For the last term in (24), by (7) in Lemma 1, we have that

$$\sum_{t=1}^{k} \mathbb{E}\left[\left\langle e^t, \theta^t - \theta^*\right\rangle\right] \leq L(P)\sum_{t=1}^{k} u_t \mathbb{E}\left[\left\|v_t\right\|_* \left\|\theta^t - \theta^*\right\|\right] \leq \frac{1}{2}M(\mu)RL(P)\sum_{t=1}^{k} u_t. \tag{26}$$

  The last step we use the relation $||\theta^t - \theta^*|| \leq \sqrt{2D_\psi(\theta^*, \theta)} \leq R$.

Proof of Theorem 1

- By combing the above inequalities, we have that

$$\begin{aligned}&\sum_{t=1}^{t} f(\theta^t) - f(\theta^*)\\ &\leq \frac{R^2}{2\alpha(k)} + s(d)G^2 \sum_{t=1}^{k} \alpha(t) + \frac{L(P)^2 M(\mu)^2}{4} \sum_{t=1}^{k} u_t^2 \alpha(t) + \frac{M(\mu) R L(P)}{2} \sum_{t=1}^{k} u_t.\end{aligned}$$

- It remains to plug-in the chosen step and perturbation sizes and to apply Jensen's inequality.

Outline

Proof of Proposition 1

- Main idea: reduce the optimization to binary hypothesis testing problems.
- First, we construct a finite set of functions, upon of which the optimality gap is lower bounded by the sign difference.
- Consequently, we lower bound the probability of sign difference after observing k random samples with total variation distance by Le Cam's inequality.
- Finally, we present a sharp bound of total variation distance for this problem.

Proof of Proposition 1: The First Part

- We consider the binary vector v in the Boolean hypercube $\mathcal{V} = \{-1, 1\}^d$.
- The objective functions are in the form $F(\theta; x) = \langle \theta, x \rangle$.
- For each v, P_v is the Gaussian distribution $\mathcal{N}(\delta v, \sigma^2 \mathbb{I})$, where $\delta > 0$ is to be chosen later.
- Now, the problem becomes that

$$\min_{\theta \in \Theta} f_v(\theta) := \mathbb{E}_{P_v}\left[F(\theta; X)\right] = \delta \langle \theta, v \rangle, \tag{27}$$

  where $\Theta = \{\theta \in \mathbb{R}^d \,\big|\, ||\theta||_q \leq R\}$.
- It's clear that the optimal solution is given by $\theta^v = -R d^{1/q} v$.

Proof of Proposition 1: The First Part

- We claim that for any $\widehat{\theta} \in \mathbb{R}^d$ the optimality gap is bounded by (see the next page).

$$f_v(\widehat{\theta}) - f_v\left(\theta^v\right) \geq \frac{1 - 1/q}{d^{1/q}} \delta R \sum_{j=1}^{d} \mathbf{1}\left\{\operatorname{sign}\left(\widehat{\theta}_j\right) \neq \operatorname{sign}\left(\theta_j^v\right)\right\}. \tag{28}$$

- To understand (28), we note that if $\operatorname{sign}\left(\widehat{\theta}_j\right) = \operatorname{sign}\left(\theta_j^v\right)$ for all j, then (28) holds trivially. Therefore, we only need to care the case where there exist some coordinates j such that $\operatorname{sign}\left(\widehat{\theta}_j\right) \neq \operatorname{sign}\left(\theta_j^v\right)$.

Proof of Proposition 1: The First Part

- Let's split θ^v the coordinates into two parts: $\mathcal{I}_+ = \{v_i = 1\}$ and $\mathcal{I}_- = \{v_i = -1\}$. Now, we represent θ^v as below (only signs are shown):

$$\theta^v = \Big(\underbrace{+, \cdots, +}_{\mathcal{I}_-} \quad \Big| \quad \underbrace{- \cdots, -}_{\mathcal{I}_+} \Big).$$

- With the same order, we can also represent the estimator $\widehat{\theta}$ as below (only signs are shown):

$$\widehat{\theta} = \Big(\underbrace{-}_{\mathcal{I}_-^-}, \underbrace{\cdots, +}_{\mathcal{I}_-^+} \quad \Big| \quad \underbrace{+}_{\mathcal{I}_+^+} \underbrace{\cdots, -}_{\mathcal{I}_+^-} \Big).$$

Here $\mathcal{I}_-^-$ and $\mathcal{I}_+^+$ are two "error" sets, in which the sign of $\widehat{\theta}$ is different from θ^v.

Proof of Proposition 1: The First Part

- We define two optimization problems to lower bound the cost due to sign difference.

$$\min v^\top \theta, \qquad \text{s.t. } \|\theta\|_q \le 1 \tag{29}$$

$$\min v^\top \theta, \qquad \text{s.t. } \|\theta\|_q \le 1;\ \theta_j \le 0, \forall j \in \mathcal{I}_-^-;\ \theta_j \ge 0, \forall j \in \mathcal{I}_+^+. \tag{30}$$

- Denote the optimal solution by θ^A and θ^B, respectively. We have

$$\theta^A = d^{-1/q}\left(\mathbf{1}_{\mathcal{I}_-} - \mathbf{1}_{\mathcal{I}_+}\right), \quad \text{and} \quad \theta^B = (d-c)^{-1/q}\left(\mathbf{1}_{\mathcal{I}_-^+} - \mathbf{1}_{\mathcal{I}_+^-}\right),$$

where $\mathbf{1}_A$ denotes the vector with 1 for coordinates are in A and 0 otherwise. In addition, c is the sum of cardinalities of $\mathcal{I}_-^-$ and $\mathcal{I}_+^+$.

Proof of Proposition 1: The First Part

- As a consequence, the objective values are given:

$$v^\top \theta^A = -d^{1-1/q}, \quad \text{and} \quad v^\top \theta^B = -(d-c)^{1-1/q}.$$

- We use the fact that the function $f(x) = -x^{1-1/q}$ is convex for $q \in [1, \infty)$:

$$\begin{aligned} -\nabla f(d)c \leq f(d-c) - f(d) \Longrightarrow \quad & \frac{1-1/q}{d^{1/q}} c \leq -(d-c)^{1-1/q} - (-d^{1-1/q}) \\ \Longrightarrow \quad & \frac{1-1/q}{d^{1/q}} c \leq v^\top \theta^B - v^\top \theta^A. \end{aligned}$$

- Note that θ^B is the "optimal" estimator among all estimators with sign differences. Hence, the above bound gives relation (28).

Proof of Proposition 1: The Second Part

- We consider the performance on the mixture distribution $\mathbb{P} := (1/|\mathcal{V}|)\sum_{v\in\mathcal{V}} P_v$, then,

$$\max_v \mathbb{E}_{P_v}\left[f_v(\widehat{\theta}) - f_v\left(\theta^v\right)\right] \geq \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \mathbb{E}_{P_v}\left[f_v(\widehat{\theta}) - f_v\left(\theta^v\right)\right]$$
$$\geq \frac{1-1/q}{d^{1/q}}\delta R \sum_{j=1}^{d} \mathbb{P}\left(\operatorname{sign}\left(\widehat{\theta}_j\right) \neq -V_j\right).$$

- As a result, the minimax error is lower bounded as

$$\epsilon_k^*\left(\mathcal{F}_{G,2},\Theta\right) \geq \frac{1-1/q}{d^{1/q}}\delta R\left\{\inf_{\widehat{v}} \sum_{j=1}^{d} \mathbb{P}\left(\widehat{v}_j\left(Y^1,\ldots,Y^k\right) \neq V_j\right)\right\}, \tag{31}$$

where $\widehat{v}$ denotes any testing function mapping $\{Y^t\}_{t=1}^k$ to $\{-1,1\}^d$.

Proof of Proposition 1: The Second Part

- In the next, we lower bound the testing error by a total variation distance. To do so, we use Le Cam's inequality that for any set A and distributions P, Q, we have

$$P(A) + Q(A^c) \geq 1 - \|P - Q\|_{\mathsf{TV}} .$$

- We split the coordinates into the positive parts and the negative parts.

$$P_{+j} := \frac{1}{2^{d-1}} \sum_{v \in \mathcal{V}: v_j = 1} P_v \quad \text{and} \quad P_{-j} := \frac{1}{2^{d-1}} \sum_{v \in \mathcal{V}: v_j = -1} P_v .$$

- That is, P_{+j} and P_{-j} corresponds to conditional distributions over Y^t given the events $\{v_j = 1\}$ and $\{v_j = -1\}$.

Proof of Proposition 1: The Second Part

- Applying Le Cam's inequality yields

$$\begin{aligned}\mathbb{P}\left(\widehat{v}_j\left(Y^{1:k}\right) \neq V_j\right) &= \frac{1}{2}P_{+j}\left(\widehat{v}_j\left(Y^{1:k}\right) \neq 1\right) + \frac{1}{2}P_{-j}\left(\widehat{v}_j\left(Y^{1:k}\right) \neq -1\right) \\ &\geq \frac{1}{2}\left(1 - \|P_{+j} - P_{-j}\|_{\mathrm{TV}}\right).\end{aligned}$$

- Applying the Cauchy-Schwartz inequality , we have an upper bound for $\|P_{+j} - P_{-j}\|_{\mathrm{TV}}$:

$$\sum_{j=1}^{d} \|P_{+j} - P_{-j}\|_{\mathrm{TV}} \leq \sqrt{d}\left(\sum_{j=1}^{d} \|P_{+j} - P_{-j}\|_{\mathrm{TV}}^2\right)^{\frac{1}{2}}.$$

Proof of Proposition 1: The Second Part

- Then, we get a lower bound for the minimax error:

$$\epsilon_k^* \left(\mathcal{F}_{G,2}, \Theta\right) \geq \left(1 - \frac{1}{q}\right) \frac{d^{1-1/q} \delta R}{2} \left(1 - \frac{1}{\sqrt{d}} \left(\sum_{j=1}^{d} \|P_{+j} - P_{-j}\|_{\mathrm{TV}}^2\right)^{\frac{1}{2}}\right). \tag{32}$$

- In the following, we present a sharp bound on $\sum_{j=1}^{d} \|P_{+j} - P_{-j}\|_{\mathrm{TV}}^2$.

Proof of Proposition 1: The Third Part

- Defined the covariance matrix:

$$\Sigma := \sigma^2 \begin{bmatrix} \|\theta\|_2^2 & \langle \theta, \tau \rangle \\ \langle \theta, \tau \rangle & \|\tau\|_2^2 \end{bmatrix} = \sigma^2 [\theta\tau]^\top [\theta\tau], \tag{33}$$

with the corresponding shorthand Σ^t for the covariance computed for the t^{th} pair (θ^t, τ^t).

Lemma 3.

For each $j \in \{1, \cdots, d\}$, the total variation norm is bounded as

$$\|P_{+j} - P_{-j}\|_{\mathrm{TV}}^2 \leq \delta^2 \sum_{t=1}^{k} \mathbb{E}\left[\begin{bmatrix} \theta_j^t \\ \tau_j^t \end{bmatrix}^\top \left(\Sigma^t\right)^{-1} \begin{bmatrix} \theta_j^t \\ \tau_j^t \end{bmatrix} \right]. \tag{34}$$

Proof of Proposition 1: The Third Part

- Note the identity:

$$\sum_{j=1}^{d} \begin{bmatrix} \theta_j \\ \tau_j \end{bmatrix} \begin{bmatrix} \theta_j \\ \tau_j \end{bmatrix}^\top = \begin{bmatrix} \|\theta\|_2^2 & \langle \theta, \tau \rangle \\ \langle \theta, \tau \rangle & \|\tau\|_2^2 \end{bmatrix}. \tag{35}$$

- By Lemma 3, we have that

$$\begin{aligned} \sum_{j=1}^{d} \|P_{+j} - P_{-j}\|_{\mathrm{TV}}^2 &\le \delta^2 \sum_{t=1}^{k} \mathbb{E}\left[\sum_{j=1}^{d} \mathrm{tr}\left(\left(\Sigma^t\right)^{-1} \begin{bmatrix} \theta_j^t \\ \tau_j^t \end{bmatrix} \begin{bmatrix} \theta_j^t \\ \tau_j^t \end{bmatrix}^\top\right)\right] \\ &= \frac{\delta^2}{\sigma^2} \sum_{t=1}^{k} \mathbb{E}\left[\mathrm{tr}\left(\left(\Sigma^t\right)^{-1} \Sigma^t\right)\right] = 2\frac{k\delta^2}{\sigma^2}. \end{aligned}$$

Proof of Proposition 1: The Third Part

- By now, we find the nearly final lower bound:

$$\epsilon_k^* \left(\mathcal{F}_{G,2}, \Theta\right) \geq \left(1 - \frac{1}{q}\right) \frac{d^{1-1/q}\delta R}{2} \left(1 - \left(\frac{2k\delta^2}{d\sigma^2}\right)^{\frac{1}{2}}\right). \tag{36}$$

- We now restrict $(F, P) \in \mathcal{F}_{G,2}$ and we need to choose parameter σ^2 and δ^2 so that $\mathbb{E}\left[||X||_2^2\right] \leq G^2$ for $X \in \mathcal{N}(\delta v, \sigma^2 \mathbb{I}_d)$.
- We show that the following parameters are sufficient:

$$\sigma^2 = \frac{8G^2}{9d} \quad \text{and} \quad \delta^2 = \frac{G^2}{9} \min\left\{\frac{1}{k}, \frac{1}{d}\right\},$$

$$\Longrightarrow \quad 1 - \left(\frac{2k\delta^2}{d\sigma^2}\right)^{\frac{1}{2}} \geq 1 - \left(\frac{18}{72}\right)^{\frac{1}{2}} = \frac{1}{2} \quad \text{and} \quad \mathbb{E}\left[\|X\|_2^2\right] = \frac{8G^2}{9} + \frac{G^2 d}{9} \min\left\{\frac{1}{k}, \frac{1}{d}\right\} \leq G^2.$$

Proof of Proposition 1: The Third Part

- Plugging the chosen parameters into (36), we get the desired lower bound:

$$\epsilon_k^*\left(\mathcal{F}_{G,2},\Theta\right) \geq \frac{1}{12}\left(1-\frac{1}{q}\right) d^{1-1/q} RG \min\left\{\frac{1}{\sqrt{k}}, \frac{1}{\sqrt{d}}\right\}$$
$$= \frac{1}{12}\left(1-\frac{1}{q}\right)\frac{d^{1-1/q}RG}{\sqrt{k}} \min\{1, \sqrt{k/d}\}.$$

Outline

Conclusion

- We focus on the stochastic, convex and zero-order optimization problems.
- Zero-order algorithms use two-point evaluations to approximate directional derivative that the first-order methods utilize.
- For smooth optimization, we show that stochastic mirror descent based on zero-order gradient estimate is only $\mathcal{O}\left(\sqrt{d}\right)$ slower than the one based on first-order information.
- For non-smooth optimization, we show that by the smoothing technique, its convergence speed is at most $\mathcal{O}\left(\sqrt{\log d}\right)$ worse than the one of smooth optimization.
- Lower bounds indicate that the proposed methods are minimax optimal.

References I

A. Agarwal, O. Dekel, and L. Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In Proceedings of the 23rd Conference on Learning Theory, pages 28–40, 2010.

P. L. Bartlett, V. Dani, T. P. Hayes, S. M. Kakade, A. Rakhlin, and A. Tewari. High-probability regret bounds for bandit online linear optimization. In Proceedings of the 21st Conference on Learning Theory, pages 335–342, 2008.

A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. Operation Research Letters, 31(3):167–175, 2003.

J. C. Duchi, P. L. Bartlett, and M. J. Wainwright. Randomized smoothing for stochastic optimization. SIAM Journal on Optimization, 22(2):674–701, 2012a.

References II

J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. Finite sample convergence rates of zero-order stochastic optimization methods. In Advances in Neural Information Processing Systems 25, pages 1448–1456, 2012b.

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

A. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms, pages 385–394, 2005.

K. G. Jamieson, R. D. Nowak, and B. Recht. Query complexity of derivative-free optimization. In Advances in Neural Information Processing Systems 25, pages 2681–2689, 2012.

References III

C. Jin, L. T. Liu, R. Ge, and M. I. Jordan. On the local minima of the empirical risk. In Advances in Neural Information Processing Systems 31, pages 4901–4910, 2018.

A. Nemirovski, A. B. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. SIAM Journal on Optimization, 19(4):1574–1609, 2009.

Y. Nesterov. Lectures on convex optimization. Springer, 2018.

Y. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. Foundations of Computational Mathematics, 17(2):527–566, 2017.

T. Salimans, J. Ho, X. Chen, and I. Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. arXiv, 1703.03864, 2017.

S. Shalev-Shwartz. Online learning and online convex optimization. Foundations and Trends in Machine Learning, 4(2):107–194, 2012.

References IV

O. Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In The Proceedings of the 26th Annual Conference on Learning Theory, pages 3–24, 2013.

J. C. Spall. Introduction to stochastic search and optimization: estimation, simulation, and control. John Wiley & Sons, 2005.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Foundations and Trends in Machine Learning, 1(1-2):1–305, 2008.

D. Wierstra, T. Schaul, T. Glasmachers, Y. Sun, J. Peters, and J. Schmidhuber. Natural evolution strategies. Journal of Machine Learning Research, 15(1):949–980, 2014.