Mirror Descent Strikes Again: Optimal Stochastic Convex Optimization under Infinite Noise Variance

Nuri Mert Vural VURAL@CS.TORONTO.EDU
Department of Computer Science at the University of Toronto, and Vector Institute.

Lu Yu STAT.YU@MAIL.UTORONTO.CA
Department of Statistical Sciences at University of Toronto, and Vector Institute.

Krishnakumar Balasubramanian KBALA@UCDAVIS.EDU
Department of Statistics at University of California, Davis.

Stanislav Volgushev STANISLAV.VOLGUSHEV@UTORONTO.CA
Department of Statistical Sciences at University of Toronto.

Murat A. Erdogdu ERDOGDU@CS.TORONTO.EDU
Department of Computer Science and Department of Statistical Sciences at the University of Toronto, and Vector Institute.

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We study stochastic convex optimization under infinite noise variance. Specifically, when the stochastic gradient is unbiased and has uniformly bounded $(1+\kappa)$-th moment, for some $\kappa \in (0,1]$, we quantify the convergence rate of the Stochastic Mirror Descent algorithm with a particular class of uniformly convex mirror maps, in terms of the number of iterations, dimensionality and related geometric parameters of the optimization problem. Interestingly this algorithm does not require any explicit gradient clipping or normalization, which have been extensively used in several recent empirical and theoretical works. We complement our convergence results with information-theoretic lower bounds showing that no other algorithm using only stochastic first-order oracles can achieve improved rates. Our results have several interesting consequences for devising online/streaming stochastic approximation algorithms for problems arising in robust statistics and machine learning.

Keywords: Mirror descent algorithm, uniformly convex functions, heavy-tailed gradient noise, oracle complexity, information-theoretic lower bounds.

1. Introduction

For a compact convex set $\mathcal{S} \subset \mathbb{R}^d$, and a convex objective function $f : \mathcal{S} \to \mathbb{R}$, we consider the optimization problem

$$\underset{x \in \mathcal{S}}{\text{minimize}}\, f(x)\,, \tag{1.1}$$

in the stochastic first-order oracle model where one has access to noisy unbiased gradients at every iteration of an algorithm. This problem naturally emerges in many statistical learning tasks, thus there has been a substantial amount of research dedicated to understanding convergence guarantees as well as information-theoretic lower bounds in the classical setting where the noise has finite variance (Bubeck, 2014; Nesterov, 2018). However, recent studies have shown empirical and

theoretical evidence that stochastic gradients arising from modern learning problems may not have finite variance, in which case the optimal convergence guarantees and computational lower bounds for solving (1.1) are not well understood.

Indeed, heavy-tailed behavior is ubiquitous in statistical learning. Such behavior may either arise through the stochastic iterative training process (Hodgkinson and Mahoney, 2021; Camuto et al., 2021; Gürbüzbalaban et al., 2021) or due to the underlying statistical model (for example, this is observed in training attention models (Zhang et al., 2020) and in training convolutional networks (Simsekli et al., 2019b; Gürbüzbalaban and Hu, 2021)). In the regime where stochastic gradients have infinite variance, while the vanilla stochastic gradient descent (SGD) algorithm converges under strong convexity-type assumptions (Wang et al., 2021), more robust methods like gradient-clipped SGD (used, for example, by Zhang et al. (2020) for attention models) turn out to have optimal rates under strong convexity when the dimension is treated as a constant. However, it is not clear if gradient-clipped SGD or any other first-order method would exhibit similar optimality guarantees in the case of convex problems or when the dimension is not treated as a constant.

In this regard, it is highly desirable to obtain a rigorous understanding of the oracle complexity of stochastic convex optimization in the infinite noise variance setting. Such an understanding boils down to two fundamental questions:

> An *information-theoretic* question: What is the best achievable lower bound in convex optimization in the stochastic first-order oracle model under infinite noise variance?
>
> An *algorithmic complexity* question: Is there an optimal optimization algorithm that achieves this information-theoretic lower bound, under the same stochastic first-order oracle model?

We provide concrete answers to both of these questions, where the optimal algorithm is, yet again, stochastic mirror descent (SMD).

Mirror descent is a first-order method which generalizes the standard gradient descent to the non-Euclidean setting by relying on a mirror map that captures the underlying geometric structure of the problem (Nemirovski and Yudin, 1983). Although originally developed for deterministic frameworks, SMD is known to achieve the information-theoretic lower bound in the classical stochastic first-order oracle model where the noise has finite variance (Agarwal et al., 2012). This is remarkable as by simply choosing the appropriate mirror map, one can design algorithms that are optimal in their respective oracle models. This property of mirror descent has been exploited in many works for establishing the algorithm's optimality in classical settings (Nemirovski et al., 2009; Sridharan, 2012), and for demonstrating its universality in the online setting (Duchi et al., 2010; Srebro et al., 2011). In this work, we show that the stochastic mirror descent with an appropriate mirror map has an *inherent robustness* to heavy-tailed gradient noise, and achieves the information-theoretic lower bound for stochastic convex optimization under infinite noise variance. Towards that we make the following contributions.

- We establish the first non-asymptotic convergence of stochastic mirror descent algorithm in the heavy-tailed case where the gradient noise has infinite variance. We provide explicit rate estimates for a class of convex optimization problems in Theorem 6 and Corollary 7 for a variety of uniformly convex mirror maps.
- We establish lower bounds for the minimax error in Theorem 9, for constrained convex optimization in the first-order stochastic oracle model under infinite gradient noise variance.

- We show that for a careful choice of mirror map (which depends on the largest defined moment order in the gradient noise), the stochastic mirror descent algorithm achieves the minimax lower bound. This result proves optimality of the mirror descent algorithm in the heavy-tailed stochastic first-order oracle setting.
- Remarkably, the stochastic mirror descent algorithm achieves optimal bounds for the heavy-tailed setting without explicit gradient clipping, To the best of our knowledge, our results provide the first example of an optimal first-order optimization method for heavy-tailed setting without gradient clipping, or normalizing the magnitude of the stochastic gradients.

1.1. Related work

Earlier works on stochastic approximation with infinite variance largely focus on investigating the asymptotic behavior of stochastic approximation methods. Krasulina (1969) first establish the almost sure and L^p convergence for the one-dimensional stochastic approximation process without variance. Anantharam and Borkar (2012) demonstrate the stability and convergence properties of multivariate stochastic approximation algorithms with the heavy-tailed noise. Recently, the works of Simsekli et al. (2019a), Zhang et al. (2020), Chen et al. (2020), and Wang et al. (2021) investigate the behavior of SGD under infinite noise variance with various types of objectives. Simsekli et al. (2019a) considers non-convex optimization and analyze the SGD as a discretization of a stochastic differential equation driven by a Lévy process. Zhang et al. (2020) and Chen et al. (2020) study the convergence of SGD with gradient clipping, and establish the dimension-free optimal bound with strongly convex and non-convex objectives. Wang et al. (2021) provide the convergence rate of SGD with a strongly convex objective function under a state-dependent and heavy-tailed noise; see also Mirek (2011). High-probability bounds under certain moment assumptions (but not infinite variance) have also recently been established in Nazin et al. (2019); Cutkosky and Mehta (2021); Davis et al. (2021); Gorbunov et al. (2021); Tsai et al. (2021); Lou et al. (2022).

There exists a vast literature on mirror descent algorithm in a stochastic optimization setting with the stochastic gradient having finite variance (Nemirovski et al., 2009; Bubeck, 2014; Beck, 2017). Another line of work (Sridharan and Tewari, 2010; Srebro et al., 2011) establishes the (near) optimal regret rate of the mirror descent with the aid of uniformly convex mirror maps in a deterministic online setting. SMD was analyzed with almost surely bounded stochastic gradient, for composite problems, in Duchi et al. (2010). Mirror descent in the non-i.i.d. setting was considered in Duchi et al. (2012). We emphasize here that these works consider the standard finite variance noise setting, and thus the uniformly convex mirror map proposed in there is inadequate to deal with the infinite variance noise that we focus on in this work. Focusing on the finite-sum setup, D'Orazio et al. (2021) investigate the convergence of SMD in (relative) smooth optimization under the finite optimal objective difference assumption (Loizou et al., 2021), which allows for convergence without bounded gradient or variance assumptions and achieves exact convergence under interpolation.

More broadly, robust statistics is a classical topic with too large a literature to summarize completely. We refer the reader to Huber (2004) for an overview. The revival of robust statistics in modern mathematical statistics and learning theory communities arguably started with the work of Catoni (2012). Since then, there has been intense work on robust mean and covariance estimation (Minsker, 2015; Cardot et al., 2017; Minsker, 2018; Lugosi and Mendelson, 2019a,b; Hopkins, 2020), and robust empirical risk minimization (Hsu and Sabato, 2016; Diakonikolas et al., 2019;

Geoffrey et al., 2020; Lecué and Lerasle, 2020; Bartl and Mendelson, 2021). However, such results are mainly statistical in nature and are not directly applicable for stochastic approximation with heavy-tailed gradients.

Outline of the paper. The rest of the paper is organized as follows. In Section 2, we provide a definition of the stochastic first-order oracle model considered in this work, and a review of the stochastic mirror descent (SMD) algorithm focusing on uniform convexity and smoothness properties. In Section 3, we establish the convergence of SMD with a particular choice of mirror map, and illustrate the effect of this choice on heavy-tailed noisy gradient updates. We then provide information-theoretic lower bounds in Section 4, proving the optimality of SMD. We conclude in Section 5 with a discussion and future directions. All proofs are deferred to the Appendix.

2. Stochastic Mirror Descent: Preliminaries

Consider a setup in which a convex function f is minimized over a convex and compact set $\mathcal{S}$, using a stochastic optimization method M, which produces the iterate $x_t \in \mathcal{S}$ at iteration t. We assume that the sequence of iterates $\{x_t\}_{t\geq 0}$ is adapted to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$ and the method M has access to the following stochastic first-order oracle (SFO).

Assumption 1 (Stochastic First-order Oracle) *For all $t \geq 0$, given the current iterate x_t, the SFO produces random variables $f_{t+1} \in \mathbb{R}$ and $g_{t+1} \in \mathbb{R}^d$ that are $\mathcal{F}_{t+1}$-measurable, satisfying the following two properties.*

1. ***Unbiasedness:*** *For every $t \geq 0$, we have*

$$\mathbb{E}[f_{t+1}|\mathcal{F}_t] = f(x_t) \;\; \textit{and} \;\; \mathbb{E}[g_{t+1}|\mathcal{F}_t] \in \partial f(x_t).$$

2. ***Finite $(1+\kappa)$-th moment:*** *For some $\kappa \in (0,1]$, $q \in [1,\infty]$, and $\sigma > 0$, we have*

$$\sup_{t\geq 0} \mathbb{E}[\|g_{t+1}\|_q^{1+\kappa}|\mathcal{F}_t] \leq \sigma^{1+\kappa}.$$

Here, $\partial f(x) := \{v \in \mathbb{R}^d \mid f(y) \geq f(x) + \langle v, y - x\rangle \text{ for all } y \in \mathbb{R}^d\}$ denotes the sub-differential set of f at the point x and $\|\cdot\|_q$ denotes the q-norm. We note that the bounded $(1+\kappa)$-th moment assumption with $\kappa = 1$ corresponds to the classical setting of having stochastic gradient with finite second-moments (Agarwal et al., 2012). In this paper, we are mainly interested in the case where $\kappa < 1$, when the variance of the stochastic gradient is undefined. Perhaps, the most popular stochastic optimization method M operating under SFO is the (projected) stochastic gradient descent (SGD) in the Euclidean setting, as given by

$$y_{t+1} = x_t - \eta g_{t+1} \quad \text{and} \quad x_{t+1} = \arg\min_{x\in\mathcal{S}} \|x - y_{t+1}\|_2^2. \tag{SGD}$$

Remark 1 *Any function f that is compatible with an SFO satisfying Assumption 1 must be Lipschitz continuous with respect to q^*-norm with Lipschitz constant $L \leq \sigma$. To see this, we note that a convex function is L-Lipschitz on $\mathcal{S}$ in $\|\cdot\|_{q^\star}$ if and only if*

$$\sup_{x\in\mathcal{S}} \max_{v\in\partial f(x)} \|v\|_q \leq L,$$

where $q, q^* \in [1, \infty]$ *satisfy* $\frac{1}{q} + \frac{1}{q^*} = 1$. *Moreover, for* $v_t \in \partial f(x_t)$ *elementary calculations imply*

$$\|v_t\|_q = \|\mathbb{E}[g_{t+1}|\mathcal{F}_t]\|_q \leq \mathbb{E}[\|g_{t+1}\|_q|\mathcal{F}_t] \leq \left(\mathbb{E}[\|g_{t+1}\|_q^{1+\kappa}|\mathcal{F}_t]\right)^{\frac{1}{1+\kappa}} \leq \sigma, \quad \forall t \geq 1\,.$$

Mirror descent, first introduced by Nemirovski and Yudin (1983), refers to a family of algorithms for first-order optimization (Beck and Teboulle, 2003; Cesa-Bianchi and Lugosi, 2006; Bubeck, 2014), which was originally developed to exploit the geometry of the problem. Compared to the classical gradient descent for which the iterates are updated along the direction of the negative gradient, in mirror descent, the updates are performed in the "mirrored" dual space determined by a transformation called the *mirror map*. The family of mirror descent algorithms extends naturally to the stochastic first-order oracle setup, which is the main focus of this paper.

For a function $\Psi : \mathbb{R}^d \to \mathbb{R}$ that is strictly-convex, continuously differentiable with a norm coercive gradient (i.e. $\lim_{\|x\|_2 \to \infty} \|\nabla \Psi(x)\|_2 = \infty$), we denote its Fenchel conjugate and Bregman divergence respectively as

$$\Psi^\star(y) \coloneqq \sup_{x \in \mathbb{R}^d} \Big\{ \langle y, x \rangle - \Psi(x) \Big\} \text{ and } D_\Psi(x, y) \coloneqq \Psi(x) - \Psi(y) - \langle \nabla \Psi(y), y - x \rangle.$$

The stochastic mirror descent (SMD) updates are defined as

$$y_{t+1} = \nabla \Psi^\star\big(\nabla \Psi(x_t) - \eta g_{t+1}\big) \quad \text{and} \quad x_{t+1} = \arg\min_{x \in \mathcal{S}} D_\Psi(x, y_{t+1}). \tag{SMD}$$

The conditions on Ψ imply that the (SMD) update is well-defined, and $\nabla \Psi$ is an invertible map that satisfies $(\nabla \Psi)^{-1} = \nabla \Psi^\star$ (Cesa-Bianchi and Lugosi, 2006). The map $\nabla \Psi$ is also referred to as the *mirror map* and makes (SMD) adapt to the geometric properties of the optimization problem.

The mirror map. In the (SMD) update, the descent is performed in the dual space which is the mirror image of the primal space under the mirror map. Different choices of the mirror maps turn out to be suitable for different optimization problems, and the *right* mirror map corresponds to understanding the geometry of the problem, the objective function we minimize as well as the noise model. Notable examples include:

- *Stochastic Gradient Descent:* For the function $\Psi(x) = \frac{1}{2}\|x\|_2^2$, the mirror map $\nabla \Psi$ reduces to the identity map, and its Bregman divergence reduces to $D_\Psi(x, y) = \frac{1}{2}\|x - y\|_2^2$. Therefore, the update rule (SMD) reduces to the well-known (SGD) update.
- *p-norms Algorithm:* For $p \in (1, 2]$ and the function $\Psi(x) = \frac{1}{2}\|x\|_p^2$, the (SMD) update reduces to the so-called p-norms algorithm (Gentile and Littlestone, 1999), which is optimal for stochastic convex optimization under finite noise variance (Agarwal et al., 2009).
- *Exponentiated Gradient Descent:* For the function[1] $\Psi(x) = \sum_j x_j \log x_j$, the Bregman divergence becomes the unnormalized relative entropy, i.e., $D_\Psi(x, y) = \sum_j x_j \log \frac{x_j}{y_j} - \sum_j x_j + \sum_j y_j$, and the update rule (SMD) corresponds to the exponentiated gradient descent, which is widely used in the prediction with expert advice setting (Cesa-Bianchi and Lugosi, 2006).

In this work, we observe that the choice of mirror map is beneficial when dealing with the particular noise model of stochastic gradients. In what follows, we will use uniformly convex mirror maps in the infinite noise variance setting.

1. The domain of mirror map can also be defined over a smaller set containing the feasible set, see e.g. Bubeck (2014).

Definition 2 (Uniform convexity) *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 with respect to p-norm if for any $x, y \in \mathbb{R}^d$, we have*

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

Uniformly convex functions with $r = 2$ are known as *strongly convex* in p-norm, and the case $p = 2$ reduces to the classical notion of strong convexity in the Euclidean setting.

Definition 3 (Uniform smoothness) *A function $\psi : \mathbb{R}^d \to \mathbb{R}$ is (K, r)-uniformly smooth with respect to p-norm if it is differentiable and if there exist a constant $K > 0$ and an exponent $r \in (1, 2]$ such that for any $x, y \in \mathbb{R}^d$, we have*

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

Similarly, uniformly smooth functions with $r = 2$ are known as *strongly smooth* and the case $p = 2$ reduces to the classical notion of first-order smoothness in the Euclidean setting.

Uniform convexity and uniform smoothness are dual properties by Fenchel conjugacy (Zalinescu, 1983; Azé and Penot, 1995), a property that is better known for their strong versions. Given the norm $\|\cdot\|_p$ with $p \in [1, \infty]$, denote its associated dual norm by $\|\cdot\|_{p^\star}$, where $1/p + 1/p^\star = 1$. We recall the statement below both for completeness and to obtain quantitative statements later in Proposition 5 for a special class of uniformly convex functions.

Proposition 4 *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 with respect to p-norm if and only if $\psi^\star$ is $\left(K^{-\frac{1}{r-1}}, \frac{r}{r-1}\right)$-uniformly smooth with respect to $p^\star$-norm.*

Next we quantify the uniform convexity and smoothness parameters of functions of the form $\frac{1}{r}\|\cdot\|_p^r$, for $p, r \in (1, \infty)$. Gradients of these functions with an appropriate choice of r and p will be used as mirror maps in the (SMD) update, which will ultimately achieve the minimax lower bound in the heavy-tailed stochastic oracle setup. It is known in the optimization literature that $\frac{1}{2}\|x\|_p^2$ is $(p-1)$-strongly smooth for $p \in [2, \infty)$ with respect to p-norm[2], see e.g. (Juditsky and Nemirovski, 2008, Ex. 3.2). The next proposition extends this result to p-norm with an arbitrary exponent.

Proposition 5 *For $\kappa \in (0, 1]$, $p \in [1 + \kappa, \infty)$ and $p^\star$ satisfying $\frac{1}{p} + \frac{1}{p^\star} = 1$, we define*

$$K_p := 10 \max\left\{1, (p-1)^{\frac{1+\kappa}{2}}\right\}, \quad \varphi(x) := \frac{1}{1+\kappa}\|x\|_p^{1+\kappa} \text{ and } \varphi^\star(y) := \frac{\kappa}{1+\kappa}\|y\|_{p^\star}^{\frac{1+\kappa}{\kappa}} . \tag{2.1}$$

Then, the following statements hold for the Fenchel conjugate functions φ and $\varphi^\star$.

1. *φ is $(K_p, 1 + \kappa)$-uniformly smooth with respect to p-norm.*
2. *$\varphi^\star$ is $\left(K_p^{-\frac{1}{\kappa}}, \frac{1+\kappa}{\kappa}\right)$-uniformly convex with respect to $p^\star$-norm.*

We emphasize that both φ and $\varphi^\star$ in (2.1) depend on the choice of p and consequently $p^\star$. We also note that when $\kappa = 1$, Proposition 5 recovers the strong convexity/smoothness parameter of $\frac{1}{2}\|x\|_p^2$ up to a constant factor (Juditsky and Nemirovski, 2008; Kakade et al., 2009).

2. Equivalently, $\frac{1}{2}\|x\|_{p^\star}^2$ is $\frac{1}{p-1}$-strongly convex for $p^\star \in (1, 2]$ with respect to. $p^\star$-norm (Kakade et al., 2009).

3. Convergence of SMD with a Uniformly Convex Potential

We now present our main convergence result for the SMD algorithm with a uniformly convex potential, under an SFO that satisfies Assumption 1.

Theorem 6 *Let Assumption 1 hold for some $q \in [1, \infty]$ and define $q^\star$ through $\frac{1}{q} + \frac{1}{q^\star} = 1$. For a function Ψ which is $(1, \frac{1+\kappa}{\kappa})$-uniformly convex with respect to $q^\star$-norm, the (SMD) algorithm with the corresponding mirror map $\nabla\Psi$, initialized at $x_0 = \arg\min_{x\in\mathcal{S}} \Psi(x)$ and run with step size*

$$\eta = \frac{R_0^{1/\kappa}}{\sigma} T^{-\frac{1}{1+\kappa}}, \quad \textit{where} \quad R_0^{\frac{1+\kappa}{\kappa}} := \frac{1+\kappa}{\kappa} \sup_{x\in\mathcal{S}} \left\{\Psi(x) - \Psi(x_0)\right\}$$

satisfies

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=0}^{T-1} x_t\right) - \min_{x\in\mathcal{S}} f(x)\right] \le R_0\, \sigma\, T^{-\frac{\kappa}{1+\kappa}}. \tag{3.1}$$

To our knowledge, the above result is the first convergence result for stochastic mirror descent under a noise model that allows infinite noise variance. In contrast to other gradient-based methods in the literature dealing with heavy-tailed noise (Zhang et al., 2020; Gorbunov et al., 2020), stochastic mirror descent does not require (explicit) gradient clipping or gradient normalization to guarantee convergence. In Section 3.1, we will present an instance to illustrate the intuition behind this result.

The initialization error R_0 in the above bound (3.1) introduces the dimension dependency to the convergence rate. To make this more explicit, in the next corollary, we fix the domain as $\mathcal{S} = \mathbb{B}_\infty(R)$, where $\mathbb{B}_\infty(R)$ is the $\|\cdot\|_\infty$-ball with radius R, centered at the origin, and use a specific uniformly convex function as the mirror map.

Corollary 7 *Let $U_p(x) := K_p^{\frac{1}{\kappa}} \varphi^*(x)$, where K_p and φ^* are defined in (2.1), and $\mathcal{S} = \mathbb{B}_\infty(R)$. Under the conditions of Theorem 6, the following statements hold.*

i) For $q \in [1, 1+\kappa]$, (SMD) with $\Psi := U_p$ for $p = 1+\kappa$ satisfies

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=0}^{T-1} x_t\right) - \min_{x\in\mathcal{S}} f(x)\right] \le 10\, R\, \sigma \left(\frac{d}{T}\right)^{\frac{\kappa}{1+\kappa}}.$$

ii) For $q \in (1+\kappa, \infty)$, (SMD) with $\Psi := U_p$ for $p = q$ satisfies

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=0}^{T-1} x_t\right) - \min_{x\in\mathcal{S}} f(x)\right] \le 10 \max\left\{1, \sqrt{q-1}\right\} R\, \sigma\, \frac{d^{1-\frac{1}{q}}}{T^{\frac{\kappa}{1+\kappa}}}.$$

iii) For $q \in (\log d, \infty]$, (SMD) with $\Psi := U_p$ for $p = 1 + \log d$ satisfies

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=0}^{T-1} x_t\right) - \min_{x\in\mathcal{S}} f(x)\right] \le 10\, R\, \sigma \sqrt{\log d}\, \frac{d^{1-\frac{1}{q}}}{T^{\frac{\kappa}{1+\kappa}}}.$$

Remark 8 *Note that part (ii) of Corollary 7 also covers the case $q \in (\log d, \infty)$; however, the result in part (iii) provides a better convergence rate in terms of dimension dependence. Part (iii) includes the boundary case $q = \infty, q^\star = 1$ at the expense of additional $\sqrt{\log d}$ factor in the rate.*

Corollary 7 provides explicit rates for SMD in the infinite noise variance case ($\kappa < 1$), with an explicit mirror map. It also recovers the known optimal rates in the finite variance case ($\kappa = 1$) (Nemirovski and Yudin, 1983; Agarwal et al., 2009, 2012), in which case it reduces to the well-known p-norms algorithm (Gentile and Littlestone, 1999).

3.1. Robustness of SMD under heavy-tailed noise

We consider a particular instance of (SMD) to provide additional intuition behind the result. Let $q = 2$, $\Psi = U_2$, where U is defined in Corollary 7. Denote the current iterate by x. Based on the noisy gradient g returned by the SFO, the (SMD) update (without projection) is given by

$$x_{\text{SMD}} = \nabla\Psi^\star\big(\tilde{x}\big) \;\text{ for }\; \tilde{x} = \nabla\Psi(x) - \eta\, g$$
$$= \frac{x\|x\|_2^{\frac{1}{\kappa}-1} - \frac{\eta}{10^{1/\kappa}} g}{\Big\|x\|x\|_2^{\frac{1}{\kappa}-1} - \frac{\eta}{10^{1/\kappa}} g\Big\|_2^{1-\kappa}}. \tag{3.2}$$

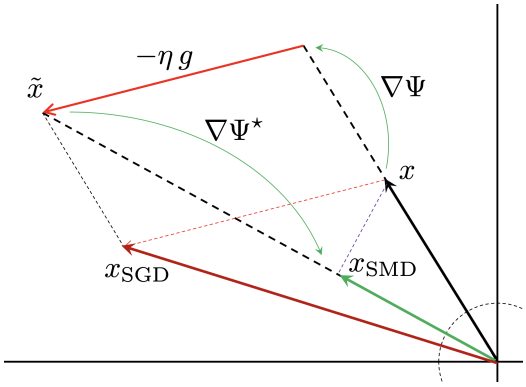

Figure 1: Illustration of SMD and SGD updates under heavy-tailed noise.

For $q = 2$, the primal and the dual spaces are both $(L^2(\mathbb{R}^d), \|\cdot\|_2)$, and Figure 1 shows the updates in the same space for simplicity, and to illustrate the robustness of (SMD) in comparison to (SGD). In the case where g is large due to heavy-tailed noise, SGD update would be significantly impacted, whereas SMD first amplifies the magnitude of the iterate x, then performs the noisy gradient update in the "dual space" to get $\tilde{x}$, and finally contracts the resulting value to x_{SMD}. This mechanism of SMD is illustrated in (3.2). The descent is performed in the dual space, and the inverse mirror map *shrinks* vectors that are larger in magnitude more when mapping it back to the primal space. This provides an inherent regularization, preventing instabilities due to heavy-tailed noise.

We formally prove in the next section that SMD remains optimal for the case $\kappa < 1$ (i.e., even for the case when the stochastic gradients have infinite noise variance).

4. Information-theoretic Lower Bounds

In this section, we prove that the rates obtained in Theorem 6 and Corollary 7 are minimax optimal in an information theoretical sense, up to constants and log(dimension) factors. To prove this result, we provide lower bounds on the convergence of any algorithm with access to an SFO satisfying Assumption 1, by extending ideas of Nemirovski and Yudin (1983) and Agarwal et al. (2009, 2012) to the infinite noise variance setting. We now give a formal definition of minimax complexity of optimization algorithms in the heavy-tailed setting.

For a convex and compact set $\mathcal{S}$, consider the function class $\mathcal{H}_{cvx}$ consisting of all convex functions $f : \mathcal{S} \to \mathbb{R}$, that are L-Lipschitz with respect to q^*-norm. That is,

$$\mathcal{H}_{cvx}(\mathcal{S}, L, q^*) := \{f : \mathcal{S} \to \mathbb{R} : f \text{ is convex and } L\text{-Lipschitz with respect to } q^* \text{ norm}\}\,.$$

Recall from Remark 1 that any oracle satisfying Assumption 1 must operate on an objective function $f \in \mathcal{H}_{cvx}(\mathcal{S}, L, q^*)$ with $L \leq \sigma$. Thus in our minimax bounds in the sequel, we will only consider such convex and Lipschitz functions.

Recall that a SFO, which we denote as ϕ, takes the current iterate x_t and returns the noisy unbiased pair (f_t, g_t) satisfying Assumption 1. We denote by $\Phi(\kappa, q, \sigma)$, the class of all such SFOs with parameters (κ, q, σ) appearing in Assumption 1. Given an oracle $\phi \in \Phi(\kappa, q, \sigma)$, let $\mathcal{M}_T$ represent the class of all optimization methods that query the oracle ϕ exactly T times and return $\bar{x}_T \in \mathcal{S}$ as an estimate of the optimum $\arg\min_{x\in\mathcal{S}} f(x)$ based on those queries. For any method $M_T \in \mathcal{M}_T$, consider the error in optimizing f after T iterations,

$$\epsilon(M_T, f, \mathcal{S}, \phi) := f(\bar{x}_T) - \min_{x\in\mathcal{S}} f(x)\,.$$

Here, $\bar{x}_T$ should be seen as the output of the method M_T after T iterations, not necessarily the t-th iterate of the optimization method. For example, Theorem 6 and Corollary 7 provide upper bounds on the expected value of $\epsilon(M_T, f, \mathcal{S}, \phi)$ for optimization method M_T corresponding to specific instances of (SMD), and $\bar{x}_T$ corresponds to the average of (SMD) iterates.

To provide lower bounds on the best possible performance, uniformly over all functions $f \in \mathcal{H}_{cvx}$, of any optimization method $M_T \in \mathcal{M}_T$, we define the minimax error as

$$\epsilon^*(\mathcal{H}_{cvx}, \mathcal{S}, \phi) := \inf_{M_T\in\mathcal{M}_T} \sup_{f\in\mathcal{H}_{cvx}} \mathbb{E}_\phi[\epsilon(M_T, f, \mathcal{S}, \phi)]\,.$$

The following theorem characterizes the minimax oracle complexity of optimization over the function class $\mathcal{H}_{cvx}$, where the constraint set $\mathcal{S}$ is convex and contains $\mathbb{B}_\infty(R)$, the $\|\cdot\|_\infty$-ball of radius R centered at the origin.

Theorem 9 *Assume that $\mathcal{S} \supseteq \mathbb{B}_\infty(R)$. We have the following minimax lower bounds*

1. *For all $q \in [1, 1+\kappa]$, we have*

$$\sup_{\phi\in\Phi(\kappa,q,\sigma)} \epsilon_T^*(\mathcal{H}_{cvx}, \mathcal{S}, \phi) \geq C_1 RL\left(\frac{d}{T}\right)^{\frac{\kappa}{1+\kappa}}.$$

2. *For all $q \in (1+\kappa, \infty]$, we have*

$$\sup_{\phi\in\Phi(\kappa,q,\sigma)} \epsilon_T^*(\mathcal{H}_{cvx}, \mathcal{S}, \phi) \geq C_2 RL\frac{d^{1-\frac{1}{q}}}{T^{\frac{\kappa}{1+\kappa}}}.$$

Here, $C_1 > 0$ and $C_2 > 0$ are universal constants.

Remark 10 *The above minimax lower bounds match the rate estimates in Corollary 7 up to constants for $q \in [1, \infty)$, and an additional $\sqrt{\log d}$ factor for the case of $q = \infty$, proving the optimality of the stochastic mirror descent in the infinite noise variance setting.*

In contrast to existing lower bounds on the oracle complexity in stochastic convex optimization (Agarwal et al., 2009, 2012; Ramdas and Singh, 2013; Iouditski and Nesterov, 2014), Theorem 9 covers a wider range of stochastic first-order oracles. It extends existing minimax lower bounds to the heavy-tailed noise setting with $\kappa < 1$. Our results recover the information-theoretic lower bound in the classical finite variance setting (Agarwal et al., 2009, 2012, Theorem 1). For the fixed dimension, those bounds can also be linked to limits of results in (Ramdas and Singh, 2013; Iouditski and Nesterov, 2014) who establish the optimal rate $\Omega\Big(T^{-\frac{\rho}{2(\rho-1)}}\Big)$ for ρ-uniformly convex functions under finite noise variance. Letting $\rho \to \infty$, which corresponds to convex functions, yields the convergence rate $\Omega(T^{-1/2})$, which is recovered by our results with $\kappa = 1$. Moreover, our lower bounds provide sharp dimension dependence. This extends the findings in (Raginsky and Rakhlin, 2009, Theorem 3) and (Nemirovski and Yudin, 1983, Section 5.3.1) who proved a rate of the form $\Omega\Big(T^{-\frac{\kappa}{1+\kappa}}\Big)$ for the first-order stochastic convex optimization under the heavy-tailed noise setting while treating the dimension d as a constant.

The proof strategy of the above oracle complexity lower bound involves a standard reduction from stochastic optimization to a hypothesis testing problem. Similar arguments appeared in earlier works (Agarwal et al., 2009, 2012; Raginsky and Rakhlin, 2009; Zhang et al., 2020). However, those works either considered finite-variance noise or treated the dimension as fixed. Covering heavy-tailed stochastic gradient noise and providing explicit dimension dependence requires a more delicate construction of the function class and the first-order oracles, which may be of independent interest. We refer to Appendix D for the details.

5. Discussion

In this work, we showed that stochastic mirror descent, with a particular choice of mirror map, achieves the information-theoretically optimal rates for stochastic convex optimization when the stochastic gradient has finite $(1+\kappa)$-th moment, for $\kappa \in (0, 1]$. To do so, on the *algorithmic side* we showed that our choice of mirror-map has an inherent regularization property to prevent instabilities that might occur due to heavy-tailed noise in the stochastic gradient. On the *information-theoretic* side, we provided minimax lower bounds that match the upper bound achieved by the stochastic mirror descent algorithm that we analyze. Our work opens up several interesting directions:

1. The current choice of our step-size parameter requires knowledge of the noise level and κ (this is true for all optimization methods that deal with the heavy-tailed noise setting). It is extremely interesting and practically relevant to develop adaptive procedures that achieve optimal rates without knowledge of the problem parameters.
2. While our current results are in expectation, establishing results that hold with high-probability in the infinite-noise variance setting would provide an interesting complement to our results.
3. Developing distributional convergence results for the iterates of (SMD), along with related statistical inferential procedures is important for uncertainty quantification.

4. Finally, examining the performance of (SMD) in the non-convex setting with infinite-noise variance, is interesting both theoretically and practically.

References

Alekh Agarwal, Martin J Wainwright, Peter Bartlett, and Pradeep Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. *Advances in Neural Information Processing Systems*, 22:1–9, 2009.

Alekh Agarwal, Peter L Bartlett, Pradeep Ravikumar, and Martin J Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

Venkat Anantharam and Vivek S. Borkar. Stochastic approximation with long range dependent and heavy tailed noise. *Queueing Systems*, 71:221–242, 2012.

Dominique Azé and Jean-Paul Penot. Uniformly convex and uniformly smooth convex functions. *Annales de la Faculté des Sciences de Toulouse*, 4:705–730, 1995.

Daniel Bartl and Shahar Mendelson. On monte-carlo methods in convex stochastic optimization. *arXiv preprint arXiv:2101.07794*, 2021.

Amir Beck. *First-order methods in optimization*. Society for Industrial and Applied Mathematics (SIAM), 2017.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, 31:167–175, 2003.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*, 2014.

Alexander Camuto, Xiaoyu Wang, Lingjiong Zhu, Chris Holmes, Mert Gürbüzbalaban, and Umut Şimşekli. Asymmetric heavy tails and implicit bias in Gaussian noise injections. In *International Conference on Machine Learning*, pages 1249–1260. PMLR, 2021.

Hervé Cardot, Peggy Cénac, and Antoine Godichon-Baggioni. Online estimation of the geometric median in Hilbert spaces: Nonasymptotic confidence balls. *The Annals of Statistics*, 2017.

Olivier Catoni. Challenging the empirical mean and empirical variance: A deviation study. In *Annales de l'IHP Probabilités et statistiques*, volume 48-4, pages 1148–1185, 2012.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Xiangyi Chen, Steven Z Wu, and Mingyi Hong. Understanding gradient clipping in private sgd: A geometric perspective. *Advances in Neural Information Processing Systems*, 33:13773–13782, 2020.

Ashok Cutkosky and Harsh Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. *Advances in Neural Information Processing Systems*, 34, 2021.

Damek Davis, Dmitriy Drusvyatskiy, Lin Xiao, and Junyu Zhang. From low probability to high confidence in stochastic convex optimization. *Journal of machine learning research*, 22(49), 2021.

Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high-dimensions without the computational intractability. *SIAM Journal on Computing*, 2019.

Ryan D'Orazio, Nicolas Loizou, Issam Laradji, and Ioannis Mitliagkas. Stochastic mirror descent: Convergence analysis and adaptive variants via the mirror stochastic Polyak stepsize. *arXiv preprint arXiv:2110.15412*, 2021.

John C Duchi, Shai Shalev-Shwartz, Yoram Singer, and Ambuj Tewari. Composite objective mirror descent. In *COLT*, volume 10, pages 14–26, 2010.

John C Duchi, Alekh Agarwal, Mikael Johansson, and Michael I Jordan. Ergodic mirror descent. *SIAM Journal on Optimization*, 22(4):1549–1578, 2012.

Claudio Gentile and Nick Littlestone. The robustness of the p-norm algorithms. In *Proceedings of Annual Conference on Learning Theory*, 1999.

Chinot Geoffrey, Lecué Guillaume, and Lerasle Matthieu. Robust high dimensional learning for lipschitz and convex losses. *Journal of Machine Learning Research*, 21(233):1–47, 2020.

Eduard Gorbunov, Marina Danilova, and Alexander Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping, 2020.

Eduard Gorbunov, Marina Danilova, Innokentiy Shibaev, Pavel Dvurechensky, and Alexander Gasnikov. Near-optimal high probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise. *arXiv preprint arXiv:2106.05958*, 2021.

Mert Gürbüzbalaban and Yuanhan Hu. Fractional moment-preserving initialization schemes for training deep neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 2233–2241. PMLR, 2021.

Mert Gürbüzbalaban, Umut Şimşekli, and Lingjiong Zhu. The heavy-tail phenomenon in sgd. *ArXiv*, abs/2006.04740, 2021.

Liam Hodgkinson and Michael W. Mahoney. Multiplicative noise and heavy tails in stochastic optimization. *ArXiv*, abs/2006.06293, 2021.

Samuel B Hopkins. Mean estimation with sub-gaussian rates in polynomial time. *Annals of Statistics*, 48(2):1193–1213, 2020.

Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research*, 17(1):543–582, 2016.

Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

Anatoli Iouditski and Yuri Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions. *arXiv preprint arXiv:1401.1792*, 2014.

Anatoli Juditsky and Arkadii S Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813*, 2008.

Sham Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. *Unpublished Manuscript, http://ttic.uchicago.edu/shai/papers/KakadeShalevTewari09.pdf*, 2009.

Tatiana P. Krasulina. On stochastic approximation processes with infinite variance. *Theory of Probability and Its Applications*, 14:522–526, 1969.

Guillaume Lecué and Matthieu Lerasle. Robust machine learning by median-of-means: theory and practice. *Annals of Statistics*, 48(2):906–931, 2020.

Nicolas Loizou, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien. Stochastic Polyak step-size for sgd: An adaptive learning rate for fast convergence. In *International Conference on Artificial Intelligence and Statistics*, pages 1306–1314. PMLR, 2021.

Zhipeng Lou, Wanrong Zhu, and Wei Biao Wu. Beyond sub-gaussian noises: Sharp concentration analysis for stochastic gradient descent. *Journal of Machine Learning Research*, 23:1–22, 2022.

Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019a.

Gábor Lugosi and Shahar Mendelson. Sub-gaussian estimators of the mean of a random vector. *Annals of Statistics*, 47(2):783–794, 2019b.

Pascal Massart. *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer, 2007.

Stanislav Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4): 2308–2335, 2015.

Stanislav Minsker. Sub-gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Annals of Statistics*, 46(6A):2871–2903, 2018.

Mariusz Mirek. Heavy tail phenomenon and convergence to stable laws for iterated lipschitz maps. *Probability Theory and Related Fields*, 151(3):705–734, 2011.

Alexander V Nazin, Arkadi S Nemirovsky, Alexandre B Tsybakov, and Anatoli B Juditsky. Algorithms of robust stochastic optimization based on mirror descent method. *Automation and Remote Control*, 80(9):1607–1627, 2019.

Arkadi Nemirovski and David Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience, 1983.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

David Pollard, Erik Torgersen, and Grace L Yang. *Festschrift for Lucien Le Cam: Research papers in probability and statistics*. Springer Science & Business Media, 2012.

Maxim Raginsky and Alexander Rakhlin. Information complexity of black-box convex optimization: A new look via feedback information theory. In *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 803–510. IEEE, 2009.

Aaditya Ramdas and Aarti Singh. Optimal rates for stochastic convex optimization under Tsybakov noise condition. In *International Conference on Machine Learning*, pages 365–373. PMLR, 2013.

R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, N. J., 1970.

Jonathan Scarlett and Volkan Cevher. An introductory guide to Fano's inequality with applications in statistical estimation. *arXiv preprint arXiv:1901.00555*, 2019.

Umut Simsekli, Mert Gürbüzbalaban, Thanh Huy Nguyen, Gaël Richard, and Levent Sagun. On the heavy-tailed theory of stochastic gradient descent for deep neural networks. *ArXiv*, abs/1912.00018, 2019a.

Umut Simsekli, Levent Sagun, and Mert Gurbuzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pages 5827–5837. PMLR, 2019b.

Nati Srebro, Karthik Sridharan, and Ambuj Tewari. On the universality of online mirror descent. *Advances in neural information processing systems*, 24, 2011.

Karthik Sridharan. Learning from an optimization viewpoint. *arXiv preprint arXiv:1204.4145*, 2012.

Karthik Sridharan and Ambuj Tewari. Convex games in Banach spaces. In *COLT*, pages 1–13, 2010.

Che-Ping Tsai, Adarsh Prasad, Sivaraman Balakrishnan, and Pradeep Ravikumar. Heavy-tailed streaming statistical estimation. *arXiv preprint arXiv:2108.11483*, 2021.

Hongjian Wang, Mert Gurbuzbalaban, Lingjiong Zhu, Umut Simsekli, and Murat A Erdogdu. Convergence rates of stochastic gradient descent under infinite noise variance. *Advances in Neural Information Processing Systems*, 34, 2021.

Constantin Zalinescu. On uniformly convex functions. *Journal of Mathematical Analysis and Applications*, 95:344–374, 1983.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020.

Appendix A. Proofs for Section 2

A.1. Proof of Proposition 4

Proof [Proof of Proposition 4] ($\Rightarrow$) Since $\psi : \mathbb{R}^d \to \mathbb{R}$ is uniformly convex and differentiable, by Proposition 20, $\psi^\star$ is differentiable. Moreover, since ψ is continuous and convex, by (Rockafellar, 1970, Corollary 23.5.1.), $(\nabla\psi^\star)^{-1} = (\nabla\psi)$. Let $y_1, y_2 \in \mathbb{R}^d$ be two arbitrary vectors, and let $\nabla\psi^\star(y_1) = x_1$. Then, we have $\nabla\psi(x_1) = y_1$ and by (Rockafellar, 1970, Theorem 23.5)

$$\psi(x_1) + \psi^\star(y_1) = \langle x_1, y_1\rangle. \tag{A.1}$$

We can write that

$$\begin{aligned}
\psi^\star(y_2) &= \sup_{x\in\mathbb{R}^d}\{\langle y_2, x\rangle - \psi(x)\} \\
&\le \sup_{x\in\mathbb{R}^d}\left\{\langle y_2, x\rangle - \left(\psi(x_1) + \langle\nabla\psi(x_1), x - x_1\rangle + \frac{K}{r}\|x - x_1\|_p^r\right)\right\} \text{ (by the uniform convexity of } \psi) \\
&= \sup_{x\in\mathbb{R}^d}\left\{\langle y_2 - y_1, x - x_1\rangle - \frac{K}{r}\|x - x_1\|_p^r\right\} - \psi(x_1) + \langle y_2, x_1\rangle \text{ (since } \nabla\psi(x_1) = y_1) \\
&= \sup_{x\in\mathbb{R}^d}\left\{\langle y_2 - y_1, x - x_1\rangle - \frac{K}{r}\|x - x_1\|_p^r\right\} + \psi^\star(y_1) + \langle\nabla\psi^\star(y_1), y_2 - y_1\rangle \qquad \text{(A.2)} \\
&= \sup_{x\in\mathbb{R}^d}\left\{\langle y_2 - y_1, K^{-\frac{1}{r-1}}x\rangle - \frac{K}{r}\|K^{-\frac{1}{r-1}}x\|_p^r\right\} + \psi^\star(y_1) + \langle\nabla\psi^\star(y_1), y_2 - y_1\rangle \\
&= \psi^\star(y_1) + \langle\nabla\psi^\star(y_1), y_2 - y_1\rangle + K^{-\frac{1}{r-1}}\frac{r-1}{r}\|y_2 - y_1\|_{p^\star}^{\frac{r}{r-1}}, \text{ (by Proposition 19)}
\end{aligned}$$

where we use $\nabla\psi^\star(y_1) = x_1$ and (A.1) to obtain (A.2). Then, for arbitrary y_1 and $y_2 \in \mathbb{R}^d$, we have

$$\psi^\star(y_2) \le \psi^\star(y_1) + \langle\nabla\psi^\star(y_1), y_2 - y_1\rangle + K^{-\frac{1}{r-1}}\frac{r-1}{r}\|y_2 - y_1\|_{p^\star}^{\frac{r}{r-1}}.$$

Therefore, $\psi^\star$ is $(K^{-\frac{1}{r-1}}, \frac{r}{r-1})$-Hölder smooth with respect to $p^\star$-norm.

($\Leftarrow$) Since ψ is continuous and convex, by (Rockafellar, 1970, Theorem 12.2), we have

$$\psi(x) = \sup_{y\in\mathbb{R}^d}\{\langle x, y\rangle - \psi^\star(y)\}.$$

Let $x_1, x_2 \in \mathbb{R}^d$ be two arbitrary vectors, and let $\nabla\psi(x_1) = y_1$. Then, we have $\nabla\psi^\star(y_1) = x_1$, and (A.1). Let $\overline{K} = K^{-\frac{1}{r-1}}$. Then,

$$\begin{aligned}
\psi(x_2) &= \sup_{y\in\mathbb{R}^d}\{\langle x_2, y\rangle - \psi^\star(y)\} \\
&\ge \sup_{y\in\mathbb{R}^d}\left\{\langle x_2, y\rangle - \left(\psi^\star(y_1) + \langle\nabla\psi^\star(y_1), y - y_1\rangle + \overline{K}\frac{r-1}{r}\|y - y_1\|_{p^\star}^{\frac{r}{r-1}}\right)\right\} \\
&= \sup_{y\in\mathbb{R}^d}\left\{\langle x_2 - x_1, y - y_1\rangle - \overline{K}\frac{r-1}{r}\|y - y_1\|_{p^\star}^{\frac{r}{r-1}}\right\} - \psi^\star(y_1) + \langle x_2, y_1\rangle \text{ (since } \nabla\psi^\star(y_1) = x_1)
\end{aligned}$$

$$
\begin{aligned}
&= \sup_{y\in\mathbb{R}^d}\left\{\langle x_2-x_1, y-y_1\rangle - \overline{K}\frac{r-1}{r}\|y-y_1\|_{p^\star}^{\frac{r}{r-1}}\right\} + \psi(x_1) + \langle\nabla\psi(x_1), x_2-x_1\rangle \qquad \text{(A.3)}\\
&= \sup_{y\in\mathbb{R}^d}\left\{\langle x_2-x_1, \overline{K}^{-(r-1)}y\rangle - \overline{K}\frac{r-1}{r}\|\overline{K}^{-(r-1)}y\|_{p^\star}^{\frac{r}{r-1}}\right\} + \psi(x_1) + \langle\nabla\psi(x_1), x_2-x_1\rangle\\
&=\psi(x_1) + \langle\nabla\psi(x_1), x_2-x_1\rangle + \frac{\overline{K}^{-(r-1)}}{r}\|x_2-x_1\|_p^r \text{ (by Proposition 19)}\\
&=\psi(x_1) + \langle\nabla\psi(x_1), x_2-x_1\rangle + \frac{K}{r}\|x_2-x_1\|_p^r,
\end{aligned}
$$

where we use $\nabla\psi(x_1) = y_1$ and (A.1) in (A.3). Therefore, ψ is (K, r)-uniformly convex with respect to p-norm. ■

A.2. Proof of Proposition 5

In this part, we use the following notation.

- For $x = (x_1, \cdots, x_d)^T \in \mathbb{R}^d$ and $p > 1$, we let

$$x^{\langle p-1\rangle} := (\text{sgn}(x_1)|x_1|^{p-1}, \cdots, \text{sgn}(x_d)|x_d|^{p-1})^T,$$

  where for $t \in \mathbb{R}$,

$$\text{sgn}(t) := \begin{cases} 1, & \text{if } t > 0\\ 0, & \text{if } t = 0\\ -1, & \text{if } t < 0. \end{cases}$$

- We note that $\|x\|_p^r$, $x \in \mathbb{R}^d$, is continuously differentiable for all $p, r > 1$, with a gradient of

$$\nabla\|x\|_p^r = \begin{cases} r\|x\|_p^{r-p}x^{\langle p-1\rangle}, & \text{if } x \neq 0\\ 0, & \text{if } x = 0. \end{cases} \qquad \text{(A.4)}$$

  In the following, for the sake of convenience, we use an abuse of notation, $\|0\|_p^{r-p}0^{\langle p-1\rangle} := 0$, for any $p, r > 1$.

A.2.1. AUXILIARY RESULTS

We start with proving some auxiliary results. Let $x \in \mathbb{R}^d - \{0\}$ and $y \in \mathbb{R}^d$ be two arbitrary vectors. We let $h \in \mathbb{R}^d$ be

$$h := \frac{\langle x^{\langle p-1\rangle}, y\rangle}{\|x\|_p^p}x, \text{ where } p \in [1+\kappa, \infty).$$

Proposition 11 *For $p \in [1+\kappa, \infty)$, we have*

i) $\langle x^{\langle p-1\rangle}, y\rangle = \langle x^{\langle p-1\rangle}, h\rangle$

ii) $\|h\|_p \leq \|y\|_p$.

Proof

(i) Note that

$$\langle x^{\langle p-1\rangle}, x\rangle = \sum_{i=1}^{d} |x_i|^{p-1}\mathrm{sgn}(x_i)x_i = \sum_{i=1}^{d} |x_i|^{p-1}\mathrm{sgn}(x_i)|x_i|\mathrm{sgn}(x_i)$$
$$= \sum_{i=1}^{d} |x_i|^p = \|x\|_p^p.$$

Hence,

$$\langle x^{\langle p-1\rangle}, h\rangle = \frac{\langle x^{\langle p-1\rangle}, y\rangle}{\|x\|_p^p}\langle x^{\langle p-1\rangle}, x\rangle = \langle x^{\langle p-1\rangle}, y\rangle.$$

(ii) Note that $\|h\|_p = \frac{|\langle x^{\langle p-1\rangle}, y\rangle|}{\|x\|_p^{p-1}}$. By using Hölder's inequality, we can write that

$$|\langle x^{\langle p-1\rangle}, y\rangle| \le \Big(\sum_{i=1}^{d} (|x_i|^{p-1})^{\frac{p}{p-1}}\Big)^{\frac{p-1}{p}} \Big(\sum_{i=1}^{d} |y_i|^p\Big)^{\frac{1}{p}} = \|x\|_p^{p-1}\|y\|_p.$$

Hence,

$$\|h\|_p = \frac{|\langle x^{\langle p-1\rangle}, y\rangle|}{\|x\|_p^{p1}} \le \frac{\|x\|_p^{p-1}\|y\|_p}{\|x\|_p^{p-1}} = \|y\|_p.$$

■

Proposition 12 *Let* $\widetilde{\varphi}(x) = \frac{1}{r}\|x\|_p^r$, *where* $p, r \in [1+\kappa, \infty)$. *Then, we have*

$$\langle \nabla\widetilde{\varphi}(x+h), y-h\rangle = 0.$$

Proof Note that if $x+h=0$, the statement is trivially correct. Therefore, without loss of generality, we assume that $x+h \neq 0$.

By (A.4), we have

$$\langle \nabla\widetilde{\varphi}(x+h), y-h\rangle = \|x+h\|_p^{r-p}\langle (x+h)^{\langle p-1\rangle}, y-h\rangle.$$

Note that

$$(x+h)^{\langle p-1\rangle} = \Bigg(\underbrace{\Big(1+\frac{\langle x^{\langle p-1\rangle}, y\rangle}{\|x\|_p^p}\Big)}_{:=a} x\Bigg)^{\langle p-1\rangle}$$
$$= (ax)^{\langle p-1\rangle}$$
$$= (|ax|^{p-1}\mathrm{sgn}(ax_1), \cdots, |ax_d|^{p-1}\mathrm{sgn}(ax_d))^T$$
$$= |a|^{p-1}\mathrm{sgn}(a)(\mathrm{sgn}(x_1)|x_1|^{q-1}, \cdots, \mathrm{sgn}(x_d)|x_d|^{q-1})^T$$
$$= |a|^{p-1}\mathrm{sgn}(a)x^{\langle p-1\rangle}.$$

Then,

$$\begin{aligned}\langle \nabla\widetilde{\varphi}(x+h), y-h\rangle &= \|\underbrace{x+h}_{=ax}\|_p^{r-p}\langle \underbrace{(x+h)^{\langle p-1\rangle}}_{=|a|^{p-1}\mathrm{sgn}(a)x^{\langle p-1\rangle}}, y-h\rangle \\ &= |a|^{r-p}\|x\|_p^{r-p}|a|^{p-1}\mathrm{sgn}(a)\langle x^{\langle p-1\rangle}, y-h\rangle = 0 \text{ (by Proposition 11).}\end{aligned}$$

■

Proposition 13 *For any $p \in [1+\kappa, \infty)$, we have*

$$\|x+h\|_p^{1+\kappa} - \|x\|_p^{1+\kappa} - (1+\kappa)\|x\|_p^{1+\kappa-p}\langle x^{\langle p-1\rangle}, h\rangle \le 2\|h\|_p^{1+\kappa}.$$

Proof We have

$$\begin{aligned}\|x+h\|_p^{1+\kappa} &= \Big|1+\frac{\langle x^{\langle p-1\rangle}, y\rangle}{\|x\|_p^p}\Big|^{1+\kappa}\|x\|_p^{1+\kappa} \\ &\le \Big(1+(1+\kappa)\frac{\langle x^{\langle p-1\rangle}, y\rangle}{\|x\|_p^p} + 2\Big|\frac{\langle x^{\langle p-1\rangle}, y\rangle}{\|x\|_p^p}\Big|^{1+\kappa}\Big)\|x\|_p^{1+\kappa} \text{ (by Proposition 18)} \\ &= \|x\|_p^{1+\kappa} + (1+\kappa)\|x\|_p^{1+\kappa-p}\langle x^{\langle p-1\rangle}, y\rangle + 2\Big|\frac{\langle x^{\langle p-1\rangle}, y\rangle}{\|x\|_p^{p-1}}\Big|^{1+\kappa} \\ &= \|x\|_p^{1+\kappa} + (1+\kappa)\|x\|_p^{1+\kappa-p}\langle x^{\langle p-1\rangle}, y\rangle + 2\|h\|_p^{1+\kappa}.\end{aligned}$$

■

Proposition 14 *For any $p \in (1,2]$ and $x, y \in \mathbb{R}^d$, we have*

$$\|x+y\|_p^p - \|x\|_p^p - p\langle x^{\langle p-1\rangle}, y\rangle \le 2\|y\|_p^p.$$

Proof Since $p \in (1,2]$, we have

$$\begin{aligned}\|x+y\|_p^p &= \sum_{i=1}^d |x_i+y_i|^p \\ &\le \sum_{i=1}^d |x_i|^p + p|x_i|^{p-1}\mathrm{sgn}(x_i)y_i + 2|y_i|^p \text{ (by Proposition 18)} \\ &= \|x\|_p^p + p\langle x^{\langle p-1\rangle}, y\rangle + 2\|y\|_p^p.\end{aligned}$$

By rearranging the terms, we can obtain the statement. ■

Proposition 15 *For any $p > 2$ and $x, y \in \mathbb{R}^d$, we have*

$$\|x+y\|_p^2 - \|x\|_p^2 - 2\|x\|_p^{2-p}\langle x^{\langle p-1\rangle}, y\rangle \le (p-1)\|y\|_p^2.$$

Proof Since $p-1>1$, the statement holds when $x=0$. Therefore, in the following, without loss of generality, we assume $x \neq 0$.
We prove the statement in two different cases, separately.

- If $\frac{x}{\|x\|_2} = \pm\frac{y}{\|y\|_2}$, then $y = tx$ for some $t \in \mathbb{R}$. In that case,

$$\begin{aligned}\|x+y\|_p^2 = (1+t)^2\|x\|_p^2 &= (1+2t+t^2)\|x\|_p^2 \\ &= \|x\|_p^2 + 2t\|x\|_p^2 + t^2\|x\|_p^2 \\ &= \|x\|_p^2 + 2\|x\|_p^{2-p}\langle x^{\langle p-1\rangle}, tx\rangle + \|tx\|_p^2 \\ &= \|x\|_p^2 + 2\|x\|_p^{2-p}\langle x^{\langle p-1\rangle}, y\rangle + \|y\|_p^2 \\ &\leq \|x\|_p^2 + 2\|x\|_p^{2-p}\langle x^{\langle p-1\rangle}, y\rangle + (p-1)\|y\|_p^2.\end{aligned}$$

- If $\frac{x}{\|x\|_2} \neq \pm\frac{y}{\|y\|_2}$, then $x \neq ty$ (i.e., $x - ty \neq 0$) for all $t \in \mathbb{R}$. Then, $g(t) = \|x+ty\|_p^2$ is twice continuously differentiable on $\mathbb{R}$, and

$$\begin{aligned}g''(t) &= -2(p-2)\|x+ty\|_p^{2-2p}(\langle (x+ty)^{\langle p-1\rangle}, y\rangle)^2 \\ &\quad + 2(p-1)\|x+ty\|_p^{2-p}\Big(\sum_{i=1}^d |x_i+ty_i|^{p-2}y_i^2\Big) \\ &\leq 2(p-1)\|x+ty\|_p^{2-p}\Big(\sum_{i=1}^d |x_i+ty_i|^{p-2}y_i^2\Big) \\ &\leq 2(p-1)\|x+ty\|_p^{2-p}\|x+ty\|_p^{p-2}\|y\|_p^2 \text{ (by Hölder's inequality)} \\ &= 2(p-1)\|y\|_p^2.\end{aligned}$$

  Then, we have

$$\begin{aligned}\|x+y\|_p^2 - \|x\|_p^2 - 2\|x\|_p^{2-p}\langle x^{\langle p-1\rangle}, y\rangle &= g(1) - g(0) - g'(0) \\ &= \int_0^1 g'(t) - g'(0)\,dt \\ &= \int_0^1 \int_0^t g''(u)\,du\,dt \\ &\leq (p-1)\|y\|_p^2.\end{aligned}$$

■

A.2.2. PROOF OF PROPOSITION 5

Proof [Proof of Proposition 5]

1. We want to show that for all $x, y \in \mathbb{R}^d$,

$$\frac{1}{1+\kappa}\|x+y\|_p^{1+\kappa} - \frac{1}{1+\kappa}\|x\|_p^{1+\kappa} - \|x\|_p^{1+\kappa-p}\langle x^{\langle p-1\rangle}, y\rangle \leq \frac{K_p}{1+\kappa}\|y\|_p^{1+\kappa}.$$

Note that since $K_p > 1$, the statement is correct when $x = 0$. Therefore, in the following, without loss of generality, we assume $x \neq 0$. Let

$$h = \frac{\langle x^{\langle p-1 \rangle}, y \rangle}{\|x\|_p^p} x, \text{ for } p \in [1+\kappa, \infty).$$

We will prove the $p \in [1+\kappa, 2]$ and $p \in (2, \infty)$ cases separately.

For $p \in [1+\kappa, 2]$, by using Proposition 14, we can write that

$$\|x+y\|_p^p - \|x+h\|_p^p - \underbrace{p\langle (x+h)^{\langle p-1 \rangle}, y-h \rangle}_{=0 \text{ (by Prop. 12)}} \leq 2\|y-h\|_p^p.$$

Therefore, we have

$$\|x+y\|_p^p \leq \|x+h\|_p^p + 2\|y-h\|_p^p.$$

Then,

$$\begin{aligned}
\|x+y\|_p^{1+\kappa} \leq & (\|x+h\|_p^p + 2\|y-h\|_p^p)^{\frac{1+\kappa}{p}} \\
\leq & \|x+h\|_p^{1+\kappa} + 2\|y-h\|_p^{1+\kappa} \text{ (since } 1+\kappa \leq p \text{, by Proposition 17)} \\
\leq & \|x\|_p^{1+\kappa} + (1+\kappa)\|x\|_p^{1+\kappa-p}\langle x^{\langle p-1 \rangle}, y \rangle + 2\|h\|_p^{1+\kappa} + 2\|y-h\|_p^{1+\kappa} \text{ (by Proposition 13)} \\
\leq & \|x\|_p^{1+\kappa} + (1+\kappa)\|x\|_p^{1+\kappa-p}\langle x^{\langle p-1 \rangle}, y \rangle + 2\|y\|_p^{1+\kappa} + 2\,2^{1+\kappa}\|y\|_p^{1+\kappa} \text{ (by Proposition 11)} \\
\leq & \|x\|_p^{1+\kappa} + (1+\kappa)\|x\|_p^{1+\kappa-p}\langle x^{\langle p-1 \rangle}, y \rangle + 10\|y\|_p^{1+\kappa} \text{ (since } \kappa \in (0,1]\text{).}
\end{aligned} \tag{A.5}$$

For $p \in (2, \infty)$, by using Proposition 15, we can write that

$$\|x+y\|_p^2 - \|x+h\|_p^2 - \underbrace{2\|x+h\|_p^{2-p}\langle (x+h)^{\langle p-1 \rangle}, y-h \rangle}_{=0 \text{ (by Prop. 12)}} \leq (p-1)\|y-h\|_p^2.$$

Therefore, we have

$$\|x+y\|_p^2 \leq \|x+h\|_p^2 + (p-1)\|y-h\|_p^2.$$

Then,

$$\begin{aligned}
\|x+y\|_p^{1+\kappa} \leq & (\|x+h\|_p^2 + (p-1)\|y-h\|_p^2)^{\frac{1+\kappa}{2}} \\
\leq & \|x+h\|_p^{1+\kappa} + (p-1)^{\frac{1+\kappa}{2}}\|y-h\|_p^{1+\kappa} \text{ (since } 1+\kappa \leq 2\text{, by Proposition 17)} \\
\leq & \|x\|_p^{1+\kappa} + (1+\kappa)\|x\|_p^{1+\kappa-p}\langle x^{\langle p-1 \rangle}, y \rangle + 2\|h\|_p^{1+\kappa} + (p-1)^{\frac{1+\kappa}{2}}\|y-h\|_p^{1+\kappa} \text{ (by Proposition 13)} \\
\leq & \|x\|_p^{1+\kappa} + (1+\kappa)\|x\|_p^{1+\kappa-p}\langle x^{\langle p-1 \rangle}, y \rangle + 2\|y\|_p^{1+\kappa} + (p-1)^{\frac{1+\kappa}{2}}2^{1+\kappa}\|y\|_p^{1+\kappa} \text{ (by Proposition 11)} \\
\leq & \|x\|_p^{1+\kappa} + (1+\kappa)\|x\|_p^{1+\kappa-p}\langle x^{\langle p-1 \rangle}, y \rangle + 10(p-1)^{\frac{1+\kappa}{2}}\|y\|_p^{1+\kappa} \text{ (since } p > 2 \text{ and } \kappa \in (0,1]\text{).}
\end{aligned} \tag{A.6}$$

By multiplying both sides in (A.5) and (A.6) with $\frac{1}{1+\kappa}$, the statement follows.

2. Let us fix an arbitrary $p \in [1+\kappa, \infty)$. Note that by Proposition 19,

$$\varphi(x) = \frac{1}{1+\kappa}\|x\|_p^{1+\kappa} \quad \text{and} \quad \varphi^\star(y) = \frac{\kappa}{1+\kappa}\|y\|_{p^\star}^{\frac{1+\kappa}{\kappa}}$$

are convex conjugate pairs. By the previous part, we know that φ is $(K_p, 1+\kappa)$-Hölder smooth with respect to p-norm. Then, by Proposition 4, $\varphi^\star$ is $(K_p^{-\frac{1}{\kappa}}, \frac{1+\kappa}{\kappa})$-uniformly convex with respect to $p^\star$-norm.

■

Appendix B. Proofs for Section 3

B.1. Proof of Theorem 6

We start with an auxiliary result, given in (Bubeck, 2014, Lemma 4.1).

Proposition 16 *(Bubeck, 2014, Lemma 4.1) Let Ψ be the mirror function defined in Theorem 6. For $y \in \mathbb{R}^d$, let $\hat{y} = \arg\min_{x \in \mathcal{S}} D_\Psi(x, y)$. Then, for any $x \in \mathcal{S}$,*

i) $\langle \nabla\Psi(\hat{y}) - \nabla\Psi(y), \hat{y} - x\rangle \leq 0$

ii) $D_\Psi(x, \hat{y}) + D_\Psi(\hat{y}, y) \leq D_\Psi(x, y)$.

Proof [Proof of Theorem 6] For notational convenience, we let $x^\star = \arg\min_{x\in\mathcal{S}} f(x)$. We start with two observations:

- $g_{t+1} = \frac{1}{\eta}(\nabla\Psi(x_t) - \nabla\Psi(y_{t+1}))$,
- $D_\Psi(x^\star, x_t) + D_\Psi(x_t, y_{t+1}) - D_\Psi(x^\star, y_{t+1}) = \langle \nabla\Psi(x_t) - \nabla\Psi(y_{t+1}), x_t - x^\star\rangle$.

Then, we write

$$\begin{aligned}
\langle g_{t+1}, x_t - x^\star\rangle &= \frac{1}{\eta}\langle \nabla\Psi(x_t) - \nabla\Psi(y_{t+1}), x_t - x^\star\rangle \\
&= \frac{1}{\eta}(D_\Psi(x^\star, x_t) + D_\Psi(x_t, y_{t+1}) - D_\Psi(x^\star, y_{t+1})) \\
&\leq \frac{1}{\eta}(D_\Psi(x^\star, x_t) + D_\Psi(x_t, y_{t+1}) - D_\Psi(x^\star, x_{t+1}) - D_\Psi(x_{t+1}, y_{t+1})) \text{ (by Proposition 16)} \\
&= \frac{1}{\eta}(D_\Psi(x^\star, x_t) - D_\Psi(x^\star, x_{t+1}) + D_\Psi(x_t, y_{t+1}) - D_\Psi(x_{t+1}, y_{t+1})). \qquad \text{(B.1)}
\end{aligned}$$

Note that $D_\Psi(x^\star, x_t) - D_\Psi(x^\star, x_{t+1})$ will lead to a telescoping sum when summing over $t = 1$ to $t = T$. Therefore, it remains to bound the other term:

$$\begin{aligned}
D_\Psi(x_t, y_{t+1}) - D_\Psi(x_{t+1}, y_{t+1}) &= \Psi(x_t) - \Psi(x_{t+1}) - \langle \nabla\Psi(y_{t+1}), x_t - x_{t+1}\rangle \\
&\leq \langle \nabla\Psi(x_t) - \nabla\Psi(y_{t+1}), x_t - x_{t+1}\rangle - \frac{\kappa}{1+\kappa}\|x_t - x_{t+1}\|_{q^\star}^{\frac{1+\kappa}{\kappa}}
\end{aligned}$$

$$= \eta\langle g_{t+1}, x_t - x_{t+1}\rangle - \frac{\kappa}{1+\kappa}\|x_t - x_{t+1}\|_{q^\star}^{\frac{1+\kappa}{\kappa}} \tag{B.2}$$

$$\leq \eta\|g_{t+1}\|_q\|x_t - x_{t+1}\|_{q^\star} - \frac{\kappa}{1+\kappa}\|x_t - x_{t+1}\|_{q^\star}^{\frac{1+\kappa}{\kappa}} \tag{B.3}$$

$$\leq \frac{1}{1+\kappa}\eta^{1+\kappa}\|g_{t+1}\|_q^{1+\kappa}. \tag{B.4}$$

where we use that Ψ is $(1, \frac{1+\kappa}{\kappa})$ uniformly convex in (B.2), and maximize the right-hand side of (B.3) to obtain (B.4).

By (B.1) and (B.4), we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\langle g_{t+1}, x_t - x^\star\rangle \leq \frac{D_\Psi(x^\star, x_0)}{\eta T} + \frac{\eta^\kappa}{1+\kappa}\frac{1}{T}\sum_{t=0}^{T-1}\|g_{t+1}\|_q^{1+\kappa}$$

$$\leq \frac{\Psi(x^\star) - \Psi(x_0)}{\eta T} + \frac{\eta^\kappa}{1+\kappa}\frac{1}{T}\sum_{t=0}^{T-1}\|g_{t+1}\|_q^{1+\kappa} \text{ (since } x_0 = \arg\min_{x\in\mathcal{S}}\Psi(x) \text{ and } \mathcal{S} \text{ is convex)}$$

$$\leq \frac{\kappa}{1+\kappa}\frac{R_0^{\frac{1+\kappa}{\kappa}}}{\eta T} + \frac{\eta^\kappa}{1+\kappa}\frac{1}{T}\sum_{t=0}^{T-1}\|g_{t+1}\|_q^{1+\kappa}.$$

Note that x_t is $\mathcal{F}_t$-measurable. Hence, we can write that

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1}\langle\mathbb{E}[g_{t+1}|\mathcal{F}_t], x_t - x^\star\rangle\right] \leq \frac{\kappa}{1+\kappa}\frac{R_0^{\frac{1+\kappa}{\kappa}}}{\eta T} + \frac{\eta^\kappa}{1+\kappa}\frac{1}{T}\mathbb{E}\Big[\sum_{i=0}^{T-1}\mathbb{E}[\|g_{t+1}\|_q^{1+\kappa}|\mathcal{F}_t]\Big]$$

$$\leq \frac{\kappa}{1+\kappa}\frac{R_0^{\frac{1+\kappa}{\kappa}}}{\eta T} + \frac{\eta^\kappa}{1+\kappa}\frac{1}{T}\sum_{i=1}^{T}\sigma^{1+\kappa},$$

which for $v_t \in \partial f(x_t)$, leads to

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1}\langle v_t, x_t - x^\star\rangle\right] \leq \frac{\kappa}{1+\kappa}\frac{R_0^{\frac{1+\kappa}{\kappa}}}{\eta T} + \frac{\eta^\kappa}{1+\kappa}\frac{1}{T}\sum_{t=0}^{T-1}\sigma^{1+\kappa}.$$

Then,

$$\frac{\kappa}{1+\kappa}\frac{R_0^{\frac{1+\kappa}{\kappa}}}{\eta T} + \frac{\eta^\kappa}{1+\kappa}\sigma^{1+\kappa} \geq \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}f(x_t) - \min_{x\in\mathcal{S}}f(x)\right] \text{ (by the convexity of } f)$$

$$\geq \mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=0}^{T-1}x_t\right) - \min_{x\in\mathcal{S}}f(x)\right] \text{ (by Jensen's inequality).} \tag{B.5}$$

By using $\eta = \frac{R_0^{\frac{1}{\kappa}}}{\sigma}\frac{1}{T^{\frac{1}{1+\kappa}}}$ in (B.5), we can obtain the statement. ■

B.2. Proof of Corollary 7

Proof [Proof of Corollary 7]

i) For $q \in [1, 1+\kappa]$, we have

$$\mathbb{E}[\|g_t\|_{1+\kappa}^{1+\kappa}] \le \mathbb{E}[\|g_t\|_q^{1+\kappa}] \le L^{1+\kappa} \text{ (since } q \le 1+\kappa).$$

Moreover, $x_0 = 0$ and

$$R_0^{\frac{1+\kappa}{\kappa}} = \frac{1+\kappa}{\kappa} \sup_{x \in \mathbb{B}_\infty(R)} (\Psi(x) - \Psi(x_0)) \le 10^{\frac{1}{\kappa}} \sup_{x \in \mathbb{B}_\infty(R)} \|x\|_{\frac{1+\kappa}{\kappa}}^{\frac{1+\kappa}{\kappa}}.$$

Then,

$$R_0 \le 10^{\frac{1}{1+\kappa}} \sup_{x \in \mathbb{B}_\infty(R)} \|x\|_{\frac{1+\kappa}{\kappa}} \le 10 R d^{\frac{\kappa}{1+\kappa}}.$$

Since $\Psi = U_p$ for $p = 1+\kappa$ is $(1, \frac{1+\kappa}{\kappa})$-uniformly convex with respect to $\frac{1+\kappa}{\kappa}$-norm (see Proposition 5), by Theorem 6, we have

$$\mathbb{E}\left[f\left(\frac{1}{T} \sum_{t=0}^{T-1} x_t \right) - \min_{x \in \mathcal{S}} f(x) \right] \le 10\, R\sigma \left(\frac{d}{T} \right)^{\frac{\kappa}{1+\kappa}}.$$

ii) For $q \in (1+\kappa, \infty)$, we have $x_0 = 0$ and

$$R_0^{\frac{1+\kappa}{\kappa}} = \frac{1+\kappa}{\kappa} \sup_{x \in \mathbb{B}_\infty(R)} (\Psi(x) - \Psi(x_0)) \le \left(10 \max\{1, (q-1)^{\frac{1+\kappa}{2}}\} \right)^{\frac{1}{\kappa}} \sup_{x \in \mathbb{B}_\infty(R)} \|x\|_{q^\star}^{\frac{1+\kappa}{\kappa}}.$$

Then,

$$R_0 \le 10^{\frac{1}{1+\kappa}} \max\{1, \sqrt{q-1}\} \sup_{x \in \mathbb{B}_\infty(R)} \|x\|_{q^\star} \le 10 \max\{1, \sqrt{q-1}\} R d^{1-\frac{1}{q}}.$$

Since $\Psi = U_p$ for $p = q$ is $(1, \frac{1+\kappa}{\kappa})$-uniformly convex with respect to $q^\star$-norm (see Proposition 5), by Theorem 6, we have

$$\mathbb{E}\left[f\left(\frac{1}{T} \sum_{t=0}^{T-1} x_t \right) - \min_{x \in \mathcal{S}} f(x) \right] \le 10 \max\{1, \sqrt{q-1}\} R\sigma \frac{d^{1-\frac{1}{q}}}{T^{\frac{\kappa}{1+\kappa}}}.$$

iii) If $q \in (\log d, \infty]$, we have

$$\begin{aligned} \mathbb{E}[\|g_t\|_{1+\log d}^{1+\kappa}] \le \mathbb{E}[(d^{\frac{1}{1+\log d} - \frac{1}{q}} \|g_t\|_q)^{1+\kappa}] &= d^{\frac{1+\kappa}{1+\log d} - \frac{1+\kappa}{q}} \mathbb{E}[\|g_t\|_q^{1+\kappa}] \text{ (since } q > \log d) \\ &\le d^{\frac{1+\kappa}{1+\log d} - \frac{1+\kappa}{q}} \sigma^{1+\kappa}. \end{aligned}$$

Moreover, $x_0 = 0$ and

$$R_0^{\frac{1+\kappa}{\kappa}} = \frac{1+\kappa}{\kappa} \sup_{x \in \mathbb{B}_\infty(R)} (\Psi(x) - \Psi(x_0)) \le \left(10(1 + \log d - 1)^{\frac{1+\kappa}{2}} \right)^{\frac{1}{\kappa}} \sup_{x \in \mathbb{B}_\infty(R)} \|x\|_{\frac{1+\log d}{\log d}}^{\frac{1+\kappa}{\kappa}}.$$

Then,

$$R_0 \leq 10^{\frac{1}{1+\kappa}}\sqrt{\log d}\sup_{x\in\mathbb{B}_\infty(R)}\|x\|_{\frac{1+\log d}{\log d}} \leq 10\sqrt{\log d}Rd^{\frac{\log d}{1+\log d}}.$$

As $\Psi = U_p$ for $p = 1+\log d$ is $(1, \frac{1+\kappa}{\kappa})$-uniformly convex with respect to $\frac{(1+\log d)}{\log d}$-norm, by Theorem 6, we have

$$\mathbb{E}\left[f\left(\frac{1}{T}\sum_{t=0}^{T-1}x_t\right) - \min_{x\in\mathcal{S}} f(x)\right] \leq 10\, R\sigma\sqrt{\log d}\frac{d^{1-\frac{1}{q}}}{T^{\frac{\kappa}{1+\kappa}}}.$$

■

Appendix C. Auxiliary Results for Sections 2 and 3

Proposition 17 *Let $x, y \geq 0$ and $\kappa \in (0, 1]$. Then,*

(i) $(x+y)^\kappa \leq x^\kappa + y^\kappa$

(ii) $x^\kappa + y^\kappa \leq 2^{1-\kappa}(x+y)^\kappa$.

Proof

(i) Without loss of generality, assume $x \geq y$. By concavity, we have

$$\begin{aligned}(x+y)^\kappa &\leq x^\kappa + \kappa x^{\kappa-1}y \\ &\leq x^\kappa + y^\kappa. \text{ (Since } \kappa \in (0,1] \text{ and } y \leq x)\end{aligned}$$

(ii) By using $p = 1/\kappa$ and $p^\star = 1/(1-\kappa)$ in Hölder's inequality, we write

$$x^\kappa + y^\kappa \leq (1^{p^\star} + 1^{p^\star})^{1/p^\star}(x+y)^{1/p} = 2^{1-\kappa}(x+y)^\kappa.$$

■

Proposition 18 *Let $x, y \in \mathbb{R}$ and $\kappa \in (0, 1]$. Then,*

$$|x+y|^{1+\kappa} - |x|^{1+\kappa} - (1+\kappa)|x|^\kappa sgn(x)y \leq 2^{1-\kappa}|y|^{1+\kappa}.$$

Proof Let $g(x) = |x|^{1+\kappa}$ for $x \in \mathbb{R}$. Note that g is convex and continuously differentiable, where $g'(x) = (1+\kappa)|x|^\kappa \text{sgn}(x)$. Then,

$$\begin{aligned}|x+y|^{1+\kappa} - |x|^{1+\kappa} - (1+\kappa)|x|^\kappa \text{sgn}(x)y &= g(x+y) - g(x) - g'(x)y \\ &= \int_x^{x+y}(g'(t) - g'(x))\,dt \\ &\leq \int_x^{x+y}|g'(t) - g'(x)|\,dt \\ &= (1+\kappa)\int_x^{x+y}\left||t|^\beta \text{sgn}(t) - |x|^\beta \text{sgn}(x)\right|dt. \qquad \text{(C.1)}\end{aligned}$$

In the following, we will find an upper-bound for the integrand in (C.1).

- If $\operatorname{sgn}(t) = \operatorname{sgn}(x)$, we have

$$\begin{aligned}\left||t|^\kappa \operatorname{sgn}(t) - |x|^\kappa \operatorname{sgn}(x)\right| &= \left||t|^\kappa - |x|^\kappa\right| \\ &\leq |t-x|^\kappa \text{ (By Proposition 17).}\end{aligned}$$

- If $\operatorname{sgn}(t) \neq \operatorname{sgn}(x)$,

$$\begin{aligned}\left||t|^\kappa \operatorname{sgn}(t) - |x|^\kappa \operatorname{sgn}(x)\right| &= |t|^\kappa + |x|^\kappa \\ &\leq 2^{1-\kappa}(|t|^\kappa + |x|^\kappa) \text{ (By Proposition 17)} \\ &= 2^{1-\kappa}|t-x|^\kappa.\end{aligned}$$

Then, we have

$$\begin{aligned}\text{(C.1)} \leq 2^{1-\kappa}\int_x^{x+y}(1+\kappa)|t-x|^\kappa\,dt &= 2^{1-\kappa}\int_0^y (1+\kappa)|t|^\kappa\,dt \\ &= 2^{1-\kappa}|y|^\kappa \\ &\leq 2|y|^\kappa.\end{aligned}$$

■

Proposition 19 *Let $r > 1$, $p \in [1,\infty]$,*

$$\widetilde{\varphi}(x) = \frac{1}{r}\|x\|_p^r \quad \textit{and} \quad \widetilde{\varphi}^\star(y) = \sup_{x\in\mathbb{R}^d}\{\langle y, x\rangle - \widetilde{\varphi}(x)\}.$$

Then, for $1/p + 1/p^\star = 1$, we have

$$\widetilde{\varphi}^\star(y) = \frac{r-1}{r}\|y\|_{p^\star}^{\frac{r}{r-1}}.$$

Proof Let us fix an arbitrary $y \in \mathbb{R}^d$. By using Hölder's inequality, for any $x \in \mathbb{R}^d$, we can write that

$$\begin{aligned}\langle y, x\rangle - \frac{1}{r}\|x\|_p^r &\leq \|y\|_{p^\star}\|x\|_p - \frac{1}{r}\|x\|_p^r \\ &\leq \frac{r-1}{r}\|y\|_{p^\star}^{\frac{r}{r-1}} \text{ (by maximizing the right-hand side).}\end{aligned}$$

Therefore, we have

$$\widetilde{\varphi}^\star(y) \leq \frac{r-1}{r}\|y\|_{p^\star}^{\frac{r}{r-1}}.$$

Moreover, since dual norm can be formulated as a supremum on a compact set, there exists a $x \in \mathbb{R}^d$ such that $\langle y, x\rangle = \|y\|_{p^\star}\|x\|_p$ and $\|y\|_{p^\star} = \|x\|_p^{r-1}$. In this case,

$$\langle y, x\rangle - \frac{1}{r}\|x\|_p^r = \frac{r-1}{r}\|y\|_{p^\star}^{\frac{r}{r-1}}.$$

Therefore, we have

$$\widetilde{\varphi}^\star(y) \geq \frac{r-1}{r}\|y\|_{p^\star}^{\frac{r}{r-1}}.$$

Consequently, $\widetilde{\varphi}^\star(y) = \frac{r-1}{r}\|y\|_{p^\star}^{\frac{r}{r-1}}$. ■

Proposition 20 *If $\psi : \mathbb{R}^d \to \mathbb{R}$ is differentiable and uniformly convex, $\psi^\star$ is everywhere differentiable.*

Proof Let us say that ψ is (K, r)-uniformly convex with respect to some p-norm. First, we will show that $\psi^\star$ is subdifferentiable by proving that it is everywhere finite. For any $y \in \mathbb{R}^d$, we have

$$\begin{aligned}
\psi^\star(y) &= \sup_{x\in\mathbb{R}^d} \{\langle y, x\rangle - \psi(x)\} \\
&\leq \sup_{x\in\mathbb{R}^d} \left\{\langle y, x\rangle - (\psi(y) + \langle \nabla\psi(y), x - y\rangle + \frac{K}{r}\|y - x\|_p^r)\right\} \\
&= \sup_{x\in\mathbb{R}^d} \left\{\langle y - \nabla\psi(y), x - y\rangle - \frac{K}{r}\|y - x\|_p^r)\right\} - \psi(y) + \|y\|_2^2 \\
&= \sup_{x\in\mathbb{R}^d} \left\{\langle y - \nabla\psi(y), K^{-\frac{1}{r-1}}x\rangle - \frac{K}{r}\|K^{-\frac{1}{r-1}}x\|_p^r)\right\} - \psi(y) + \|y\|_2^2 \\
&= K^{-\frac{1}{r-1}}\frac{r-1}{r}\|y - \nabla\psi(y)\|_{p^\star}^{\frac{r}{r-1}} - \psi(y) + \|y\|_2^2.
\end{aligned}$$

Therefore, for any $y \in \mathbb{R}^d$, $\psi^\star(y)$ is finite. By (Rockafellar, 1970, Theorem 23.4), $\psi^\star$ is a subdifferentiable convex function.

Next, we prove an intermediate result. Since ψ is differentiable and uniformly convex, we have

$$\psi(y) \geq \psi(x) + \langle \nabla\psi(x), y - x\rangle + \frac{K}{r}\|y - x\|_p^r, \quad \forall x, y \in \mathbb{R}^d, \tag{C.2}$$

and

$$\psi(x) \geq \psi(y) - \langle \nabla\psi(y), y - x\rangle + \frac{K}{r}\|y - x\|_p^r \quad \forall x, y \in \mathbb{R}^d. \tag{C.3}$$

By summing (C.2) and (C.3), we can write that

$$\langle \nabla\psi(x) - \nabla\psi(y), x - y\rangle \geq \frac{2K}{r}\|y - x\|_p^r, \quad \forall x, y \in \mathbb{R}^d. \tag{C.4}$$

We show that $\psi^\star$ is differentiable by using proof by contradiction. Choose an arbitrary $y_0 \in \mathbb{R}^d$. Since $\psi^\star$ is subdifferentiable, we know that $\partial\psi(y_0) \neq \emptyset$. Let us assume that $x_1, x_2 \in \partial\psi^\star(y_0)$, and $x_1 \neq x_2$. Since ψ is continuous and convex, by (Rockafellar, 1970, Corollary 23.5.1.),

$$\nabla\psi(x_1) = \nabla\psi(x_2) = y_0. \tag{C.5}$$

However, (C.4) contradicts with (C.5). Since $\psi^\star$ is subdifferentiable, there must be a unique element in $\partial\psi^\star(y_0)$. Therefore, by (Rockafellar, 1970, Theorem 25.1), $\psi^\star$ is differentiable at y_0. Since y_0 was chosen arbitrarily, $\psi^\star$ is everywhere differentiable. ■

Appendix D. Proofs for Section 4

D.1. Auxiliary lemmas

To prove Theorem 9, we need the following lemmas.

Lemma 21 (KL-divergence between Bernoulli distributions) *The Kullback-Leibler divergence between Bernoulli distributions* $\text{BERNOULLI}(1-\frac{2+\alpha}{4}p)$ *and* $\text{BERNOULLI}(1-\frac{2-\alpha}{4}p)$ *is bounded by* p*, i.e.,*

$$D_{\mathrm{KL}}\Big(\text{BERNOULLI}\big(1-\frac{2+\alpha}{4}p\big)||\text{BERNOULLI}\big(1-\frac{2-\alpha}{4}p\big)\Big) \le p$$
$$D_{\mathrm{KL}}\Big(\text{BERNOULLI}\big(1-\frac{2-\alpha}{4}p\big)||\text{BERNOULLI}\big(1-\frac{2+\alpha}{4}p\big)\Big) \le p\,,$$

where $\alpha \in \{-1,+1\}, p \in \Big(0,\frac{1}{2}\Big)$.

Proof [Proof of Lemma 21] Denote the Bernoulli distributions $\text{BERNOULLI}(1-\frac{3}{4}p)$ and $\text{BERNOULLI}(1-\frac{1}{4}p)$ by $\mathbb{P}^+$ and $\mathbb{P}^-$, respectively. By the definition KL divergence, it holds that

$$D_{\mathrm{KL}}(\mathbb{P}^+||\mathbb{P}^-) = \Big(1-\frac{3}{4}p\Big)\log\Big(\frac{1-\frac{3}{4}p}{1-\frac{1}{4}p}\Big) + \frac{3}{4}p\log\Big(\frac{\frac{3}{4}p}{\frac{1}{4}p}\Big) \le \frac{3}{4}p\log 3 \le p$$

We now prove

$$D_{\mathrm{KL}}(\mathbb{P}^-||\mathbb{P}^+) = (1-\frac{1}{4}p)\log\Big(\frac{1-\frac{1}{4}p}{1-\frac{3}{4}p}\Big) + \frac{1}{4}p\log\Big(\frac{\frac{1}{4}p}{\frac{3}{4}p}\Big) \le p\,.$$

Define the function $h(p) = p - \frac{1}{4}p\log(\frac{1}{3}) - (1-\frac{1}{4}p)\log(\frac{4-p}{4-3p})$. Then, we obtain

$$\nabla h(p) = 1 - \frac{1}{4}\log\Big(\frac{1}{3}\Big) + \frac{1}{4}\log\Big(\frac{4-p}{4-3p}\Big) - \frac{2}{4-3p}\,,$$

and

$$\nabla^2 h(p) = -\frac{16}{(4-p)(4-3p)^2}\,.$$

When $p \in (0,\frac{1}{2}]$, it follows that

$$\nabla^2 h(p) < 0, \quad \nabla h\big(\frac{1}{2}\big) \ge 0.5 \quad \text{and} \quad h(0) = 0\,,$$

which implies $h(p) \ge 0$, that is

$$p \ge \frac{1}{4}p\log\Big(\frac{1}{3}\Big) + \Big(1-\frac{1}{4}p\Big)\log\Big(\frac{4-p}{4-3p}\Big)\,.$$

■

Lemma 22 (Lower bound 1 with $d \ge 2$) *Suppose a vector* $\alpha^* = (\alpha_1^*,\ldots,\alpha_d^*)^\top$ *is chosen uniformly at random from the set* $\mathcal{V}$*, where* $\mathcal{V}$ *is a subset of the hypercube* $\{-1,+1\}^d$ *such that* $\Delta_{\mathrm{H}}(\alpha,\tilde{\alpha}) = \sum_{i=1}^d \mathbb{1}\{\alpha_i \ne \tilde{\alpha}_i\} \ge \frac{d}{4}$ *for any* $\alpha,\tilde{\alpha} \in \mathcal{V}$*. Given the vector* α^**,* $\kappa \in (0,1]$*, and* $\delta \in (0,\frac{1}{8}]$*, set the parameter*

$$\tilde{\alpha}^* = \Big(1-\frac{2+\alpha_1^*}{4}(4\delta)^{\frac{\kappa+1}{\kappa}},\ldots,1-\frac{2+\alpha_d^*}{4}(4\delta)^{\frac{\kappa+1}{\kappa}}\Big)^\top.$$

Suppose the oracle ϕ tosses a set of d coins with bias $\tilde{\alpha}^*$ a total of T times, and the outcome of only one coin chosen uniformly at random is given at each round. When $d \geq 2$, it holds for any holds for any estimator $\hat{\alpha} \in \mathcal{V}$ that

$$\mathbb{P}(\hat{\alpha} \neq \alpha^*) \geq 1 - \frac{(4\delta)^{\frac{\kappa+1}{\kappa}} T + \log 2}{d/8} .$$

Here, the probability is taken over the randomness of α^* and ϕ.

Proof [Proof of Lemma 22] Let $U_t \in \{1, \ldots, d\}$ be the variable indicating the U_t-th coin revealed at time t, and let $X_t \in \{0, 1\}$ denote its outcome. By (Pollard et al., 2012, Sec 15.3.2, Lemma 4) and (Scarlett and Cevher, 2019, Theorem 1), if the parameter α^* is uniform on $\mathcal{V}$, it holds for any estimator $\hat{\alpha} \in \mathcal{V}$ that

$$\mathbb{P}(\hat{\alpha} \neq \alpha^*) \geq 1 - \frac{I\big(\{U_t, X_t\}_{t=1}^T; \alpha^*\big) + \log 2}{\log |\mathcal{V}|} ,$$

where $I(\{U_t, X_t\}_{t=1}^T; \alpha^*)$ denotes the mutual information between the data sequence $\{U_t, X_t\}_{t=1}^T$ and α^*. By the Varshamov-Gilbert bound (Massart, 2007, Lemma 4.7), there exists such a packing set $\mathcal{V} \subseteq \{-1, +1\}^d$ with $|\mathcal{V}| \geq \exp(\frac{d}{8})$ satisfies $\Delta_{\mathrm{H}}(\alpha, \tilde{\alpha}) = \sum_{i=1}^d \mathbb{1}\{\alpha_i \neq \tilde{\alpha}_i\} \geq \frac{d}{4}$ for any $\alpha, \tilde{\alpha} \in \mathcal{V}$. It suffice to show that $I(\{U_t, X_t\}_{t=1}^T; \alpha^*) \leq (4\delta)^{\frac{\kappa+1}{\kappa}} T$. By the independent and identically distributed the sampling, we have

$$I(\{U_t, X_t\}_{t=1}^T; \alpha^*) = \sum_{t=1}^T I\big((U_1, X_1); \alpha^*\big) = T I\big((U_1, X_1); \alpha^*\big) .$$

By chain rule of mutual information and the sampling scheme, it holds that

$$I\big((U_1, X_1); \alpha^*\big) = I(X_1; \alpha^* | U_1) + I(\alpha^*; U_1) .$$

Note that U_1 is sampled independent of α^*, this implies $I(\alpha^*; U_1) = 0$. It remains to show that $I(X_1; \alpha^* | U_1) \leq (4\delta)^{\frac{\kappa+1}{\kappa}}$. By definition of the conditional mutual information, and the factorization $\mathbb{P}_{X_1, \alpha^* | U_1} = \mathbb{P}_{\alpha^* | U_1} \mathbb{P}_{X_1 | \alpha^*, U_1}$, it holds that

$$I(X_1; \alpha^* | U_1) = \mathbb{E}_{U_1}\big[D_{\mathrm{KL}}(\mathbb{P}_{X_1 | \alpha^*, U_1} || \mathbb{P}_{X_1 | U_1})\big] .$$

Assume a random vector α is uniform on $\mathcal{V}$, by the convexity of KL divergence, it then follows that

$$D_{\mathrm{KL}}(\mathbb{P}_{X_1 | \alpha^*, U_1} || \mathbb{P}_{X_1 | U_1}) \leq \frac{1}{|\mathcal{V}|} \sum_{\alpha \in \mathcal{V}} D_{\mathrm{KL}}(\mathbb{P}_{X_1 | \alpha^*, U_1} || \mathbb{P}_{X_1 | \alpha, U_1}) .$$

For any pair $\alpha^*, \alpha \in \mathcal{V}$, the KL divergence $D_{\mathrm{KL}}(\mathbb{P}_{X_1 | \alpha^*, U_1} || \mathbb{P}_{X_1 | \alpha, U_1})$ can be at most the KL divergence between a pair of Bernoulli variables with parameters

$$1 - \frac{2 + \alpha_i}{4} (4\delta)^{\frac{\kappa+1}{\kappa}}, \quad \text{and} \quad 1 - \frac{2 + \alpha_j}{4} (4\delta)^{\frac{\kappa+1}{\kappa}}, \quad \forall \alpha_i, \alpha_j \in \mathcal{V}.$$

By Lemma 21 (setting $p = (4\delta)^{\frac{\kappa+1}{\kappa}}$), we have $D_{\mathrm{KL}}(\mathbb{P}_{X_1 | \alpha^*, U_1} || \mathbb{P}_{X_1 | \alpha, U_1}) \leq (4\delta)^{\frac{\kappa+1}{\kappa}}$. This complete the proof.

■

Lemma 23 (Lower bound 1 with $d = 1$) *Given a constant $\kappa \in (0, 1]$ and a parameter $\alpha^* \in \mathcal{V}$, where $\mathcal{V} = \{-1, +1\}$, the oracle ϕ generates the data sequence $\{X_t\}_{t=1}^T$ where X_t are i.i.d random variables following from the Bernoulli distribution with parameter $1 - \frac{2+\alpha^*}{4}(4\delta)^{\frac{\kappa+1}{\kappa}}$. Then, for any $\delta \in (0, \frac{1}{8}]$, it holds for any estimator $\hat{\alpha} \in \mathcal{V}$ based on the data sequence $\{X_t\}_{t=1}^T$ that*

$$\max_{\alpha^* \in \mathcal{V}} \mathbb{P}(\hat{\alpha} \neq \alpha^*) \geq \frac{1}{2}\left(1 - \sqrt{\frac{(4\delta)^{\frac{\kappa+1}{\kappa}} T}{2}}\right).$$

Proof [Proof of Lemma 23] Set $p = (4\delta)^{\frac{\kappa+1}{\kappa}}$. Define $\hat{\alpha}' := 1 - \frac{2+\hat{\alpha}}{4}p, \alpha^{*\prime} := 1 - \frac{2+\alpha^*}{4}p$. It then follows that $\mathbb{P}(\hat{\alpha} \neq \alpha^*) = \mathbb{P}(\hat{\alpha}' \neq \alpha^{*\prime})$. Note that

$$\mathbb{E}[|\hat{\alpha}' - \alpha^{*\prime}|] = \frac{1}{2}p\mathbb{P}(\hat{\alpha}' \neq \alpha^{*\prime}).$$

Based on the proof of Lemma 4 in Agarwal et al. (2012), we have

$$\max_{\alpha^{*\prime} \in \{1-\frac{1}{4}p, 1-\frac{3}{4}p\}} \mathbb{E}[|\hat{\alpha}' - \alpha^{*\prime}|] \geq \frac{1}{4}p\left(1 - \frac{1}{2}\sqrt{2TD_{\mathrm{KL}}\left(\mathbb{P}^+ || \mathbb{P}^-\right)}\right),$$

where $\mathbb{P}^+, \mathbb{P}^-$ denote the Bernoulli distributions $\textsc{bernoulli}(1 - \frac{3}{4}p)$ and $\textsc{bernoulli}(1 - \frac{1}{4}p)$, respectively. Combining these two displays with Lemma 21 gives

$$\begin{aligned}\max_{\alpha^{*\prime} \in \{1-\frac{1}{4}p, 1-\frac{3}{4}p\}} \mathbb{P}(\hat{\alpha}' \neq \alpha^{*\prime}) &= \frac{\max_{\alpha^{*\prime} \in \{1-\frac{1}{4}p, 1-\frac{3}{4}p\}} \mathbb{E}[|\hat{\alpha}' - \alpha^{*\prime}|]}{\frac{1}{2}p} \\ &\geq \frac{1}{2}\left(1 - \sqrt{\frac{(4\delta)^{\frac{\kappa+1}{\kappa}} T}{2}}\right).\end{aligned}$$

as desired. ■

Lemma 24 (Lower bound 2 with $d \geq 2$) *Suppose the vector $\alpha^* = (\alpha_1^*, \ldots, \alpha_d^*)^\top$ is chosen uniformly at random from the set $\mathcal{V}$, where $\mathcal{V}$ is a subset of the hypercube $\{-1, +1\}^d$ such that $\Delta_{\mathrm{H}}(\alpha, \tilde{\alpha}) = \sum_{i=1}^d \mathbb{1}\{\alpha_i \neq \tilde{\alpha}_i\} \geq \frac{d}{4}$ for any $\alpha, \tilde{\alpha} \in \mathcal{V}$. Set the parameter*

$$\tilde{\alpha}^* = \left(\frac{1}{2} + \alpha_1^*\delta, \ldots, \frac{1}{2} + \alpha_d^*\delta\right)^\top.$$

Given the parameter $\tilde{\alpha}^*$, a constant $\delta \in (0, \frac{1}{100}]$, and the time horizon T, at each round $t = 1, \ldots, T$, the oracle ϕ flips a coin with bias $\frac{1}{T}$ (the probability of the coin landing heads up is $\frac{1}{T}$) at first. If the coin has a head, the oracle tosses set of d coins with bias $\tilde{\alpha}^*$, and then reveal the outcomes of the d coins. If the coin has a tail, the oracle reveals nothing. When $d \geq 2$, it holds for any estimator $\hat{\alpha} \in \mathcal{V}$ that

$$\mathbb{P}(\hat{\alpha} \neq \alpha^*) \geq 1 - \frac{16d\delta^2 + \log 2}{d/8}.$$

Here, the probability is taken over the randomness of α^* and ϕ.

Proof [Proof of Lemma 24] Let $U_t \in \{0,1\}$ following the Bernoulli distribution with parameter $\frac{1}{T}$ be the random variable indicating whether the oracle reveals the information. Let $X_t := (X_{t,1}, \ldots, X_{t,d})^\top$ denote the outcome of oracle's coin toss at time t with the components $X_{t,i} \in \{0,1\}$ denote the outcome for coordinate i. When $U_t = 0$, set $X_{t,i} = -1, i = 1, \ldots, d$. By (Pollard et al., 2012, Sec 15.3.2, Lemma 4) and (Scarlett and Cevher, 2019, Theorem 1), if the parameter α^* is uniform on $\mathcal{V}$, it holds for any estimator $\hat{\alpha} \in \mathcal{V}$ that

$$\mathbb{P}(\hat{\alpha} \neq \alpha^*) \geq 1 - \frac{I(\{U_t, X_t\}_{t=1}^T; \alpha^*) + \log 2}{\log |\mathcal{V}|},$$

where $I(\{U_t, X_t\}_{t=1}^T; \alpha^*)$ denotes the mutual information between the data sequence $\{U_t, X_t\}_{t=1}^T$ and α^*. By the Varshamov-Gilbert bound, there exists such a packing set $\mathcal{V} \subseteq \{-1, +1\}^d$ with $|\mathcal{V}| \geq \exp(\frac{d}{8})$ satisfies $\Delta_{\mathrm{H}}(\alpha, \tilde{\alpha}) = \sum_{i=1}^d \mathbb{1}\{\alpha_i \neq \tilde{\alpha}_i\} \geq \frac{d}{4}$ for any $\alpha, \tilde{\alpha} \in \mathcal{V}$. It suffice to show that $I(\{U_t, X_t\}_{t=1}^T; \alpha^*) \leq 16 d\delta^2$. By the independent and identically distributed the sampling, we have

$$I(\{U_t, X_t\}_{t=1}^T; \alpha^*) = \sum_{t=1}^T I\big((U_1, X_1); \alpha^*\big) = T I\big((U_1, X_1); \alpha^*\big).$$

By chain rule of mutual information and the sampling scheme, it holds that

$$I\big((U_1, X_1); \alpha^*\big) = I(X_1; \alpha^* | U_1) + I(\alpha^*; U_1).$$

Note that U_1 is sampled independent of α^*, this implies $I(\alpha^*; U_1) = 0$. It remains to show that $I(X_1; \alpha^* | U_1) \leq \frac{1}{T} 16 d\delta^2$. By definition of the conditional mutual information, and the factorization $\mathbb{P}_{X_1, \alpha^* | U_1} = \mathbb{P}_{\alpha^* | U_1} \mathbb{P}_{X_1 | \alpha^*, U_1}$, it holds that

$$I(X_1; \alpha^* | U_1) = \mathbb{E}_{U_1}\left[D_{\mathrm{KL}}(\mathbb{P}_{X_1 | \alpha^*, U_1} || \mathbb{P}_{X_1 | U_1}) \right].$$

Assume a random vector α is uniform on $\mathcal{V}$, by the convexity of KL divergence, it then follows that

$$D_{\mathrm{KL}}(\mathbb{P}_{X_1 | \alpha^*, U_1} || \mathbb{P}_{X_1 | U_1}) \leq \frac{1}{|\mathcal{V}|} \sum_{\alpha \in \mathcal{V}} D_{\mathrm{KL}}(\mathbb{P}_{X_1 | \alpha^*, U_1} || \mathbb{P}_{X_1 | \alpha, U_1}).$$

Combing these two displays with fact that $U_1 \sim \textsc{bernoulli}(\frac{1}{T})$ gives

$$\begin{aligned}
I(X_1; \alpha^* | U_1) \leq & \frac{1}{|\mathcal{V}|} \sum_{\alpha \in \mathcal{V}} \mathbb{E}_{U_1} D_{\mathrm{KL}}(\mathbb{P}_{X_1 | \alpha^*, U_1} || \mathbb{P}_{X_1 | \alpha, U_1})\big] \\
\leq & \frac{1}{T} \frac{1}{|\mathcal{V}|} \sum_{\alpha \in \mathcal{V}} D_{\mathrm{KL}}(\mathbb{P}_{X_1 | \alpha^*, U_1 = 1} || \mathbb{P}_{X_1 | \alpha, U_1 = 1}) \\
& + \Big(1 - \frac{1}{T}\Big) \frac{1}{|\mathcal{V}|} \sum_{\alpha \in \mathcal{V}} D_{\mathrm{KL}}(\mathbb{P}_{X_1 | \alpha^*, U_1 = 0} || \mathbb{P}_{X_1 | \alpha, U_1 = 0}) \\
= & \frac{1}{T} \frac{1}{|\mathcal{V}|} \sum_{\alpha \in \mathcal{V}} D_{\mathrm{KL}}(\mathbb{P}_{X_1 | \alpha^*, U_1 = 1} || \mathbb{P}_{X_1 | \alpha, U_1 = 1}).
\end{aligned}$$

For any pair $\alpha^*, \alpha \in \mathcal{V}$, the KL divergence $D_{\mathrm{KL}}(\mathbb{P}_{X_1 | \alpha^*, U_1} || \mathbb{P}_{X_1 | \alpha, U_1})$ can be at most the KL divergence between d independent pairs of Bernoulli variables with parameters $\frac{1}{2} + \delta$ and $\frac{1}{2} - \delta$. By Lemma 3 in Agarwal et al. (2012), it holds that

$$D_{\mathrm{KL}}(\mathbb{P}_{X_1 | \alpha^*, U_1 = 1} || \mathbb{P}_{X_1 | \alpha, U_1 = 1}) \leq 16 d\delta^2. \tag{D.1}$$

Thus, we have

$$I(X_1; \alpha^* | U_1) \leq \frac{1}{T} 16 d \delta^2$$

as desired.

■

Lemma 25 (Lower bound 2 with $d = 1$) *Given a parameter $\alpha^* \in \mathcal{V}$, where $\mathcal{V} = \{-1, +1\}$, a constant $\delta \in (0, \frac{1}{100}]$, and the time horizon T. At each round $t = 1, \ldots, T$, the oracle ϕ flips a coin with probability of getting heads being $\frac{1}{T}$. If the coin lands on heads, the oracle tosses a coin with bias $\frac{1}{2} + \alpha^* \delta$ and then reveal the outcome. If the coin has a tail, the oracle reveals nothing. Then, it holds for any estimator $\hat{\alpha} \in \mathcal{V}$ that*

$$\max_{\alpha^* \in \mathcal{V}} \mathbb{P}(\hat{\alpha} \neq \alpha^*) \geq 1 - \sqrt{8\delta^2}\,.$$

Proof [Proof of Lemma 25] Define $\hat{\alpha}' := \frac{1}{2} + \hat{\alpha}\delta, \alpha^{*\prime} := \frac{1}{2} + \alpha^* \delta$. It then follows that $\mathbb{P}(\hat{\alpha} \neq \alpha^*) = \mathbb{P}(\hat{\alpha}' \neq \alpha^{*\prime})$. Note that

$$\mathbb{E}[|\hat{\alpha}' - \alpha^{*\prime}|] = 2\delta \mathbb{P}(\hat{\alpha}' \neq \alpha^{*\prime})\,.$$

Based on the proof of (Agarwal et al., 2012, Lemma 4) and display (D.1), we have

$$\max_{\alpha^{*\prime} \in \{\frac{1}{2}+\delta, \frac{1}{2}-\delta\}} \mathbb{E}[|\hat{\alpha}' - \alpha^{*\prime}|] \geq 2\delta\big(1 - \sqrt{8\delta^2}\big)\,,$$

Combining these two displays gives

$$\max_{\alpha^{*\prime} \in \{\frac{1}{2}+\delta, \frac{1}{2}-\delta\}} \mathbb{P}(\hat{\alpha}' \neq \alpha^{*\prime}) \geq 1 - \sqrt{8\delta^2}$$

as desired.

■

D.2. Proofs of minimax lower bounds

We are now ready to prove the minimax lower bounds. In this section, we use the subscript i to denote the i-th digit of a vector and use the superscript t to denote the time index. For instance, given the t-th iteration $x^t \in \mathbb{R}^d$, x_i^t represents the i-th element of x^t.

Proof [Proof of Theorem 9]

Proof of lower bound (1)
At first, we consider the special case $\mathcal{S} = S_\infty(R)$. The proof consists four steps. We first construct a subclass of functions parametrized by a subset of the vertices of a d-dimensional hypercube with finite cardinality. Then, we construct a stochastic oracle based on Bernoulli random variables, each of which corresponds to the parameters of the constructed function in the previous step. Next, we convert the parameter estimation to the stochastic optimization problem by showing that optimizing any function in this subclass to certain tolerance requires identifying the hypercube vertices. Finally,

we employ Fano types of inequality to lower bound the probability of misspecification error, along with the results obtained in the previous steps, to finish the proof. The four mentioned steps now read in detail.

1. Construct a subclass of functions

Assume $\mathcal{V} \subseteq \{-1,+1\}^d$ is a subset of the hypercube such that

$$\Delta_{\mathrm{H}}(\alpha,\tilde{\alpha}) = \sum_{i=1}^{d} \mathbb{1}\{\alpha_i \neq \tilde{\alpha}_i\} \geq \frac{d}{4},$$

for any $\alpha, \tilde{\alpha} \in \mathcal{V}$. Given a vector $\alpha = (\alpha_1, \ldots, \alpha_d)^\top \in \mathcal{V}$, consider the convex function $g_\alpha(x) : \mathcal{S} \to \mathbb{R}$ defined via

$$g_\alpha(x) := \frac{L}{d} \sum_{i=1}^{d} \frac{2+\alpha_i}{4} \delta \Big\{ (1+\alpha_i)|x_i + R| + (1-\alpha_i)|x_i - R| \Big\}$$

with $\delta \in (0, \frac{1}{8}]$. Define the function $h(\alpha, x)$ via

$$\begin{aligned} h : \{-1,+1\} \times \mathcal{S} &\to [0,\infty) \\ (\alpha, x) &\mapsto \frac{1}{2}\big[(1+\alpha)|x+R| + (1-\alpha)|x-R|\big]\,. \end{aligned}$$

We then have $|\nabla h(\alpha, x)| \leq 1$. Hence, it holds for any $q \in [1, 1+\kappa]$ that

$$\|\nabla g_\alpha(x)\|_q \leq \frac{L}{d} \Big(\sum_{i=1}^{d} \big(\frac{2+\alpha_i}{2} \delta |\nabla h_i(\alpha_i, x_i)| \big)^q \Big)^{\frac{1}{q}} \leq L\,.$$

This implies $g_\alpha(x)$ is L-Lipschitz with respect to q^* norm, where q^* satisfies $\frac{1}{q} + \frac{1}{q^*} = 1$. It follows that $g_\alpha \in \mathcal{H}_{cvx}, \forall \alpha \in \mathcal{V}$. Define the function class $\mathcal{G}(\delta) := \{g_\alpha : \alpha \in \mathcal{V}\}$. Set

$$p := \big(4\delta\big)^{\frac{\kappa+1}{\kappa}}, \quad \text{and} \quad \Lambda := \frac{1}{2} p^{-\frac{1}{1+\kappa}}, \quad \text{where} \quad \kappa \in (0,1].$$

It then follows that

$$p \in (0, 1/2]\,, \quad p\Lambda = 2\delta \in (0, 1/4] \quad \text{and} \quad g_\alpha(x) = \frac{L}{d} \sum_{i=1}^{d} \frac{2+\alpha_i}{4} p\Lambda h(\alpha_i, x_i)\,.$$

2. Construct an oracle

Now, we describe the stochastic first order oracle ϕ which satisfies the conditions stated in Assumption 1. Given a vector $\alpha \in \mathcal{V}$, consider the oracle ϕ that returns noisy value and gradient sample as following for $t = 1. \ldots .T$:

1). Pick an index $i_t \in \{1, \ldots, d\}$ uniformly.
2). Draw $b_{i_t} \in \{0,1\}$ according to $\textsc{bernoulli}\big(1 - \frac{2+\alpha_{i_t}}{4} p\big)$.
3). For the given input $x \in \mathcal{S}$, return the function value $\hat{g}_\alpha(x) = L(1-b_{i_t})\Lambda h(\alpha_{i_t}, x)$ and its subgradient.

Now, we verify the constructed oracle satisfies the conditions stated in Assumption 1. Note that

$$\mathbb{E}[\hat{g}_\alpha(x^t)|\mathcal{F}_t] = \frac{L}{d}\sum_{i=1}^{d}\frac{2+\alpha_i}{4}p\Lambda h(\alpha_i, x_i^t) = g_\alpha(x^t)\,.$$

Moreover, note that

$$\frac{\partial}{\partial x_i}L(1-b_i)\Lambda h(\alpha_i, x_i) = L(1-b_i)\Lambda\nabla h(\alpha_i, x_i)\,.$$

We then find

$$\mathbb{E}[\nabla\hat{g}_\alpha(x^t)|\mathcal{F}_t] = \nabla g_\alpha(x^t)\,,$$

and

$$\mathbb{E}[\|\nabla\hat{g}_\alpha(x^t)\|_q^{1+\kappa}|\mathcal{F}_t] \le \frac{L^{1+\kappa}}{d}\sum_{i=1}^{d}\Lambda^{\kappa+1}\frac{2+\alpha_i}{4}p \le L^{1+\kappa}, \quad \forall q\in[1,1+\kappa]\,.$$

3. Optimizing well is equivalent to function identification

In this step, we employ the same quantification of the function separation as in Agarwal et al. (2012). Define the discrepancy measure between two functions f, g over the same domain $\mathcal{S}$ as

$$\rho(f,g) := \inf_{x\in\mathcal{S}}[f(x)+g(x)-f(x_f^*)-g(x_g^*)]\,.$$

Given the function class $\mathcal{G}(\delta)$, define $\psi(\mathcal{G}(\delta)) := \min_{\alpha\neq\beta\in\mathcal{V}}\rho(g_\alpha, g_\beta)$.

Given an vector $\alpha^*\in\mathcal{V}$, we have corresponding function g_{α^*}. Suppose the method M_T makes T queries to the oracle ϕ, and thus obtains the information sequence $\{\phi(x^1; g_{\alpha^*}), \ldots, \phi(x^T; g_{\alpha^*})\}$, denoted by $\phi(x_T^1; g_{\alpha^*})$. By (Agarwal et al., 2012, Lemma 2), for any method $M_T\in\mathcal{M}_T$ one can construct a hypothesis test $\hat{\alpha}: \phi(x_T^1; g_{\alpha^*}) \to \mathcal{V}$ such that

$$\mathbb{P}_\phi\big(\hat{\alpha}(M_T)\neq\alpha^*\big) \le \mathbb{P}_\phi\Big(\epsilon(M_T, g_{\alpha^*}, \mathcal{S}, \phi) \ge \frac{\psi(\mathcal{G}(\delta))}{3}\Big), \forall\alpha^*\in\mathcal{V}\,.$$

This implies

$$\frac{1}{|\mathcal{V}|}\sum_{\alpha^*\in\mathcal{V}}\mathbb{P}_\phi\big(\hat{\alpha}(M_T)\neq\alpha^*\big) \le \frac{1}{|\mathcal{V}|}\sum_{\alpha^*\in\mathcal{V}}\mathbb{P}_\phi\Big(\epsilon(M_T, g_{\alpha^*}, \mathcal{S}, \phi) \ge \frac{\psi(\mathcal{G}(\delta))}{3}\Big)\,. \tag{D.2}$$

Moreover, by the definition of $\epsilon^*(\mathcal{H}_{cvx}, \mathcal{S}, \phi)$, we have

$$\epsilon^*(\mathcal{H}_{cvx}, \mathcal{S}, \phi) \ge \inf_{M_T\in\mathcal{M}_T}\sup_{\alpha^*\in\mathcal{V}}\mathbb{E}_\phi[\epsilon(M_T, g_{\alpha^*}, \mathcal{S}, \phi)]\,.$$

By Markov's inequality, we then find

$$\mathbb{E}_\phi[\epsilon(M_T, g_{\alpha^*}, \mathcal{S}, \phi)] \ge \frac{\psi(\mathcal{G}(\delta))}{3}\mathbb{P}_\phi\Big(\epsilon(M_T, g_{\alpha^*}, \mathcal{S}, \phi) > \frac{\psi(\mathcal{G}(\delta))}{3}\Big)\,.$$

Combining this with previous display provides us with

$$\begin{aligned}\epsilon^*(\mathcal{H}_{cvx}, \mathcal{S}, \phi) \geq & \inf_{M_T \in \mathcal{M}_T} \sup_{\alpha^* \in \mathcal{V}} \frac{\psi(\mathcal{G}(\delta))}{3} \mathbb{P}_\phi\Big(\epsilon(M_T, g_{\alpha^*}, \mathcal{S}, \phi) > \frac{\psi(\mathcal{G}(\delta))}{3}\Big) \\ \geq & \frac{\psi(\mathcal{G}(\delta))}{3} \inf_{M_T \in \mathcal{M}_T} \frac{1}{|\mathcal{V}|} \sum_{\alpha^* \in \mathcal{V}} \mathbb{P}_\phi\Big(\epsilon(M_T, g_{\alpha^*}, \mathcal{S}, \phi) > \frac{\psi(\mathcal{G}(\delta))}{3}\Big)\,.\end{aligned}$$

Plugging inequality (D.2) into it gives

$$\epsilon^*(\mathcal{H}_{cvx}, \mathcal{S}, \phi) \geq \frac{\psi(\mathcal{G}(\delta))}{3} \inf_{M_T \in \mathcal{M}_T} \frac{1}{|\mathcal{V}|} \sum_{\alpha^* \in \mathcal{V}} \mathbb{P}_\phi\big(\hat{\alpha}(M_T) \neq \alpha^*\big)\,,$$

which implies

$$\epsilon^*(\mathcal{H}_{cvx}, \mathcal{S}, \phi) \geq \frac{\psi(\mathcal{G}(\delta))}{3} \inf_{\hat{\alpha} \in \mathcal{V}} \frac{1}{|\mathcal{V}|} \sum_{\alpha^* \in \mathcal{V}} \mathbb{P}_\phi\big(\hat{\alpha}(M_T) \neq \alpha^*\big)\,. \tag{D.3}$$

In the next step, we will finish the proof by providing the lower bounds for the discrepancy $\psi(\mathcal{G}(\delta))$ and the probability $\inf_{\hat{\alpha} \in \mathcal{V}} \frac{1}{|\mathcal{V}|} \sum_{\alpha^* \in \mathcal{V}} \mathbb{P}_\phi\big(\hat{\alpha}(M_T) \neq \alpha^*\big)$ with some specific choice of δ.

4. Complete the proof

Note that the minimizer of $g_\alpha(x)$ is $x^*_\alpha = -R\alpha$, and $\min_{x \in \mathcal{S}} g_\alpha(x) = 0$. Then, it holds that

$$\begin{aligned}& g_\alpha(x) + g_\beta(x) - g_\alpha(x^*_\alpha) - g_\beta(x^*_\beta) \\ =& \frac{L}{d} \sum_{i=1}^{d} \left\{ \frac{2 + \alpha_i}{4} p\Lambda h(\alpha_i, x_i) + \frac{2 + \beta_i}{4} p\Lambda h(\beta_i, x_i) \right\} \\ =:& \sum_{i=1}^{d} I(x_i; \alpha_i, \beta_i)\,,\end{aligned}$$

where $I(x_i; \alpha_i, \beta_i) := \frac{L}{d}\Big\{ \frac{2+\alpha_i}{4} p\Lambda h(\alpha_i, x_i) + \frac{2+\beta_i}{4} p\Lambda h(\beta_i, x_i) \Big\}$. When $\alpha_i = \beta_i$, it holds that $\min_{x \in \mathcal{S}} I(x; \alpha_i, \beta_i) = 0$. When $\alpha_i \neq \beta_i$, it holds that

$$I(x; \alpha_i, \beta_i) = \frac{L\delta}{d}\Big\{ \frac{3}{2}|x_i + R| + \frac{1}{2}|x_i - R| \Big\}\,,$$

it then follows that $\min_{x \in \mathcal{S}} I(x; \alpha_i, \beta_i) = \frac{L\delta}{d} R$. Thus, we obtain

$$\rho(g_\alpha, g_\beta) = \frac{RL\delta}{d} \Delta_{\mathrm{H}}(\alpha, \beta) \geq \frac{RL\delta}{d} \frac{d}{4} = \frac{RL\delta}{4}\,,$$

which implies

$$\psi(\mathcal{G}(\delta)) \geq \frac{RL\delta}{4}\,.$$

Recall that we obtain the following in step 3

$$\epsilon^*(\mathcal{H}_{cvx}, \mathcal{S}, \phi) \geq \frac{\psi(\mathcal{G}(\delta))}{3} \inf_{\hat{\alpha} \in \mathcal{V}} \frac{1}{|\mathcal{V}|} \sum_{\alpha^* \in \mathcal{V}} \mathbb{P}_\phi\big(\hat{\alpha}(M_T) \neq \alpha^*\big)\,.$$

Combining the previous two displays gives

$$\epsilon^*(\mathcal{H}_{cvx}, \mathcal{S}, \phi) \geq \frac{RL\delta}{12} \inf_{\hat{\alpha} \in \mathcal{V}} \frac{1}{|\mathcal{V}|} \sum_{\alpha^* \in \mathcal{V}} \mathbb{P}_\phi\big(\hat{\alpha}(M_T) \neq \alpha^*\big)\,. \tag{D.4}$$

When $d > 8$, invoking Lemma 22 yields

$$\epsilon^*(\mathcal{H}_{cvx}, \mathcal{S}, \phi) \geq \frac{RL\delta}{12}\Big(1 - \frac{(4\delta)^{\frac{\kappa+1}{\kappa}} T + \log 2}{d/8}\Big)\,. \tag{D.5}$$

Let $T \geq d$ with $d \geq 9$, and set $\delta := \frac{1}{32}\Big(\frac{d}{T}\Big)^{\frac{\kappa}{1+\kappa}}$. It then follows that

$$0 < \delta \leq \frac{1}{8}\,,$$

and

$$\frac{(4\delta)^{\frac{\kappa+1}{\kappa}} T + \log 2}{d/8} \leq \frac{8}{8^{\frac{\kappa+1}{\kappa}}} + \frac{8\log 2}{d} \leq \frac{3}{4}\,.$$

Plugging these into display (D.5) then gives

$$\epsilon^*(\mathcal{H}_{cvx}, \mathcal{S}, \phi) \geq \frac{1}{1536} RL\Big(\frac{d}{T}\Big)^{\frac{\kappa}{1+\kappa}}\,.$$

When $d < 9$, we restrict to the case where $d = 1$. The lower bounds corresponding $1 < d \leq 8$ can be established based on the case of $d = 1$. Combining the lower bound in Lemma 23 with the display (D.4) gives

$$\epsilon^*(\mathcal{H}_{cvx}, \mathcal{S}, \phi) \geq \frac{RL\delta}{12}\frac{1}{2}\left(1 - \sqrt{\frac{(4\delta)^{\frac{\kappa+1}{\kappa}} T}{2}}\right).$$

Set $\delta := \frac{1}{32} T^{-\frac{\kappa}{1+\kappa}}$. Then we have $\delta \in (0, 1/8]$ and

$$\sqrt{\frac{(4\delta)^{\frac{\kappa+1}{\kappa}} T}{2}} \leq \sqrt{\Big(\frac{1}{8}\Big)^{\frac{1+\kappa}{\kappa}} \frac{1}{2}} \leq \frac{1}{10}\,.$$

Combing these two displays yields

$$\epsilon^*(\mathcal{H}_{cvx}, \mathcal{S}, \phi) \geq \frac{9}{7680} RL\Big(\frac{1}{T}\Big)^{\frac{\kappa}{1+\kappa}}\,.$$

This completes the proof for the special case $\mathcal{S} = S_\infty(R)$. Note that the Lipschitz constant of g_α does not depend on $\mathcal{S}$, $x^*_\alpha = \arg\min_{x \in S} g_\alpha(x) \in \mathcal{S}$, and thus the preceding proof goes through when $\mathcal{S} \supseteq S_\infty(R)$. Hence, the desired general claim follows.

Proof of lower bound (2)

The proof strategy is similar to the proof of lower bound (1), but with a different function class and the first-order oracle. At first, we consider the special case $\mathcal{S} = S_\infty(R)$. The proof consists

four steps as follows.

1. Construct a subclass of functions
Assume $\mathcal{V} \subseteq \{-1,+1\}^d$ is a subset of the hypercube such that

$$\Delta_{\mathrm{H}}(\alpha,\tilde{\alpha}) = \sum_{i=1}^{d} \mathbb{1}\{\alpha_i \neq \tilde{\alpha}_i\} \geq \frac{d}{4},$$

for any $\alpha, \tilde{\alpha} \in \mathcal{V}$. Given the time horizon T, and a vector $\alpha = (\alpha_1, \ldots, \alpha_d)^\top \in \mathcal{V}$, we consider the convex function $g_\alpha(x) : \mathcal{S} \to \mathbb{R}$ defined via

$$g_\alpha(x) := \frac{L}{T^{\frac{\kappa}{1+\kappa}} d^{\frac{1}{q}}} \sum_{i=1}^{d} \Big\{ \Big(\frac{1}{2} + \alpha_i \delta\Big)|x_i + R| + \Big(\frac{1}{2} - \alpha_i \delta\Big)|x_i - R| \Big\}$$

with $\delta \in (0, 1/100]$. Define the function $h(\alpha, x)$ via

$$\begin{aligned} h : \{-1,+1\} \times \mathcal{S} &\to [0,\infty) \\ (\alpha, x) &\mapsto \big(\frac{1}{2} + \alpha\delta\big)|x + R| + \big(\frac{1}{2} - \alpha\delta\big)|x - R| \,. \end{aligned}$$

We then have $|\nabla h(\alpha, x)| \leq 1$. Hence, it holds for any $q \in [1, \infty]$ that

$$\|\nabla g_\alpha(x)\|_q \leq \frac{L}{T^{\frac{\kappa}{1+\kappa}} d^{\frac{1}{q}}} \Big(\sum_{i=1}^{d} |\nabla h_i|^q \Big)^{\frac{1}{q}} \leq L \,.$$

This implies $g_\alpha(x)$ is L-Lipschitz with respect to q^* norm, where q^* satisfies $\frac{1}{q} + \frac{1}{q^*} = 1$. It follows that $g_\alpha \in \mathcal{H}_{cvx}, \forall \alpha \in \mathcal{V}$. Define the function class $\mathcal{G}(\delta) := \{g_\alpha : \alpha \in \mathcal{V}\}$.

2. Construct an oracle

Now, we describe the stochastic first order oracle ϕ which satisfies the conditions stated in Assumption 1. Given the time horizon T, and a vector $\alpha \in \mathcal{V}$, consider the oracle ϕ that returns noisy value and gradient sample as following for $t = 1, \ldots, T$:

1). Draw $Y_t \in \{0, 1\}$ according to $\textsc{bernoulli}\Big(\frac{1}{T}\Big)$.
2a). When $Y_t = 1$, draw $b_i \in \{0, 1\}$ according to $\textsc{bernoulli}\big(\frac{1}{2} + \alpha_i\delta\big), i = 1, \ldots, d$. For the given input $x \in \mathcal{S}$, return the function value

$$\hat{g}_\alpha(x) = LT^{\frac{1}{1+\kappa}} d^{-\frac{1}{q}} \sum_{i=1}^{d} \big\{ b_i|x_i + R| + (1 - b_i)|x_i - R| \big\}$$

and its subgradient.
2b). When $Y_t = 0$, for any input $x \in \mathcal{S}$, return $\hat{g}_\alpha(x) = 0$ and its subgradient.

Now, we verify the conditions in Assumption 1 for the constructed oracle. It is obvious that

$$\mathbb{E}[\hat{g}_\alpha(x^t)|\mathcal{F}_t] = g_\alpha(x^t) \,.$$

and

$$\mathbb{E}[\nabla \hat{g}_\alpha(x^t)|\mathcal{F}_t] = \nabla g_\alpha(x^t)\,.$$

Moreover, it holds that

$$\frac{\partial}{\partial x}\Big(b_i|x+R| + (1-b_i)|x-R|\Big) \le 1\,.$$

It then follows that

$$\mathbb{E}\big[\|\nabla \hat{g}_\alpha(x^t)\|_q^{1+\kappa}|\mathcal{F}_t\big] \le \frac{1}{T}L^{1+\kappa}Td^{-\frac{1+\kappa}{q}}d^{\frac{1+\kappa}{q}} = L^{1+\kappa}\,.$$

3. Optimizing well is equivalent to function identification
In this step, we employ the same quantification of the function separation as in step 3 of the proof of Theorem 9, where the discrepancy measure between two functions f, g over the same domain $\mathcal{S}$ is

$$\rho(f,g) = \inf_{x\in\mathcal{S}}[f(x)+g(x)-f(x_f^*)-g(x_g^*)]\,.$$

Given the function class $\mathcal{G}(\delta)$, define $\psi(\mathcal{G}(\delta)) := \min_{\alpha\neq\beta\in\mathcal{V}}\rho(g_\alpha, g_\beta)$. Invoking display (D.3), we have

$$\epsilon^*(\mathcal{H}_{cvx},\mathcal{S},\phi) \ge \frac{\psi(\mathcal{G}(\delta))}{3}\inf_{\hat{\alpha}\in\mathcal{V}}\frac{1}{|\mathcal{V}|}\sum_{\alpha^*\in\mathcal{V}}\mathbb{P}_\phi\big(\hat{\alpha}(M_T)\neq\alpha^*\big)\,.$$

In the next step, we will finish the proof by providing the lower bounds for the discrepancy $\psi(\mathcal{G}(\delta))$ and the probability $\inf_{\hat{\alpha}\in\mathcal{V}}\frac{1}{|\mathcal{V}|}\sum_{\alpha^*\in\mathcal{V}}\mathbb{P}_\phi\big(\hat{\alpha}(M_T)\neq\alpha^*\big)$ with some specific choice of δ.

4. Complete the proof
We note that the function $g_\alpha(x)$ is a specification of the function class considered in part (a) of in (Agarwal et al., 2012, Theorem 1)

$$g_\alpha(x) := \frac{c}{d}\sum_{i=1}^{d}\Big\{\Big(\frac{1}{2}+\alpha_i\delta\Big)|x_i+R| + \Big(\frac{1}{2}-\alpha_i\delta\Big)|x_i-R|\Big\}$$

by setting $c = \frac{Ld}{T^{\frac{\kappa}{1+\kappa}}d^{\frac{1}{q}}}$. By the last display in the proof of Theorem 1 of Agarwal et al. (2012), it holds that $\rho(g_\alpha, g_\beta) \ge \frac{cR\delta}{2}, \forall\alpha\neq\beta\in\mathcal{V}$. We then have

$$\psi(\mathcal{G}(\delta)) \ge \frac{1}{2}R\delta LT^{-\frac{\kappa}{1+\kappa}}d^{1-\frac{1}{q}}\,.$$

Recall that we obtain the following in step 3

$$\epsilon^*(\mathcal{H}_{cvx},\mathcal{S},\phi) \ge \frac{\psi(\mathcal{G}(\delta))}{3}\inf_{\hat{\alpha}\in\mathcal{V}}\frac{1}{|\mathcal{V}|}\sum_{\alpha^*\in\mathcal{V}}\mathbb{P}_\phi\big(\hat{\alpha}(M_T)\neq\alpha^*\big)\,.$$

Combining the previous two displays gives

$$\epsilon^*(\mathcal{H}_{cvx}, \mathcal{S}, \phi) \geq \frac{1}{6} R\delta L T^{-\frac{\kappa}{1+\kappa}} d^{1-\frac{1}{q}} \inf_{\hat{\alpha} \in \mathcal{V}} \frac{1}{|\mathcal{V}|} \sum_{\alpha^* \in \mathcal{V}} \mathbb{P}_\phi\big(\hat{\alpha}(M_T) \neq \alpha^*\big)\,. \tag{D.6}$$

When $d > 8$, invoking Lemma 24 yields

$$\epsilon^*(\mathcal{H}_{cvx}, \mathcal{S}, \phi) \geq \frac{1}{6} R\delta L T^{-\frac{\kappa}{1+\kappa}} d^{1-\frac{1}{q}} \Big(1 - \frac{16d\delta^2 + \log 2}{d/8}\Big)\,. \tag{D.7}$$

Note that when $d \geq 9$, and set $\delta = \frac{1}{100}$, it holds that

$$1 - \frac{16d\delta^2 + \log 2}{d/8} = 1 - 128\delta^2 - 8\frac{\log 2}{d} = 1 - \frac{128}{10000} - \log 2 \geq \frac{1}{4}\,.$$

Plugging these into display (D.7) then gives

$$\epsilon^*(\mathcal{H}_{cvx}, \mathcal{S}, \phi) \geq \frac{1}{2400} R L T^{-\frac{\kappa}{1+\kappa}} d^{1-\frac{1}{q}}\,.$$

When $d < 9$, we restrict to the case where $d = 1$. Combining the lower bound derived in Lemma 25 with display (D.6) gives

$$\epsilon^*(\mathcal{H}_{cvx}, \mathcal{S}, \phi) \geq \frac{1}{6} R\delta L T^{-\frac{\kappa}{1+\kappa}} \big(1 - \sqrt{8\delta^2}\big)\,.$$

When $\delta = \frac{1}{100}$, it holds that

$$\epsilon^*(\mathcal{H}_{cvx}, \mathcal{S}, \phi) \geq \frac{1}{1200} R L \Big(\frac{1}{T}\Big)^{\frac{\kappa}{1+\kappa}}\,.$$

This completes the proof for the special case $\mathcal{S} = S_\infty(R)$. Note that the Lipschitz constant of g_α does not depend on $\mathcal{S}$, $x^*_\alpha = \arg\min_{x \in S} g_\alpha(x) \in \mathcal{S}$, and thus the preceding proof goes through when $\mathcal{S} \supseteq S_\infty(R)$. Hence, the desired general claim follows.

■