Graphical Convergence of Subgradients in Nonconvex Optimization and Learning

Damek Davis *   Dmitriy Drusvyatskiy †

Abstract

We investigate the stochastic optimization problem of minimizing population risk, where the loss defining the risk is assumed to be weakly convex. Compositions of Lipschitz convex functions with smooth maps are the primary examples of such losses. We analyze the estimation quality of such nonsmooth and nonconvex problems by their sample average approximations. Our main results establish dimension-dependent rates on subgradient estimation in full generality and dimension-independent rates when the loss is a generalized linear model. As an application of the developed techniques, we analyze the nonsmooth landscape of a robust nonlinear regression problem.

Key words: subdifferential, stability, population risk, sample average approximation, weak convexity, Moreau envelope, graphical convergence

1 Introduction

Traditional machine learning theory quantifies how well a decision rule, learned from a limited data sample, generalizes to the entire population. The decision rule itself may enable the learner to correctly classify (as in image recognition) or predict the value of continuous statistics (as in regression) of previously unseen data samples. A standard mathematical formulation of this problem associates to each decision rule x and each sample z, a loss $f(x, z)$, which may for example penalize misclassification of the data point by the decision rule. Then the learner seeks to minimize the *regularized population risk*:

$$\min_x \; \varphi(x) = f(x) + r(x) \qquad \text{where} \qquad f(x) = \mathbb{E}_{z \sim P}\left[f(x, z)\right]. \qquad (1.1)$$

Here, $r \colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is an auxiliary function defined on $\mathbb{R}^d$ that may encode geometric constraints or promote low-complexity structure (e.g. sparsity or low-rank) on x. The main assumption is that the only access to the population data is by drawing i.i.d. samples from

P. Numerical methods then seek to obtain a high-quality solution estimate for (1.1) using as few samples as possible. Algorithmic strategies for (1.1) break down along two lines: streaming strategies and regularized empirical risk minimization (ERM).

Streaming algorithms in each iteration update a solution estimate of (1.1) based on drawing a relatively small batch of samples. Streaming algorithms deviate from each other in precisely how the sample is used in the update step. The proximal stochastic subgradient method [17, 29, 47] is one popular streaming algorithm, although there are many others, such as the stochastic proximal point and Gauss-Newton methods [16, 25, 70]. In contrast, algorithms based on regularized ERM, draw a large sample $S = \{z_1, z_2, \ldots, z_m\}$ at the onset, and then output the solution of the deterministic problem

$$\min_{x \in \mathbb{R}^d} \ \varphi_S(x) := f_S(x) + r(x) \qquad \text{where} \qquad f_S(x) := \frac{1}{m} \sum_{i=1}^m f(x, z_i). \qquad (1.2)$$

Solution methodologies for (1.2) depend on the structure of the loss function. One generic approach, often used in practice, is to apply a streaming algorithm directly to (1.2) by interpreting $f_S(\cdot)$ as an expectation over the discrete distribution over the samples $\{z_i\}_{i=1}^n$ and performing multiple passes through the sampled data. Our current work focuses on the ERM strategy, though it is strongly influenced by recent progress on streaming algorithms.

The success of the ERM approach rests on knowing that the minimizer of the surrogate problem (1.2) is nearly optimal for the true learning task (1.1). Quantitative estimates of this type are often based on a uniform convergence principle. For example, when the functions $f(\cdot, z)$ are L-Lipschitz continuous for a.e. $z \sim P$, then with probability $1 - \gamma$, the estimate holds [60, Theorem 5]:

$$\sup_{x: \|x\| \leq R} |f(x) - f_S(x)| \leq \widetilde{\mathcal{O}}\left(\sqrt{\frac{L^2 R^2 d}{m} \cdot \log\left(\frac{d}{\gamma}\right)}\right). \qquad (1.3)$$

An important use of the bound in (1.3) is to provide a threshold beyond which algorithms for the surrogate problem (1.2) should terminate, since further accuracy on the ERM problem does not improve the accuracy on the true learning task. It is natural to ask if under stronger assumptions, learning is possible with sample complexity that is independent of the ambient dimension d. In the landmark paper [60], the authors showed that the answer is indeed yes when the functions $f(\cdot, z)$ are convex and one incorporates further strongly convex regularization. Namely, under an appropriate choice of the parameter $\lambda > 0$, the solution of the quadratically regularized problem

$$\hat{x}_S := \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ \varphi_S(x) + \lambda \|x\|^2 \right\}, \qquad (1.4)$$

satisfies

$$\varphi(\hat{x}_S) - \inf \varphi \leq \mathcal{O}\left(\sqrt{\frac{L^2 R^2}{\gamma m}}\right) \qquad (1.5)$$

with probability $1 - \gamma$, where R is the diameter of the domain of r. In contrast to previous work, the proof of this estimate is not based on uniform convergence. Indeed, uniform

2

convergence in function values may fail in infinite dimensions even for convex learning problems. Instead, the property underlying the dimension independent bound (1.5) is that the solution $\hat{x}_S$ of the quadratically regularized ERM (1.4) is stable in the sense of Bousquet and Elisseff [12]. That is, the solution $\hat{x}_S$ does not change much under an arbitrary perturbation of a single sample z_i. Stability of quadratically regularized ERM will also play a central role in our work for reasons that will become clear shortly.

The aforementioned bounds on the accuracy of regularized ERM are only meaningful if one can globally solve the deterministic problems (1.2) or (1.4). Convexity certainly facilitates global optimization techniques. Many problems of contemporary interest, however, are nonconvex, thereby making ERM-based learning rules intractable. When the functions $f(\cdot, z)$ are not convex but smooth, the most one can hope for is to find points that are critical for the problem (1.2). Consequently, it may be more informative to estimate the deviation in the gradients, $\sup_{x:\|x\|\leq R} \|\nabla F(x) - \nabla F_S(x)\|$, along with deviations in higher-order derivatives when they exist. Such uniform bounds have recently appeared in [27, 44].

When the loss $f(\cdot, z)$ is neither smooth nor convex, the situation becomes less clear. Indeed, one should reassess what "uniform convergence of gradients" should mean in light of obtaining termination criteria for algorithms on the regularized ERM problem. As the starting point, one may replace the gradient by a generalized subdifferential $\partial \varphi(x)$ in the sense of nonsmooth and variational analysis [45, 57]. Then the minimal norms, $\text{dist}(0, \partial \varphi(x))$ and $\text{dist}(0, \partial \varphi_S(x))$, could serve as stationarity measures akin to the norm of the gradient in smooth minimization. One may then posit that the stationarity measures, $\text{dist}(0, \partial \varphi(x))$ and $\text{dist}(0, \partial \varphi_S(x))$, are uniformly close with high probability when the sample size is large. Pointwise convergence is indeed known to hold (e.g. [64, Theorem 7.54]). On the other hand, to our best knowledge, all results on uniform convergence of the stationarity measure are asymptotic and require extra assumptions, such as polyhedrality for example [52]. The main obstacle is that the function $x \mapsto \text{dist}(0, \partial \varphi(x))$ is highly discontinuous. We refer the reader to [64, pp. 380] for a discussion. Indeed, the need to look beyond pointwise uniform convergence is well-documented in optimization and variational analysis [2, 4]. One remedy is to instead focus on graphical convergence concepts. Namely, one could posit that the Hausdorff distance between the subdifferential graphs, $\text{gph}\,\partial \varphi$ and $\text{gph}\,\partial \varphi_S$, tends to zero. Here, we take a closely related approach, but aiming for finite-sample bounds.

In this work, we aim to provide tight threshold estimates beyond which algorithms on (1.2) should terminate. In contrast to previous literature, however, we will allow the loss function to be both nonconvex and nonsmooth. The only serious assumption we make is that $f(\cdot, z)$ is a ρ-weakly convex function for a.e. $z \sim P$, by which we mean that the assignment $x \mapsto f(x, z) + \frac{\rho}{2}\|x\|^2$ is convex. The class of weakly convex functions is broad and its importance in optimization is well documented [1, 48, 49, 56, 58].[1] It trivially includes all convex functions and all C^1-smooth functions with Lipschitz gradient. More broadly, it includes all compositions $f(x, z) = h(c(x, z), z)$, where $h(\cdot, z)$ is convex and Lipschitz, and $c(\cdot, z)$ is C^1-smooth with Lipschitz Jacobian. Robust principal component analysis, phase retrieval, blind deconvolution, sparse dictionary learning, and minimization of risk measures naturally lead to stochastic weakly convex minimization problems. We refer the interested

---

[1]Weakly convex functions also go by other names such as lower-C^2, uniformly prox-regularity, t^2-paraconvex, and semiconvex.

reader to [16, Section 2.1] and [22] for detailed examples.

The approach we take is based on a smoothing technique, well-familiar to optimization specialists. For any function g, define the Moreau envelope and the proximal map:

$$g_t(x) := \min_y \left\{ g(y) + \frac{1}{2t}\|y - x\|^2 \right\}, \qquad \text{prox}_{tg}(x) := \operatorname*{argmin}_y \left\{ g(y) + \frac{1}{2t}\|y - x\|^2 \right\}.$$

It is well-known from [46], where both constructions were introduced, that if g is ρ-weakly convex and $t < \frac{1}{\rho}$, then the envelope g_t is C^1-smooth with gradient

$$\nabla g_t(x) = t^{-1}(x - \text{prox}_{tg}(x)).$$

Note that $\nabla g_t(x)$ is in principle computable by solving a convex optimization problem in the definition of the proximal point $\text{prox}_{tg}(x)$.

Our main result (Theorem 4.4) shows that with probability $1 - \gamma$, the estimate holds

$$\sup_{x:\,\|x\|\leq R} \|\nabla\varphi_{1/2\rho}(x) - \nabla(\varphi_S)_{1/2\rho}(x)\| \leq \mathcal{O}\left(\sqrt{\frac{L^2 d}{m} \log\left(\frac{R\rho m}{\gamma}\right)} \right). \qquad (1.6)$$

The guarantee (1.6) is appealing: even though the subgradients of φ and φ_S may be far apart pointwise, the gradients of the smooth approximations $\varphi_{1/2\rho}$ and $(\varphi_S)_{1/2\rho}$ are uniformly close at a controlled rate governed by the sample size. Moreover, (1.6) directly implies estimates on the Hausdorff distance between subdifferential graphs, $\text{gph}\,\partial\varphi$ and $\text{gph}\,\partial\varphi_S$, as we alluded to above. Indeed, the subdifferential graph is related to the graph of the proximal map by a linear isomorphism. The guarantee (1.6) is also perfectly in line with the recent progress on streaming algorithms [16, 18, 20, 76]. These works showed that a variety of popular streaming algorithms (e.g. stochastic subgradient, Gauss-Newton, and proximal point) drive the gradient of the Moreau envelope to zero at a controlled rate. Consequently, the estimate (1.6) provides a tight threshold beyond which such streaming algorithms on the regularized ERM problem (1.2) should terminate. The proof of (1.6) uses only the most elementary techniques: stability of quadratically regularized ERM [60], McDiarmid's inequality [41], and a covering argument.

It is intriguing to ask when the dimension dependence in the bound (1.6) can be avoided. For example, for certain types of losses (e.g. modeling a linear predictor) there are well-known dimension independent bounds on uniform convergence in function values. Can we therefore obtain dimension independent bounds in similar circumstances, but on the deviations $\|\nabla\varphi_{1/2\rho} - \nabla(\varphi_S)_{1/2\rho}\|$? The main tool we use to address this question is entirely deterministic. We will show that if φ and φ_S are uniformly δ close, then the gradients $\nabla\varphi_{1/2\rho}$ and $\nabla(\varphi_S)_{1/2\rho}$ are uniformly $O(\sqrt{\delta})$ close, as well as their subdifferential graphs in the Hausdorff distance. We illustrate the use of such bounds with two examples.

As the first example, consider the loss f modeling a generalized linear model (GLM):

$$f(x, z) = \ell(\langle x, \phi(z)\rangle, z).$$

Here ϕ is some feature map and $\ell(\cdot, z)$ is a loss function. It is well-known that if $\ell(\cdot, z)$ is Lipschitz, then the empirical function values $f_S(x)$ converge uniformly to the population

4

values $f(x)$ at a dimension-independent rate that scales as $m^{-1/2}$ in the sample size. We thus deduce that the gradient $\nabla(\varphi_S)_{1/2\rho}$ converges uniformly to $\nabla\varphi_{1/2\rho}$ at the rate $m^{-1/4}$. We leave it as an intriguing open question whether this rate can be improved to $m^{-1/2}$. The second example analyzes the landscape of a robust nonlinear regression problem, wherein we observe a series of nonlinear measurements $\sigma(\langle \bar{x}, z \rangle)$ of input data $\bar{x}$, possibly with adversarial corruption. Using the aforementioned techniques, we will show that under mild distributional assumptions on z, every stationary point of the associated nonsmooth nonconvex empirical risk is within a small ball around $\bar{x}$.

Though, in the discussion above, we focused on the norm that is induced by an inner product, the techniques we present apply much more broadly to Bregman divergences. In particular, any Bregman divergence generates an associated smoothing of the empirical and population risks, making our techniques applicable under high order growth of the loss function and under non-euclidean geometries. The outline of the paper is as follows. In Section 2, we introduce our notation as well as several key concepts including Bregman divergences and subdifferentials. In Section 3, we describe the problem setting. In Section 4, we describe a general procedure, based on algorithmic stability, for obtaining dimension dependent rates on the error between the gradients of the Moreau envelopes of the population and subsampled objectives. In Section 5, we illustrate the techniques of the previous section by obtaining dimension independent rates for generalized linear models and analyzing the landscape of a robust nonlinear regression problem.

Related literature

This paper builds on the vast literature on sample average approximations found in the stochastic programming and statistical learning literature. The results in these communities are similar in many respects, but differ in their focus on convergence criteria. In the stochastic programming literature, much attention has been given to the convergence of (approximate) minimizers and optimal values both in the distributional and almost sure limiting sense [28, 33, 34, 50, 53, 54, 61–63]. In contrast, the statistical learning community puts a greater emphasis on excess risk bounds that hold with high probability, often with minimal or no dependence on dimension [13, 30, 32, 37, 42, 43, 51, 60, 66, 67, 71, 77].

Several previous works have studied (sub)gradient based convergence, as we do here. For example, [74] proves nonasymptotic, dimension dependent high probability bounds on the distance between the empirical and population subdifferential under the Hausdorff metric. The main assumption in this work, however, essentially requires smoothness of the population objective. The work [75] takes a different approach, directly smoothing the empirical losses $f(x, z)$. They show that the limit of the gradients of a certain smoothing of the empirical risk converges to an element of the population subdifferential. No finite-sample bounds are developed in [75]. The most general asymptotic convergence result that we are aware of is presented in [65]. There, the authors show that with probability one, the limit of a certain enlarged subdifferential of the empirical lose converges to an enlarged subdifferential of the population risk under the Hausdorff metric.

The two works most closely related to this paper are more recent. The paper [44] proves high probability uniform convergence of gradients for smooth objectives under the assumption that the gradient $\nabla f(x, z)$ is sub-Gaussian with respect to the population data. The

bounds presented in [44] are all dimension dependent and rely on covering arguments. The more recent paper [27], on the other hand, provides dimension independent high probability uniform rates of convergence of gradients for smooth Lipschitz generalized linear models. The main technical tool developed [27] is a "chain rule" for Radamacher complexity. We note that, in contrast to the $m^{-1/4}$ rates developed in this paper, [27] obtains rates of the form $m^{-1/2}$ for smooth generalized linear models.

2 Preliminaries

Throughout, we follow standard notation from convex analysis, as set out for example by Rockafellar [55]. The symbol $\mathbb{R}^d$ will denote a d-dimensional Euclidean space with inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\|x\|_2 = \sqrt{\langle x, x \rangle}$. Given any other norm $\| \cdot \|$, the symbol $\| \cdot \|_*$ will denote the dual norm. For any set $Q \subset \mathbb{R}^d$, we let $\operatorname{int} Q$ and $\operatorname{cl} Q$ stand for the interior and closure of Q, respectively. The symbol $\operatorname{ri} Q$ will refer to the interior of Q relative to its affine hull. The closed unit ball and the unit simplex in $\mathbb{R}^d$ will be denoted by $\mathbf{B}$ and Δ, respectively. The effective domain of any function $f \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, denoted by $\operatorname{dom} f$, consists of all points where f is finite. The indicator function of any set $Q \subset \mathbb{R}^d$, denoted ι_Q, is defined to be zero on Q and $+\infty$ on it.

2.1 Bregman divergence

In this work, we will use techniques based on the Bregman divergence, as is now standard in optimization and machine learning literature. For more details on the topic, we refer the reader to the expository articles [14, 31, 69]. To this end, henceforth, we fix a *Legendre function* $\Phi \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, meaning:

1. (Convexity) Φ is proper, closed, and strictly convex.

2. (Essential smoothness) The domain of Φ has nonempty interior, Φ is differentiable on $\operatorname{int}(\operatorname{dom} \Phi)$ and for any sequence $\{x_k\} \subset \operatorname{int}(\operatorname{dom} \Phi)$ converging to a boundary point of $\operatorname{dom} \Phi$, it must be the case that $\|\nabla \Phi(x_k)\|_2 \to \infty$.

Typical examples of Legendre functions are the squared Euclidean norm $\Phi(x) = \frac{1}{2}\|x\|_2^2$, convex polynomials $\Phi(x) = \|x\|_2 + \|x\|_2^r$, the Shannon entropy $\Phi(x) = \sum_{i=1}^d x_i \log(x_i)$ with $\operatorname{dom} \Phi = \mathbb{R}_+^d$, and the Burge function $\Phi(x) = -\sum_{i=1}^d \log(x_i)$ with $\operatorname{dom} \Phi = \mathbb{R}_{++}^d$. For more examples, see the articles [5, 9, 26, 68] or the recent survey [69].

The function Φ induces the *Bregman divergence* defined by

$$D_\Phi(y, x) := \Phi(y) - \Phi(x) - \langle \nabla \Phi(x), y - x \rangle,$$

for all $x \in \operatorname{int}(\operatorname{dom} \Phi)$, $y \in \operatorname{dom} \Phi$. In the classical setting $\Phi(\cdot) = \frac{1}{2}\| \cdot \|_2^2$, the divergence reduces to the square deviation $D_\Phi(y, x) = \frac{1}{2}\|y - x\|_2^2$. Notice that since Φ is strictly convex, equality $D_\Phi(y, x) = 0$ holds for some $x, y \in \operatorname{int}(\operatorname{dom} \Phi)$ if and only if $y = x$. Therefore, the quantity $D_\Phi(y, x)$ measures the proximity between the two points x, y. The divergence typically is asymmetric in x and y, and therefore we define the symmetrization

$$D_\Phi^{\mathrm{sym}}(x, y) := D_\Phi(x, y) + D_\Phi(y, x).$$

A function g is called *compatible* with Φ if the inclusion $\operatorname{ri}(\operatorname{dom} g) \subset \operatorname{int}(\operatorname{dom} \Phi)$ holds. Compatibility will be useful for guaranteeing that "proximal points" of g induced by Φ lie in $\operatorname{int}(\operatorname{dom} \Phi)$; see the forthcoming Theorem 3.1. Our focus will be primarily on those functions that can be convexified by adding a sufficiently large multiple of Φ. Formally, we will say that a function $g \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is *ρ-weakly convex relative to* Φ, for some $\rho \in \mathbb{R}$, if the perturbed function $x \mapsto g(x) + \rho\Phi(x)$ is convex. Similarly g is *α-strongly convex relative to* Φ if the function $x \mapsto g(x) - \alpha\Phi(x)$ is convex. The notions of relative weak/strong convexity are closely related to the generalized descent lemma introduced in [8] and the recent work on relative smoothness in [38, 39]. We postpone discussing examples of weakly convex functions to section 3.

2.2 Subdifferentials

First-order optimality conditions for nonsmooth and nonconvex problems are often most succinctly stated using subdifferentials. The *subdifferential* of a function g at a point $x \in \operatorname{dom} g$ is denoted by $\partial g(x)$ and consists of all vectors $v \in \mathbb{R}^d$ satisfying

$$g(y) \geq g(x) + \langle v, y - x \rangle + o(\|y - x\|) \qquad \text{as } y \to x.$$

When g is differentiable at x, the subdifferential reduces to the singleton $\partial g(x) = \{\nabla g(x)\}$, while for convex functions it reduces to the subdifferential in the sense of convex analysis.

We will call a point x *critical* for g if the inclusion $0 \in \partial g(x)$ holds. When g is ρ-weakly convex, the subdifferential automatically satisfies the seemingly stronger property [18, Lemma 2.2]:

$$g(y) \geq g(x) + \langle v, y - x \rangle - \rho D_\Phi(y, x). \tag{2.1}$$

for any $x \in \operatorname{int}(\operatorname{dom} \Phi), y \in \operatorname{dom} \Phi$, and any $v \in \partial g(x)$. It is often convenient to interpret the assignment $x \mapsto \partial g(x)$ as a set-valued map, and as such, it has a graph defined by

$$\operatorname{gph} \partial g(x) := \{(x, v) \in \mathbb{R}^d \times \mathbb{R}^d : v \in \partial g(x)\}.$$

3 Problem setting

Fix a probability space $(\Omega, \mathcal{F}, P)$. In this paper, we focus on the optimization problem

$$\min_{x \in \mathbb{R}^d} \varphi(x) = f(x) + r(x) \qquad \text{where} \qquad f(x) = \mathbb{E}_{z \sim P}\left[f(x, z)\right], \tag{3.1}$$

under the following assumptions on the functional components:

(A1) (**Bregman Divergence**) The Legendre function $\Phi \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ satisfies the compatibility condition, $\operatorname{ri}(\operatorname{dom} r) \subset \operatorname{int}(\operatorname{dom} \Phi)$. Moreover Φ is α-strongly convex with respect to some norm $\|\cdot\|$ on the set $\operatorname{dom} r \cap \operatorname{int}(\operatorname{dom} \Phi)$, meaning:

$$\frac{\alpha}{2}\|y - x\|^2 \leq D_\Phi(x, y) \qquad \text{for all } x, y \in \operatorname{dom} r \cap \operatorname{int}(\operatorname{dom} \Phi).$$

(A2) (**Weak Convexity**) The functions $f(\cdot, z) + r(\cdot)$ are ρ-weakly convex relative to Φ, for a.e. $z \in \Omega$, and are bounded from below.

(A3) (**Lipschitzian Property**) There exists a square integrable function $L\colon \Omega \to \mathbb{R}_+$ such that for all $x, y \in \text{int}(\text{dom}\,\Phi) \cap \text{dom}\,r$ and $z \in \Omega$, we have

$$|f(x, z) - f(y, z)| \leq L(z)\sqrt{D_\Phi^{\text{sym}}(x, y)},$$
$$\sqrt{\mathbb{E}\left[L(z)^2\right]} \leq \sigma.$$

The stochastic optimization problem (3.1) is the standard task of minimizing the regularized population risk. The function $f(x, z)$ is called the loss while $r\colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is a structure promoting regularizer. Alternatively r can encode feasibility constraints as an indicator function. Assumptions $(A1)$ and $(A2)$ are self-explanatory. Assumption $(A3)$ asserts control on the variation in the loss $f(\cdot, z)$. In the simplest setting when $\Phi = \frac{1}{2}\|\cdot\|_2^2$, Assumption $(A3)$ simply amounts to Lipschitz continuity of the loss $f(\cdot, z)$ on $\text{dom}\,r$ with a square integrable Lipschitz constant $L(z)$. The use of the Bregman divergence allows much more flexibility, as highlighted in the recent works [8, 18, 38, 39]. On one hand, it allows one to consider losses that are Lipschitz in a non-Euclidean norm, as long as Φ is strongly convex in that norm—a standard technique in machine learning and optimization. On the other hand, the Bregman divergence could also accommodate losses that that grow faster than linear. For example, suppose that the Lipschitz constant of the loss on bounded sets is polynomially controlled:

$$\frac{f(x, z) - f(y, z)}{\|x - y\|_2} \leq L(z)\sqrt{\frac{p(\|x\|_2) + p(\|y\|_2)}{2}} \qquad \text{for all distinct } x, y \in \mathbb{R}^d, z \in \Omega,$$

where $p(u) = \sum_{i=0}^r a_i u^i$ is some degree r univariate polynomial with nonnegative coefficients and $L(\cdot)$ is square integrable. Then the result [18, Proposition 3.2] shows that (A3) holds for the Legendre function

$$\Phi(x) := \sum_{i=0}^r a_i \left(\frac{3i + 7}{i + 2}\right) \|x\|_2^{i+2}. \tag{3.2}$$

We could in principle state and prove all the results of the paper in the Euclidean setting $\Phi = \frac{1}{2}\|\cdot\|_2^2$. On the other, the use of the Bregman divergence adds no complications whatsoever, and therefore we work in this greater generality. The reader should keep in mind, however, that all the results are of interest even in the Euclidean setting.

3.1 Convex compositions

The most important example of the problem class (3.1) corresponds to the setting when $r(\cdot)$ is convex and the loss has the form:

$$f(x, z) = h(c(x, z), z),$$

where $h(\cdot, z)$ is convex and $c(\cdot, z)$ is C^1-smooth. To see this, let us first consider the setting $\Phi = \frac{1}{2}\|\cdot\|_2^2$. Then provided that $h(\cdot, z)$ is ℓ-Lipschitz and the Jacobian $\nabla c(\cdot, z)$ is β-Lipschitz, a quick argument [23, Lemma 4.2] shows that the loss $f(\cdot, z)$ is $\ell\beta$-weakly convex relative to

8

Φ; therefore (A2) holds with $\rho = \ell\beta$. Moreover, if there exists a square integrable function $M(\cdot)$ satisfying $\|\nabla c(x, z)\|_{\mathrm{op}} \leq M(z)$ for all $x \in \mathrm{dom}\, r$ and $z \in \Omega$, then (A3) holds with $L(z) = \ell M(z)$.

The assumption that $h(\cdot, z)$ is Lipschitz continuous is mild and is often true in applications. On the other hand, Lipschitz continuity and boundedness of $\nabla c(\cdot, z)$ are strong assumptions. They can be relaxed by switching to a different Bregman divergence. Indeed, suppose that $h(\cdot, z)$ is convex and ℓ-Lipschitz as before, while $c(\cdot, z)$ now satisfies the two growth properties:

$$\frac{\|\nabla c(x, z) - \nabla c(y, z)\|_{\mathrm{op}}}{\|x - y\|_2} \leq p(\|x\|_2) + p(\|y\|_2) \qquad \forall x \neq y, \forall z \in \Omega$$

$$\|\nabla c(x, z)\|_{\mathrm{op}} \leq L_1(z) \cdot \sqrt{q(\|x\|_2)} \qquad \forall x, \forall z \in \Omega,$$

for some polynomials $p(u) = \sum_{i=}^{r} a_i u^i$ and $q(u) = \sum_{i=}^{r} b_i u^i$, with $a_i, b_i \geq 0$, and some square integrable function $L_1(\cdot)$. Define the Legendre function

$$\Phi(x) = \sum_{i=0}^{r} \frac{b_i + a_i(3i + 7)}{i + 2} \|x\|_2^{i+2}.$$

Then the result [18, Propositions 3.4, 3.5] shows that assumption (A2) and (A3) hold with $\rho = \ell$ and $L(z) = \sqrt{2}\ell L_1(z)$.

The class of composite problems is broad and has attracted some attention lately [16, 18–21, 23–25, 36, 76] as an appealing setting for nonsmooth and nonconvex optimization. The following table summarizes a few interesting problems of this type; details can all be found in the aforementioned works.

Problem	Loss function	Regularizer
Phase retrieval	$f(x, (a, b)) = \lvert\langle a, x\rangle^2 - b\rvert$	$r(x) = 0, \|x\|_1$
Blind deconvolution	$f((x, y), (u, v, b)) = \lvert\langle u, x\rangle\langle v, y\rangle - b\rvert$	—
Covariance estimation	$f(x, (U, b)) = \lvert\|Ux\|^2 - b\rvert$	—
Censored block model	$f(x, (ij, b)) = \lvert x_i x_j - b\rvert$	—
Conditional Value-at-Risk	$f((x, \gamma), z) = (\ell(x, z) - \gamma)^+$	$r(x, \gamma) = (1 - \alpha)\gamma$
Trimmed estimation	$f((x, w), i) = w_i f_i(x)$	$r(x, w) = \iota_{[0,1]^n \cap k\Delta}(w)$
Robust PCA	$f((U, V), (ij, b)) = \lvert\langle u_i, v_j\rangle - b\rvert$	—
Sparse dictionary learning	$f((D, x), z) = \|z - Dx\|_2$	$r(D, x) = \iota_{\mathbf{B}}(D) + \lambda\|x\|_1$

Table 1: Common stochastic weakly convex optimization problems.

3.2 The stationarity measure and implicit smoothing

Since the problem (3.1) is nonconvex and nonsmooth, typical algorithms can only be guaranteed to find critical points of the problem, meaning those satisfying $0 \in \partial\varphi(x)$. Therefore, one of our main goals is to estimate the Hausdorff distance between the subdifferential graphs, $\mathrm{gph}\,\partial\varphi$ and $\mathrm{gph}\,\partial\varphi_S$. We employ an indirect strategy based on a smoothing technique.

Setting the formalism, for any function $g \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, we define the Φ-*envelope* and the Φ-*proximal map*:

$$g_\lambda^\Phi(x) := \inf_y \left\{ g(y) + \frac{1}{\lambda} D_\Phi(y, x) \right\}, \qquad \mathrm{prox}_{\lambda g}^\Phi(x) := \operatorname*{argmin}_y \left\{ g(y) + \frac{1}{\lambda} D_\Phi(y, x) \right\},$$

respectively. Note that in the Euclidean setting $\Phi = \frac{1}{2} \| \cdot \|_2^2$, these two constructions reduce to the widely used Moreau envelope and the proximity map introduced in [46]. In this case, we will omit the subscript Φ from g_λ^Φ and $\mathrm{prox}_{\lambda g}^\Phi$. Note that in the Euclidean setting, the graphs of the proximal map and the subdifferential are closely related:

$$y = \mathrm{prox}_{\lambda g}(x) \iff (y, \lambda^{-1}(x - y)) \in \mathrm{gph}\, \partial g.$$

Consequently the graph of the proximal map $\mathrm{prox}_{\lambda g}$ is linearly isomorphic to the graph of the subdifferential ∂g by the linear map $(x, y) \mapsto (y, \lambda^{-1}(y - x))$. It is this observation that will allow us to pass from uniform estimates on the deviations $\|\mathrm{prox}_{\varphi/\lambda}(x) - \mathrm{prox}_{\varphi_S/\lambda}(x)\|$ to estimates on the Hausdorff distance between subdifferential graphs, $\mathrm{gph}\, \partial\varphi$ and $\mathrm{gph}\, \partial\varphi_S$.

It will be important to know when the set $\mathrm{prox}_{\lambda g}^\Phi(x)$ is a singleton lying in $\mathrm{int}(\mathrm{dom}\,\Phi)$. The following theorem follows quickly from [10], with a self-contained argument given in [18, Theorem 4.1]. The Euclidean version of the result was already proved in [46]. We will often appeal to this theorem without explicitly referencing it to shorten the exposition.

Theorem 3.1 (Smoothness of the Φ-envelope). *Consider a closed and lower-bounded function $g \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ that is ρ-weakly convex and compatible with some Legendre function $\Phi \colon \mathbb{R}^d \to \{\infty\}$. Then the following are true for any positive $\lambda < \rho^{-1}$ and any $x \in \mathrm{int}(\mathrm{dom}\,\Phi)$.*

- *The proximal point $\mathrm{prox}_{\lambda g}^\Phi(x)$ is uniquely defined and lies in $\mathrm{dom}\, g \cap \mathrm{int}(\mathrm{dom}\,\Phi)$.*

- *If Φ is twice differentiable on the interior of its domain, then the envelope g_λ^Φ is differentiable at x with gradient given by*

$$\nabla g_\lambda^\Phi(x) := \frac{1}{\lambda} \nabla^2 \Phi(x) \left(x - \mathrm{prox}_{\lambda g}^\Phi(x) \right). \tag{3.3}$$

The main application of Theorem 3.1 is to the functions φ and φ_S under the assumptions (A1)-(A3). Looking at the expression (3.3), it is clear that the deviation between x and $\mathrm{prox}_{\lambda g}^\Phi(x)$ provides an estimate on the size of the gradient $\nabla g_\lambda^\Phi(x)$. To make this observation precise, for any point $x \in \mathrm{int}(\mathrm{dom}\,\Phi)$, define the primal-dual pair of local norms:

$$\|y\|_x := \left\| \nabla^2 \Phi(x) y \right\|_*, \qquad \|v\|_x^* = \left\| \nabla^2 \Phi(x)^{-1} v \right\|.$$

Thus for any positive $\lambda < \rho^{-1}$ and $x \in \mathrm{int}(\mathrm{dom}\,\Phi)$, using (3.3), we obtain the estimate

$$\sqrt{D_\Phi \left(\mathrm{prox}_{\lambda g}^\Phi(x), x \right)} \geq \lambda \sqrt{\frac{\alpha}{2}} \cdot \|\nabla g_\lambda^\Phi(x)\|_x^*.$$

Consequently, following the recent literature on the subject, we will treat the quantity

$$D_\Phi(\mathrm{prox}_{\lambda g}^\Phi(x), x), \tag{3.4}$$

as a measure of approximate stationarity of g at x. In particular, when specializing to the target setting $g = \varphi$, it is this quantity (3.4) that streaming algorithms drive to zero at a controlled rate, as shown in [16, 18, 76].

4 Dimension Dependent Rates

In this section, we prove the uniform convergence bound (1.6), appropriately generalized using the Bregman divergence. The proof outline is as follows. First, in Theorem 4.1 we will estimate the expected error between the true proximal point and it's sampled version,

$$\mathbb{E}_S \|\mathrm{prox}_{\varphi/\lambda}(y) - \mathrm{prox}_{\varphi_S/\lambda}(y)\|,$$

where y is fixed. The same theorem also shows $\mathrm{prox}_{\varphi_S/\lambda}(y)$ is stable in the sense that $\mathrm{prox}_{\varphi_S/\lambda}(y)$ does not vary much when one sample $z_i \in S$ is changed arbitrarily. In the Euclidean setting, this is precisely the main result of [60] on stability of quadratically regularized ERM in stochastic convex optimization. Using McDiarmid's inequality [41] in Theorem 4.4, we will then deduce that the quantity $\|\mathrm{prox}_{\varphi/\lambda}(y) - \mathrm{prox}_{\varphi_S/\lambda}(y)\|$ concentrates for a fixed y. A covering argument over y in an appropriate norm will then complete the proof.

We begin following the outlined strategy with the following theorem.

Theorem 4.1 (Stability of regularized ERM)**.** *Consider a set $S = (z_1, \ldots, z_m)$ and define $S^i := (z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots z_m)$, where both the index i and the point $z_i' \in \Omega$ are arbitrary. Fix an arbitrary point y and set*

$$\mathcal{A}^* := \underset{y \in \mathbb{R}^d}{\mathrm{argmin}} \; \{\varphi(y) + \bar{\rho} D_\Phi(x, y)\} \qquad and \qquad \mathcal{A}(S) := \underset{y \in \mathbb{R}^d}{\mathrm{argmin}} \; \{\varphi_S(y) + \bar{\rho} D_\Phi(x, y)\}.$$

Then for any $\bar{\rho} > \rho$, the estimates hold:

$$\sqrt{D_\Phi^{\mathrm{sym}}(\mathcal{A}(S), \mathcal{A}(S^i))} \leq \frac{L(z_i) + L(z_i')}{(\bar{\rho} - \rho)m} \tag{4.1}$$

$$\mathbb{E}_S \left[D_\Phi(\mathcal{A}(S), \mathcal{A}^*) \right] \leq \frac{2\sigma^2}{(\bar{\rho} - \rho)^2 m} \tag{4.2}$$

$$0 \leq \mathbb{E}_S[\varphi_{1/\bar{\rho}}(y) - (\varphi_S)_{1/\bar{\rho}}(y)] \leq \frac{2\sigma^2}{(\bar{\rho} - \rho)m}. \tag{4.3}$$

Proof. We first verify (4.1). To this end, a quick computation yields for any points u and v the equation:

$$f_S(v) - f_S(u) = f_{S^i}(v) - f_{S^i}(u) + \frac{f(v, z_i) - f(u, z_i)}{m} + \frac{f(u, z_i') - f(v, z_i')}{m}. \tag{4.4}$$

Define now the regularized functions

$$\widehat{\varphi}(x) := \varphi(x) + \bar{\rho} D_\Phi(x, y) \qquad and \qquad \widehat{\varphi}_S(x) := \varphi_S(x) + \bar{\rho} D_\Phi(x, y).$$

Then adding $[r(v) + \bar{\rho} D_\Phi(v, y)] - [r(u) + \bar{\rho} D_\Phi(u, y)]$ to both sides of (4.4), we obtain

$$\widehat{\varphi}_S(v) - \widehat{\varphi}_S(u) = \widehat{\varphi}_{S^i}(v) - \widehat{\varphi}_{S^i}(u) + \frac{f(v, z_i) - f(u, z_i)}{m} + \frac{f(u, z_i') - f(v, z_i')}{m}.$$

Henceforth, set $v := \mathcal{A}(S^i)$ and $u := \mathcal{A}(S)$. Thus v is the minimizer of $\widehat{\varphi}_{S^i}$ and u is the minimizer of $\widehat{\varphi}_S$. Taking into account that both $\widehat{\varphi}_S(\cdot)$ and $\widehat{\varphi}_{S^i}(\cdot)$ are $(\bar{\rho} - \rho)$-strongly convex relative to Φ, we deduce

$$(\bar{\rho} - \rho)D_\Phi(v, u) \leq \widehat{\varphi}_S(v) - \widehat{\varphi}_S(u) \leq \frac{f(v, z_i) - f(u, z_i)}{m} + \frac{f(u, z_i') - f(v, z_i')}{m} - (\bar{\rho} - \rho)D_\Phi(u, v).$$

Rearranging, we arrive at the estimate

$$D_\Phi^{\mathrm{sym}}(u, v) \leq \frac{1}{\bar{\rho} - \rho} \left[\frac{f(v, z_i) - f(u, z_i)}{m} + \frac{f(u, z_i') - f(v, z_i')}{m} \right] \leq \frac{L(z_i) + L(z_i')}{(\bar{\rho} - \rho)m} \sqrt{D_\Phi^{\mathrm{sym}}(u, v)}.$$

Dividing through by $\sqrt{D_\Phi^{\mathrm{sym}}(u, v)}$, we obtain the claimed stability guarantee (4.1).

To establish (4.3), observe first

$$(\varphi_S)_{1/\bar{\rho}}(y) = \varphi_S(\mathcal{A}(S)) + \bar{\rho}D_\Phi(\mathcal{A}(S), y) \leq \varphi_S(x) + \bar{\rho}D_\Phi(x, y) \qquad \text{for all } x.$$

Taking expectations, we conclude $\mathbb{E}_S[(\varphi_S)_{1/\bar{\rho}}(y)] \leq \varphi_{1/\bar{\rho}}(y)$, which is precisely the left-hand-side of (4.3). Next, it is standard to verify the expression [59, Theorem 13.2]:

$$\begin{aligned}
\mathbb{E}_S[f(A(S))] &= \mathbb{E}_S[f_S(A(S))] + \mathbb{E}_S[f(A(S)) - f_S(A(S))] \\
&= \mathbb{E}_S[f_S(A(S))] + \mathbb{E}_{(S, z') \sim P, i \sim U(m)}[f(A(S^i), z_i) - f(A(S), z_i)],
\end{aligned} \tag{4.5}$$

where $U(m)$ denotes the discrete uniform distribution. Taking into account (4.1), we obtain

$$\begin{aligned}
|\mathbb{E}_S[\widehat{\varphi}(A(S)) - \widehat{\varphi}_S(A(S))]| &\leq \mathbb{E}\left[L(z)\sqrt{D_\Phi^{\mathrm{sym}}(\mathcal{A}(S), \mathcal{A}(S^i))} \right] \\
&\leq \sqrt{\mathbb{E}_z[L(z)^2]} \sqrt{\mathbb{E}_S[D_\Phi^{\mathrm{sym}}(\mathcal{A}(S), \mathcal{A}(S^i))]} \leq \frac{2\sigma^2}{(\bar{\rho} - \rho)m}. \tag{4.6}
\end{aligned}$$

Using the expressions, $\widehat{\varphi}(A(S)) \geq \varphi_{1/\bar{\rho}}(y)$ and $(\varphi_S)_{1/\bar{\rho}}(y) = \widehat{\varphi}_S(A(S))$, yields the right-hand-side of (4.3).

Finally noting that $\widehat{\varphi}$ is $(\bar{\rho} - \rho)$-strongly convex relative to Φ, we deduce

$$(\bar{\rho} - \rho)D_\Phi(\mathcal{A}(S), \mathcal{A}^*) \leq \widehat{\varphi}(\mathcal{A}(S)) - \min \widehat{\varphi}.$$

Taking expectation, and using the inequalities $\mathbb{E}_S[\widehat{\varphi}_S(A(S))] \leq \min \widehat{\varphi}$ and (4.6), we arrive at

$$(\bar{\rho} - \rho)\mathbb{E}_S[D_\Phi(\mathcal{A}(S), \mathcal{A}^*)] \leq \mathbb{E}_S[\widehat{\varphi}(\mathcal{A}(S)) - \min \widehat{\varphi}] \leq \mathbb{E}_S[\widehat{\varphi}(\mathcal{A}(S)) - \widehat{\varphi}_S(\mathcal{A}(S))] \leq \frac{2\sigma^2}{(\bar{\rho} - \rho)m}.$$

Thus we have established (4.2), and the proof is complete. $\qquad \square$

Next, we pass to high probability bounds on the deviation $\|\mathrm{prox}_{\varphi/\lambda}(y) - \mathrm{prox}_{\varphi_S/\lambda}(y)\|$ by means of McDiarmid's inequality [41]. The basic result reads a follows. Suppose that a function g satisfies the bounded difference property:

$$|g(z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_m) - g(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_m)| \leq c_i,$$

for all $i, z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_m, z_i'$, where c_i are some constants. Then for any independent random variables $X_1, \ldots, X_m$, the random variable $Y = g(X_1, \ldots, X_m)$ is unlikely to exceed its mean by much:

$$\mathbb{P}(Y - \mathbb{E}Y \geq t) \leq \exp\left(\frac{-2t^2}{\|c\|_2^2}\right).$$

A direct application of this inequality to $\|\mathrm{prox}_{\varphi/\lambda}(y) - \mathrm{prox}_{\varphi_S/\lambda}(y)\|$ using (4.1) would require the Lipschitz constant $L(\cdot)$ to be globally bounded. This could be a strong assumption, as it essentially requires the population data to be bounded. We will circumvent this difficulty by extending the McDiarmid's inequality to the setting where the constants c_i are replaced by some functions $\omega_i(\cdot, \cdot)$ of the data, z_i and z_i'. Let ε_i be a Rademacher random variable, meaning a random random variable taking value ± 1 with equal probability. Then as long as the symmetric random variables $\varepsilon_i \omega_i(Z_i, Z_i')$ have sufficiently light tails, a McDiarmid type bound will hold. In particular, we will be able to derive high probability upper bounds on the deviations $\|\mathrm{prox}_{\varphi/\lambda}(y) - \mathrm{prox}_{\varphi_S/\lambda}(y)\|$ only assuming that the random variable εL is sub-Gaussian. The proof follows known techniques for establishing McDiarmid's inequality, and in particular is essentially the same as that in [35, Theorem 1], though there the theorem was stated when ω_i is a metric and $\varepsilon_i \omega(z_i, z_i)$ is sub-Gaussian.

Henceforth, given a random variable X, we will let $\psi(\lambda) := \log(\mathbb{E}e^{\lambda X})$ denote the logarithm of the moment generating function. The symbol $\psi^\star \colon \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ will stand for the Fenchel conjugate of ψ, defined by $\psi^\star(t) = \sup_\lambda \{t\lambda - \psi(\lambda)\}$.

Theorem 4.2 (McDiarmid extended). *Let $z_1, \ldots, z_m$ be independent random variables with ranges $z_i \in \mathcal{Z}_i$. Suppose that there exist functions $\omega_i \colon \mathcal{Z}_i \times \mathcal{Z}_i \to \mathbb{R}_+$ such that the inequality*

$$|g(z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots) - g(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots)| \leq \omega(z_i, z_i'),$$

holds for all $z_j \in \mathcal{Z}_j$, $z_i, z_i' \in \mathcal{Z}_i$, and all $i, j \in \{1, \ldots, m\}$. Then the estimate holds:

$$\psi_{g(z) - \mathbb{E}[g(z)]}(\lambda) \leq \sum_{i=1}^m \psi_{\varepsilon_i \omega_i}(\lambda) \qquad \forall \lambda, \tag{4.7}$$

where ω_i denotes the random variable $\omega_i(z_i, z_i')$ and ε_i are i.i.d Rademacher random variables. In particular if $\omega_i = \omega_j$ for all indices i and j, then we have

$$\mathbb{P}(g(z) - \mathbb{E}[g(z)] \geq t) \leq \exp\left(-m\psi_{\varepsilon\omega}^\star\left(\frac{t}{m}\right)\right) \qquad \forall t \geq 0. \tag{4.8}$$

Proof. Define the Doob martingale sequence:

$$Y_0 := \mathbb{E}[g(z)] \qquad \text{and} \qquad Y_i := \mathbb{E}[g(z) \mid z_1, \ldots z_i] \text{ for } i = 1, \ldots, m,$$

and consider the martingale differences

$$V_i := Y_i - Y_{i-1} \qquad \text{for } i = 1, \ldots, m.$$

13

We aim to bound the moment generating function of $Y_m = g(z)$. To this end, observe

$$\mathbb{E}[e^{\lambda Y_i}] = \mathbb{E}\left[e^{\lambda Y_{i-1}}\mathbb{E}[e^{\lambda(Y_i - Y_{i-1})} \mid z_1, \ldots, z_{i-1}]\right]. \tag{4.9}$$

Thus, the crux of the proof is to bound the conditional expectation $\mathbb{E}[e^{\lambda V_i} \mid z_1, \ldots, z_{i-1}]$.

Form a vector z' from z by replacing z_i by an identical distributed copy z_i'. Clearly then

$$\mathbb{E}[g(z') \mid z_1, \ldots, z_i] = \mathbb{E}[g(z) \mid z_1, \ldots, z_{i-1}] = Y_{i-1}.$$

Therefore we may write $V_i = Y_i - Y_{i-1} = \mathbb{E}[g(z') - g(z) \mid z_1, \ldots, z_i]$. Hence, we deduce

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda V_i} \mid z_1, \ldots, z_{i-1}\right] &= \mathbb{E}\left[e^{\lambda \mathbb{E}[g(z) - g(z') \mid z_1 \ldots, z_i]} \mid z_1, \ldots, z_{i-1}\right] \\
&\leq \mathbb{E}\left[e^{\lambda(g(z) - g(z'))} \mid z_1, \ldots, z_{i-1}\right] \\
&= \mathbb{E}\left[\frac{1}{2}e^{\lambda(g(z) - g(z'))} + \frac{1}{2}e^{\lambda(g(z') - g(z))} \mid z_1, \ldots, z_{i-1}\right] \\
&= \mathbb{E}\left[\cosh(\lambda(g(z) - g(z'))) \mid z_1, \ldots, z_{i-1}\right] \\
&= \mathbb{E}\left[\cosh(\lambda|g(z) - g(z')|) \mid z_1, \ldots, z_{i-1}\right] \\
&\leq \mathbb{E}\left[\cosh(\lambda(\omega_i(z_i, z_i'))) \mid z_1, \ldots, z_{i-1}\right] \\
&= \mathbb{E}\left[\frac{1}{2}e^{\lambda(\omega_i(z_i, z_i'))} + \frac{1}{2}e^{-\lambda(\omega_i(z_i, z_i'))}\right] \\
&= \mathbb{E}\left[e^{\lambda \varepsilon_i \omega_i(z_i, z_i')}\right] = e^{\psi_{\varepsilon_i \omega_i}(\lambda)},
\end{aligned}$$

where the line follows by Jensen's inequality and the tower rule. Appealing to (4.9), and using induction, we therefore conclude

$$\mathbb{E}[e^{\lambda Y_m}] \leq e^{\psi_{\varepsilon_m \omega_m}(\lambda)}\mathbb{E}[e^{\lambda Y_{m-1}}] \leq \ldots \leq e^{\lambda \mathbb{E}[g(z)] + \sum_{i=1}^m \psi_{\varepsilon_i \omega_i}(\lambda)}.$$

Thus (4.7) is proved. The estimate (4.8) then follows by the standard Cramér-Chernoff bounding method. Namely, $\omega_i = \omega_j$ for all indices i and j. Then for every $t \geq 0$, Chernoff's inequality [11, Page 21] together with (4.7) implies

$$\mathbb{P}(g(z) - \mathbb{E}[g(z)] \geq t) \leq e^{-(m\psi_{\varepsilon\omega})^\star(t)}. \tag{4.10}$$

Noting the equality $(m\psi_{\varepsilon\omega})^\star(t) = m\psi_{\varepsilon\omega}^\star(\frac{t}{m})$ completes the proof. $\square$

The final ingredient is to perform a covering argument over the points y. To this end, we will need the following theorem that guarantees Lipschitz continuity of the proximal map in a metric $d_\Phi(x, y) := \|\nabla\Phi(x) - \nabla\Phi(y)\|_*$.

Lemma 4.3 (Lipschitz continuity). *Consider a closed and lower-bounded function $g: \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ that is ρ-weakly convex and compatible with some Legendre function $\Phi: \mathbb{R}^d \to \{\infty\}$. Assume in addition that Φ is α-strongly convex on $\operatorname{dom} g \cap \operatorname{int}(\operatorname{dom} \Phi)$. Then for any $\bar\rho > \rho$ and $x, y \in \operatorname{int}(\operatorname{dom} \Phi)$, we have*

$$\|\operatorname{prox}_{g/\bar\rho}(x) - \operatorname{prox}_{g/\bar\rho}(y)\| \leq \frac{\bar\rho}{\alpha(\bar\rho - \rho)}\|\nabla\Phi(x) - \nabla\Phi(y)\|_*.$$

Proof. For any x, y, set $\hat{x} = \text{prox}_{g/\bar{\rho}}(x)$ and $\hat{y} = \text{prox}_{g/\bar{\rho}}(y)$. Taking into account that $g + \bar{\rho}\Phi$ is $(\bar{\rho} - \rho)$ strongly convex relative to Φ, we deduce

$$(\bar{\rho} - \rho)D_\Phi(\hat{y}, \hat{x}) \leq (g(\hat{y}) + \bar{\rho}D_\Phi(\hat{y}, x)) - (g(\hat{x}) + \bar{\rho}D_\Phi(\hat{x}, x))$$
$$(\bar{\rho} - \rho)D_\Phi(\hat{x}, \hat{y}) \leq (g(\hat{x}) + \bar{\rho}D_\Phi(\hat{x}, y)) - (g(\hat{y}) + \bar{\rho}D_\Phi(\hat{y}, y)).$$

Adding these estimates together, we obtain

$$\begin{aligned}
\alpha(\bar{\rho} - \rho)\|\hat{x} - \hat{y}\|^2 &\leq (\bar{\rho} - \rho)(D_\Phi(\hat{y}, \hat{x}) + D_\Phi(\hat{x}, \hat{y}) \\
&\leq \bar{\rho}\left(D_\Phi(\hat{y}, x) - D_\Phi(\hat{x}, x) + D_\Phi(\hat{x}, y) - D_\Phi(\hat{y}, y)\right) \\
&= \bar{\rho}\langle \nabla\Phi(x) - \nabla\Phi(y), \hat{x} - \hat{y} \rangle \\
&\leq \bar{\rho}\|\nabla\Phi(x) - \nabla\Phi(y)\|^*\|\hat{x} - \hat{y}\|.
\end{aligned}$$

The result follows by dividing both sides by $\|\hat{x} - \hat{y}\|$. $\qquad\square$

We now have all the ingredients to prove the main result of this section. To this end, for any set $C \subseteq \text{int}(\text{dom}\,\Phi)$, we will let $\mathcal{N}(C, \Phi, \delta)$ denote the covering number of C in the metric $(x, y) \mapsto \|\nabla\Phi(x) - \nabla\Phi(y)\|_*$.

Theorem 4.4 (Concentration of the stationarity measure). *Let $C \subseteq \text{int}(\text{dom}\,\Phi)$ be any set. Then with probability*

$$1 - \mathcal{N}(C, \Phi, \delta)\exp\left(-m \cdot \psi^*_{\varepsilon L}\left(\frac{s}{\sqrt{m}}\right)\right),$$

we have

$$\sup_{y \in C} \|\text{prox}^\Phi_{\varphi_S/\bar{\rho}}(y) - \text{prox}^\Phi_{\varphi/\bar{\rho}}(y)\| \leq \sqrt{\frac{4(\sigma+s)^2}{\alpha(\bar{\rho}-\rho)^2 m}} + \frac{2\bar{\rho}\delta}{\alpha(\bar{\rho}-\rho)}.$$

If Φ is in addition twice differentiable on $\text{int}(\text{dom}\,\Phi)$, then with the same probability, the estimate holds:

$$\sup_{y \in C} \|\nabla(\varphi_S)^\Phi_{1/2\rho}(y) - \nabla\varphi^\Phi_{1/2\rho}(y)\|^*_x \leq \sqrt{\frac{16(\sigma+s)^2}{\alpha m}} + \frac{8\rho\delta}{\alpha}.$$

Proof. Following the notation of Theorem 4.1, set

$$\mathcal{A}(y, S) := \text{prox}_{\varphi_S/\bar{\rho}}(y) \qquad \text{and} \qquad \mathcal{A}^*(y) := \text{prox}_{\varphi/\bar{\rho}}(y).$$

Define the function

$$g(y, S) := \|\mathcal{A}(y, S) - \mathcal{A}^*(y)\|.$$

We will first apply Theorem 4.2 to the function $g(y, \cdot)\colon \Omega^m \to \mathbb{R}$, with y fixed. To verify the bounded difference property, we compute

$$\begin{aligned}
|g(y, S) - g(y, S^i)| &= \left|\|\mathcal{A}(y, S) - \mathcal{A}^*(y)\| - \|\mathcal{A}(y, S^i) - \mathcal{A}^*(y)\|\right|, \\
&\leq \|\mathcal{A}(y, S) - \mathcal{A}(y, S^i)\|, && (4.11) \\
&\leq \frac{L(z_i) + L(z_i')}{\sqrt{\alpha}(\bar{\rho} - \rho)m}, && (4.12)
\end{aligned}$$

where (4.11) uses the reverse triangle inequality, while (4.12) follows from the estimate (4.1) and strong convexity of Φ. Setting $\omega(z_i, z_i') = \frac{L(z_i) + L(z_i')}{\sqrt{\alpha}(\bar{\rho} - \rho)m}$, we deduce

$$\psi_{\varepsilon\omega}(\lambda) = \psi_{\varepsilon L}\left(\frac{2\lambda}{\sqrt{\alpha}(\bar{\rho} - \rho)m}\right) \qquad \text{and} \qquad \psi_{\varepsilon\omega}^{\star}(t) = \psi_{\varepsilon L}^{\star}\left(\sqrt{\alpha}(\bar{\rho} - \rho)mt/2\right).$$

Note moreover from (4.2) the bound $\mathbb{E}g(y, S) = \mathbb{E}\|\mathcal{A}(y, S) - \mathcal{A}^*(y)\| \leq \sqrt{\frac{4\sigma^2}{\alpha(\bar{\rho} - \rho)^2 m}}$. Thus, applying Theorem 4.2, we conclude

$$\mathbb{P}\left(g(y, S) \geq \sqrt{\frac{4\sigma^2}{\alpha(\bar{\rho} - \rho)^2 m}} + t\right) \leq \exp\left(-m\psi_{\varepsilon L}^{\star}\left(\sqrt{\alpha}(\bar{\rho} - \rho)t/2\right)\right).$$

Next we show using Lemma 4.3 that $g(\cdot, S)$ is $\frac{2\bar{\rho}}{\alpha(\bar{\rho} - \rho)}$-Lipschitz with respect to the metric $d_{\Phi}(x, y)$. Indeed, observe

$$|g(y, S) - g(y', S)| \leq \left|\|\mathcal{A}(y, S) - \mathcal{A}^*(y)\| - \|\mathcal{A}(y', S) - \mathcal{A}^*(y')\|\right|$$
$$\leq \|\mathcal{A}(y, S) - \mathcal{A}(y', S)\| + \|\mathcal{A}^*(y) - \mathcal{A}^*(y')\| \leq \frac{2\bar{\rho}}{\alpha(\bar{\rho} - \rho)}d_{\Phi}(y, y'),$$

where we used the triangle inequality and Lipschitz continuity of the proximal operator (Lemma 4.3). Let $\{y_i\}$ be a δ-net of C in the metric d_{Φ}. Thus for every y in a δ-ball of y_i, we have $g(y, S) \leq g(y_i, S) + \frac{2\bar{\rho}\delta}{\alpha(\bar{\rho} - \rho)}$. Taking a union bound over the cover, we therefore deduce

$$\mathbb{P}\left(g(y, S) \leq \sqrt{\frac{4\sigma^2}{\alpha(\bar{\rho} - \rho)^2 m}} + \frac{2\bar{\rho}\delta}{\alpha(\bar{\rho} - \rho)} + t\right) \geq 1 - \mathcal{N}(C, \Phi, \delta)\exp\left(-m\psi_{\varepsilon L}^{\star}\left(\sqrt{\alpha}(\bar{\rho} - \rho)t/2\right)\right).$$

Setting $t = \frac{2s}{\sqrt{\alpha}m(\bar{\rho} - \rho)}$ completes the proof. $\qquad\qquad\square$

Let us instantiate the Theorem 4.4 in the Euclidean setting $\Phi = \frac{1}{2}\|\cdot\|_2^2$ with C be being the unit ball of radius B. To this end, we will use the estimate, $\mathcal{N}(C, \Phi, \delta) \leq \left(1 + \frac{2B}{\delta}\right)^d$; see e.g. [73, Corollary 4.2.13]. We will also require the use of the sub-Gaussian norm of any random variable X, which is defined to be $\|X\|_{sg} := \inf\{t > 0 : \mathbb{E}\exp(X^2/t^2) \leq 2\}$, along with the sub-exponential norm $\|X\|_{se} := \inf\{t > 0 : \mathbb{E}\exp(|X|/t) \leq 2\}$.

Sub-Gaussian Lipschitz constant. Suppose that $L - \mathbb{E}L$ is a sub-Gaussian random variable with parameter $\nu = \|L - \mathbb{E}[L]\|_{sg}$. Using the triangle inequality, we therefore deduce

$$\|\varepsilon L\|_{sg} = \|L\|_{sg} \leq \|L - \mathbb{E}[L]\|_{sg} + \|\mathbb{E}[L]\|_{sg} \leq \nu + \frac{\sigma}{\sqrt{\ln 2}}.$$

Appealing to [73, Equation 2.16], we conclude $\psi_{\varepsilon L}(\lambda) \leq \frac{c}{2}(\nu + \sigma)^2 \lambda^2$ for all $\lambda \in \mathbb{R}$, where c is a numerical constant. Taking conjugates yields the relation $\psi_{\varepsilon L}^{\star}(t) \geq \frac{t^2}{c(\nu + \sigma)^2}$. Appealing to Theorem 4.4, while setting $s = \sqrt{c(\nu + \sigma)^2 \ln\left(\frac{\mathcal{N}(C, \Phi, \delta)}{\gamma}\right)}$ and $\delta = \frac{1}{\rho}\sqrt{\frac{d}{m}}$, we deduce that with probability $1 - \gamma$, the estimate holds:

$$\sup_{y \in C}\|\nabla(\varphi_S)_{1/2\rho}(y) - \nabla\varphi_{1/2\rho}(y)\| \lesssim \sqrt{\frac{\sigma^2 + d}{m} + \frac{(\nu + \sigma)^2 d}{m}\ln\left(\frac{R}{\gamma}\right)},$$

where we set $R := 1 + 2\rho B\sqrt{\frac{m}{d}}$.

16

Globally bounded Lipschitz constant As the next example, suppose that there exists a constant L satisfying $L(z) \leq \mathsf{L}$ for a.e. $z \in \Omega$. Then clearly we have $\sigma \leq \mathsf{L}$ and $\nu := \|L - \mathbb{E}[L]\|_{sg} \lesssim \mathsf{L}$. Consequently, we deduce that with $1 - \gamma$, the estimate holds:

$$\sup_{y \in C} \|\nabla(\varphi_S)_{1/2\rho}(y) - \nabla\varphi_{1/2\rho}(y)\| \lesssim \sqrt{\frac{L^2 + d}{m} + \frac{L^2 d}{m} \ln\left(\frac{R}{\gamma}\right)}.$$

where we set $R := 1 + 2\rho B\sqrt{\frac{m}{d}}$.

Sub-exponential Lipschitz constant As the final example, we suppose that $L - \mathbb{E}[L]$ is sub-exponential and set $\nu = \|L - \mathbb{E}[L]\|_{se}$. A completely analogous argument as in the sub-Gaussian case implies $\|\varepsilon L\|_{se} \leq \nu + \frac{\sigma}{\ln(2)}$. Appealing to [73, Proposition 2.7.1], we deduce $\psi_{\varepsilon L}(\lambda) \leq c^2(\nu + \sigma)^2\lambda^2$ for all $|\lambda| \leq \frac{1}{c(\nu+\sigma)}$. To simplify notation set $\eta := c(\nu + \sigma)$. Taking conjugates, we therefore deduce

$$\psi_{\varepsilon L}^\star(t) \geq \begin{cases} \frac{t^2}{4\eta^2} & \text{if } |t| \leq 2\eta \\ \frac{|t|}{\eta} - 1 & \text{otherwise} \end{cases}.$$

Consequently, we deduce the usual Bernstein-type bound $\psi_{\varepsilon L}^\star(t) \geq \min\{\frac{t^2}{4\eta^2}, \frac{t}{2\eta}\}$. Setting $s = 2\eta \cdot \max\left\{\sqrt{\ln\left(\frac{\mathcal{N}(C,\Phi,\delta)}{\gamma}\right)}, \frac{1}{\sqrt{m}}\ln\left(\frac{\mathcal{N}(C,\Phi,\delta)}{\gamma}\right)\right\}$ and $\delta = \frac{1}{\rho}\sqrt{\frac{d}{m}}$ in Theorem 4.4, we deduce that with probability $1 - \gamma$, we have

$$\sup_{y \in C} \|\nabla(\varphi_S)_{1/2\rho}(y) - \nabla\varphi_{1/2\rho}(y)\| \lesssim \sqrt{\frac{\sigma^2 + d}{m} + (\nu + \sigma)^2 \max\left\{\frac{d}{m}\log\left(\frac{R}{\gamma}\right), \frac{d^2}{m^2}\log^2\left(\frac{R}{\gamma}\right)\right\}},$$

where we set $R := 1 + 2\rho B\sqrt{\frac{m}{d}}$.

We end the section by showing how Theorem 4.4 directly implies analogous bounds on a localized Hausdorff distance between the subdifferential graphs, $\mathrm{gph}\,\partial\varphi$ and $\mathrm{gph}\,\partial\varphi_S$.

Theorem 4.5 (Concentration of subdifferential graphs). *Let $C \subseteq \mathbb{R}^d$ be any set and let $r > 0$ and $\bar{\rho} > \rho$ be arbitrary. Then with probability*

$$1 - \mathcal{N}\left(C + r\bar{\rho}\mathbf{B}, \Phi, \delta\right) \exp\left(-m \cdot \psi_{\varepsilon L}^*\left(\frac{s}{\sqrt{m}}\right)\right),$$

the estimates hold

$$(C \times r\mathbf{B}) \cap \mathrm{gph}\,\partial\varphi_S \subset \mathrm{gph}\,\partial\varphi + \left(\sqrt{\frac{4(\sigma+s)^2}{\alpha(\bar{\rho}-\rho)^2 m}} + \frac{2\bar{\rho}\delta}{\alpha(\bar{\rho}-\rho)}\right)(\mathbf{B} \times \bar{\rho}\mathbf{B}), \qquad (4.13)$$

$$(C \times r\mathbf{B}) \cap \mathrm{gph}\,\partial\varphi \subset \mathrm{gph}\,\partial\varphi_S + \left(\sqrt{\frac{4(\sigma+s)^2}{\alpha(\bar{\rho}-\rho)^2 m}} + \frac{2\bar{\rho}\delta}{\alpha(\bar{\rho}-\rho)}\right)(\mathbf{B} \times \bar{\rho}\mathbf{B}). \qquad (4.14)$$

17

Proof. Fix a pair of points $x, y \in \mathbb{R}^d$ and observe the equivalence

$$y = \mathrm{prox}_{\varphi_S/\bar{\rho}}(x) \quad \Longleftrightarrow \quad \bar{\rho}^{-1}(x - y) \in \partial\varphi_S(y).$$

Let $y \in C$ and $v \in r\mathbf{B} \cap \partial\varphi_S(y)$ be arbitrary. Define the point $x := y + \bar{\rho}v$. Clearly then we may write $y = \mathrm{prox}_{\varphi_S/\bar{\rho}}(x)$ and the inclusion $x \in C + r\bar{\rho}\mathbf{B}$ holds. Appealing to Theorem 4.4, we therefore deduce that with probability

$$1 - \mathcal{N}(C + r\bar{\rho}\mathbf{B}, \Phi, \delta) \exp\left(-m \cdot \psi_{\varepsilon L}^*\left(\frac{s}{\sqrt{m}}\right)\right),$$

simultaneously for all $y \in C$ and $v \in r\mathbf{B} \cap \partial\varphi_S(y)$ and $\delta > 0$, we have

$$\|y - \mathrm{prox}_{\varphi/\bar{\rho}}(x)\| \leq \sqrt{\frac{4(\sigma+s)^2}{\alpha(\bar{\rho}-\rho)^2 m} + \frac{2\bar{\rho}\delta}{\alpha(\bar{\rho}-\rho)}}.$$

Set $y' := \mathrm{prox}_{\varphi/\bar{\rho}}(x)$ and $v' := \bar{\rho}^{-1}(x - y') \in \partial\varphi(z)$, and observe

$$\frac{1}{\bar{\rho}}\|v - v'\| = \|y - y'\| \leq \sqrt{\frac{4(\sigma+s)^2}{\alpha(\bar{\rho}-\rho)^2 m} + \frac{2\bar{\rho}\delta}{\alpha(\bar{\rho}-\rho)}}.$$

Thus we showed

$$(y, v) \in (y', v') + \left(\sqrt{\frac{4(\sigma+s)^2}{\alpha(\bar{\rho}-\rho)^2 m} + \frac{2\bar{\rho}\delta}{\alpha(\bar{\rho}-\rho)}}\right)(\mathbf{B} \times \bar{\rho}\mathbf{B}).$$

The inclusion (4.13) follows immediately, while (4.14) follows by a symmetric argument. $\quad\square$

5 Dimension Independent Rates

In this section, we introduce a technique for obtaining bounds on the graphical distance between subdifferentials from estimates on the closeness of function values. The main result we use is a quantitative version of the Attouch convergence theorem from variational analysis [2, 3]. A variant of this theorem was recently used by the authors to analyze the landscape of the phase retrieval problem [21, Theorem 6.1]. For the sake of completeness, we present a short argument, which significantly simplifies the original exposition in [21].

The approach of this section has benefits and drawbacks. The main benefit is that because we obtain subdifferential distance bounds via closeness of values, whenever function values uniformly converge at a dimension independent rate, so do the subdifferentials. This type of result is in stark contrast to the results in Section 4, which scale with the dimension. The main drawback is that for losses that uniformly converge at a rate of δ, we can only deduce subdifferential bounds on the order of $O(\sqrt{\delta})$, yielding what appear to be suboptimal rates. Nevertheless, the very existence of dimension independent bounds is notable. We will illustrate the use of the techniques on two examples: learning with generalized linear models (Section 5.1) and robust nonlinear regression (Section 5.2).

Theorem 5.1 (Closeness of subdifferential graphs). *Consider two closed and lower-bounded functions $g, h \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, having identical domain $\mathcal{D}$, and which are ρ-weakly convex and compatible with a Legendre function $\Phi \colon \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$. Suppose moreover that Φ is α-strongly convex relative to some norm $\|\cdot\|$ on $\mathcal{D} \cap \operatorname{int}(\operatorname{dom}\Phi)$ and for some real $l, u \in \mathbb{R}$, the inequalities hold:*

$$l \leq h(x) - g(x) \leq u, \qquad \forall x \in \mathcal{D} \cap \operatorname{int}(\operatorname{dom}\Phi). \tag{5.1}$$

Then for any $\bar{\rho} > \rho$ and any point $x \in \operatorname{int}(\operatorname{dom}\Phi)$, the estimate holds:

$$D_\Phi^{\mathrm{sym}}\left(\operatorname{prox}_{g/\bar{\rho}}^\Phi(x), \operatorname{prox}_{h/\bar{\rho}}^\Phi(x)\right) \leq \frac{u-l}{\bar{\rho} - \rho}, \tag{5.2}$$

and therefore, provided that Φ is twice differentiable on $\operatorname{int}(\operatorname{dom}\Phi)$, we have

$$\|\nabla g_{1/\bar{\rho}}^\Phi(x) - \nabla h_{1/\bar{\rho}}^\Phi(x)\|_x^* \leq \bar{\rho}\sqrt{\frac{u-l}{\alpha(\bar{\rho}-\rho)}}. \tag{5.3}$$

Moreover, in the Euclidean setting $\Phi = \frac{1}{2}\|\cdot\|_2^2$, we obtain the estimate:

$$\operatorname{dist}_{1/\bar{\rho}}(\operatorname{gph}\partial g, \operatorname{gph}\partial h) \leq \sqrt{\frac{u-l}{\bar{\rho}-\rho}}, \tag{5.4}$$

where the Hausdorff distance $\operatorname{dist}_{1/\bar{\rho}}(\cdot, \cdot)$ is induced by the norm $(x, v) \mapsto \max\left\{\|x\|_2, \frac{1}{\bar{\rho}}\|v\|_2\right\}$.

Proof. Fix a point $x \in \operatorname{int}(\operatorname{dom}\Phi)$ and define $x_g := \operatorname{prox}_{g/\bar{\rho}}^\Phi(x)$ and $x_h := \operatorname{prox}_{h/\bar{\rho}}^\Phi(x)$. We successively deduce

$$g(x_g) + \bar{\rho}D_\Phi(x_g, x) \leq (g(x_h) + \bar{\rho}D_\Phi(x_h, x)) - (\bar{\rho} - \rho)D_\Phi(x_h, x_g) \tag{5.5}$$

$$\leq h(x_h) + \bar{\rho}D_\Phi(x_h, x) - (\bar{\rho} - \rho)D_\Phi(x_h, x_g) - l \tag{5.6}$$

$$\leq h(x_g) + \bar{\rho}D_\Phi(x_g, x) - (\bar{\rho} - \rho)D_\Phi^{\mathrm{sym}}(x_h, x_g) - l \tag{5.7}$$

$$\leq g(x_g) + \bar{\rho}D_\Phi(x_g, x) - (\bar{\rho} - \rho)D_\Phi^{\mathrm{sym}}(x_h, x_g) + (u - l), \tag{5.8}$$

where (5.5) and (5.7) follow from strong convexity of $g(\cdot) + \bar{\rho}D_\Phi(\cdot, x)$ and $h(\cdot) + \bar{\rho}D_\Phi(\cdot, x)$, respectively, while (5.6) and (5.8) follow from the assumption (5.1). Rearranging immediately yields (5.2). The inequality (5.3) follows directly from (5.2) and Theorem 3.1.

Consider now the Euclidean setting $\Phi = \frac{1}{2}\|\cdot\|_2^2$ and fix an arbitrary pair $(x, v) \in \operatorname{gph}\partial g$. A quick computation shows then $x = \operatorname{prox}_{g/\bar{\rho}}(x + \frac{1}{\bar{\rho}}v)$. Define now $x' := \operatorname{prox}_{h/\bar{\rho}}(x + \frac{1}{\bar{\rho}}v)$ and $v' = \bar{\rho}(x - x' + \frac{1}{\bar{\rho}}v)$, and note the inclusion $v' \in \partial h(x')$. Appealing to (5.2), we therefore deduce $\|x' - x\|_2 \leq \sqrt{\frac{u-l}{\bar{\rho}-\rho}}$ and $\|v' - v\|_2 = \bar{\rho}\|x - x'\|_2 \leq \bar{\rho}\sqrt{\frac{u-l}{\bar{\rho}-\rho}}$. We have thus shown $\operatorname{dist}_{1/\bar{\rho}}((x, v), \operatorname{gph}\partial h) \leq \sqrt{\frac{u-l}{\bar{\rho}-\rho}}$. A symmetric argument reversing the roles of f and g completes the proof of (5.4). $\square$

Note that simple examples of uniformly close functions, such as $h(x) = \delta \sin(\delta^{-1/2}x)$ and $g(x) = 0$, show that Theorem 5.1 is tight.

In a typical application of Theorem 5.1 to subgradient estimation, one might set h to be the population risk and g to be the empirical risk or vice versa. The attractive feature of this approach is that it completely decouples probabilistic arguments (for proving functional convergence) from variational analytic arguments (for proving graphical convergence of subdifferentials). The following two sections illustrate the use of Theorem 5.1 on two examples: learning with generalized linear models (Section 5.1) and robust nonlinear regression (Section 5.2).

5.1 Illustration I: Dimension Independent Rates for Generalized Linear Models

In this section, we develop *dimension-independent* convergence guarantees for a wide class of generalized linear models. We consider a loss functions $f: \mathbb{R}^d \times \Omega \to \Omega$ over a bounded set $\mathcal{X}$, where $f(x, z)$ has the parametric form

$$f(x, z) = \ell(\langle x, \phi_1(z)\rangle, \ldots, \langle x, \phi_K(z)\rangle, z),$$

Here $\ell: \mathbb{R}^K \times \Omega \to \mathbb{R}$ is a loss function and $\phi_1, \ldots, \phi_K$ are feature maps. For simplicity, the results of this section are only derived in the ℓ_2 norm, though some of the results hold in greater generality. We make the following assumptions:

(C1) (**Norms**) We equip $\mathbb{R}^d$ and $\mathbb{R}^K$ with the ℓ_2 norms $\|\cdot\|_2$.

(C2) (**Region of Convergence**) We assume that $\mathcal{X} \subseteq \mathbb{R}^d$ is a closed set containing a point $x_0 \in \mathcal{X}$. We assume that $\sup_{x \in \mathcal{X}} \|x - x_0\|_2 \leq B$ for a constant $B > 0$.

(C3) (**Feature Mapping**) The feature maps $\phi_k: \Omega \to \mathbb{R}^d$ are measurable for $k = 1, \ldots, K$.

(C4) (**Loss Function and Regularizer**) $\ell: \mathbb{R}^K \times \Omega \to \mathbb{R}$ is a measurable function. We assume that for each $z \in \Omega$, the function $\ell(\cdot, z)$ is $L(z)$-Lipschitz over the set

$$\{(\langle x, \phi_1(z)\rangle, \ldots, \langle x, \phi_K(z)\rangle) \mid x \in \mathcal{X}\}$$

for a measurable map $L: \Omega \to \mathbb{R}_+$. The function $r: \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is lower semi-continuous.

Then we have the following theorem, whose proof is Presented in Appendix A

Theorem 5.2 (Dimension Independent Functional Concentration). *Let $z_1, \ldots, z_n, z, z'$ be an i.i.d sample from P and define the random variable*

$$Y = \left[|f(x_0, z) - f(x_0, z')| + BL(z)\sqrt{\sum_{k=1}^{K} \|\phi_k(z)\|^2} + BL(z')\sqrt{\sum_{k=1}^{K} \|\phi_k(z')\|^2} \right].$$

Then under assumptions (C1)-(C4), with probability

$$1 - 2\exp\left(-m\psi_{\varepsilon Y}^\star(t)\right),$$

20

we have following bound:

$$\sup_{x \in \mathcal{X}} \left| \frac{1}{m} \sum_{i=1}^{m} f(x, z_i) - \mathbb{E}\left[f(x, z)\right] \right| \leq 2\sqrt{\frac{2B^2 K \max_{k=1,\ldots,K} \mathbb{E}_z \left[L(z)^2 \|\phi_k(z)\|_2^2\right]}{m}} + t.$$

Thus far we have not assumed any weak convexity of the function $f(\cdot, z)$. In order to prove concentration of the subdifferential graphs, we now explicitly make this assumption:

(C5) (**Weak Convexity With High Probability**) There exists a constant $\rho > 0$ and a probability $p_m \in [0, 1]$ such that with probability $1 - p_m$ over the sample $S = \{z_1, \ldots, z_m\}$, the functions

$$\varphi(x) := \mathbb{E}\left[f(x, z)\right] + r(x) + \iota_{\mathcal{X}}(x) \qquad \text{and} \qquad \varphi_S(x) := f_S(x) + r(x) + \iota_{\mathcal{X}}(x).$$

are ρ-weakly convex relative to $\Phi(x) = \frac{1}{2}\|x\|_2^2$.

Given these assumptions, we may deduce subdifferential convergence with Theorem 5.1—the main result of this section.

Corollary 5.3 (Dimension Independent Rates for GLMs). *Assume the setting of Theorem 5.2 and Assumptions (C1)-(C5). Let $z_1, \ldots, z_m, z, z'$ be an i.i.d. sample from P and define the random variable*

$$Y = 2L(z)B\sqrt{\sum_{k=1}^{K} \|\phi_k(z)\|_2^2}.$$

Then with probability

$$1 - 2\exp\left(-m\psi_{\varepsilon Y}^{\star}(t)\right) - p_m,$$

we have the following bound:

$$\sup_{x \in \mathcal{X}} \|\nabla \varphi_{1/2\rho} - \nabla(\varphi_S)_{1/2\rho}(x)\| \leq \sqrt{\frac{\bar{\rho}}{\bar{\rho} - \rho}} \cdot \sqrt{\sqrt{\frac{32B^2 K \max_{k=1,\ldots,K} \mathbb{E}_z \left[L(z)^2 \|\phi_k(z)\|_2^2\right]}{m}} + 2t},$$

$$\operatorname{dist}_{1/\bar{\rho}}(\operatorname{gph} \partial\varphi, \operatorname{gph} \partial\varphi_S) \leq \frac{1}{\sqrt{\bar{\rho} - \rho}} \cdot \sqrt{\sqrt{\frac{32B^2 K \max_{k=1,\ldots,K} \mathbb{E}_z \left[L(z)^2 \|\phi_k(z)\|_2^2\right]}{m}} + 2t},$$

where the Hausdorff distance $\operatorname{dist}_{1/\bar{\rho}}(\cdot, \cdot)$ *is induced by the norm* $(x, v) \mapsto \max\left\{\|x\|_2, \frac{1}{\bar{\rho}}\|v\|_2\right\}$.

Proof. We will apply Theorem 5.2 after shift. Namely set

$$\bar{l}(s, z) = l(s, z) - l(\langle x_0, \phi_1(z)\rangle, \ldots, \langle x_0, \phi_K(z)\rangle)$$

and define the loss $\bar{f}(x, z) = \bar{l}(\langle x, \phi_1(z)\rangle, \ldots, \langle x, \phi_K(z)\rangle, z)$. Define now the functions $\bar{\varphi}(x) = \varphi(x) - \mathbb{E}\left[f(x_0, z)\right]$ and $\bar{\varphi}_S = \varphi_S(x) - \frac{1}{m}\sum_{z \in S} f(x_0, z)$. Applying Theorem 5.2 to $\bar{f}(x, z)$, we deduce that with probability $1 - 2\exp\left(-m\psi_{\varepsilon Y}^{\star}(t)\right)$, we have

$$\sup_{x \in \mathcal{X}} |\bar{\varphi}_S(x) - \bar{\varphi}(x)| \leq 2\sqrt{\frac{2B^2 K \max_{k=1,\ldots,K} \mathbb{E}\left[L(z)^2 \|\phi_k(z)\|_2^2\right]}{m}} + t.$$

21

Thus, due to assumption (C5), we may apply Theorem 5.1 to the functions $\bar{\varphi}(x)$ and $\bar{\varphi}_S$, noticing that $\partial\bar{\varphi}(x) = \partial\varphi(x)$ and $\partial\bar{\varphi}_S(x) = \partial\varphi_S(x)$, as desired. $\qquad\square$

If the random variable $2L(z)B\sqrt{\sum_{k=1}^{K}\|\phi_k(z)\|_2^2}$ is subgaussian, we immediately obtain a dimension independent $m^{-1/4}$ rate of convergence. This is in stark contrast to all other results obtained in this paper.

5.2   Illustration II: Landscape of Robust Nonlinear Regression

In this section, we investigate a robust nonlinear regression problem in $\mathbb{R}^d$, using the techniques we have developed. Setting the stage, consider a function $\sigma\colon \mathbb{R}^d \times \Omega \to \mathbb{R}$ that is differentiable in its first component and let $\bar{x}$ be the ground truth. Our observation model is

$$b(z, \delta, \xi) = \sigma\left(\langle \bar{x}, z \rangle, z\right) + \delta\xi,$$

where z, δ and ξ are random variables. One should think of z as the population data, δ as encoding presence or absence of an outlier, and ξ as the size of the outlying measurement. Seeking to recover $\bar{x}$, we consider the formulation

$$\min_{x \in \mathcal{X}} \ f(x, z) := \mathbb{E}_{z,\delta,\xi}[|\sigma\left(\langle x, z \rangle, z\right) - b(z, \delta, \xi)|]$$

where the set $\mathcal{X}$ will soon be determined. We make the following assumptions on the problem data.

(D1) (**Sufficient Support**) There exist constants $c, C > 0$, such for all $x \in \mathbb{R}^d$, we have

$$C^2\|x\|_2^2 \geq \mathbb{E}\left[|\langle x, z \rangle|^2\right], \qquad \mathbb{E}\left[|\langle x, z \rangle|\right] \geq c\|x\|_2 \qquad \text{and} \qquad P(\langle x, z \rangle \neq 0) = 1.$$

(D2) (**Corruption Frequency**) δ is a $\{0, 1\}$-valued random variable. We define

$$p_{\text{fail}} := P(\delta = 1),$$

which is independent from z and ξ.

(D3) (**Finite Moment**) ξ is a random variable with finite first moment.

(D4) (**Lipschitz, Smooth, and Monotonic Link**) There exist constants $a > 1$ and $c_\sigma, C_\sigma > 0$ satisfying $c_\sigma \leq \sigma'(u, z) \leq C_\sigma$ for all $u \in \{\langle x, z \rangle \mid \|x\|_2 \leq a\|\bar{x}\|_2\}$ and $z \in \Omega$. In addition, for every $z \in \Omega$ the function $\sigma'(\cdot, z)$ is L-Lipschitz continuous.

(D5) (**Concentration**) Let $p_m \in [0, 1]$ and $\tau_m > 0$ be sequences satisfying

$$\mathbb{P}_S\left(\left\|\frac{1}{m}\sum_{z \in S}^{m} zz^T\right\|_{\text{op}} \leq \tau_m\right) \geq 1 - p_m.$$

where $S = \{z_1, \ldots, z_m\}$ is an i.i.d. sample from P.

The noise model considered above allows for adversarial corruption, meaning that ξ may take the form $\xi = \sigma(\langle x_0, z \rangle, z) - \sigma(\langle \bar{x}, z \rangle, z)$ for an arbitrary point x_0. This allows us to "plant" a completely different signal in the measurements. The rest of the assumptions serve to make $\bar{x}$ identifiable from the measurements $\sigma(\langle \bar{x}, z \rangle, z)$, as we will soon show.

The goal of this section is to prove the following theorem, which shows that the empirical risk is well-behaved. In particular, the empirical risk is weakly convex and its stationary points cluster around $\bar{x}$. Henceforth we only consider the Euclidean setting $\Phi(x) = \frac{1}{2}\|x\|_2^2$.

Theorem 5.4 (Stationary Points of the Empirical Risk). *Define* $\mathcal{X} = a\|\bar{x}\|_2 \mathbf{B}$. *For any sample* $S \subseteq \Omega$ *of size* m, *set*

$$\varphi(x) := f(x) + \iota_{\mathcal{X}}(x) \qquad and \qquad \varphi_S(x) := f_S(x) + \iota_{\mathcal{X}}(x).$$

Then φ *is* $2LC^2$*-weakly convex and with probability* $1 - p_m$ *the function* φ_S *is* $2L\tau_m$*-weakly convex. Set* $\rho := \max\{2LC^2, 2L\tau_m\}$ *Suppose now* $p_{p_\text{fail}} < \frac{c_\sigma c}{2C_\sigma C}$ *and set*

$$\rho = \max\{2LC^2, 2L\tau_m\} \qquad and \qquad D = c_\sigma c - 2p_\text{fail} C_\sigma C.$$

Then whenever t *and* m *satisfy*

$$t \leq \frac{1}{256\rho} D^2 \qquad and \qquad m \geq \frac{2^{21} \rho^2 C_\sigma^2 a^2 \|\bar{x}\|_2^2 \mathbb{E}\left[\|z\|_2^2\right]}{D^4},$$

we have, with probability

$$1 - 2\exp\left(-m\psi^\star_{\varepsilon\|z\|_2}\left(\frac{t}{2a\|\bar{x}\|_2 C_\sigma}\right)\right) - p_m,$$

that any pair $(x, v) \in \operatorname{gph} \varphi_S$ *satisfies at least one of the following:*

1. (***Near global optimality***)

$$\|x - \bar{x}\|_2 \leq \frac{16}{D} \cdot \left(\sqrt{\frac{8a^2\|\bar{x}\|^2 C_\sigma^2 \mathbb{E}\left[\|z\|_2^2\right]}{m}} + t\right).$$

2. (***Large Subgradient***)

$$\|v\|_2 \geq \frac{1}{2}D.$$

Let us briefly examine Assumptions (D1)-(D5) and the conclusion of the theorem in the case of a Gaussian population $z \sim N(0, I_{d\times d})$. Assumption (D1) holds true with $C = 1$ and $c = \sqrt{2/\pi}$. Assumption (D2)-(D4) are independent of the distribution of z. Assumption (D5) holds true with

$$\tau_m = 4 + \frac{d}{m} + 4\sqrt{\frac{d}{m}} \qquad and \qquad p_m = 2\exp(-m/2),$$

by Corollary [72, Corollary 5.35]. Thus, assumption (D1)-(D5) are satisfied. Now we examine the various quantities included in the theorem.

The expected squared norm of a gaussian is $\mathbb{E}\left[\|z\|_2^2\right] = d$. One can also show, using standard probabilistic techniques, that the moment generating function satisfies the bound

$$\psi_{\varepsilon\|z\|_2}(t) \leq \frac{d\kappa t^2}{2},$$

for a numerical constant $\kappa > 0$. Thus, we find that $\psi^\star_{\varepsilon\|z\|_2}(t) \geq \frac{t^2}{2d\kappa}$. Therefore, by equating

$$\frac{\delta}{2} = \exp\left(\frac{-mt^2}{2d\kappa(2a\|\bar{x}\|_2 C_\sigma)^2}\right)$$

and solving for t, we find that with probability $1 - \delta - p_m$, every pair $(x, v) \in \operatorname{gph}\varphi_S$ satisfies

$$\|x - \bar{x}\|_2 = O\left(\sqrt{\frac{a^2\|\bar{x}\|_2^2 C_\sigma^2 d}{m}\log\left(\frac{1}{\delta}\right)}\right) \qquad \text{or} \qquad \|v\|_2 \geq \frac{1}{2}\left(c_\sigma\sqrt{\frac{2}{\pi}} - 2p_{\text{fail}}C_\sigma\right).$$

Interestingly, although Theorem 5.1 in general provides rates of convergence that scale as $m^{-1/4}$ as shown in Corollary 5.3, we obtain standard statistical rates of convergence for $\|\bar{x} - x\|_2$. This would not be possible with a direct application of Theorem 4.4, as we would obtain rates that scale as $\sqrt{d^2/m}$. Finally, we note that for this bound to be useful, we must have corruption frequency p_{fail} strictly less than $\frac{c_\sigma}{C_\sigma}\sqrt{\frac{1}{2\pi}}$.

We now present the proof of Theorem 5.4.

Proof of Theorem 5.4. Although φ is nonsmooth and nonconvex, it is fairly well-behaved. We first show that φ and φ_S are both weakly convex.

Claim 1 (Weak Convexity). The functions f and φ are $2LC^2$-weakly convex. Moreover with probability $1 - p_m$ the functions f_S and φ_S are $2L\tau_m$-weakly convex.

Proof of Claim 1. For any fixed x, z, ξ, δ, by the mean value theorem, there exists η in the interval $[\langle x, z\rangle, \langle y, z\rangle]$, so that for all $y \in \mathbb{R}^d$, we have

$$\begin{aligned}
&|\sigma(\langle y, z\rangle, z) - \sigma(\langle\bar{x}, z\rangle, z) + \xi \cdot \delta| \\
&= |\sigma(\langle x, z\rangle, z) + \sigma'(\eta, z)\langle y - x, z\rangle - \sigma(\langle\bar{x}, z\rangle, z) + \xi \cdot \delta| \\
&\geq |\sigma(\langle x, z\rangle, z) + \sigma'(\langle x, z\rangle, z)\langle y - x, z\rangle - \sigma(\langle\bar{x}, z\rangle, z) + \xi \cdot \delta| \qquad (5.9) \\
&\quad - |\sigma'(\eta, z) - \sigma'(\langle x, z\rangle, z)||\langle y - x, z\rangle| \\
&\geq |\sigma(\langle x, z\rangle, z) + \sigma'(\langle x, z\rangle, z)\langle y - x, z\rangle - \sigma(\langle\bar{x}, z\rangle, z) + \xi \cdot \delta| - L|\langle y - x, z\rangle|^2.
\end{aligned}$$

Therefore, taking expectations we deduce

$$\begin{aligned}
f(y) &\geq \mathbb{E}\left[|\sigma(\langle x, z\rangle, z) + \sigma'(\langle x, z\rangle, z)\langle y - x, z\rangle - \sigma(\langle\bar{x}, z\rangle, z) + \xi \cdot \delta|\right] - L\mathbb{E}\left[|\langle y - x, z\rangle|^2\right] \\
&\geq \mathbb{E}\left[|\sigma(\langle x, z\rangle, z) + \sigma'(\langle x, z\rangle, z)\langle y - x, z\rangle - \sigma(\langle\bar{x}, z\rangle, z) + \xi \cdot \delta|\right] - LC^2\|y - x\|_2^2.
\end{aligned}$$

Notice that the right rand-side is a $2LC^2$-weakly convex function in y. We have thus deduced that for every x, there is a $2LC^2$-weakly convex function that globally lower bounds $f(\cdot)$ while agreeing with it at x. Therefore f is $2LC^2$-weakly convex, as claimed.

Next, using (5.9) yields the inequality:

$$f_S(y) \geq \frac{1}{m} \sum_{z \in S} |\sigma(\langle x, z \rangle, z) + \sigma'(\langle x, z \rangle, z)\langle y - x, z \rangle - \sigma(\langle \bar{x}, z \rangle, z) + \xi \cdot \delta| - \frac{L}{m} \sum_{z \in S} |\langle y - x, z \rangle|^2.$$

Finally, notice with probability $\tau_m > 0$ we get the upper bound:

$$\frac{L}{m} \sum_{z \in S} |\langle y - x, z \rangle|^2 \leq L \left\| \frac{1}{m} \sum_{z \in S} zz^T \right\|_{\mathrm{op}} \cdot \|y - x\|_2^2 \leq \tau_m \cdot L \|y - x\|_2^2.$$

By the same reasoning as for the population objective, we deduce that f_S is $2L\tau_m$ weakly convex with probability p_m, as claimed. $\qquad\square$

Having established weak convexity, we now lower bound the subgradients of f and show that for all $x \neq \bar{x}$, the subgradients of f always point toward $\bar{x}$. In particular, the point $\bar{x}$ is the unique stationary point of f.

Claim 2 (Stationarity conditions for f). For every $x \neq \bar{x}$ and $v \in \partial f(x)$, we have

$$(c_\sigma c - 2p_{\mathrm{fail}} C_\sigma C) \cdot \|x - \bar{x}\|_2 \leq \langle v, x - \bar{x} \rangle,$$

and consequently

$$\|v\|_2 \geq c_\sigma c - 2p_{\mathrm{fail}} C_\sigma C.$$

Proof of Claim 2. For every $x \in \mathbb{R}^d$, define the measurable mapping $\zeta_0(x, \cdot) : \Omega \to \mathbb{R}^d$ by

$$\zeta_0(x, z) := \sigma'(\langle x, z \rangle, z)\mathrm{sign}\left(\sigma(\langle x, z \rangle, z) - \sigma(\langle \bar{x}, z \rangle, z)\right) \cdot z.$$

Now, observe that

$$
\begin{aligned}
\mathbb{E}\left[\langle \zeta_0(x, z), x - \bar{x} \rangle\right] &= \mathbb{E}\left[\sigma'(\langle x, z \rangle, z)\mathrm{sign}(\sigma(\langle x, z \rangle, z) - \sigma(\langle \bar{x}, z \rangle, z))\langle z, x - \bar{x} \rangle\right] \\
&= \mathbb{E}\left[\sigma'(\langle x, z \rangle, z)|\langle z, x - \bar{x} \rangle|\right] \\
&\geq c_\sigma \mathbb{E}\left[|\langle z, x - \bar{x} \rangle|\right] \\
&\geq c_\sigma c \cdot \|x - \bar{x}\|_2,
\end{aligned}
$$

where the second equality follows from monotonicity of $\sigma(\cdot, z)$.

As each term $|\sigma(\langle x, z \rangle, z) - \sigma(\langle \bar{x}, z \rangle, z) + \delta \cdot \xi|$ is subdifferentially regular (each term is Lipschitz and weakly convex by Claim 1), it follows that

$$\partial f(x) = \left\{ \mathbb{E}\left[\zeta(x, (z, \xi, \delta))\right] \mid \zeta(x, (z, \xi, \delta)) \in \partial_x(|\sigma(\langle \cdot, z \rangle, z) - \sigma(\langle \bar{x}, z \rangle, z) + \delta \cdot \xi|)(x) \text{ a.e.} \right\},$$

where the set definition ranges over all possible $\zeta(x, \cdot) : \Omega \to \mathbb{R}^d$ that are also measurable [15, Theorem 2.7.2]. Next, we claim that for any such measurable mapping, we have

$$\mathbb{E}\left[\zeta(x, (z, \xi, \delta)) - \zeta_0(x, z)\right) \mid \delta = 0\right] = 0. \tag{5.10}$$

To see this, observe that $\mathbb{E}|\sigma(\langle \cdot, z \rangle, z) - \sigma(\langle \bar{x}, z \rangle, z)|$ is differentiable at any $x \neq \bar{x}$, since $\mathbb{P}(\langle y, z \rangle = 0) = 1$. It follows that the subdifferential of this function consists only of the expectation of the measurable selection ζ_0. The claimed equality (5.10) follows.

Thus, by the linearity of expectation as well as the inclusion $\zeta(x,(z,\xi,\delta)), \zeta_0(x,z) \in \sigma'(\langle x,z \rangle, z)[-1,1]z$, we have

$$
\begin{aligned}
\langle \mathbb{E}\left[\zeta(x,(z,\xi,\delta))\right] - \zeta_0(x,z)], x - \bar{x} \rangle &= (1 - p_{\text{fail}})\langle \mathbb{E}\left[\zeta(x,(z,\xi,\delta)) - \zeta_0(x,z)\right) \mid \delta = 0], x - \bar{x} \rangle \\
&\quad + p_{\text{fail}}\langle \mathbb{E}\left[\zeta(x,(z,\xi,\delta)) - \zeta_0(x,z) \mid \delta = 1\right], x - \bar{x} \rangle \\
&= p_{\text{fail}}\langle \mathbb{E}\left[\zeta(x,(z,\xi,\delta)) - \zeta_0(x,z) \mid \delta = 1\right], x - \bar{x} \rangle \\
&\geq -p_{\text{fail}}\mathbb{E}\left[2\sigma'(\langle x,z \rangle, z)|\langle z, x - \bar{x} \rangle|\right] \\
&\geq -2p_{\text{fail}}C_\sigma C \cdot \|x - \bar{x}\|_2.
\end{aligned}
$$

Therefore, we arrive at the bound:

$$
\begin{aligned}
\langle \mathbb{E}\left[\zeta(x,(z,\xi,\delta))\right], x - \bar{x} \rangle &= \langle \mathbb{E}\left[\zeta_0(x,z)\right], x - \bar{x} \rangle + \langle \mathbb{E}\left[\zeta(x,(z,\xi,\delta)) - \zeta_0(x,z)\right], x - \bar{x} \rangle \\
&\geq c_\sigma c \|x - \bar{x}\|_2 - 2p_{\text{fail}}C_\sigma C \cdot \|x - \bar{x}\|_2 \\
&= (c_\sigma c - 2p_{\text{fail}}C_\sigma C) \cdot \|x - \bar{x}\|_2.
\end{aligned}
$$

As every element of the subdifferential $\partial F(x)$ is of the form $\mathbb{E}\left[\zeta(x,(z,\xi,\delta))\right]$, the proof is complete. $\qquad\square$

While the only stationary point of f is $\bar{x}$, it is as yet unclear where the (random) stationary points of f_S lie, since we can only guarantees that that the functional deviation $|F - F_S|$ is small on bounded sets. Thus, we first show that constraining f to a ball containing $\bar{x}$ does not create any extraneous stationary points at the boundary of the ball.

Claim 3 (Constrained Stationary Conditions of F). Let $a > 1$ be a fixed constant. Let $x \in a\|\bar{x}\|_2 \mathbf{B}$ be such that $x \neq \bar{x}$. Then for every $v \in \partial f(x) + N_{a\|\bar{x}\|_2\mathbf{B}}(x)$, we have

$$
(c_\sigma c - 2p_{\text{fail}}C_\sigma C) \cdot \|x - \bar{x}\|_2 \leq \langle v, x - \bar{x} \rangle
$$

and consequently

$$
\|v\|_2 \geq c_\sigma c - 2p_{\text{fail}}C_\sigma C.
$$

Proof of Claim 3. By Claim 2, we must only consider the case when $\|x\|_2 = a\|\bar{x}\|_2$ since otherwise $N_{a\|\bar{x}\|_2\mathbf{B}}(x) = \{0\}$ and $v \in \partial f(x)$. In this case, we have

$$
v = v_F + \lambda x
$$

where $v_F \in \partial f(x)$ and $\lambda > 0$. Therefore, we find that

$$
\begin{aligned}
\langle v, x - \bar{x} \rangle &= \langle v_F, x - \bar{x} \rangle + \langle \lambda x, x - \bar{x} \rangle \\
&\geq \langle v_F, x - \bar{x} \rangle + \lambda\|x\|_2^2 - \lambda\langle x, \bar{x} \rangle \\
&\geq \langle v_F, x - \bar{x} \rangle + \lambda a^2\|\bar{x}\|_2^2 - \lambda a\|\bar{x}\|_2^2 \geq \langle v_F, x - \bar{x} \rangle.
\end{aligned}
$$

Thus, applying Claim 2 completes the proof. $\qquad\square$

Finally, we may now examine the stationary points of f_S constrained to a ball. We show that every nearly stationary point of $f_S + \delta_{\mathcal{X}}$ must be within a small ball around $\bar{x}$. To that end, we define

$$
\eta = \sqrt{\frac{32a^2\|\bar{x}\|_2^2 C_\sigma^2 \mathbb{E}\left[\|z\|_2^2\right]}{m}} + 2t.
$$

Notice that for all $z \in \Omega$, the function $\sigma(\cdot, z)$ is $L(z) = C_\sigma$ Lipschitz. In addition, every point in $\mathcal{X} = a\|\bar{x}\|_2 \mathbf{B}$ is bounded in norm by $a\|\bar{x}\|_2$. Therefore, by Corollary 5.3 with $x_0 = 0$, we have that with probability

$$1 - 2\exp\left(-m\psi^\star_{\varepsilon\|z\|_2}\left(\frac{t}{2a\|\bar{x}\|_2 C_\sigma}\right)\right) - p_m,$$

the bound holds:

$$\mathrm{dist}_{1/\bar{\rho}}(\mathrm{gph}\,\partial\varphi, \mathrm{gph}\,\partial\varphi_S) \leq \sqrt{\frac{\eta}{\bar{\rho} - \rho}},$$

where we set $\rho := \max\{2LC^2, 2L\tau_m\}$ and $\bar{\rho} > \rho$ is arbitrary.

In particular, for any $\gamma > 0$, setting $\bar{\rho} = \frac{\gamma^2}{\eta} + \rho$, we deduce that for any pair $(x, v) \in \mathrm{gph}\,\partial\varphi_S$ there exists a point $\hat{x} \in \mathcal{X}$ and a subgradient $\hat{v} \in \partial\varphi(\hat{x})$ satisfying

$$\|x - \hat{x}\|_2 \leq \eta/\gamma \qquad \text{and} \qquad \|v - \hat{v}\|_2 \leq \bar{\rho} \cdot \eta/\gamma = \gamma + \rho\eta/\gamma.$$

Let us choose $\gamma > 0$ so that

$$\gamma + \rho\eta/\gamma \leq \frac{1}{2}D,$$

which may be accomplished by finding a root of the polynomial

$$\gamma^2 - \frac{1}{2}D\gamma + \rho\eta = 0.$$

Thus by the quadratic formula, we have

$$\gamma = \frac{\frac{1}{2}D + \sqrt{\frac{1}{4}D^2 - 8\rho\eta}}{2}.$$

Notice that by our assumptions on t and m, we have $8\rho\eta \leq \frac{1}{8}D^2$, and therefore we deduce $\frac{D}{4} \leq \gamma$. Thus by Claim 3, if $\hat{x} \neq \bar{x}$, there exists $\hat{v} \in \partial\varphi(\hat{x}) = \partial f(\hat{x}) + N_\mathcal{X}(\hat{x})$ such that

$$\|v\|_2 \geq \|\hat{v}\|_2 - \|v - \hat{v}\|_2 \geq (c_\sigma c - 2p_{\mathrm{fail}}C_\sigma C) - \frac{1}{2}(c_\sigma c - 2p_{\mathrm{fail}}C_\sigma C) = \frac{1}{2}(c_\sigma c - 2p_{\mathrm{fail}}C_\sigma C).$$

Otherwise, $\hat{x} = \bar{x}$ and $\|x - \bar{x}\|_2 \leq 2\eta/\gamma \leq \frac{8\eta}{D}$, as desired. $\qquad\square$

Acknowledgement

References

[1] Paolo Albano and Piermarco Cannarsa. Singularities of semiconcave functions in Banach spaces. In *Stochastic analysis, control, optimization and applications*, Systems Control Found. Appl., pages 171–190. Birkhäuser Boston, Boston, MA, 1999.

[2] H. Attouch. *Variational convergence for functions and operators.* Applicable Mathematics Series. Pitman (Advanced Publishing Program), Boston, MA, 1984.

[3] Hédy Attouch. Convergence de fonctions convexes, des sous-différentiels et semi-groupes associés. *C. R. Acad. Sci. Paris Sér. A-B*, 284(10):A539–A542, 1977.

[4] Hédy Attouch and Roger J.-B. Wets. Epigraphical processes: laws of large numbers for random LSC functions. *Sém. Anal. Convexe*, 20:Exp. No. 13, 29, 1990.

[5] A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM J. Optim.*, 16(3):697–725, 2006.

[6] Peter L Bartlett, Olivier Bousquet, Shahar Mendelson, et al. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

[7] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[8] H.H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Math. Oper. Res.*, 42(2):330–348, 2017.

[9] H.H. Bauschke and J.M. Borwein. Legendre functions and the method of random Bregman projections. *J. Convex Anal.*, 4(1):27–67, 1997.

[10] H.H. Bauschke, M.N. Dao, and S.B. Lindstrom. Regularizing with Bregman-Moreau envelopes. *arXiv:1705.06019*, 2017.

[11] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence.* Oxford university press, 2013.

[12] Olivier Bousquet and André Elisseeff. Stability and generalization. *J. Mach. Learn. Res.*, 2(3):499–526, 2002.

[13] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

[14] S. Bubeck. *Convex Optimization: Algorithms and Complexity.* Foundations and Trends in Machine Learning. Now Publishers, 2015.

[15] Frank H Clarke. *Optimization and nonsmooth analysis*, volume 5. Siam, 1990.

[16] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *To appear in SIAM J. Optim., arXiv:1803.06523*, 2018.

[17] D. Davis, D. Drusvyatskiy, S.M. Kakade, and J.D. Lee. Stochastic subgradient method converges on tame functions. In *To appear in Found. Compt. Math., arXiv:1804.07795*, 2018.

[18] D. Davis, D. Drusvyatskiy, and K.J. MacPhee. Stochastic model-based minimization under high-order growth. *arXiv:1807.00255*, 2018.

[19] D. Davis, D. Drusvyatskiy, K.J. MacPhee, and C. Paquette. Subgradient methods for sharp weakly convex functions. *To appear in J. Optim. Theory App., arXiv:1803.02461*, 2018.

[20] D. Davis and B. Grimmer. Proximally guided stochastic method for nonsmooth, nonconvex problems. *Preprint arXiv:1707.03505*, 2017.

[21] Damek Davis, Dmitriy Drusvyatskiy, and Courtney Paquette. The nonsmooth landscape of phase retrieval. *arXiv preprint arXiv:1711.03247*, 2017.

[22] D. Drusvyatskiy. The proximal point method revisited. *SIAG/OPT Views and News*, 26(1):1–8, 2018.

[23] D. Drusvyatskiy and C. Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Preprint arXiv:1605.00125*, 2016.

[24] J.C. Duchi and F. Ruan. Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Preprint arXiv:1705.02356*, 2017.

[25] J.C. Duchi and F. Ruan. Stochastic methods for composite optimization problems. *Preprint arXiv:1703.08570*, 2017.

[26] J. Eckstein. Nonlinear proximal point algorithms using Bregman functions, with applications to convex programming. *Math. Oper. Res.*, 18(1):202–226, 1993.

[27] Dylan Foster, Ayush Sekhari, and Karthik Sridharan. Uniform convergence of gradients for non-convex learning and optimization. In *Accepted to Advances in Neural Information Processing Systems*, 2018.

[28] Charles J. Geyer. On the asymptotics of constrained m-estimation. *Ann. Statist.*, 22(4):1993–2010, 12 1994.

[29] S. Ghadimi, G. Lan, and H. Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Math. Program.*, 155(1-2, Ser. A):267–305, 2016.

[30] Peter D Grünwald and Nishant A Mehta. Fast rates for general unbounded loss functions: from erm to generalized bayes. *arXiv preprint arXiv:1605.00252*, 2016.

[31] A. Juditsky and A.S. Nemirovski. First order methods for nonsmooth convex large-scale optimization, II: Utilizing problem's structure. In Stephen J. Wright Suvrit Sra, Sebastian Nowozin, editor, *Optimization for Machine Learning*, pages 29–63. MIT Press, August 2010.

[32] Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in neural information processing systems*, pages 793–800, 2009.

[33] Yuri M. Kaniovski, Alan J. King, and Roger J-B Wets. Probabilistic bounds (via large deviations) for the solutions of stochastic programming problems. *Annals of Operations Research*, 56(1):189–208, Dec 1995.

[34] Alan J. King and R. Tyrrell Rockafellar. Asymptotic theory for solutions in statistical estimation and stochastic programming. *Mathematics of Operations Research*, 18(1):148–162, 1993.

[35] Aryeh Kontorovich. Concentration in unbounded metric spaces and algorithmic stability. In *International Conference on Machine Learning*, pages 28–36, 2014.

[36] X. Li, Z. Zhihui, A.M.-C So, and R. Vidal. Nonconvex robust low-rank matrix recovery. *arXiv:1809.09237*, 2018.

[37] Mingrui Liu, Xiaoxuan Zhang, Lijun Zhang, Rong Jin, and Tianbao Yang. Fast rates of erm and stochastic approximation: Adaptive to error bound conditions. *arXiv preprint arXiv:1805.04577*, 2018.

[38] H. Lu. Relative continuity for non-lipschitz non-smooth convex optimization using stochastic (or deterministic) mirror descent. *arXiv:1710.04718*, 2017.

[39] H. Lu, R. Freund, and Y. Nesterov. Relatively-smooth convex optimization by first-order methods, and applications. *arXiv:1610.05708*, 2016.

[40] Andreas Maurer. A vector-contraction inequality for rademacher complexities. In Ronald Ortner, Hans Ulrich Simon, and Sandra Zilles, editors, *Algorithmic Learning Theory*, pages 3–17, Cham, 2016. Springer International Publishing.

[41] Colin McDiarmid. On the method of bounded differences. In *Surveys in combinatorics, 1989 (Norwich, 1989)*, volume 141 of *London Math. Soc. Lecture Note Ser.*, pages 148–188. Cambridge Univ. Press, Cambridge, 1989.

[42] Nishant A Mehta. Fast rates with high probability in exp-concave statistical learning. *arXiv preprint arXiv:1605.01288*, 2016.

[43] Nishant A. Mehta and Robert C. Williamson. From stochastic mixability to fast rates. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, pages 1197–1205, Cambridge, MA, USA, 2014. MIT Press.

[44] Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for nonconvex losses. *Ann. Statist.*, 46(6A):2747–2774, 2018.

[45] B.S. Mordukhovich. *Variational Analysis and Generalized Differentiation I: Basic Theory*. Grundlehren der mathematischen Wissenschaften, Vol 330, Springer, Berlin, 2006.

[46] J.-J. Moreau. Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France*, 93:273–299, 1965.

[47] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim.*, 19(4):1574–1609, 2008.

[48] E. A. Nurminskii. The quasigradient method for the solving of the nonlinear programming problems. *Cybernetics*, 9(1):145–150, Jan 1973.

[49] R.A. Poliquin and R.T. Rockafellar. Prox-regular functions in variational analysis. *Trans. Amer. Math. Soc.*, 348:1805–1838, 1996.

[50] Svetlozar T. Rachev and Werner Rmisch. Quantitative stability in stochastic programming: The method of probability metrics. *Mathematics of Operations Research*, 27(4):792–818, 2002.

[51] A. Rakhlin, S. Mukherjee, and T. Poggio. Stability results in learning theory. *Analysis and Applications*, 03(04):397–417, 2005.

[52] D. Ralph and H. Xu. Convergence of stationary points of sample average two-stage stochastic programs: a generalized equation approach. *Math. Oper. Res.*, 36(3):568–592, 2011.

[53] W. Rmisch and R. Wets. Stability of ε-approximate solutions to convex stochastic programs. *SIAM Journal on Optimization*, 18(3):961–979, 2007.

[54] Stephen M. Robinson. Analysis of sample-path optimization. *Mathematics of Operations Research*, 21(3):513–528, 1996.

[55] R. T. Rockafellar. *Convex analysis.* Princeton Mathematical Series, No. 28. Princeton University Press, Princeton, N.J., 1970.

[56] R.T. Rockafellar. Favorable classes of Lipschitz-continuous functions in subgradient optimization. In *Progress in nondifferentiable optimization*, volume 8 of *IIASA Collaborative Proc. Ser. CP-82*, pages 125–143. Int. Inst. Appl. Sys. Anal., Laxenburg, 1982.

[57] R.T. Rockafellar and R.J-B. Wets. *Variational Analysis.* Grundlehren der mathematischen Wissenschaften, Vol 317, Springer, Berlin, 1998.

[58] S. Rolewicz. On paraconvex multifunctions. In *Third Symposium on Operations Research (Univ. Mannheim, Mannheim, 1978), Section I*, volume 31 of *Operations Res. Verfahren*, pages 539–546. Hain, Königstein/Ts., 1979.

[59] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014.

[60] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic Convex Optimization. In *Proceedings of the Conference on Learning Theory (COLT)*, 2009.

[61] A. Shapiro and T. Homem-de Mello. On the rate of convergence of optimal solutions of monte carlo approximations of stochastic programs. *SIAM Journal on Optimization*, 11(1):70–86, 2000.

[62] Alexander Shapiro. On the asymptotics of constrained local m-estimators. *Ann. Statist.*, 28(3):948–960, 05 2000.

[63] Alexander Shapiro. *Stochastic programming by Monte Carlo simulation methods.* Humboldt-Universitt zu Berlin, Mathematisch-Naturwissenschaftliche Fakultt II, Institut fr Mathematik, 2000.

[64] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming*, volume 9 of *MOS-SIAM Series on Optimization.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Optimization Society, Philadelphia, PA, second edition, 2014. Modeling and theory.

[65] Alexander Shapiro and Huifu Xu. Uniform laws of large numbers for set-valued mappings and subdifferentials of random functions. *Journal of Mathematical Analysis and Applications*, 325(2):1390 – 1399, 2007.

[66] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2199–2207. Curran Associates, Inc., 2010.

[67] Karthik Sridharan, Shai Shalev-shwartz, and Nathan Srebro. Fast rates for regularized objectives. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1545–1552. Curran Associates, Inc., 2009.

[68] M. Teboulle. Entropic proximal mappings with applications to nonlinear programming. *Math. Oper. Res.*, 17(3):670–690, 1992.

[69] M. Teboulle. A simplified view of first order methods for optimization. *Mathematical Programming*, May 2018.

[70] P. Toulis and E.M. Airoldi. Asymptotic and finite-sample properties of estimators based on stochastic gradients. *Ann. Statist.*, 45(4):1694–1727, 2017.

[71] Tim van Erven, Peter D. Grünwald, Nishant A. Mehta, Mark D. Reid, and Robert C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16:1793–1861, 2015.

[72] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[73] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

[74] Huifu Xu. Uniform exponential convergence of sample average random functions under general sampling with applications in stochastic programming. *Journal of Mathematical Analysis and Applications*, 368(2):692 – 710, 2010.

[75] Huifu Xu and Dali Zhang. Smooth sample average approximation of stationary points in nonsmooth stochastic optimization and applications. *Mathematical Programming*, 119(2):371–401, Jul 2009.

[76] S. Zhang and N. He. On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization. *arXiv:1806.04781*, 2018.

[77] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pages 928–935. AAAI Press, 2003.

A   Rademacher Complexity and Functional Bounds

In this section, we use the well-known technique for bounding the suprema of empirical processes, based on Rademacher complexities (see e.g., [6, 7].We will use these bounds to obtain concentration inequalities for multi-class generalized linear models. Although such arguments have become standard in the literature, we present a proof that explicitly uses Theorem 4.2 in order to obtain a slightly more general result for unbounded classes. None of the results or techniques here are new; rather, the purpose of this section is to keep the paper self-contained. We begin with the following standard definition.

Definition A.1. The *Rademacher complexity* of a set $A \subset \mathbb{R}^m$ is the quantity

$$\mathcal{R}(A) = \frac{1}{m} \mathbb{E}_\varepsilon \left[\sup_{a \in A} \langle \varepsilon, a \rangle \right].$$

where the coordinates of $\varepsilon \in \mathbb{R}^m$ are i.i.d Rademacher random variables.

Given a collection of functions $\mathcal{G}$ from Ω to $\mathbb{R}$ and a set $S = \{z_1, \ldots, z_m\} \subset \Omega$, we define

$$\mathcal{G} \circ S := \{(g(z_1), \ldots, g(z_m)) : g \in \mathcal{G}\}.$$

The following theorem shows that the Rademacher complexity directly controls uniform convergence of the sample average approximation.

Theorem A.2. *Consider a countable class $\mathcal{G}$ of measurable functions from Ω to $\mathbb{R}$ and let $S = \{z_1, \ldots, z_m\}$ be an i.i.d. sample from P. Define the random variable*

$$Y = \sup_{g \in \mathcal{G}} |g(z) - g(z')|,$$

for independent copies $z, z' \sim P$ and let ε be a Radamacher random variable. Then for all $t > 0$, with probability

$$1 - 2 \exp\left(-m \psi_{\varepsilon Y}^\star(t)\right),$$

we have the following bound:

$$\sup_{g \in \mathcal{G}} \left| \mathbb{E}_z \left[g(z) \right] - \frac{1}{m} \sum_{i=1}^m g(z_i) \right| \le 2 \mathbb{E}_S \mathcal{R}(\mathcal{G} \circ S) + t.$$

Proof. Define the two random variables $X^+ = \sup_{g \in \mathcal{G}} \left\{ \mathbb{E}_z\left[g(z)\right] - \frac{1}{m}\sum_{i=1}^m g(z_i) \right\}$ and $X^- = \sup_{g \in \mathcal{G}} \left\{ \frac{1}{m}\sum_{i=1}^m g(z_i) - \mathbb{E}_z\left[g(z)\right] \right\}$. We first bound the expectations of X^+ and X^-. Appealing to [59, Lemma 26.2] we deduce $\mathbb{E}[X^+] \leq 2\mathbb{E}_S\mathcal{R}(\mathcal{G} \circ S)$. Replacing $\mathcal{G}$ with $-\mathcal{G}$ and using [59, Lemma 26.2], we also learn $\mathbb{E}[X^-] \leq 2\mathbb{E}_S\mathcal{R}(-\mathcal{G} \circ S) = 2\mathbb{E}_S\mathcal{R}(\mathcal{G} \circ S)$. Next, a quick computation shows

$$|X^+(z_1, \ldots, z_m) - X^+(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_m)| \leq \frac{1}{m}\sup_{g \in \mathcal{G}}|g(z_i) - g(z_i')| = \frac{1}{m}Y,$$

as well as the analogous inequality for X^-. Thus using Theorem 4.2, we conclude that with probability $1 - 2\exp(-m\psi^\star_{m^{-1}\varepsilon Y}(t/m))$, we have $\max\{X^+, X^-\} \leq 2\mathbb{E}_S\mathcal{R}(\mathcal{G} \circ S) + t$. Noting the equality $\psi^\star_{m^{-1}\varepsilon Y}(t/m) = \psi^\star_{\varepsilon Y}(t)$ completes the proof. $\qquad\square$

The following theorem provides an upper bound on the Rademacher complexity of linear classes; see the original article [32] or the monograph [59, Lemma 26.10].

Lemma A.3 (Rademacher complexity of linear classes).
Consider the set $A = \{(\langle w, z_1\rangle, \ldots, \langle w, z_m\rangle) : \|w\|_2 \leq 1\}$. Then the estimate holds:

$$\mathcal{R}(A) \leq \frac{\sqrt{\sum_{i=1}^m \|z_i\|^2}}{m}.$$

The class of loss functions $\mathcal{G}$ considered below will be formed from compositions of functions with linear classes. A useful result for unraveling such compositions is the following vector-valued contraction inequality, recently proved by Maurer [40].

Theorem A.4 (Contraction Inequality [40, Theorem 3]). *Let $\mathcal{X}$ denote a countable set. For $i = 1, \ldots, m$, let $F_i \colon \mathcal{S} \to \mathbb{R}$ and $G_i \colon \mathcal{S} \to \mathbb{R}^K$ be functions satisfying*

$$F_i(s) - F_i(u) \leq \|G_i(s) - G_i(u)\|_2 \qquad \text{for all } s, u \in \mathcal{S}.$$

Define the two sets

$$F \circ \mathcal{S} = \{(F_1(s), \ldots, F_m(s)) : s \in \mathcal{S}\} \qquad \text{and} \qquad G \circ \mathcal{S} = \left\{\left(G_i^k(s)\right)_{i,k} : s \in \mathcal{S}\right\},$$

where $G_i^k(s)$ denotes the k'th coordinate of $G_i(s)$. Then the estimate holds:

$$\mathcal{R}(F \circ \mathcal{S}) \leq \sqrt{2}K \cdot \mathcal{R}(G \circ \mathcal{S}).$$

We are now ready to prove Theorem 5.2.

Proof of Theorem 5.2. We will apply Theorem A.2, to the function class

$$\mathcal{G} = \{z \mapsto f(x, z) \mid x \in \mathcal{X}\}.$$

We note that, due to the separability of $\mathbb{R}^d$ and the continuity of the integrands, any supremum over all $x \in \mathcal{X}$ may be replaced by a supremum over a countable dense subset of $\mathcal{X}$, without affecting its value. We ignore this technicality throughout the proof.

As the first step in applying Theorem A.2, we compute

$$\sup_{x \in \mathcal{X}} |f(x, z) - f(x, z')|$$

$$\leq |f(x_0, z) - f(x_0, z')| + L(z) \sup_{x \in \mathcal{X}} \sqrt{\sum_{k=1}^{K} \langle x - x_0, \phi_k(z) \rangle^2} + L(z') \sup_{x \in \mathcal{X}} \sqrt{\sum_{k=1}^{K} \langle x - x_0, \phi_k(z') \rangle^2}$$

$$\leq |f(x_0, z) - f(x_0, z')| + BL(z) \sqrt{\sum_{k=1}^{K} \|\phi_k(z)\|^2} + BL(z') \sqrt{\sum_{k=1}^{K} \|\phi_k(z')\|^2},$$

where the last inequality uses the bound $\|x - x_0\|_2 \leq B$ twice. Notice the wight-hand-side is precisely the random variable Y.

Next we upper bound the expected Rademacher complexity $\mathbb{E}_S \mathcal{R}(\mathcal{G} \circ S)$ by using Theorem A.4. To this end, fix a sample set $S = \{z_1, \ldots, z_m\}$ and define

$$\mathcal{S} = \{ (\langle x, \phi_k(z_i) \rangle)_{i,k} : x \in \mathcal{X} \}.$$

For every index i, set $s_i := (s_{i1}, \ldots, s_{iK})$ and define the functions $F_i(s) = \ell(s_i, z_i)$ and $G_i(s) = L(z_i) s_i$. We successively compute

$$\mathcal{R}(\mathcal{G} \circ S) = \frac{1}{m} \sup_{x \in \mathcal{X}} \sum_{i=1}^{m} \sigma_i f(x, z_i)$$

$$= \frac{1}{m} \sup_{x \in \mathcal{X}} \sum_{i=1}^{m} \sigma_i l \left((\langle x, \phi_1(z_i) \rangle, \ldots, \langle x, \phi_K(z_i) \rangle), z_i \right) \qquad (A.1)$$

$$= \frac{1}{m} \sup_{s \in \mathcal{S}} \sum_{i=1}^{m} \sigma_i F_i(s) = \mathcal{R}(F \circ \mathcal{S}) \leq \sqrt{2} K \cdot \mathcal{R}(G \circ \mathcal{S}),$$

where the last inequality follows from Theorem A.4.

Next, unraveling notation, observe $G \circ \mathcal{S} = \{ (\langle x, L(z_i) \phi_k(z_i) \rangle)_{i,k} : x \in \mathcal{X} \}$. Moreover, shifting and shrinking $\mathcal{X}$, it follows directly from the definition of Rademacher complexity that $\mathcal{R}(\mathcal{G} \circ \mathcal{S}) = B \cdot \mathcal{R}(A')$ where we set $A' = \{ (\langle x, L(z_i) \phi_k(z_i) \rangle)_{i,k} : \|x\|_2 \leq 1 \}$. Thus applying Lemma A.3, we deduce $\mathcal{R}(\mathcal{G} \circ \mathcal{S}) \leq \frac{\sqrt{\sum_{i,k} B^2 L(z_i)^2 \|\phi_k(z_i)\|^2}}{mK}$. Combining this estimate with (A.1) and taking expectations yields

$$\mathbb{E}_S \mathcal{R}(\mathcal{G} \circ S) \leq \frac{\sqrt{2 \sum_{i,k} B^2 \mathbb{E}_{z_i}[L(z_i)^2 \|\phi_k(z_i)\|^2]}}{m} = \sqrt{\frac{2B^2 K \max_{k=1,\ldots,K} \mathbb{E}_z[L(z)^2 \|\phi_k(z)\|^2]}{m}}.$$

Appealing to Theorem A.2 completes the proof. $\qquad \square$