Mathematics of Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Graphical Convergence of Subgradients in Nonconvex Optimization and Learning

Damek Davis, Dmitriy Drusvyatskiy

To cite this article:
Damek Davis, Dmitriy Drusvyatskiy (2022) Graphical Convergence of Subgradients in Nonconvex Optimization and Learning. Mathematics of Operations Research 47(1):209-231. https://doi.org/10.1287/moor.2021.1126

http://pubsonline.informs.org/journal/moor

MATHEMATICS OF OPERATIONS RESEARCH
Vol. 47, No. 1, February 2022, pp. 209–231
ISSN 0364-765X (print), ISSN 1526-5471 (online)

Graphical Convergence of Subgradients in Nonconvex Optimization and Learning

Damek Davis,[a] **Dmitriy Drusvyatskiy**[b]

[a] School of Operations Research and Information Engineering, Cornell University, Ithaca, New York 14850; [b] Department of Mathematics, University of Washington, Seattle, Washington 98195

Contact: dsd95@cornell.edu, https://orcid.org/0000-0003-2105-4641 (DaD); ddrusv@uw.edu (DmD)

Received: November 18, 2018
Revised: December 30, 2019
Accepted: September 26, 2020
Published Online in Articles in Advance: April 20, 2021

MSC2020 Subject Classification: Primary: 90C15; secondary: 68Q32, 65K10

https://doi.org/10.1287/moor.2021.1126

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.

Funding: D. Davis was supported by an Alfred P. Sloan Research Fellowship. D. Drusvyatskiy was supported by the Division of Mathematical Sciences [Grant 1651851], Air Force Office of Scientific Research [Grant FA9550-15-1-0237], and Division of Computing and Communication Foundations [Grant 1740551].

Keywords: 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*:

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

Here, $r:\mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ 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 (Davis et al. [12], Ghadimi et al. [23], Nemirovski et al. [41]) is one popular streaming algorithm, although there are many others, such as the stochastic proximal point and Gauss-Newton methods (Davis and Drusvyatskiy [11], Duchi and Ruan [20], Toulis and Airoldi [60]). In contrast, ERM-based algorithms draw a large sample $S = \{z_1, z_2, \ldots, z_m\}$ at the onset and output the solution of the deterministic problem

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

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 on 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.

Published in Mathematics of Operations Research on April 20, 2021 as DOI: 10.1287/moor.2021.1126.
This article has not been copyedited or formatted. The final version may differ from this version.

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 (Shalev-Shwartz et al. [54], theorem 5):

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

Here, and throughout the paper, the symbol $\|\cdot\|$ denotes the ℓ_2-norm on $\mathbb{R}^d$.

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 may fail to 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 (Shalev-Shwartz et al. [54]), 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

$$\hat{x}_S := \operatorname*{arg\,min}_{x\in\mathbb{R}^d}\{\varphi_S(x) + \lambda\|x\|^2\}, \tag{1.4}$$

satisfies

$$\varphi(\hat{x}_S) - \inf\varphi \le \sqrt{\frac{8L^2R^2}{\gamma m}} \tag{1.5}$$

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 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 Elisseeff [8]. That is, the solution $\hat{x}_S$ does not vary much under an arbitrary perturbation of a single sample z_i. It is worthwhile to note that stability arguments have also been used to establish generalization bounds for multipass streaming methods in Hardt et al. [25]. 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, $\sup_{x:\|x\|\le R}\|\nabla f(x) - \nabla f_S(x)\|$, along with deviations in higher-order derivatives when they exist. Indeed, then in the simplest setting $r = 0$, the standard decomposition

$$\|\nabla f(x)\| \le \underbrace{\|\nabla f(x) - \nabla f_S(x)\|}_{\text{generalization error}} + \underbrace{\|\nabla f_S(x)\|}_{\text{optimization error}},$$

relates near-stationarity for the empirical risk to near-stationarity for the population risk. Such uniform bounds have recently appeared in (Foster et al. [21], Mei et al. [37]).

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 (Mordukhovich [38], Rockafellar and Wets [50]). Then the minimal norms, $\operatorname{dist}(0, \partial\varphi(x))$ and $\operatorname{dist}(0, \partial\varphi_S(x))$, could serve as stationarity measures akin to the norm of the gradient in smooth minimization. One may then posit that the stationarity measures, $\operatorname{dist}(0, \partial\varphi(x))$ and $\operatorname{dist}(0, \partial\varphi_S(x))$, are uniformly close with high probability when the sample size is large. Pointwise convergence is indeed known to hold (e.g., Shapiro et al. [59], 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 (Ralph and Xu [46]). The main obstacle is that the function $x \mapsto \operatorname{dist}(0, \partial\varphi(x))$ is highly discontinuous. We refer the reader to

Shapiro et al. [59, p. 380] for a discussion. Indeed, the need to look beyond pointwise uniform convergence is well-documented in optimization and variational analysis (Attouch [3], Attouch and Wets [4]). One remedy is to instead focus on graphical convergence concepts. Namely, one could posit that the Hausdorff distance between the subdifferential graphs, $\operatorname{gph}\partial\varphi$ and $\operatorname{gph}\partial\varphi_S$, tends to zero. Here, we take a closely related approach, while aiming for finite-sample bounds.

1.1. Contributions

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 (Albano and Cannarsa [1], Nurminskii [42], Poliquin and Rockafellar [43], Rockafellar [49], Rolewicz [51]).[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 problems. We refer the interested reader to Davis and Drusvyatskiy [11] (section 2.1) and Drusvyatskiy [17] for detailed examples.

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

$$g_\lambda(x) := \min_y \left\{ g(y) + \frac{1}{2\lambda}\|y-x\|^2 \right\}, \qquad \operatorname{prox}_{\lambda g}(x) := \arg\min_y \left\{ g(y) + \frac{1}{2\lambda}\|y-x\|^2 \right\}.$$

It is well-known that if g is ρ-weakly convex and $\lambda < \frac{1}{\rho}$, then the envelope g_λ is C^1-smooth with gradient

$$\nabla g_\lambda(x) = \lambda^{-1}\left(x - \operatorname{prox}_{\lambda g}(x)\right).$$

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

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

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

where L is a Lipschitz constant of the losses $f(\cdot,z)$ on the ball $B_R(0)$ for almost every $z \sim P$. The bound (1.6) is stated here for simplicity with a deterministic Lipschitz constant L; our full result allows L to be a random variable. 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, $\operatorname{gph}\partial\varphi$ and $\operatorname{gph}\partial\varphi_S$, 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 (Davis and Drusvyatskiy [11], Davis et al. [13], Davis and Grimmer [16], Zhang and He [67]). 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 we present of (1.6) uses only the most elementary techniques: stability of quadratically regularized ERM (Shalev-Shwartz et al. [54]), McDiarmid's inequality (McDiarmid [34]), 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.

Published in Mathematics of Operations Research on April 20, 2021 as DOI: 10.1287/moor.2021.1126.
This article has not been copyedited or formatted. The final version may differ from this version.

We illustrate the use of such bounds with two examples. As the first example, consider the loss f modeling a generalized linear model:

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

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 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 $\bar{x}$.

1.2. 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 (Geyer [22], Kaniovski et al. [27], King and Rockafellar [28], Rachev and Römisch [44], Römisch and Wets [52], Robinson [47], Shapiro [55], Shapiro [56], Shapiro and Homem-de Mello [57]). 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 (Bousquet and Elisseeff [8], Grünwald and Mehta [24], Kakade et al. [26], Liu et al. [32], Mehta [35], Mehta and Williamson [36], Rakhlin et al. [45], Shalev-Shwartz et al. [54], Zemel and Culotta [66], Koller et al. [29], van Erven et al. [61], Zinkevich [68]).

Several previous works have studied (sub)gradient-based convergence, as we do here. For example, Xu [64] 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 by Xu and Zhang [65] 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 Xu and Zhang [65]. The most general asymptotic convergence result that we are aware of is presented in Shapiro and Xu [58]. There, the authors show that with probability one, the limit of a certain enlarged subdifferential of the empirical loss 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 by Mei et al. [37] 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 Mei et al. [37] are all dimension dependent and rely on covering arguments. The more recent paper by Foster et al. [21], 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 in Foster et al. [21] is a "chain rule" for Rademacher complexity. We note that, in contrast to the $m^{-1/4}$ rates developed in this paper, Foster et al. [21] obtain rates of the form $m^{-1/2}$ for smooth generalized linear models.

1.3. Outline

In Section 2, we introduce our notation. Section 3 describes the problem setting and the smoothing technique. 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.

Though the current paper focuses 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 regularization of the empirical and population risks, making our techniques applicable under non-Euclidean geometries and under high order growth of the loss function. We have found that such generalizations add significant notational overhead, and as a result we have placed the details in the arXiv version of the paper (Davis and Drusvyatskiy [10]).

Published in Mathematics of Operations Research on April 20, 2021 as DOI: 10.1287/moor.2021.1126.
This article has not been copyedited or formatted. The final version may differ from this version.

2. Preliminaries

Throughout, we follow standard notation from convex and variational analysis, as set out for example in the classical monographs of Rockafellar [48] and Rockafellar and Wets [50], and the recent book by Mordukhovich [39]. The symbol $\mathbb{R}^d$ will denote a d-dimensional Euclidean space with inner product $\langle \cdot, \cdot \rangle$ and the induced norm $\|x\| = \sqrt{\langle x, x \rangle}$. 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 $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, 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 $+\infty$ off it. Our focus will be primarily on those functions that can be convexified by adding a sufficiently large multiple of the squared norm $\frac{1}{2}\|\cdot\|^2$. Formally, we will say that a function $g : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is *ρ-weakly convex*, for some $\rho \in \mathbb{R}$, if the perturbed function $x \mapsto g(x) + \frac{\rho}{2}\|x\|^2$ is convex.

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[2]

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

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 (Davis et al. [13], lemma 2.2):

$$g(y) \geq g(x) + \langle v, y - x \rangle - \frac{\rho}{2}\|y - x\|^2. \tag{2.1}$$

for any $x, y \in \operatorname{dom} g$ and $v \in \partial g(x)$. 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

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

3. Problem Setting

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

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

under the following assumptions on the functional components:

Assumption A1 (Weak Convexity). The functions $f(\cdot, z) + r(\cdot)$ are closed and ρ-weakly convex for a.e. $z \in \Omega$.

Assumption A2 (Lipschitzian Property). There exists a square integrable function $L : \Omega \to \mathbb{R}_+$ such that for all $x, y \in \operatorname{dom} r$ and $z \in \Omega$, we have

$$|f(x, z) - f(y, z)| \leq L(z)\|x - y\| \qquad \text{and} \qquad \sqrt{\mathbb{E}\big[L(z)^2\big]} \leq \sigma.$$

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 $r : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ is a structure promoting regularizer. Alternatively, r can encode feasibility constraints as an indicator function. Assumption $(A1)$ is self-explanatory, while assumption $(A2)$ simply amounts to Lipschitz continuity of the loss $f(\cdot, z)$ on $\operatorname{dom} r$ with a square integrable Lipschitz constant $L(z)$.

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. Indeed, provided that $h(\cdot, z)$ is ℓ-Lipschitz and the Jacobian $\nabla c(\cdot, z)$ is β-Lipschitz, a quick argument (Drusvyatskiy and Paquette [18], lemma 4.2) shows that the loss $f(\cdot, z)$ is $\ell\beta$-weakly convex; therefore (A1) 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 $x \in \operatorname{dom} r$ and $z \in \Omega$, then (A2) holds with $L(z) = \ell M(z)$. The class of

Published in Mathematics of Operations Research on April 20, 2021 as DOI: 10.1287/moor.2021.1126.
This article has not been copyedited or formatted. The final version may differ from this version.

composite problems is broad and has attracted some attention lately (Davis and Drusvyatskiy [11], Davis et al. [13], Davis et al. [14], Davis et al. [15], Davis and Grimmer [16], Drusvyatskiy and Paquette [18], Duchi and Ruan [19], Duchi and Ruan [20], Li et al. [31], Zhang and He [67]) as an appealing setting for nonsmooth nonconvex optimization. Table 1 summarizes a few interesting problems of this type; details can be found for example in Davis and Drusvyatskiy [11], Davis and Grimmer [16], and Drusvyatskiy [17].

Because 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, $\operatorname{gph}\partial\varphi$ and $\operatorname{gph}\partial\varphi_S$. We employ an indirect strategy based on a smoothing technique.

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

$$g_\lambda(x) := \inf_y \left\{ g(y) + \frac{1}{2\lambda}\|y-x\|^2 \right\}, \qquad \operatorname{prox}_{\lambda g}(x) := \operatorname*{argmin}_y \left\{ g(y) + \frac{1}{2\lambda}\|y-x\|^2 \right\},$$

respectively. These two constructions were introduced by Moreau [40], and are now routinely used in optimization literature. The following result establishes the smoothing properties of the Moreau envelope (Moreau [40]). We will often appeal to this theorem without explicitly referencing it.

Theorem 3.1 (Smoothness of the Moreau Envelope). *Consider a closed and ρ-weakly convex function $g:\mathbb{R}^d \to \mathbb{R}\cup\{\infty\}$. Then for any positive $\lambda < \rho^{-1}$, the envelope g_λ is differentiable with gradient given by*

$$\nabla g_\lambda(x) := \frac{1}{\lambda}\left(x - \operatorname{prox}_{\lambda g}(x)\right), \tag{3.2}$$

and the equivalence holds:

$$y = \operatorname{prox}_{\lambda g}(x) \iff \left(y, \lambda^{-1}(x-y)\right) \in \operatorname{gph}\partial g. \tag{3.3}$$

This theorem has been instrumental in recent work establishing convergence guarantees for algorithms in nonsmooth and nonconvex optimization (Davis and Drusvyatskiy [11], Davis et al. [13], Zhang and He [67]). These works argue that the natural measure of convergence for such problems is the gradient norm $\|\nabla\varphi_\lambda(x)\|$ and show that streaming algorithms drive this measure in expectation to zero at a controlled rate. Simple examples show that one cannot expect similar guarantees for the quantity $\operatorname{dist}(0,\partial\varphi(x))$, which may be bounded below by a fixed constant at all nonstationary points (e.g., $\varphi(x) = |x|$).

The equivalence (3.3) has two important consequences. First it shows that the gradient norm of the Moreau envelope is closely related to minimal norm subgradients at nearby points. Namely, setting $\hat{x} := \operatorname{prox}_{\lambda g}(x)$, Equations (3.2) and (3.3) immediately yield the estimates:

$$\|x-\hat{x}\| = \lambda\|\nabla g_\lambda(x)\| \quad \text{and} \quad \operatorname{dist}\left(0,\partial g(\hat{x})\right) \le \|\nabla g_\lambda(x)\|.$$

Hence if the norm $\|\nabla g_\lambda(x)\|$ is small, then x must be near some point (namely $\hat{x}$) that is nearly stationary for g. Figure 1(b) illustrates this phenomenon, whereas Figure 1(a) plots the Moreau envelope of the weakly convex loss $|x^2-1|$, which appears in the phase retrieval problem (Davis et al. [15], Duchi and Ruan [19]).

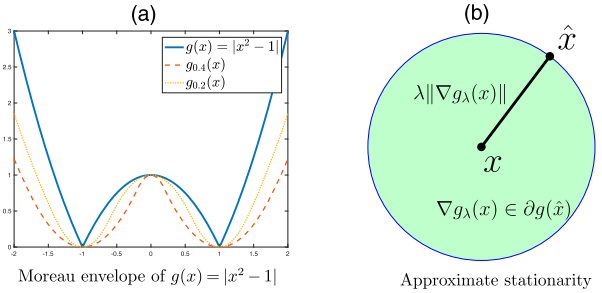
Secondly, note that (3.3) shows that the graph of the proximal map $\operatorname{prox}_{\lambda g}$ is linearly isomorphic to the graph of the subdifferential ∂g by the linear map $(x,y) \mapsto (y, \lambda^{-1}(y-x))$. It is this observation that will allow us to pass from uniform estimates on the deviations $\|\operatorname{prox}_{\varphi/\lambda}(x) - \operatorname{prox}_{\varphi_S/\lambda}(x)\|$ to estimates on the Hausdorff distance between subdifferential graphs, $\operatorname{gph}\partial\varphi$ and $\operatorname{gph}\partial\varphi_S$.

Table 1. Common stochastic weakly convex optimization problems.

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	$f((D,x),z) = \|z-Dx\|$	$r(D,x) = \iota_{\mathbf{B}}(D) + \lambda\|x\|_1$

Published in Mathematics of Operations Research on April 20, 2021 as DOI: 10.1287/moor.2021.1126.
This article has not been copyedited or formatted. The final version may differ from this version.

Figure 1. (Color online) An illustration of the Moreau envelope.



4. Dimension Dependent Rates

In this section, we prove the uniform convergence bound (1.6). The proof outline is as follows. First, in Theorem 4.1 we will estimate the expected error between the population and empirical proximal points,

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

where y is fixed. A key ingredient is leave-one-out stability of the proximal map, in the sense that $\operatorname{prox}_{\varphi_S/\lambda}(y)$ does not vary much when a single sample is changed—the main result of Shalev-Shwartz et al. [54]. Using McDiarmid's inequality (McDiarmid [34]) in Theorem 4.4, we will then deduce that the quantity $\|\operatorname{prox}_{\varphi/\lambda}(y) - \operatorname{prox}_{\varphi_S/\lambda}(y)\|$ concentrates around its mean for a fixed y. A covering argument over the points y will then complete the proof.

We begin following the outlined strategy with Theorem 4.1, which extracts the relevant conclusions that we need from Shalev-Shwartz et al. [54]. For the sake of completeness, we provide a complete proof, which parallels that in Shalev-Shwartz et al. [54].

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 $y \in \mathbb{R}^d$ and a real $\bar{\rho} > \rho$, and set*

$$\mathcal{A}^* := \operatorname*{argmin}_{y\in\mathbb{R}^d}\left\{\varphi(x) + \frac{\bar{\rho}}{2}\|x-y\|^2\right\} \quad \text{and} \quad \mathcal{A}(S) := \operatorname*{argmin}_{y\in\mathbb{R}^d}\left\{\varphi_S(x) + \frac{\bar{\rho}}{2}\|x-y\|^2\right\}.$$

Then the estimates hold:

$$\|\mathcal{A}(S) - \mathcal{A}(S^i)\| \le \frac{L(z_i) + L(z_i')}{(\bar{\rho}-\rho)m}, \tag{4.1}$$

$$\mathbb{E}_S\left[\|\mathcal{A}(S) - \mathcal{A}^*\|^2\right] \le \frac{4\sigma^2}{(\bar{\rho}-\rho)^2 m}, \tag{4.2}$$

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

Proof. We first verify (4.1). 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

$$\widehat{\varphi}(x) := \varphi(x) + \frac{\bar{\rho}}{2}\|x-y\|^2 \qquad \text{and} \qquad \widehat{\varphi}_S(x) := \varphi_S(x) + \frac{\bar{\rho}}{2}\|x-y\|^2.$$

Then adding $[r(v) + \frac{\bar{\rho}}{2}\|v-y\|^2] - [r(u) + \frac{\bar{\rho}}{2}\|u-y\|^2]$ to both sides of (4.4), we obtain

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

Published in Mathematics of Operations Research on April 20, 2021 as DOI: 10.1287/moor.2021.1126.
This article has not been copyedited or formatted. The final version may differ from this version.

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 $\widehat{\varphi}_S(\cdot)$ and $\widehat{\varphi}_{S^i}(\cdot)$ are $(\bar{\rho}-\rho)$-strongly convex, we deduce

$$\frac{\bar{\rho}-\rho}{2}\|v-u\|^2 \le \widehat{\varphi}_S(v)-\widehat{\varphi}_S(u) \le \frac{f(v,z_i)-f(u,z_i)}{m}+\frac{f(u,z_i')-f(v,z_i')}{m}-\frac{\bar{\rho}-\rho}{2}\|u-v\|^2.$$

Rearranging, we arrive at the estimate

$$\|u-v\|^2 \le \frac{1}{\bar{\rho}-\rho}\left[\frac{f(v,z_i)-f(u,z_i)}{m}+\frac{f(u,z_i')-f(v,z_i')}{m}\right] \le \frac{L(z_i)+L(z_i')}{(\bar{\rho}-\rho)m}\cdot\|u-v\|.$$

Dividing through by $\|u-v\|$, we obtain the claimed stability guarantee (4.1).

To establish (4.3), observe first

$$(\varphi_S)_{1/\bar{\rho}}(y) = \varphi_S(\mathcal{A}(S))+\frac{\bar{\rho}}{2}\|\mathcal{A}(S)-y\|^2 \le \varphi_S(x)+\frac{\bar{\rho}}{2}\|x-y\|^2 \qquad \text{for all } x.$$

Taking expectations, we conclude $\mathbb{E}_S[(\varphi_S)_{1/\bar{\rho}}(y)] \le \varphi_{1/\bar{\rho}}(y)$, which is precisely the left-hand side of (4.3). Next, it is standard to verify the expression (Shalev-Shwartz and Ben-David [53], theorem 13.2):

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

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

$$\begin{aligned}\Big|\mathbb{E}_S\big[\widehat{\varphi}(A(S))-\widehat{\varphi}_S(A(S))\big]\Big| &\le \mathbb{E}\big[L(z)\cdot\|\mathcal{A}(S)-\mathcal{A}(S^i)\|\big]\\ &\le \sqrt{\mathbb{E}_z\big[L(z)^2\big]}\sqrt{\mathbb{E}_S\big[\|\mathcal{A}(S)-\mathcal{A}(S^i)\|^2\big]} \le \frac{2\sigma^2}{(\bar{\rho}-\rho)m}.\end{aligned} \tag{4.6}$$

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

Finally taking into account that $\widehat{\varphi}$ is $(\bar{\rho}-\rho)$-strongly convex, we deduce

$$\frac{\bar{\rho}-\rho}{2}\big\|\mathcal{A}(S)-\mathcal{A}^*\big\|^2 \le \widehat{\varphi}(\mathcal{A}(S))-\min\widehat{\varphi}.$$

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

$$\frac{\bar{\rho}-\rho}{2}\mathbb{E}_S\big[\|\mathcal{A}(S)-\mathcal{A}^*\|^2\big] \le \mathbb{E}_S\big[\widehat{\varphi}(\mathcal{A}(S))-\min\widehat{\varphi}\big] \le \mathbb{E}_S\big[\widehat{\varphi}(\mathcal{A}(S))-\widehat{\varphi}_S(\mathcal{A}(S))\big] \le \frac{2\sigma^2}{(\bar{\rho}-\rho)m}.$$

Thus, we have established (4.2), and the proof is complete. □

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 (McDiarmid [34]). The basic result reads as follows. Suppose that a function g satisfies the bounded difference property:

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

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 $Z_1,\ldots,Z_m$, the random variable $Y = g(Z_1,\ldots,Z_m)$ satisfies:

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

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 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 or subexponential. The proof follows

Published in Mathematics of Operations Research on April 20, 2021 as DOI: 10.1287/moor.2021.1126.
This article has not been copyedited or formatted. The final version may differ from this version.

known techniques for establishing McDiarmid's inequality, and in particular is essentially the same as that in Kontorovich [30] (theorem 1), though there the statement of the theorem assumed that 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 $\psi^\star : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ 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 $\omega_i : \mathcal{Z}_i \times \mathcal{Z}_i \to \mathbb{R}_+$ such that the inequality*

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

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:

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

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

$$\mathbb{P}\big(g(z) - \mathbb{E}\big[g(z)\big] \ge t\big) \le \exp\left(-m\psi^\star_{\varepsilon\omega}\left(\frac{t}{m}\right)\right) \qquad \forall t \ge 0. \tag{4.8}$$

Proof. Define the Doob martingale sequence:

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

and consider the martingale differences

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

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

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

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

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

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

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

Published in Mathematics of Operations Research on April 20, 2021 as DOI: 10.1287/moor.2021.1126.
This article has not been copyedited or formatted. The final version may differ from this version.

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

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

Thus (4.7) is proved. The estimate (4.8) then follows by the standard Cramér-Chernoff bounding method. Namely, assume $\omega_i = \omega_j$ for all indices i and j. Then for every $t \ge 0$, Chernoff's inequality (Boucheron [7], p. 21) together with (4.7) implies

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

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

The final ingredient is to perform a covering argument over the points x. To this end, we will need the following lemma on Lipschitz continuity of the proximal map of weakly convex functions.

Lemma 4.3 (Lipschitz Continuity)**.** *Consider a closed and ρ-weakly convex function $g : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$. Then for any $\bar{\rho} > \rho$ and $x, y \in \mathbb{R}^d$, we have*

$$\left\|\operatorname{prox}_{g/\bar{\rho}}(x) - \operatorname{prox}_{g/\bar{\rho}}(y)\right\| \le \frac{\bar{\rho}}{\bar{\rho}-\rho}\|x-y\|.$$

Proof. For any points $x, y \in \mathbb{R}^d$, set $\hat{x} := \operatorname{prox}_{g/\bar{\rho}}(x)$ and $\hat{y} := \operatorname{prox}_{g/\bar{\rho}}(y)$. Taking into account that $g + \frac{\bar{\rho}}{2}\|\cdot\|^2$ is $(\bar{\rho}-\rho)$ strongly convex, we deduce

$$\begin{aligned}\frac{\bar{\rho}-\rho}{2}\|\hat{y}-\hat{x}\|^2 &\le \left(g(\hat{y}) + \frac{\bar{\rho}}{2}\|\hat{y}-x\|^2\right) - \left(g(\hat{x}) + \frac{\bar{\rho}}{2}\|\hat{x}-x\|^2\right)\\ \frac{\bar{\rho}-\rho}{2}\|\hat{x}-\hat{y}\|^2 &\le \left(g(\hat{x}) + \frac{\bar{\rho}}{2}\|\hat{x}-y\|^2\right) - \left(g(\hat{y}) + \frac{\bar{\rho}}{2}\|\hat{y}-y\|^2\right).\end{aligned}$$

Adding these estimates together, we obtain

$$(\bar{\rho}-\rho)\|\hat{x}-\hat{y}\|^2 \le \frac{\bar{\rho}}{2}\left(\|\hat{y}-x\|^2 - \|\hat{x}-x\|^2 + \|\hat{x}-y\|^2 - \|\hat{y}-y\|^2\right) = \bar{\rho}\langle x-y, \hat{x}-\hat{y}\rangle.$$

Using the Cauchy-Schwartz inequality and dividing both sides by $\|\hat{x}-\hat{y}\|$, the result follows. □

We now have all the ingredients to prove the main result of this section. To this end, for any set $C \subset \mathbb{R}^d$, we will let $\mathcal{N}(C,\delta)$ denote the covering number of C in the ℓ_2 norm, that is, the minimal number of balls of radius δ needed to completely cover C.

Theorem 4.4 (Concentration of the Stationarity Measure)**.** *Let $C \subseteq \mathbb{R}^d$ be any set and let $\bar{\rho} > \rho$ be arbitrary. Fix tolerances $\delta, s > 0$. Then with probability*

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

we have

$$\sup_{y\in C}\left\|\nabla(\varphi_S)_{1/\bar{\rho}}(y) - \nabla\varphi_{1/\bar{\rho}}(y)\right\| \le \frac{\bar{\rho}}{\bar{\rho}-\rho}\left(\frac{2(\sigma+s)}{\sqrt{m}} + 2\bar{\rho}\delta\right),$$

where the second moment $\sigma > 0$ is defined in Assumption (A2).

Proof. Following the notation of Theorem 4.1, set

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

Define the function

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

Published in Mathematics of Operations Research on April 20, 2021 as DOI: 10.1287/moor.2021.1126.
This article has not been copyedited or formatted. The final version may differ from this version.

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

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

where (4.11) uses the reverse triangle inequality, whereas (4.12) follows from the estimate (4.1). Setting $\omega(z_i, z_i') = \frac{L(z_i)+L(z_i')}{(\bar\rho-\rho)m}$, we deduce

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

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

$$\mathbb{P}\left(g(y,S) \ge \sqrt{\frac{4\sigma^2}{(\bar\rho-\rho)^2 m}} + t\right) \le \exp\big(-m\psi^\star_{\varepsilon L}\big((\bar\rho-\rho)t/2\big)\big).$$

Next, we show using Lemma 4.3 that $g(\cdot,S)$ is $\frac{2\bar\rho}{\bar\rho-\rho}$-Lipschitz. Indeed, observe

$$\begin{aligned} \big|g(y,S) - g(y',S)\big| &\le \Big|\|\mathcal{A}(y,S) - \mathcal{A}^*(y)\| - \|\mathcal{A}(y',S) - \mathcal{A}^*(y')\|\Big| \\ &\le \|\mathcal{A}(y,S) - \mathcal{A}(y',S)\| + \|\mathcal{A}^*(y) - \mathcal{A}^*(y')\| \le \frac{2\bar\rho}{\bar\rho-\rho}\|y-y'\|,\end{aligned}$$

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. Thus, for every y in a δ-ball of y_i, we have $g(y,S) \le g(y_i,S) + \frac{2\bar\rho\delta}{\bar\rho-\rho}$. Taking a union bound over the net, we therefore deduce

$$\mathbb{P}\left(g(y,S) \le \sqrt{\frac{4\,\sigma^2}{(\bar\rho-\rho)^2 m}} + \frac{2\bar\rho\delta}{\bar\rho-\rho} + t\right) \ge 1 - \mathcal{N}(C,\delta)\exp\big(-m\psi^\star_{\varepsilon L}\big((\bar\rho-\rho)t/2\big)\big).$$

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

In particular, under the natural choice $\bar\rho = 2\rho$, we deduce that with probability $1 - \mathcal{N}(C,\delta)\exp(-m\cdot\psi^\star_{\varepsilon L}(\frac{s}{\sqrt{m}}))$, the estimate holds:

$$\sup_{y\in C}\ \|\nabla(\varphi_S)_{1/\bar\rho}(y) - \nabla\varphi_{1/\bar\rho}(y)\| \le \sqrt{\frac{16(\sigma+s)^2}{m}} + 8\rho\delta.$$

We next instantiate Theorem 4.4 (with $\bar\rho = 2\rho$ for simplicity) under various distributional assumptions on $L(z)$. We will require the use of the sub-Gaussian norm of any random variable X, which is defined to be $\|X\|_{sg} := \inf\{t>0 : \mathbb{E}\exp(X^2/t^2) \le 2\}$, along with the subexponential norm $\|X\|_{se} := \inf\{t>0 : \mathbb{E}\exp(|X|/t) \le 2\}$. Henceforth let $C \subset \mathbb{R}^d$ be a set whose diameter is bounded by B. Consequently, the estimate $\mathcal{N}(C,\delta) \le (1+\frac{4B}{\delta})^d$, holds; see, for example, Vershynin [63] (corollary 4.2.13).

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

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

Published in Mathematics of Operations Research on April 20, 2021 as DOI: 10.1287/moor.2021.1126.
This article has not been copyedited or formatted. The final version may differ from this version.

Appealing to Vershynin [63] (equation 2.16), we conclude $\psi_{\varepsilon L}(\lambda) \le \frac{c}{2}(\nu+\sigma)^2\lambda^2$ for all $\lambda \in \mathbb{R}$, where c is a numerical constant. Taking conjugates yields the relation $\psi^\star_{\varepsilon L}(t) \ge \frac{t^2}{c(\nu+\sigma)^2}$. Appealing to Theorem 4.4, while setting $s = \sqrt{c(\nu+\sigma)^2 \ln(\frac{\mathcal{N}(C,\delta)}{\gamma})}$ 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 $R := 1 + 2\rho B\sqrt{\frac{m}{d}}$.

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

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

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

Subexponential Lipschitz Constant. As the final example, we suppose that $L - \mathbb{E}[L]$ is subexponential and set $\nu = \|L - \mathbb{E}[L]\|_{se}$. A completely analogous argument as in the sub-Gaussian case implies $\|\varepsilon L\|_{se} \le \nu + \frac{\sigma}{\ln(2)}$. Appealing to Vershynin [63] (proposition 2.7.1), we deduce $\psi_{\varepsilon L}(\lambda) \le c^2(\nu+\sigma)^2\lambda^2$ for all $|\lambda| \le \frac{1}{c(\nu+\sigma)}$. To simplify notation set $\eta := c(\nu+\sigma)$. Taking conjugates, we therefore deduce

$$\psi^\star_{\varepsilon L}(t) \ge \begin{cases} \frac{t^2}{4\eta^2} & \text{if } |t| \le 2\eta \\ \frac{|t|}{\eta} - 1 & \text{otherwise} \end{cases}.$$

Consequently, we deduce the usual Bernstein-type bound $\psi^\star_{\varepsilon L}(t) \ge \min\{\frac{t^2}{4\eta^2}, \frac{t}{2\eta}\}$. Setting $s = 2\eta \cdot \max\{\sqrt{\ln(\frac{\mathcal{N}(C,\delta)}{\gamma})}, \frac{1}{\sqrt{m}}\ln(\frac{\mathcal{N}(C,\delta)}{\gamma})\}$ 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, $\operatorname{gph}\partial\varphi$ and $\operatorname{gph}\partial\varphi_S$.

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*

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

the estimates hold

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

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

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

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

Published in Mathematics of Operations Research on April 20, 2021 as DOI: 10.1287/moor.2021.1126.
This article has not been copyedited or formatted. The final version may differ from this version.

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 $y = \operatorname{prox}_{\varphi_S/\bar{\rho}}(x)$ and the inclusion $x \in C + r\bar{\rho}\mathbf{B}$ holds. Appealing to Theorem 4.4, we therefore deduce that with probability

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

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

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

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

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

Thus, we showed

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

The inclusion (4.13) follows immediately, while (4.14) follows by a symmetric argument. □

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 (Attouch [2], [3]). A variant of this theorem was recently used by the authors to analyze the landscape of the phase retrieval problem (Davis et al. [15], theorem 6.1). For the sake of completeness, we present a short argument, which incidentally simplifies the original exposition in Davis et al. [15].

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 (Subdifferential Graphs)**.** *Consider two closed and ρ-weakly convex functions $g, h : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, having identical domain $\mathcal{D}$. Suppose moreover that for some real $l, u \in \mathbb{R}$, the inequalities hold:*

$$l \le h(x) - g(x) \le u, \qquad \forall x \in \mathcal{D}. \tag{5.1}$$

Then for any $\bar{\rho} > \rho$ and any point $x \in \mathbb{R}^d$, the estimate holds:

$$\|\nabla g_{1/\bar{\rho}}(x) - \nabla h_{1/\bar{\rho}}(x)\| \le \bar{\rho}\sqrt{\frac{u-l}{\bar{\rho}-\rho}}. \tag{5.2}$$

Consequently, we obtain the estimate:

$$\operatorname{dist}_{1/\bar{\rho}}\big(\operatorname{gph}\partial g, \operatorname{gph}\partial h\big) \le \sqrt{\frac{u-l}{\bar{\rho}-\rho}}, \tag{5.3}$$

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

Published in Mathematics of Operations Research on April 20, 2021 as DOI: 10.1287/moor.2021.1126.
This article has not been copyedited or formatted. The final version may differ from this version.

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

$$g(x_g) + \frac{\bar{\rho}}{2}\|x_g - x\|^2 \leq \left(g(x_h) + \frac{\bar{\rho}}{2}\|x_h - x\|^2\right) - \frac{\bar{\rho}-\rho}{2}\|x_h - x_g\|^2 \tag{5.4}$$

$$\leq h(x_h) + \frac{\bar{\rho}}{2}\|x_h - x\|^2 - \frac{\bar{\rho}-\rho}{2}\|x_h - x_g\|^2 - l \tag{5.5}$$

$$\leq h(x_g) + \frac{\bar{\rho}}{2}\|x_g - x\|^2 - (\bar{\rho}-\rho)\|x_h - x_g\|^2 - l \tag{5.6}$$

$$\leq g(x_g) + \frac{\bar{\rho}}{2}\|x_g - x\|^2 - (\bar{\rho}-\rho)\|x_h - x_g\|^2 + (u-l), \tag{5.7}$$

where (5.4) and (5.6) follow from strong convexity of $g(\cdot) + \frac{\bar{\rho}}{2}\|\cdot - x\|^2$ and $h(\cdot) + \frac{\bar{\rho}}{2}\|\cdot - x\|^2$, respectively, whereas (5.5) and (5.7) follow from the assumption (5.1). Rearranging yields (5.2).

Fix now 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 $\|x' - x\| \leq \sqrt{\frac{u-l}{\bar{\rho}-\rho}}$ and $\|v' - v\| = \bar{\rho}\|x - x'\| \leq \bar{\rho}\sqrt{\frac{u-l}{\bar{\rho}-\rho}}$. 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 (5.3). □

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

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

Here $\ell : \mathbb{R}^K \times \Omega \to \mathbb{R}$ is a loss function and $\phi_1, \ldots, \phi_K$ are feature maps. We make the following assumptions:

Assumption (C1) (Region of Convergence). We assume that $\mathcal{X} \subseteq \mathbb{R}^d$ is a closed set containing a point $x_0 \in \mathcal{X}$, and that the estimate $\sup_{x\in\mathcal{X}} \|x - x_0\| \leq B$ holds for some $x_0 \in \mathcal{X}$ and some $B > 0$.

Assumption (C2) (Feature Mapping). The feature maps $\phi_k : \Omega \to \mathbb{R}^d$ are measurable for $k = 1, \ldots, K$.

Assumption (C3) (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

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

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

Then we have the following theorem, whose proof is presented in the appendix. The argument we present follows well known techniques, pioneered in Bartlett and Mendelson [6] and Kakade et al. [26].

Theorem 5.2 (Dimension Independent Function 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)–(C3), with probability

$$1 - 2\exp\big(-m\psi^{\star}_{\varepsilon Y}(t)\big),$$

Published in Mathematics of Operations Research on April 20, 2021 as DOI: 10.1287/moor.2021.1126.
This article has not been copyedited or formatted. The final version may differ from this version.

we have the following bound:

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

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:

Assumption (C4) (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 $S=\{z_1,\dots,z_m\}$, the functions

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

are ρ-weakly convex.

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)–(C4). Let $z_1,\dots,z_m,z,z'$ be an i.i.d. sample from P and define the random variable

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

Then with probability

$$1-2\exp\big(-m\psi^{\star}_{\varepsilon Y}(t)\big)-p_m,$$

we have the following bounds:

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

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

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

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

$$\bar l(s,z)=l(s,z)-l\big(\langle x_0,\phi_1(z)\rangle,\dots,\langle x_0,\phi_K(z)\rangle\big),$$

and define the loss $\bar f(x,z)=\bar l(\langle x,\phi_1(z)\rangle,\dots,\langle x,\phi_K(z)\rangle,z)$. Define now the functions $\bar\varphi(x)=\varphi(x)-\mathbb{E}[f(x_0,z)]$ 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 $1-2\exp(-m\psi^{\star}_{\varepsilon Y}(t))$, we have

$$\sup_{x\in\mathcal{X}}\big|\bar\varphi_S(x)-\bar\varphi(x)\big|\le 2\sqrt{\frac{2B^2K\max_{k=1,\dots,K}\mathbb{E}\big[L(z)^2\|\phi_k(z)\|^2\big]}{m}}+t.$$

Thus, due to Assumption (C4), 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. □

If the random variable $2L(z)B\sqrt{\sum_{k=1}^{K}\|\phi_k(z)\|^2}$ is sub-Gaussian, we immediately obtain a dimension independent rate of convergence on the order of $m^{-1/4}$.

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 $\sigma:\mathbb{R}^d\times\Omega\to\mathbb{R}$ that is differentiable in its first component and let $\bar x$ be the ground truth. Our observation model is

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

Published in Mathematics of Operations Research on April 20, 2021 as DOI: 10.1287/moor.2021.1126.
This article has not been copyedited or formatted. The final version may differ from this version.

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

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

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

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

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

Assumption (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.

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

Assumption (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 $u \in \{\langle x,z\rangle \mid \|x\| \leq a\|\bar{x}\|\}$ and $z \in \Omega$. In addition, for every $z \in \Omega$ the function $\sigma'(\cdot,z)$ is L-Lipschitz continuous.

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

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

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

The noise model considered 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. We note that a smooth variant of the robust nonlinear regression problem was also considered in Foster et al. [21] and Mei et al. [37]. To the best of our knowledge, we are unaware of any prior work that addresses the stationary points of the *nonsmooth* problem considered here.

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 $\bar{x}$.

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

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

Then φ is $2LC^2$-weakly convex and with probability $1 - p_m$ the function φ_S is $2L\tau_m$-weakly convex.

Suppose now $p_{p_{\text{fail}}} < \frac{c_\sigma c}{2C_\sigma C}$ and set

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

Then whenever $t > 0$ and m satisfy

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

we have, with probability

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

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

1. (Near Global Optimality)

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

Published in Mathematics of Operations Research on April 20, 2021 as DOI: 10.1287/moor.2021.1126.
This article has not been copyedited or formatted. The final version may differ from this version.

2. (Large Subgradient)

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

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 $c = \sqrt{2/\pi}$. Assumptions (D2)–(D4) are independent of the distribution of z. Assumption (D5) holds true with

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

by Corollary (Vershynin [62], corollary 5.35). Thus, Assumptions (D1)–(D5) are satisfied. Now we examine the various quantities included in the theorem.

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

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

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

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

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

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

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 $\|\bar{x} - x\|$. 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 5.1 (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 5.1. 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

$$\begin{aligned}
&|\sigma(\langle y, z\rangle, z) - \sigma(\langle \bar{x}, z\rangle, z) + \xi\cdot\delta| \\
&= |\sigma(\langle x, z\rangle, z) + \sigma'(\eta, z)\langle y - x, z\rangle - \sigma(\langle \bar{x}, z\rangle, z) + \xi\cdot\delta| \\
&\geq |\sigma(\langle x, z\rangle, z) + \sigma'(\langle x, z\rangle, z)\langle y - x, z\rangle - \sigma(\langle \bar{x}, z\rangle, z) + \xi\cdot\delta| \\
&\quad - |\sigma'(\eta, z) - \sigma'(\langle x, z\rangle, z)\|\langle y - x, z\rangle| \\
&\geq |\sigma(\langle x, z\rangle, z) + \sigma'(\langle x, z\rangle, z)\langle y - x, z\rangle - \sigma(\langle \bar{x}, z\rangle, z) + \xi\cdot\delta| - L|\langle y - x, z\rangle|^2. \qquad (5.8)
\end{aligned}$$

Therefore, taking expectations we deduce

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

Notice that the right-hand 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.

Published in Mathematics of Operations Research on April 20, 2021 as DOI: 10.1287/moor.2021.1126.
This article has not been copyedited or formatted. The final version may differ from this version.

Next, using (5.8) 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:

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

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. □

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

Claim 5.2 (Stationarity Conditions for f)**.** For every $x \neq \bar{x}$ and $v \in \partial f(x)$, we have

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

and consequently

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

Proof of Claim 5.2. For every $x \in \mathbb{R}^d$, define a 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

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

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 5.1), it follows that

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

where the set definition ranges over all possible $\zeta(x,\cdot):\Omega\to\mathbb{R}^d$ that are also measurable (Clarke [9], theorem 2.7.2). Next, we claim that for any such measurable mapping, we have

$$\mathbb{E}[\zeta(x,(z,\xi,\delta)) - \zeta_0(x,z) \mid \delta = 0] = 0. \tag{5.9}$$

To see this, observe that the function $\mathbb{E}|\sigma(\langle\cdot,z\rangle,z) - \sigma(\langle\bar{x},z\rangle,z)|$ 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 at any $x\neq\bar{x}$ consists only of the expectation of the measurable selection ζ_0. The claimed equality (5.9) follows.

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

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

Published in Mathematics of Operations Research on April 20, 2021 as DOI: 10.1287/moor.2021.1126.
This article has not been copyedited or formatted. The final version may differ from this version.

Therefore, we arrive at the bound:

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

As every element of $\partial f(x)$ is of the form $\mathbb{E}[\zeta(x,(z,\xi,\delta))]$, the proof is complete. □

Although the only stationary point of f is $\bar{x}$, it is as yet unclear where the (random) stationary points of f_S lie, because we can only guarantee 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 5.3 (Constrained Stationary Conditions of f)**.** Let $a>1$ be a fixed constant. Let $x\in a\|\bar{x}\|\mathbf{B}$ be such that $x\neq\bar{x}$. Then for every $v\in\partial f(x)+N_{a\|\bar{x}\|\mathbf{B}}(x)$, we have

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

and consequently

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

Proof of Claim 5.3. By Claim 5.2, we must only consider the case when $\|x\| = a\|\bar{x}\|$ because otherwise $N_{a\|\bar{x}\|\mathbf{B}}(x)=\{0\}$ and $v\in\partial f(x)$. In this case, we have

$$v = v_f + \lambda x$$

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

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

Thus, applying Claim 5.2 completes the proof. □

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

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

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

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

the bound holds:

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

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 $(x,v)\in\operatorname{gph}\partial\varphi_S$ there exists a point $\hat{x}\in\mathcal{X}$ and a subgradient $\hat{v}\in\partial\varphi(\hat{x})$ satisfying

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

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 $\gamma=\frac{\frac{1}{2}D+\sqrt{\frac{1}{4}D^2-4\rho\eta}}{2}$. Notice that by our assumptions on

Published in Mathematics of Operations Research on April 20, 2021 as DOI: 10.1287/moor.2021.1126.
This article has not been copyedited or formatted. The final version may differ from this version.

t and m, we have $4\rho\eta \le \frac{1}{8}D^2$, and therefore we deduce $\frac{D}{4} \le \gamma$. Thus, by Claim 5.3, 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

$$\|v\| \ge \|\hat{v}\| - \|v - \hat{v}\| \ge \big(c_\sigma c - 2p_{\text{fail}}C_\sigma C\big) - \frac{1}{2}\big(c_\sigma c - 2p_{\text{fail}}C_\sigma C\big) = \frac{1}{2}\big(c_\sigma c - 2p_{\text{fail}}C_\sigma C\big).$$

Otherwise, $\hat{x} = \bar{x}$ and $\|x - \bar{x}\| \le \eta/\gamma \le \frac{4\eta}{D}$, as desired. □

Acknowledgments
The authors thank Yudong Chen for the simple example showing the tightness of Theorem 5.1.

Appendix. 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., Bartlett et al. [5], Bartlett and Mendelson [6]). We will use these bounds to obtain concentration inequalities for multiclass 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. 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

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

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 Rademacher random variable. Then for all $t > 0$, with probability

$$1 - 2\exp\big(-m\psi^\star_{\varepsilon Y}(t)\big),$$

we have the following bound:

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

Proof. Define the two random variables $X^+ = \sup_{g\in\mathcal{G}}\{\mathbb{E}_z[g(z)] - \frac{1}{m}\sum_{i=1}^m g(z_i)\}$ and $X^- = \sup_{g\in\mathcal{G}}\{\frac{1}{m}\sum_{i=1}^m g(z_i) - \mathbb{E}_z[g(z)]\}$. We first bound the expectations of X^+ and X^-. Appealing to Shalev-Shwartz and Ben-David [] (lemma 26.2), we deduce $\mathbb{E}[X^+] \le 2\mathbb{E}_S\mathcal{R}(\mathcal{G}\circ S)$. Replacing $\mathcal{G}$ with $-\mathcal{G}$ and using Shalev-Shwartz and Ben-David [53] (lemma 26.2), we also learn $\mathbb{E}[X^-] \le 2\mathbb{E}_S\mathcal{R}(-\mathcal{G}\circ S) = 2\mathbb{E}_S\mathcal{R}(\mathcal{G}\circ S)$. Next, a quick computation shows

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

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 $\max\{X^+, X^-\} \le 2\mathbb{E}_S\mathcal{R}(\mathcal{G}\circ S) + t$. Noting the equality $\psi^\star_{m^{-1}\varepsilon Y}(t/m) = \psi^\star_{\varepsilon Y}(t)$ completes the proof. □

Published in Mathematics of Operations Research on April 20, 2021 as DOI: 10.1287/moor.2021.1126.
This article has not been copyedited or formatted. The final version may differ from this version.

The following theorem provides an upper bound on the Rademacher complexity of linear classes; see the original article (Kakade et al. [26]) or the monograph (Shalev-Shwartz and Ben-David [53], lemma 26.10).

Lemma A.3 (Rademacher Complexity of Linear Classes)**.** *Consider the set* $A = \{(\langle w, z_1\rangle, \ldots, \langle w, z_m\rangle) : \|w\| \le 1\}$, *where* $z_1 \ldots, z_m$ *are arbitrary points. Then the estimate holds*:

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

The class of loss functions $\mathcal{G}$ considered 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 [33].

Theorem A.4 (Contraction Inequality [Maurer [33], theorem 3])**.** *Let* $\mathcal{X}$ *denote a countable set. For* $i = 1, \ldots, m$, *let* $F_i : \mathcal{S} \to \mathbb{R}$ *and* $G_i : \mathcal{S} \to \mathbb{R}^K$ *be functions satisfying*

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

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:*

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

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

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

where the last inequality uses the bound $\|x - x_0\| \le B$ twice. Notice the right-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

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

For every index $s \in \mathcal{S}$ and $i \in \{1, \ldots, m\}$, 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

$$\begin{aligned}
\mathcal{R}(\mathcal{F} \circ S) &= \frac{1}{m} \sup_{x\in\mathcal{X}} \sum_{i=1}^m \sigma_i f(x, z_i) \\
&= \frac{1}{m} \sup_{x\in\mathcal{X}} \sum_{i=1}^m \sigma_i l\big(\big(\langle x, \phi_1(z_i)\rangle, \ldots, \langle x, \phi_K(z_i)\rangle\big), z_i\big) \\
&= \frac{1}{m} \sup_{s\in\mathcal{S}} \sum_{i=1}^m \sigma_i F_i(s) = \mathcal{R}(F \circ \mathcal{S}) \le \sqrt{2}K \cdot \mathcal{R}(G \circ \mathcal{S}),
\end{aligned} \tag{A.1}$$

where the last inequality follows from Theorem A.4.

Published in Mathematics of Operations Research on April 20, 2021 as DOI: 10.1287/moor.2021.1126.
This article has not been copyedited or formatted. The final version may differ from this version.

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 $A' = \{(\langle x, L(z_i)\phi_k(z_i)\rangle)_{i,k} : \|x\| \le 1\}$. Thus, applying Lemma A.3, we deduce $\mathcal{R}(\mathcal{G} \circ \mathcal{S}) \le \frac{\sqrt{\sum_{i,k} B^2 L(z_i)^2 \|\phi_k(z_i)\|^2}}{mK}$. Combining this estimate with (A.1) and taking expectations yields

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

Appealing to Theorem A.2 completes the proof. □

Endnotes

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

[2] The symbol ∂g usually refers to an "f-attentive closure" of the construction defined in (2.1). The closure, however, is superfluous for all functions we consider here, and therefore the abuse of notation should cause no confusion.

References

[1] Albano P, Cannarsa P (1999) Singularities of semiconcave functions in Banach spaces.McEneaney WM, Yin GG, Zhang Q, eds. *Stochastic analysis, control, optimization and application. Systems & Control: Foundations & Applications* (Birkhäuser, Boston), 171–190.
[2] Attouch H (1977) Convergence de fonctions convexes, des sous-différentiels et semi-groupes associés. *C. R. Acad. Sci. Paris Sér. A-B.* 284(10):A539–A542.
[3] Attouch H (1984) *Variational Convergence for Functions and Operators*. Applicable Mathematics Series (Pitman Advanced Publishing Program, Boston).
[4] Attouch H, Wets RJB (1990) Epigraphical processes: laws of large numbers for random LSC functions. *Sém. Anal. Convexe 20. Exp*. 13:29.
[5] Bartlett PL, Bousquet O, Mendelson S (2005) Local Rademacher complexities. *Ann. Statist.* 33(4):1497–1537.
[6] Bartlett PL, Mendelson S (2002) Rademacher and gaussian complexities: Risk bounds and structural results. *J. Mach. Learn. Res.* 3(Nov):463–482.
[7] Boucheron S, Lugosi G, Massart P (2013) *Concentration Inequalities: A Nonasymptotic Theory of Independence* (Oxford University Press, Oxford, UK).
[8] Bousquet O, Elisseeff A (2002) Stability and generalization. *J. Mach. Learn. Res*. 2(3):499–526.
[9] Clarke F (1983) *Optimization and Nonsmooth Analysis* (Wiley Interscience, New York).
[10] Davis D, Drusvyatskiy D (2018) Uniform graphical convergence of subgradients in nonconvex optimization and learning. Preprint, submitted December 17, https://arxiv.org/pdf/1810.07590.pdf.
[11] Davis D, Drusvyatskiy D (2019) Stochastic model-based minimization of weakly convex functions. *SIAM J. Optim.* 29(1):207–239.
[12] Davis D, Drusvyatskiy D, Kakade S, Lee J (2018) Stochastic subgradient method converges on tame functions. Preprint, submitted April 20, https://arxiv.org/abs/1804.07795.
[13] Davis D, Drusvyatskiy D, MacPhee K (2018) Stochastic model-based minimization under high-order growth. Preprint, submitted July 1, https://arxiv.org/abs/1807.00255.
[14] Davis D, Drusvyatskiy D, MacPhee KJ, Paquette C (2018) Subgradient methods for sharp weakly convex functions. *J. Optim. Theory Appl.* 179(3):962–982.
[15] Davis D, Drusvyatskiy D, Paquette C (2017) The nonsmooth landscape of phase retrieval. Preprint, submitted November 9, https://arxiv.org/abs/1711.03247.
[16] Davis D, Grimmer B (2019) Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM J. Optim.* 29(3):1908–1930.
[17] Drusvyatskiy D (2018) The proximal point method revisited. *SIAG/OPT Views News*. 26(1):1–8.
[18] Drusvyatskiy D, Paquette C (2019) Efficiency of minimizing compositions of convex functions and smooth maps. *Math. Program.* 178(1-2):503–558.
[19] Duchi JC, Ruan F (2018) Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval. *Inform. Inference* 8(3):471–529.
[20] Duchi JC, Ruan F (2018) Stochastic methods for composite and weakly convex optimization problems. *SIAM J. Optim.* 28(4):3229–3259.
[21] Foster D, Sekhari A, Sridharan K (2018) Uniform convergence of gradients for non-convex learning and optimization. Preprint, submitted October 25, https://arxiv.org/abs/1810.11059.
[22] Geyer CJ (1994) On the asymptotics of constrained M-estimation. *Ann. Statist*. 22(4):1993–2010.
[23] Ghadimi S, Lan G, Zhang H (2016) Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Math. Programming* 155(1-2, Ser. A):267–305.
[24] Grünwald PD, Mehta NA (2016) Fast rates for general unbounded loss functions: from ERM to generalized Bayes. Preprint, submitted May 1, https://arxiv.org/abs/1605.00252.
[25] Hardt M, Recht B, Singer Y (2015) Train faster, generalize better: Stability of stochastic gradient descent. Preprint, submitted September 3, https://arxiv.org/abs/1509.01240.
[26] Kakade SM, Sridharan K, Tewari A (2009) On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *Adv. Neural Inform. Process. Syst.* 21:793–800.
[27] Kaniovski YM, King AJ, Wets RJB (1995) Probabilistic bounds (via large deviations) for the solutions of stochastic programming problems. *Ann. Oper. Res*. 56(1):189–208.
[28] King AJ, Rockafellar RT (1993) Asymptotic theory for solutions in statistical estimation and stochastic programming. *Math. Oper. Res.* 18(1):148–162.

[29] Koller D, Schuurmans D, Bengio Y, Bottou L, eds. (2009) *Fast Rates for Regularized Objectives*, vol. 21, Advances in Neural Information Processing Systems (Curran Associates, Inc., Red Hook, NY), 1545–1552.
[30] Kontorovich A (2014) Concentration in unbounded metric spaces and algorithmic stability. *Internat. Conf. Mach. Learn. 32*(2):28–36.
[31] Li X, Zhihui Z, So AC, Vidal R (2018) Nonconvex robust low-rank matrix recovery. Preprint, submitted September 24, https://arxiv.org/abs/1809.09237.
[32] Liu M, Zhang X, Zhang L, Jin R, Yang T (2018) Fast rates of erm and stochastic approximation: Adaptive to error bound conditions. Preprint, submitted May 11, https://arxiv.org/abs/1805.04577.
[33] Maurer A (2016) A vector-contraction inequality for Rademacher complexities. Ortner R, Simon HU, Zilles S, eds. *Algorithmic Learning Theory* (Springer International Publishing, Cham, Switzerland), 3–17.
[34] McDiarmid C (1989) On the method of bounded differences. Surveys in Combinatorics, 1989 (Norwich, 1989), London Mathematical Society Lecture Note Series, vol. 141 (Cambridge University Press, Cambridge), 148–188.
[35] Mehta NA (2016) Fast rates with high probability in exp-concave statistical learning. Preprint, submitted May 4, https://arxiv.org/abs/1605.01288.
[36] Mehta NA, Williamson RC (2014) From stochastic mixability to fast rates. *Proc. 27th Internat. Conf. Neural Inform. Processing Sys.* (MIT Press, Cambridge, MA).
[37] Mei S, Bai Y, Montanari A (2018) The landscape of empirical risk for nonconvex losses. *Ann. Statist*. 46(6A):2747–2774.
[38] Mordukhovich B (2006) *Variational Analysis and Generalized Differentiation I: Basic Theory*, vol. 330 (Grundlehren der mathematischen Wissenschaften, Springer, Berlin).
[39] Mordukhovich BS (2018) *Variational Analysis and Applications*. Springer Monographs in Mathematics (Springer, Cham, Switzerland).
[40] Moreau JJ (1965) Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France* 93:273–299.
[41] Nemirovski A, Juditsky A, Lan G, Shapiro A (2008) Robust stochastic approximation approach to stochastic programming. *SIAM J. Optim*. 19(4):1574–1609.
[42] Nurminskii EA (1973) The quasigradient method for the solving of the nonlinear programming problems. *Cybernetics* 9(1):145–150.
[43] Poliquin R, Rockafellar R (1996) Prox-regular functions in variational analysis. *Trans. Amer. Math. Soc.* 348:1805–1838.
[44] Rachev ST, Römisch W (2002) Quantitative stability in stochastic programming: The method of probability metrics. *Math. Oper. Res*. 27(4):792–818.
[45] Rakhlin A, Mukherjee S, Poggio T (2005) Stability results in learning theory. *Anal. Appl.* 3(4):397–417.
[46] Ralph D, Xu H (2011) Convergence of stationary points of sample average two-stage stochastic programs: a generalized equation approach. *Math. Oper. Res*. 36(3):568–592.
[47] Robinson SM (1996) Analysis of sample-path optimization. *Math. Oper. Res*. 21(3):513–528.
[48] Rockafellar RT (1970) *Convex Analysis*. Princeton Mathematical Series, No. 28 (Princeton University Press, Princeton, NJ).
[49] Rockafellar R (1982) Favorable classes of Lipschitz-continuous functions in subgradient optimization. *Progress in Nondifferentiable Optimization*, vol. 8, IIASA Collaborative Proc. Ser. CP-82 (International Institute for Applied Systems Analysis, Laxenburg, Austria), 125–143.
[50] Rockafellar R, Wets RB (1998) *Variational Analysis*, vol. 317 (Grundlehren der mathematischen Wissenschaften, Springer, Berlin).
[51] Rolewicz S (1979) On paraconvex multifunctions. *Third Symposium on Operations Research* (Univ. Mannheim, Mannheim, 1978), section I, vol. 31, Oper. Res. Verfahren (Hain, Königstein/Ts.), 539–546.
[52] Römisch W, Wets R (2007) Stability of ϵ-approximate solutions to convex stochastic programs. *SIAM J. Optim*. 18(3):961–979.
[53] Shalev-Shwartz S, Ben-David S (2014) *Understanding Machine Learning: From Theory to Algorithms*, 1st ed. (Cambridge University Press, Cambridge, UK).
[54] Shalev-Shwartz S, Shamir O, Srebro N, Sridharan K (2009) Stochastic convex optimization. *Proc. Conf. Learn. Theory* (*COLT*).
[55] Shapiro A (2000) On the asymptotics of constrained local M-estimators. *Ann. Statist*. 28(3):948–960.
[56] Shapiro A (2000) *Stochastic Programming by Monte Carlo Simulation Methods* (Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät II, Institut für Mathematik, Berlin).
[57] Shapiro A, Homem-de Mello T (2000) On the rate of convergence of optimal solutions of Monte Carlo approximations of stochastic programs. *SIAM J. Optim*. 11(1):70–86.
[58] Shapiro A, Xu H (2007) Uniform laws of large numbers for set-valued mappings and subdifferentials of random functions. *J. Math. Anal. Appl*. 325(2):1390–1399.
[59] Shapiro A, Dentcheva D, Ruszczyński A (2014) Lectures on stochastic programming, vol. 9, MOS-SIAM Series on Optimization (Society for Industrial and Applied Mathematics (SIAM), Mathematical Optimization Society, Philadelphia, PA).
[60] Toulis P, Airoldi E (2017) Asymptotic and finite-sample properties of estimators based on stochastic gradients. *Ann. Statist*. 45(4):1694–1727.
[61] van Erven T, Grünwald PD, Mehta NA, Reid MD, Williamson RC (2015) Fast rates in statistical and online learning. *J. Mach. Learn. Res.* 16:1793–1861.
[62] Vershynin R (2010) Introduction to the non-asymptotic analysis of random matrices. Preprint, submitted November 12, https://arxiv.org/abs/1011.3027.
[63] Vershynin R (2018) *High-Dimensional Probability: An Introduction with Applications in Data Science*, vol. 47 (Cambridge University Press, Cambridge, UK).
[64] Xu H (2010) Uniform exponential convergence of sample average random functions under general sampling with applications in stochastic programming. *J. Math. Anal. Appl.* 368(2):692–710.
[65] Xu H, Zhang D (2009) Smooth sample average approximation of stationary points in nonsmooth stochastic optimization and applications. *Math. Programming* 119(2):371–401.
[66] Zemel RS, Culotta A (2010) Smoothness, low noise and fast rates. Lafferty JD,Williams CKI, Shawe-Taylor J, eds. vol. 23, *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Red Hook, NY), 2199–2207.
[67] Zhang S, He N (2018) On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization. Preprint, submitted June 12, https://arxiv.org/abs/1806.04781.
[68] Zinkevich M (2003) Online convex programming and generalized infinitesimal gradient ascent. *Proc. 20th Internat. Conf. Mach. Learn.* (AAAI Press), 928–935.