Generalization Bounds for (Wasserstein) Robust Optimization

Anonymous Author(s)
Affiliation
Address
email

Abstract

(Distributionally) robust optimization has gained momentum in machine learning community recently, due to its promising applications in developing generalizable learning paradigms under stochastic or adversarial environments. In this paper, we derive finite-sample generalization bounds for robust optimization and Wasserstein distributionally robust optimization. We consider a broad class of piecewise Hölder smooth loss functions, under both stochastic setting—i.i.d. or weakly dependent data—and adversarial setting, assuming that the underlying data-generating distribution satisfies transportation-information inequalities. The proofs are built on a general connection between robustness and variation regularization (including Lipschitz and gradient regularization among others), as well as new local Rademacher complexity results for variation regularization. Our theory is illustrated for various machine learning tasks, including supervised learning, principal component analysis, learning with Markovian data and risk-averse optimization.

1 Introduction

Consider the following stochastic optimization problem of seeking a function $f : \mathscr{X} \to \mathbb{R}$ from the hypothesis space $\mathscr{F}$ so as to minimize the expected loss:

$$\min_{f \in \mathscr{F}} \left\{ \mathcal{L}^{\mathsf{true}}(f) := \mathbb{E}_{X \sim \mathbb{P}_{\mathsf{true}}}[f(X)] \right\}.$$

Here the data X follows some underlying distribution $\mathbb{P}_{\mathsf{true}}$ on a space $\mathscr{X} \subset (\mathbb{R}^d, \|\cdot\|)$. In practice, the underlying distribution is hardly known exactly, but instead one solves empirical risk minimization by replacing $\mathbb{P}_{\mathsf{true}}$ with its empirical estimate $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\delta}_{\hat{x}^i}$, where $\boldsymbol{\delta}$ denotes the Dirac measure:

$$\min_{f \in \mathscr{F}} \left\{ \mathcal{L}_n^{\mathsf{erm}}(f) := \mathbb{E}_{X \sim \mathbb{P}_n}[f(X)] \right\}.$$

The discrepancy between $\mathbb{P}_{\mathsf{true}}$ and $\mathbb{P}_n$ may be resulted from (but not limited to):

(i) The sampling error of $\mathbb{P}_n$ when data are independently drew from $\mathbb{P}_{\mathsf{true}}$;

(ii) The empirical distribution $\mathbb{P}_n$ consists of non-i.i.d. data, for instance, when x^i's are trajectory samples from a Markov chain and $\mathbb{P}_{\mathsf{true}}$ is interpreted as the stationary distribution;

(iii) The environment where the decision will be deployed may be different from (albeit related to) the data-collecting environment, and such difference may be adversarial.

Such discrepancy is one of the main reasons for the poor generalization behavior of the resulting solution to the empirical risk minimization. To tackle this issue, *robust optimization* (Ben-Tal et al., 2009; Xu et al., 2009, 2010) considers the formulation

$$\min_{f \in \mathscr{F}} \left\{ \mathcal{L}_{n,p}^{\mathsf{ro}}(f; \varrho) := \max_{\{x^i\}_{i=1}^n \in \mathscr{X}} \left\{ \frac{1}{n}\sum_{i=1}^n f(x^i) : \left(\frac{1}{n}\sum_{i=1}^n \|x^i - \hat{x}^i\|^p\right)^{\frac{1}{p}} \le \varrho \right\} \right\}. \tag{RO}$$

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Here the uncertainty set, denoted by $\mathscr{X}_p(\varrho)$, describes the possible data perturbations that the model tries to hedge against; the radius $\varrho > 0$ specifies the level of robustness and $p \in [1, \infty]$. In a similar spirit, *Wasserstein distributionally robust optimization* (Esfahani & Kuhn, 2018; Blanchet & Murthy, 2019; Gao & Kleywegt, 2016) considers the formulation

$$\min_{f \in \mathscr{F}} \left\{ \mathcal{L}^{\mathsf{wo}}_{n,p}(f; \varrho) := \max_{\mathbb{P}} \left\{ \mathbb{E}_{x \sim \mathbb{P}}[f(x)] : \mathcal{W}_p(\mathbb{P}, \mathbb{P}_n) \leq \varrho \right\} \right\}. \tag{WO}$$

Here the distributional uncertainty set, denoted by $\mathscr{P}_p(\varrho)$, consists of all probability distributions that are close to $\mathbb{P}_n$ in Wasserstein distance of order p. When $p = \infty$, (WO) and (RO) coincide. Both problems have received increasing attentions in machine learning; see Kuhn et al. (2019) for a survey.

In the stochastic setting, the empirical data is usually modelled to be generated in an i.i.d. fashion or according to a Markov process (see (i)(ii) above). In this case, we would like to choose the radius $\varrho = \varrho_n$ adaptive to the sample size, so that the robust loss $\mathcal{L}^{\mathsf{ro}}_{n,p}(\varrho_n)$ or $\mathcal{L}^{\mathsf{wo}}_{n,p}(\varrho_n)$ can balance between performance and robustness. More precisely, ϱ_n should be large enough so that the robust loss dominates the true loss with high probability (up to negligible error), while at the same time, it should not be too large to produce a solution with inferior value.

In the adversarial setting (see (iii) above), the testing environment can be adversarially different from the training environment. More specifically, consider the *adversarial loss* minimization problem

$$\min_{f \in \mathscr{F}} \left\{ \mathcal{L}^{\mathsf{adv}}_{p}(f; \varrho) := \max_{\mathbb{P} \in \mathscr{P}_p(\varrho)} \left\{ \mathbb{E}_{x \sim \mathbb{P}}[f(x)] : \mathcal{W}_p(\mathbb{P}, \mathbb{P}_{\mathsf{true}}) \leq \varrho \right\} \right\}, \tag{Adv}$$

where the radius ϱ captures the difference between the training and testing environments. When $\mathbb{P}_{\mathsf{true}}$ is not available, a surrogate problem (WO) based on $\mathbb{P}_n$ is often considered. In this case, the goal is to find a robust solution against adversarial data perturbations with a fixed radius ϱ.

The goal of this paper is to provide a comprehensive analysis of the generalization properties of the two robust learning paradigms for stochastic optimization with generic loss functions. Specifically, in the stochastic setting, under proper choice of the radius ϱ_n, we derive finite-sample guarantees for (RO) and (WO) that apply to non-i.i.d. data and non-smooth losses. In the adversarial setting, we derive generalization bounds for (Adv) with fixed small radius ϱ. We exemplify our results for various machine learning tasks beyond the empirical risk minimization, including principal component analysis, learning with Markovian data, risk-averse optimization, and adversarial learning.

Related Literature

The most related work to ours is Gao (2020), but we differ in several important ways. First, Gao (2020) studies only (WO) while we also consider (RO) and establish new generalization bounds for (RO) in a general setting. Second, Gao (2020) studies i.i.d. data while we also consider non-i.i.d. and adversarial data. For non-i.i.d. data, we exploit different notions of transport inequality and Rademacher complexity than the i.i.d. case. For adversarial data, the analysis is completely different. Third, even in the i.i.d. setting, Gao (2020) studies only smooth losses and Wasserstein order $p = 1, 2$. We consider non-smooth losses and $p \in [1, 2]$, which includes important cases like $|\theta^\top x - y|^p$ and CVaR minimization. Moreover, for $p = 1$, we tighten the remainder from $O(1/\sqrt{n})$ to $O(1/n)$; for $p = 2$, we obtain a cleaner expression using a different set of assumptions; and for $p \in (1, 2)$, we generalize the sub-root functions which require careful control on local Radamecher complexity. All these extensions require in-depth analysis on worst-case scenarios.

For (RO) in the stochastic setting, its generalization properties were first studied in the pioneer work Xu & Mannor (2012), which established algorithmic generalization generalization bounds for various learning tasks with i.i.d. and Markovian data. Nonetheless, many bounds developed therein were suboptimal. This framework is then extended to metric learning (Bellet & Habrard, 2015) and ensemble models (Zahavy et al., 2016). Our results circumvent the curse of dimensionality of the bounds developed in Xu & Mannor (2012).

For (WO) with i.i.d. data, Esfahani & Kuhn (2018) originally proposed to choose the radius ϱ_n such that $\mathbb{P}_{\mathsf{true}}$ is contained in the uncertainty set with high probability, but the resulting performance bound is overly conservative and suffers from the curse of dimensionality. This issue has been mitigated asymptotically for smooth loss functions (Blanchet et al., 2019a,b; Si et al., 2020), and non-asymptotically for linear models (Shafieezadeh-Abadeh et al., 2019; Chen & Paschalidis, 2018) and for smooth loss functions (Gao, 2020). As mentioned above, our results apply to non-i.i.d. data and non-smooth losses, and yield optimal rate for various problems.

In the adversarial setting, generalization properties of (WO) with $p \in [1, \infty)$ and fixed ϱ was considered first in Lee & Raginsky (2018) under the context of domain adaptation. Their bounds are mostly useful when the radius is large while our focus is on small radius. Generalization bounds of (WO) with $p = \infty$ (or equivalently (RO)) with small fixed radius have been studied extensively in adversarial learning (Szegedy et al., 2013; Goodfellow et al., 2014; Schmidt et al., 2018; Tsipras et al., 2018; Bubeck et al., 2019; Zhang et al., 2019), and generalization bounds have been developed for linear hypotheses and neural networks (Khim & Loh, 2018; Yin et al., 2019; Awasthi et al., 2020).

Part of our analysis is based on the connection between robustness and regularization. The connection between (RO)(WO) and norm regularization was established for various machine learning tasks (Xu et al., 2010, 2009; Bertsimas & Copenhaver, 2018; Blanchet et al., 2019a; Chen & Paschalidis, 2018), and the connection between (WO) and Lipschitz/gradient regularization has been established for smooth loss functions (Shafieezadeh-Abadeh et al., 2019; Sinha et al., 2018; Finlay et al., 2018; Cranko et al., 2018; Blanchet et al., 2019b; Bartl et al., 2020) and non-smooth loss functions (Gao et al., 2017). Our results improve the sandwich bounds in Gao et al. (2017) for $p = 1$, and generalizes the results therein to piecewise Hölder smooth loss functions for $p \in (1, \infty]$.

The rest of the paper proceeds as follows. In Section 2, we briefly review several important concepts in learning theory that serve as building blocks of our results. In Section 3, we relate robust optimization with variation regularization, based on which we derived the generalization bounds for stochastic setting and for adversarial setting respectively in Sections 4 and 5. As the results for robust optimization (RO) are very similar to (WO) in the stochastic setting, we only present (RO) in the main text and defer the counterpart results of (WO) in the supplementary material. Finally, we illustrate our theory via examples in Section 6. All proofs are provided in the supplementary material.

2 Preliminaries

In this section, we prepare several technical tools that are useful in deriving our main results.

Notations. For $p \in [1, \infty]$, we denote by q its Hölder conjugate number, i.e. $\frac{1}{p} + \frac{1}{q} = 1$. Let $\mathscr{P}(\mathscr{X})$ be the set of Borel probability distributions on $\mathscr{X}$, and let $\mathscr{P}_p(\mathscr{X}) = \{\mathbb{P} \in \mathscr{P}(\mathscr{X}) : \mathbb{E}_{\mathbb{P}}[\|x\|^p] < \infty\}$. For a given probability measure $\mathbb{Q}$, denote by $(L^p_{\mathbb{Q}}(\mathscr{X}), \|\cdot\|_{p,\mathbb{Q}})$ the set of L^p functions on $\mathscr{X}$ with respect to $\mathbb{Q}$. Denote by $\|f\|_{\text{Lip}}$ the Lipschitz norm of f, whenever it exists. We will not assume the samples are i.i.d. Instead, we use $S = \{\hat{x}^i\}_{i=1}^n$ to represent a sample of size n, and denote by μ_S the sampling distribution and use $\mathbb{E}_S$ to indicate the expectation with respect to the μ_S. Furthermore, we assume the sample $\{\hat{x}^i\}_{i=1}^n$ converges to stationary distribution, and $\mathbb{E}_{\mathbb{P}_{\mathsf{true}}}[\cdot]$ is interpreted as the expectation taken with respect to the stationary distribution. For a fixed set of n points, we denote $\mathbb{P}_n$ as the empirical uniform distribution on these n points. For two real numbers, a and b, we denote $a \wedge b = \min(a, b)$, and $a \vee b = \max(a, b)$.

Variation regularization. Following Gao et al. (2017), we define the *local* and *global slope* of a continuous function $f : \mathscr{X} \to \mathbb{R}$ at $x \in \mathscr{X}$, denoted by $|\partial f|(x)$ and $G(f)(x)$ respectively, as

$$|\partial f|(x) = \limsup_{\|\tilde{x}-x\| \to 0} \frac{(f(\tilde{x}) - f(x))_+}{\|\tilde{x} - x\|}, \quad G(f)(x) = \sup_{\tilde{x} \in \mathscr{X}} \frac{(f(\tilde{x}) - f(x))_+}{\|\tilde{x} - x\|}.$$

When f is differentiable at x, we have $|\partial f|(x) = \|\nabla f(x)\|_*$. When f is L-Lipschitz, we have $G(f)(x) \le L$ for all $x \in \mathscr{X}$. We define the *variation regularization* problem as

$$\min_{f \in \mathscr{F}} \mathcal{L}^{\mathsf{vr}}_{n,q}(f; \varrho) := \begin{cases} \frac{1}{n}\sum_{i=1}^n f(\hat{x}^i) + \varrho \||\partial f|\|_{q,\mathbb{P}_n}, & q < \infty, \\ \frac{1}{n}\sum_{i=1}^n f(\hat{x}^i) + \varrho \|G(f)\|_{\infty,\mathbb{P}_n}, & q = \infty. \end{cases} \tag{VR}$$

Denote by $\mathscr{D}_f$ the set of non-differentiable points of f. For $x \in \mathscr{X}$, we denote by $d(x, \mathscr{D}_f) := \inf_{y \in \mathscr{X}} \|x - y\|$ the distance between x and $\mathscr{D}_f$ When f is differentiable, we adopt the convention that $d(x, \mathscr{D}_f) = \infty$ for all $x \in \mathscr{X}$. We introduce several classes of functions that will play an important role in our analysis:

$$\begin{aligned} &\partial\mathscr{F}_q := \left\{|\partial f|^q : f \in \mathscr{F}\right\}, \quad \mathscr{N}_q := \left\{\frac{|\partial f|^q}{\||\partial f|\|^q_{q,\mathbb{P}_{\mathsf{true}}}} : f \in \mathscr{F}\right\}, \\ &\mathscr{J}_\varrho := \left\{x \mapsto 1\{d(x, \mathscr{D}_f) < \varrho\} : f \in \mathscr{F}, \mathscr{D}_f \neq \varnothing\right\}. \end{aligned} \tag{1}$$

Transport inequalities. In the stochastic setting, we will work with distributions that satisfy the following condition. Let $p \in [1, \infty)$ and $\tau > 0$. A distribution $\mathbb{Q} \in \mathscr{P}_p(\mathscr{X})$ satisfies a *transport inequality*, denoted as $T_p(\tau)$, if for all $\mathbb{P} \in \mathscr{P}_p(\mathscr{X})$ it holds that

$$\mathcal{W}_p^2(\mathbb{P}, \mathbb{Q}) \leq \tau \int_{\mathscr{X}} \log(d\mathbb{P}/d\mathbb{Q}) d\mathbb{P},$$

where $d\mathbb{P}/d\mathbb{Q}$ denotes the Radon-Nikodym derivative. The integral on the right-hand side of the inequality above represents the relative entropy between $\mathbb{P}$ and $\mathbb{Q}$. The transportation-information inequality $T_p(\tau)$ covers a variety of useful classes of distributions. For example, if $\mathscr{X}$ is bounded by B, then any distribution on $\mathscr{X}$ satisfies $T_1(8B^2)$. Gaussian distributions with covariance Σ satisfies $T_2(2\lambda_{\max}(\Sigma))$. Interested readers are referred to Gozlan & Léonard (2010) for a recent survey.

Local Rademacher complexity. The empirical Rademacher complexity of a class of functions $f \in \mathscr{F}$, with respect to a set $S = \{x^i\}_{i=1}^n$ is defined as

$$\widehat{\mathfrak{R}}_S(\mathscr{F}) := \mathbb{E}_\epsilon \big[\sup_{f \in \mathscr{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x^i) \big],$$

where $\mathbb{P}(\epsilon_i = \pm 1) = \frac{1}{2}$ for all i. The Rademacher complexity of a class of functions of sample size n is defined as $\mathfrak{R}_n(\mathscr{F}) := \mathbb{E}_S[\widehat{\mathfrak{R}}_S(\mathscr{F})]$. Properties of Rademacher complexity, particularly those for non-i.i.d. sequences (Mohri & Rostamizadeh, 2008) are provided in the supplementary material. The *localized Rademacher complexity* of level $r > 0$ of sample size n is defined as $\mathfrak{R}_n\big(\{cf : f \in \mathscr{F}, 0 \leq c \leq 1, K(cf) \leq r\}\big)$, where $K : \mathscr{F} \to \mathbb{R}^+$ is a given functional such as expectation or variance. In the literature, one often assumes that the localized Rademacher complexity can be bounded by a sub-root function. For our purpose, we define a more general *sub-q-root* function. A function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ is sub q-root, if it is non-constant, non-decreasing, and the mapping $r \mapsto \psi(r)/r^{\frac{1}{q}}$ is non-increasing. When $q = 2$, this reduces to the sub-root function.

3 Robustness and Variation Regularization

In this section, we establishes connection between robust optimization (RO), Wasserstein distributionally robust optimization (WO) and variation regularization (VR).

We first consider $p > 1$, whose worst-case behavior is fundamentally different from the case of $p = 1$. The following two assumptions cover most loss functions in machine learning that are amenable to optimize efficiently using first-order methods.

Assumption 1. Every $f \in \mathscr{F}$ is L-Lipschitz and piecewise differentiable.

Assumption 2. There exists $h, \alpha > 0$ such that for every $f \in \mathscr{F}$ and every $\tilde{x}, x$ in the same piece of f, it holds that $\|\nabla f(\tilde{x}) - \nabla f(x)\|_* \leq h\|\tilde{x} - x\|^\alpha$.

In the sequel, to ease the notation, we introduce the a function $e_n : \mathbb{R}_+ \to \mathbb{R}_+$, defined as

$$e_n(\rho) = \sup_{f \in \mathscr{F}} \Big\{ \mathbb{E}_{X \sim \mathbb{P}_n} \Big[(\rho - d(X, \mathscr{D}_f))_+ \Big] \Big\}. \tag{2}$$

Our first result in this section bounds the difference between (WO) and (VR).

Theorem 1. *Let $p \in (1, \infty]$ and $\varrho \geq 0$. Assume Assumptions 1 and 2 hold. Set $\eta = \inf_{f \in \mathscr{F}} \|\|\partial f\|\|_{q, \mathbb{P}_n}$. Then it holds that*

$$|\mathcal{L}_{n,p}^{\mathsf{wo}}(f; \varrho) - \mathcal{L}_{n,q}^{\mathsf{vr}}(f; \varrho)| \leq C_0 \varrho^{(\alpha+1)\wedge p} + e_n\big((pL\eta)^{q-1}\varrho\big), \quad \forall f \in \mathscr{F},$$

where $C_0 = 1$ if $p = \infty$, $C_0 = h \max(1, Lp/\eta)$ if $p \in (\alpha+1, \infty)$, and $C_0 = L(h/L)^{\frac{p-1}{\alpha}}$ otherwise.

When $e_n(\varrho) = o(\varrho)$, as is often the case (see Section 6), the the right hand side of the bound is $o(\varrho)$. When every $f \in \mathscr{F}$ is differentiable, the second term on the right-hand side of the bound vanishes. The proof is based on involve Taylor expansion and the duality for (WO). Since we consider different assumptions from Gao et al. (2017), we have to make different arguments and constructions of feasible distributions, dependent on whether $p > \alpha + 1$ or not.

Assumption 3. $e_n(\rho)/\rho \to 0$ as $\rho \to 0$ and $\mathfrak{R}_n(\partial \mathscr{F}_q) \to 0$ as $n \to \infty$.

Our second result establishes a relationship between (WO) and (RO).

Theorem 2. *Assume Assumptions 1, 2, 3 hold. Suppose $\varrho_n \to 0$ as $n \to \infty$. Then for any $t > 0$, and $p \in (1, \infty)$, there exits an integer N such that for all $n > N$, with probability at least $1 - e^{-t}$,*

$$0 \le \mathcal{L}^{\mathsf{wo}}_{n,p}(f; \varrho_n) - \mathcal{L}^{\mathsf{ro}}_{n,p}(f; \varrho_n) \le \tfrac{2L^q}{n}, \quad \forall f \in \mathscr{F}.$$

This uniform upper bound on the gap between (WO) and (RO) is novel in the literature. Its proof is based on analysis on the deviation of the worst-case scenarios from the empirical data. Combining Theorems 1 and 2 yields

$$|\mathcal{L}^{\mathsf{ro}}_{n,p}(f; \varrho_n) - \mathcal{L}^{\mathsf{vr}}_{n,q}(f; \varrho_n)| \le C_0 \varrho_n^{(\alpha+1)\wedge p} + e_n\big((pL\eta)^{q-1}\varrho_n\big) + \tfrac{2L^q}{n}, \quad \forall f \in \mathscr{F}.$$

Existence of a positive η in defining C_0 can be ensured by the following assumption, with the help of a uniform version of Bennett's inequality (Lemma 1 in the supplemental).

Assumption 4. $\inf_{f \in \mathscr{F}} \||\partial f|\|_{q, \mathbb{P}_{\mathsf{true}}} > 0$.

Thus far, we have shown under Assumptions 1-4, $\mathcal{L}^{\mathsf{vr}}_{n,q}(f; \varrho_n)$, $\mathcal{L}^{\mathsf{ro}}_{n,p}(\varrho_n)$ and $\mathcal{L}^{\mathsf{wo}}_{n,p}(\varrho_n)$ only differ by a higher order term that depends on the smoothness of the loss function class. For smooth loss function classes, we have $e_n(\rho) = 0$. For non-smooth losses, in the supplemental material, we show that $e_n(\varrho) = O_p(\varrho^2 + \varrho/\sqrt{n} + \mathfrak{R}_n(\mathscr{J}_\varrho))$. As shown in Section 6, for many interesting examples we have $e_n(\varrho_n) = O(1/n)$ under the canonical radius $\varrho_n = O(1/\sqrt{n})$.

Next, we consider the case $p = 1$. In light of Example 3.3.1 in Gao et al. (2017), it is not always possible to have that $\mathcal{L}^{\mathsf{vr}}_{n,\infty}(\varrho) - \mathcal{L}^{\mathsf{wo}}_{n,1}(\varrho)$ is of a higher order than ϱ. Below we identify conditions in order to bound the gap. For $f \in \mathscr{F}$, we define cumulative distribution functions $H_f, \tilde{H}_f$ as

$$H_f(a) := \mathbb{P}_{\mathsf{true}}\big\{\|G(f)\|_{\infty, \mathbb{P}_{\mathsf{true}}} - G(f)(X) \le a\big\}, \;\; a \ge 0,$$
$$\tilde{H}_f(a) := \mathbb{P}_{\mathsf{true}}\big\{\|G(f)\|_{\infty, \mathbb{P}_{\mathsf{true}}} - |\partial f|(X) \le a\big\}, \;\; a \ge 0,$$

where the random variable X has distribution $\mathbb{P}_{\mathsf{true}}$, and recall $G(f), |\partial f|$ are the global and local slope functions of f defined in Section 2.

Assumption 5. Assume there exists $\beta, \bar{a}, c > 0$ such that for all $a \in [0, \bar{a})$ and all $f \in \mathscr{F}$, it holds that $H_f(a^\beta) \ge ca$ for (WO), and $\tilde{H}_f(a^\beta) \ge ca$ for (RO),

This assumption imposes the concentration of maximal slope, and large β indicates that $G(f)(X)$ or $|\partial f|(X)$ concentrates more on the maximal slope $\|G(f)\|_{\infty, \mathbb{P}_{\mathsf{true}}}$. If H_f (or $\tilde{H}_f$) is continuous, then Assumption 5 means that it is $1/\beta$-Hölder smooth at 0. For example, if f is a linear function, then the set of global slopes is a singleton and $H_f(a) = \tilde{H}_f(a) = 1$ for all $a \ge 0$, hence Assumption 5 is trivially satisfied. More generally, under the assumption that $\limsup_{\|x - \hat{x}\| \to \infty} \frac{(f(x) - f(\hat{x}))_+}{\|x - \hat{x}\|} = \|f\|_{\mathrm{Lip}}$, which is imposed by Gao (2020), the global slope $G(f)(x) = \|f\|_{\mathrm{Lip}}$ for all $x \in \mathscr{X}$, thereby $H_f(a) = 1$ for all $a \ge 0$.

Theorem 3. *Let $p = 1$ and $\varrho > 0$. Assume Assumptions 1, 2, 5 hold. Let $\delta \in [0, \frac{1}{2})$ and $t > 0$. Then for every large n satisfying $(\frac{t + 2n^{\frac{1}{2} - \delta}}{\sqrt{n}}) \le \bar{a}$, with probability at least $1 - e^{-2t^2}$, it holds that*

$$\mathcal{L}^{\mathsf{wo}}_{n,1}(f; \varrho), \mathcal{L}^{\mathsf{ro}}_{n,1}(f; \varrho) \in [\mathcal{L}^{\mathsf{vr}}_{n,\infty}(f; \varrho) - \epsilon_n, \, \mathcal{L}^{\mathsf{vr}}_{n,\infty}(f; \varrho)], \quad \forall f \in \mathscr{F},$$

where $\epsilon_n = c(\frac{t + 2n^{\frac{1}{2} - \delta}}{\sqrt{n}})^\beta L\varrho + 4hn^{\alpha\delta}\varrho^{\alpha+1} + 4Le_n(\varrho n^\delta)$.

Theorem 3 shows that (RO), (WO) and (VR) are equivalent up to a higher order error term ϵ_n, which depends on the smoothness of the loss function, indicated by α, and on the concentration of slopes, indicated by β. Ideally, we would like to have a large $\beta \ge 1$, in which case H_f has a Hölder constant less than or equal to 1, and in particular, it has a density at 0 when $\beta = 1$. When $\varrho_n = O(1/\sqrt{n})$ and $e_n(\varrho) = O(\varrho^2)$, we have $\epsilon_n = O(n^{(-\frac{1}{2} - \beta\delta) \vee (-\frac{\alpha+1}{2} + \alpha\delta) \vee (2\delta - 1)})$. The smoother the function and the more concentrated on the maximal global slope, the smaller the gap between (WO), (RO) and (VR) is. Particularly, when $G_f(X)$ has a positive mass on the maximal global slope and f has piecewise Lipschitz gradient, by setting $\delta = 0$, we have $\epsilon_n = O(1/n)$ under the canonical radius $\varrho_n = O(1/\sqrt{n})$, provided that $e_n(\varrho_n^2) = O(1/n)$. This improves the result of Gao et al. (2017, Section 3.3.2) in which $\epsilon_n = O(1/\sqrt{n})$. Our finer analysis is based on construction of feasible solutions for (WO) and (RO) obtained from perturbation of empirical data points as well as careful control on the distances of perturbation.

4 Generalization Bounds: Stochastic Setting

In this section, we derive performance guarantees for robust optimization when the data are generated stochastically, by making the following assumptions on the data-generating mechanism that facilitate us to deal with non-i.i.d. data.

Assumption 6. The sampling distribution μ_S satisfies $T_p(\tau_n)$.

When S contains i.i.d. samples from some underlying distribution that satisfies $T_p(\tau)$, then by tensorization of transport inequalities (Gozlan & Léonard, 2010), μ_S satisfies $T_p(\tau_n)$ with $\tau_n = \tau n^{\frac{2}{p}-1}$. In Section 6.3, we will consider Markovian data that satisfies $T_1(\tau_n)$ with $\tau_n = O(n)$.

Consistent with the literature on local Rademacher complexity, we make the following assumption.

Assumption 7. There is a sub q-root function ψ_n such that $\mathfrak{R}_n\big(\{cf : f \in \mathscr{F}, 0 \le c \le 1, K(cf) \le r\}\big) \le \psi_n(r)$. We denote the fixed point of ψ_n as $r_{n,\star}$.

This assumption generalizes the typical sub-root property in the literature, which facilitate us to consider loss functions Hölder smooth gradient.

We separate the case of $p = 1$ and $p > 1$.

Theorem 4. *Let $p = 1$. Assume Assumptions 1, 2 and 6 are in force, and Assumption 7 holds with $q = 2$ and $K(f) = \|f\|^2_{\mathrm{Lip}}$. Let $t > 0$, set $\varrho_n = \frac{2\sqrt{\tau_n t}}{n} + 4\sqrt{r_{n,\star}} + \frac{2}{n\sqrt{r_{n,\star}}}$. Then with probability at least $1 - e^{-2t^2} - \log_2(L\sqrt{\tau_n t})e^{-t}$, it holds that*

$$\mathcal{L}^{\mathsf{true}}(f) \le \mathcal{L}^{\mathsf{vr}}_{n,\infty}(f;\varrho_n) + 4r_{n,\star} + \tfrac{2}{n}, \quad \forall f \in \mathscr{F}.$$

Let $\delta \in [0, \frac{1}{2})$, further assume Assumption 5 holds. Then for every n satisfying $(\frac{t+2n^{\frac{1}{2}-\delta}}{\sqrt{n}}) \le \bar{a}$, with probability at least $1 - 2e^{-2t^2} - \log_2(L\sqrt{\tau_n t})e^{-t}$, it holds that

$$\mathcal{L}^{\mathsf{true}}(f) \le \mathcal{L}^{\mathsf{wo}}_{n,1}(f;\varrho_n) \wedge \mathcal{L}^{\mathsf{ro}}_{n,1}(f;\varrho_n) + 4r_{n,\star} + \tfrac{2}{n} + \epsilon_n, \quad \forall f \in \mathscr{F},$$

where ϵ_n is defined in Theorem 3.

The first part of Theorem 4 shows that by choosing ϱ_n properly, (VR) upper bounds the true risk with high probability up to a higher order term $4r_{n,\star} + \frac{2}{n}$, which is often $O(1/n)$, as will be shown by various examples in Section 6. The radius parameter ϱ_n depends on the sampling distribution parameter τ_n, the confidence level t, and the function class complexity $r_{n,\star}$. When $\tau_n = O(n)$ and $r_{n,\star} = O(1/n)$, we have $\varrho_n = O(1/\sqrt{n})$, corresponding to the canonical choice. The second part of Theorem 4 shows that with the same choice of the radius parameter, (WO) and (RO) upper bound the true risk with high probability up to a higher order term. We remark that our bound is tight, as can be seen in the case of linear prediction (cf. Section 6.1). The proof of Theorem 4 follows a similar flow as that of Gao (2020, Theorem 2), but our result extends it to non-i.i.d. settings and to (RO) and possess a tighter choice of the radius, thanks to our new regularization results developed in Section 3.

Theorem 5. *Let $p \in (1, 2]$, assume Assumptions 1, 2, 3, 4, 6 are in force, and Assumption 7 holds with $K(f) = \||\partial f|\|^q_{q,\mathbb{P}_{\mathsf{true}}}$. For $t > 0$, set*

$$\varrho_n = \Big(1 + 2\mathfrak{R}_n(\mathscr{N}_q) + (\tfrac{2L}{\eta})^2\sqrt{\tfrac{t}{2n}}\Big) \cdot \Big(2\sqrt{\tfrac{\tau_n t}{n^{2/p}}}\big[1 + \tfrac{2}{q}\mathfrak{R}_n(\mathscr{N}_q)\big] + \tilde{\epsilon}_n r_{n\star}^{-\frac{1}{q}} + 2^p r_{n\star}^{\frac{1}{p}} + \tfrac{p}{n} r_{n\star}^{-\frac{1}{q}}\Big).$$

Then there exists an integer N such that for all $n > N$, with probability at least $1 - (3 + \log_2(L\sqrt{\tau_n n^{2-\frac{2}{p}} t}))e^{-t}$, for every $f \in \mathscr{F}$, $\mathcal{L}^{\mathsf{true}}(f)$ is upper bounded by any of the following:

$$\mathcal{L}^{\mathsf{vr}}_{n,q}(f;\varrho_n) + \epsilon_{n,1}, \;\; \mathcal{L}^{\mathsf{wo}}_{n,p}(f;\varrho_n) + \epsilon_{n,2}, \;\; \mathcal{L}^{\mathsf{ro}}_{n,p}(f;\varrho_n) + \epsilon_{n,2} + \frac{2L^q}{n},$$

where $\tilde{\epsilon}_n = \epsilon_n + \frac{2}{qn}\mathfrak{R}_n(\mathscr{N}_q)$, and

$$\epsilon_n = 2Le_n\big((2Lp/\eta)^{\frac{1}{p-1}}\sqrt{\tfrac{\tau_n t}{n^{2/p}}}\big) + \big(L(h/L)^{\frac{p-1}{\alpha}}(\tfrac{\tau_n t}{n^{2/p}})^{\frac{p}{2}}\big)\mathbf{1}_{\{p \le \alpha+1\}} + \big(h(\tfrac{2Lp}{\eta})^{\frac{\alpha+1}{p-1}}(\tfrac{\tau_n t}{n^{2/p}})^{\frac{\alpha+1}{2}}\big)\mathbf{1}_{\{p > \alpha+1\}},$$

$$\epsilon_{n,1} = 2Le_n\big((\tfrac{2L}{\eta})^{q-1}\varrho\big) + \epsilon_n + \tfrac{2\mathfrak{R}_n(\mathscr{N}_q)}{n} + 2^p r_{n\star} + \tfrac{p}{n}, \;\; \epsilon_{n,2} = \epsilon_{n,1} + e_n\big((\tfrac{2L}{\eta})^{q-1}\varrho\big) + C\varrho^{(\alpha+1)\wedge p} + \tfrac{C}{n}.$$

Comparing with Theorem 4, for $p \in (1, 2]$, additional terms $\mathfrak{R}_n(\partial \mathscr{F}_q)$—related to the gradient norm (sharpness of losses), and $\mathfrak{R}_n(\mathscr{N}_q)$—related to the non-smoothness of losses, appear in the radius and the high-order error term. In Section 6 we will show that $\varrho_n = \tilde{O}(1/\sqrt{n})$ and $\tilde{\epsilon}_n, \epsilon_{n,1}.\epsilon_{n,2} = O(1/n)$ for various applications. Theorem 5 generalizes Gao (2020) which considers only for $p = 1, 2$ with smooth loss functions. Our result holds for piecewise smooth loss functions and any Wasserstein order $p \in [1, 2]$, which includes important cases of $|\theta^\top x - y|^p$ (Section 6.1) and $\mathbb{C}$V@R minimization (Section 6.5). The proof relies on new notions of sub-q-root properties of the function ψ_n, and hence requires different treatment on the fixed point of ψ_n and on the local Rademacher complexity.

5 Generalization Bounds: Adversarial Setting

In the adversarial setting (Adv), we are trying to minimize the expected loss under the worst-case scenario when the learning algorithm is being purposely attacked by an adversary. By replacing $\mathbb{P}_{\mathsf{true}}$ with $\mathbb{P}_n$ in (Adv), we obtain the empirical adversarial robustness problem, which is the same as the Wasserstein robust optimization problem (WO). One of the most fundamental questions in adversarial robustness problem is to estimate the generalization gap between (Adv) and its empirical counterpart (WO). Most literature has been focused on $p = \infty$ (Khim & Loh, 2018; Yin et al., 2019; Awasthi et al., 2020). Following the literature, we assume $\mathbb{P}_n$ consists of i.i.d. samples. We aim to derive generalization bounds for any $p \in [1, \infty]$ by virtue of Theorems 1 and 3.

Theorem 6. *Let $p = 1$. Assume Assumptions 1, 2, 5 hold, and that $\|f\|_\infty \le M$ for all $f \in \mathscr{F}$. Let $\delta \in [0, \frac{1}{2})$ and $t > 0$. Then there exists $C > 0$ such that for every n satisfying $(\frac{t+2n^{\frac{1}{2}-\delta}}{\sqrt{n}}) \le \bar{a}$, with probability at least $1 - 2e^{-2t^2}$, it holds that for every $f \in \mathscr{F}$,*

$$|\mathcal{L}^{\mathsf{wo}}_{n,1}(f;\varrho) - \mathcal{L}^{\mathsf{adv}}_1(f;\varrho)| \le 2\mathfrak{R}_n(\mathscr{F}) + M\sqrt{\tfrac{t}{2n}} + c\big(\tfrac{t}{\sqrt{n}} + \tfrac{1}{n}\big)^\beta L\varrho + c\big(\tfrac{t+2n^{\frac{1}{2}-\delta}}{\sqrt{n}}\big)^\beta L\varrho + 4hn^{\alpha\delta}\varrho^{\alpha+1} + 4Le_n(\varrho n^\delta) + C\varrho^2.$$

On the right-hand side of the inequality in Theorem 6, the first two terms are identical to the generalization bound of empirical risk minimization. The third and forth term are linear in the adversarial perturbation level ϱ, which represents the inflation of generalization bound in the adversarial setting as opposed to the stochastic setting; for large n, it is $O(\varrho n^{-\delta\beta})$ and the constant upfront is proportional to the Lipschitz norm of the loss function. The rest are higher-order terms for small adversarial perturbation level ϱ. It depends on the smoothness of the function: when the loss function has piecewise Lipschitz gradient and $e_n(\varrho) = O(\varrho^2)$, it is $O(n^{\alpha\delta}\varrho^{\alpha+1} + \varrho^2 n^{2\delta})$; when the loss function has Lipschitz gradient, it becomes $C\varrho^2$.

Theorem 7. *Let $p \in (1, \infty]$. Assume Assumptions 1, 2, 3, 4 hold, and that $\|f\|_\infty \le M$ for all $f \in \mathscr{F}$. Let $t > 0$. Then there exists $C > 0$ and an integer $N > 0$ such that for any $n > N$, with probability at least $1 - 3e^{-t}$, it holds that for every $f \in \mathscr{F}$,*

$$|\mathcal{L}^{\mathsf{wo}}_{n,p}(f;\varrho) - \mathcal{L}^{\mathsf{adv}}_p(f;\varrho)| \le 2\mathfrak{R}_n(\mathscr{F}) + M\sqrt{\frac{t}{2n}} + \frac{2\mathfrak{R}_n(\partial\mathscr{F}_q) + L^q\sqrt{\frac{t}{2n}}}{q\cdot(\eta/2)^{\frac{1}{p}}}\varrho + C_0\varrho^{(\alpha+1)\wedge p} + e_n\Big((\frac{2pL}{\eta})^{q-1}\varrho\Big) + C\varrho^2,$$

where C_0 is defined in Theorem 1.

Similar to the case of $p = 1$, the first two terms in the above bound is identical to the bound of empirical risk minimization, and the high-order term in the second row depends on the smoothness of the loss functions. The major difference from $p = 1$ is that, the inflation term in the adversarial setting depends on $\mathfrak{R}_n(\partial\mathscr{F}_q)$, the Rademacher complexity of the class of gradient norm functions. This term appears to be new in the literature. Intuitively, if the complexity of the variation is small, the model is more robust to the adversarial perturbations and thus generalizes better. We will illustrate this result in Section 6.5.

Theorems 6 and 7 are proved based on a different strategy than that in the stochastic setting. Set $\mathcal{L}^{\mathsf{vr}}_q(\varrho) := \min_{f\in\mathscr{F}}\{\mathbb{E}_{\mathbb{P}_{\mathsf{true}}}[f] + \varrho\||\partial f|\|_{q,\mathbb{P}_{\mathsf{true}}}\}$. Using the triangle inequality we decompose the difference $|\mathcal{L}^{\mathsf{wo}}_{n,p}(f;\varrho) - \mathcal{L}^{\mathsf{adv}}_p(f;\varrho)|$ into three terms $|\mathcal{L}^{\mathsf{wo}}_{n,p}(f;\varrho) - \mathcal{L}^{\mathsf{vr}}_{n,q}(f;\varrho)|$, $|\mathcal{L}^{\mathsf{vr}}_{n,q}(f;\varrho) - \mathcal{L}^{\mathsf{vr}}_q(f;\varrho)|$, $|\mathcal{L}^{\mathsf{vr}}_q(f;\varrho) - \mathcal{L}^{\mathsf{adv}}_p(f;\varrho)|$ and then bound each term individually.

6 Applications

In this section, we apply the theorems in previous sections to various machine learning tasks. We present the problem setup and a sketch of the verification of assumptions, but defer the details to the supplemental. We consider Markovian data in Section 6.3 and i.i.d. samples in other subsections.

6.1 Lipschitz and Gradient Regularization

Consider a supervised learning problem with feature vector $x \in \mathscr{X} \subset \mathbb{R}^d$ and response variable $y \in \mathscr{Y} \subset \mathbb{R}$. Let Θ be a bounded set in $\mathbb{R}^d$ and consider a linear predictor with polynomial loss $\mathscr{F} = \{(x, y) \mapsto f_\theta(x, y) = |\theta^\top x - y|^p, \theta \in \Theta\}$. Assume $\mathscr{X}, \mathscr{Y}$ and Θ are bounded. It is easy to verify that every $f_\theta, \theta \in \Theta$ is Lipschitz, piecewise differentiable and Hölder smooth, thus Assumptions 1,2 hold. Suppose $\mathbb{P}_{\mathsf{true}}$ satisfies $T_p(\tau)$, which implies Assumption 6 with $\tau_n = \tau n^{\frac{2}{p}-1}$.

When $p = 1$, we compute the constants in Assumption 7 in the supplemental, and Assumption 5 holds trivially with $\bar{a} = \beta = \infty$, because $G(f)(x) = |\partial f|(x) = \|(\theta, -1)\|_*$. Therefore conditions in Theorem 4 hold. When $p \in (1, 2]$, we assume that there exists $\lambda > 0$ such that $\mathbb{E}_{\mathbb{P}_{\mathsf{true}}}[|\theta^\top x - y|^p] \geq \lambda\|(\theta, -1)\|_2^p$ for all $\theta \in \Theta$. This is a mild condition, which holds as long as the true distribution does not concentrate on any hyperplane. Under this assumption, we verify Assumptions 3, 4 and compute the constants in Assumption 7 in the supplemental, and thereby the conditions in Theorem 5 hold.

6.2 Unsupervised learning

Principal component analysis seeks for the d principal directions, along which the data has maximized variance. More precisely, consider

$$\max_{W \in \mathbb{R}^{d \times k}} \left\{\|W^\top X\|_F^2 : \ W^\top W = I_k\right\}$$

where X is a $d \times n$ sample matrix consisting of n samples $\{X_i\}_{i=1}^n$ from $\mathbb{P}_{\mathsf{true}}$ on a set $\mathscr{X} \subset \{x \in \mathbb{R}^d : \ \|x\|_2 \leq B\}$; $\|\cdot\|_F$ denotes the Frobenius norm; I_k denotes the k-dimensional identity matrix; and $W = [w_1, \dots w_k]$ is the set of k orthonormal principal directions in $\mathbb{R}^d$. Assume that $\mathbb{E}_{x \sim \mathbb{P}_{\mathsf{true}}}[x] = 0$, and the smallest eigenvalue λ of the covariance matrix $\mathbb{E}_{x \sim \mathbb{P}_{\mathsf{true}}}[xx^\top]$ is positive.

Consider $p = 2$ and $\mathscr{F} = \{x \mapsto f_W(x) = -\|W^\top x\|_2^2 : W^\top W = I_k\}$. This is a smooth family, thus Assumptions 1,2 hold immediately and Assumption 3 is not needed. Assume that $\mathbb{P}_{\mathsf{true}}$ has mean zero and a non-singular covariance matrix, which implies Assumption 4, and satisfies $T_2(\tau)$, which implies Assumption 6 with $\tau_n = \tau$. We compute the constants in Assumption 7 in the supplemental. Hence, by applying Theorem 5 with $\varrho_n = \left(\frac{4\sqrt{\lambda}}{B^2} + \frac{2B^2}{\sqrt{\lambda}} + 2\sqrt{\tau t} + \frac{4t\tau\sqrt{d\lambda}}{B^2}\right)\frac{1}{\sqrt{n}} + O(\frac{1}{n})$, we have with high probability for every W it holds that

$$\mathcal{L}^{\mathsf{true}}(f_W) \leq \mathcal{L}^{\mathsf{ro}}_{n,p}(f_W; \varrho_n) + O(1/n) \leq \mathcal{L}^{\mathsf{erm}}_n(f_W) + \varrho_n \cdot \left(\frac{1}{n}\sum_{i=1}^n \|2WW^\top X_i\|_2^2\right)^{\frac{1}{2}} + O(1/n).$$

which shows that the generalization gap between $\mathcal{L}^{\mathsf{true}}(f_W)$ and $\mathcal{L}^{\mathsf{erm}}_n(f_W)$ is $O(\sqrt{d/n})$, thus improves the bound in Xu & Mannor (2012, Example 8) that depends exponentially on d.

6.3 Markovian Data

In this example, we show that our Assumption 6 about μ_S can be satisfied for the Markovian data that are useful in applications such as time series forecasts and reinforcement learning.

Assumption 8. Assume the Markov chain is homogeneous on $\mathscr{X}$, whose transition kernel is denoted as $P(dy|x)$. Assume the existence of $(x_0, y_0) \in \mathscr{X} \times \mathscr{X}, \alpha_0 > 0, \bar{\beta} > \alpha_0$ and $C < \infty$, such that

$$\forall x \in \mathscr{X}, \quad \int_{\mathscr{X}} e^{\alpha_0\|y_0 - y\|^2} P(dy|x) \geq C e^{\bar{\beta}\|x_0 - x\|^2}.$$

Let $\{x^i\}_{i=1}^n$ be n consecutive outputs of a Markov chain satisfying Assumption 8. Then by Lemma 6 in the supplemental, under Assumption 8, there exists $\tau > 0$ such that μ_S satisfies $T_1(\frac{2}{\tau n})$ and thus Assumption 6 is satisfied. Then for the function class that satisfies Assumption 1, 2, and 7 with

$K(f) = \|f\|_{\text{Lip}}^2$, by Theorem 4, setting $\varrho_n = \sqrt{\frac{8t(1+\log C)}{(\alpha_0-\bar{\beta})n}} + 4r_{n,\star} + \frac{2}{n\sqrt{r_{n,\star}}}$ we have with high probability

$$\mathcal{L}^{\mathsf{true}}(f) \le \mathcal{L}_{n,\infty}^{\mathsf{vr}}(f;\varrho_n) + O(1/n) \le \mathcal{L}_n^{\mathsf{erm}}(f) + \varrho_n\|f\|_{\text{Lip}} + O(1/n), \quad \forall f \in \mathscr{F},$$

where the second inequality follows from Theorem 3. This improves the covering number bound in Xu et al. (2009, Theorem 10) that results in suboptimal bound even for linear regression.

6.4 Risk-Averse Minimization

Conditional Value-at-Risk ($\mathbb{C}$V@R) is a popular risk measure that takes risk attitude into account in decision making. The empirical $\mathbb{C}$V@R at level ν is defined as $\mathbb{C}\text{V@R}_{\nu,n}[Z] = \min_{u\in\mathbb{R}}\left\{u + \frac{1}{\nu n}\sum_{i=1}^n (Z_i - u)_+\right\}$. Consider the following robust $\mathbb{C}$V@R minimization

$$\min_{\|w\|_2\le B, u\in\mathbb{R}} \max_{\{x^i\}_{i=1}^n \in \mathscr{X}_1(\varrho)} \left\{u + \frac{1}{\nu n}\sum_{i=1}^n (w^\top x^i - u)_+\right\}. \tag{CVaR}$$

Define

$$\mathscr{F} = \left\{f_\theta(x,y) = u + \tfrac{1}{\nu}(w^\top x - uy) : \ \theta = (w,u) \in \mathbb{R}^{d+1}, \|w\|_2 \le B\right\}.$$

The objective is piecewise linear, thus Assumptions 1, 2 holds. Assume $\mathbb{P}_{\mathsf{true}}$ satisfies $T_1(\tau)$, then Assumption 6 holds with $\tau_n = \tau n$, and set $B_1 = \mathbb{E}_{\mathbb{P}_{\mathsf{true}}}[\|x\|_2]$. In addition, it can be shown that the minimizer of (CVaR) is of the order $O(\log n)$ with high probability. Thereby we can compute the constants in Assumption 7 in the supplemental, and using Theorem 4, we can set $\varrho_n = \left(2\sqrt{\tau t} + 4(B_1+1) + \frac{2}{(B_1+1)}\right)\frac{1}{\sqrt{n}}$ to obtain that for every θ,

$$\mathbb{C}\text{V@R}_{\nu,\mathbb{P}_{\mathsf{true}}}(w^\top x) \le \mathcal{L}_{n,1}^{\mathsf{ro}}(f_\theta;\varrho_n) + O(\tfrac{1}{n^{1-2\delta}}) \le \mathcal{L}_n^{\mathsf{erm}}(f_\theta) + \varrho_n \cdot \|\nabla f_\theta\|_{2,\mathbb{P}_n} + O(\tfrac{1}{n^{1-2\delta}}),$$

where δ is an arbitrarily small positive number. Our generalization bound has a similar order compared to Theorem 4 in Lee et al. (2020) which uses Rademacher complexity (while Mhammedi et al. (2020) uses PAC-Bayes), but we do not require boundedness on the data nor the loss function.

6.5 Adversarial Learning

Consider a classification problem with loss function class

$$\mathscr{F}^\phi := \{(x,y) \mapsto \phi(yf(x)) : \ f \in \mathscr{F}\},$$

where $\phi : \mathbb{R} \to [0,1]$ is monotonically non-increasing, L_ϕ-Lipschitz and has h_ϕ-Lipschitz gradient, $y \in \{\pm 1\}$, and $f : \mathscr{X} \to [0,1]$ is a family of L-Lipschitz hypotheses with h-Lipschitz gradient. Suppose $\mathscr{X} \subset \{x \in \mathbb{R}^d : \|x\| \le 1\}$. Let $t > 0$. Then by Theorem 7, with probability at least $1 - e^{-t}$, for every $f \in \mathscr{F}$,

$$|\mathcal{L}_{n,\infty}^{\mathsf{wo}}(f;\varrho) - \mathcal{L}_\infty^{\mathsf{adv}}(f;\varrho)| \le 2L_\phi\mathfrak{R}_n(\mathscr{F}) + \sqrt{\tfrac{t}{2n}} + \left(2\mathfrak{R}_n(\partial\mathscr{F}_1^\phi) + LL_\phi\sqrt{\tfrac{t}{2n}}\right)\varrho + 2hh_\phi\varrho^2,$$

where $\partial\mathscr{F}_1^\phi = \{|\phi'(yf(x))| \cdot \|\nabla f(x)\|_* : f \in \mathscr{F}\}$. When ϱ is small, the Rademacher complexity of the gradient norm function class $\mathfrak{R}_n(\partial\mathscr{F}_1^\phi)$, together with $LL_\phi\sqrt{\frac{t}{2n}}$, is the dominating term of the difference between the generalization bounds for adversarial setting and stochastic setting. Our bound applies to general loss function classes. Particularly, when $\mathscr{F} = \{f_\theta(x) = \theta^\top x : \theta \in \Theta\}$, where $\Theta \subset \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le B\}$, then as shown in the supplemental, $\mathfrak{R}_n(\partial\mathscr{F}_1^\phi) \le (L_\phi + h_\phi B)\frac{B}{\sqrt{n}}$, which is a dimension-independent bound consistent with Awasthi et al. (2020).

7 Concluding Remarks

In this paper, we thoroughly analyzed the finite-sample performance bounds for robust optimization, drawing connection with transport inequality and local Rademacher complexity. Our results generalize existing results from various perspectives, and particularly, to non-smooth loss functions and to non-i.i.d. data, both of which requires non-trivial analysis. Our results help to better understand the regularization effect of robust learning and explain their superior empirical performances. We hope our results can inspire regularization schemes for other applications as well. Our bounds involve the Rademacher complexities of (normalized) gradient norm function $\partial\mathscr{F}_q, \mathscr{N}_q$ and margin to discontinuity $\mathscr{I}_\varrho$, which are non-standard; we have provided various examples and will investigate in more depth for future work.

References

Awasthi, P., Frank, N., and Mohri, M. Adversarial learning guarantees for linear hypotheses and neural networks. *arXiv preprint arXiv:2004.13617*, 2020.

Bartl, D., Drapeau, S., Obloj, J., and Wiesel, J. Robust uncertainty sensitivity analysis. *arXiv preprint arXiv:2006.12022*, 2020.

Bartlett, P. L., Bousquet, O., Mendelson, S., et al. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Bellet, A. and Habrard, A. Robustness and generalization for metric learning. *Neurocomputing*, 151: 259–267, 2015.

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. *Robust optimization*. Princeton university press, 2009.

Bertsimas, D. and Copenhaver, M. S. Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European Journal of Operational Research*, 270(3): 931–942, 2018.

Blanchet, J. and Murthy, K. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

Blanchet, J., Kang, Y., and Murthy, K. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019a.

Blanchet, J., Murthy, K., and Si, N. Confidence regions in wasserstein distributionally robust estimation. *arXiv preprint arXiv:1906.01614*, 2019b.

Bolley, F. and Villani, C. Weighted csiszár-kullback-pinsker inequalities and applications to transportation inequalities. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 14, pp. 331–352, 2005.

Bubeck, S., Lee, Y. T., Price, E., and Razenshteyn, I. Adversarial examples from computational constraints. In *International Conference on Machine Learning*, pp. 831–840. PMLR, 2019.

Chen, R. and Paschalidis, I. C. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19(13), 2018.

Cranko, Z., Kornblith, S., Shi, Z., and Nock, R. Lipschitz networks and distributional robustness. *arXiv preprint arXiv:1809.01129*, 2018.

Esfahani, P. M. and Kuhn, D. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1): 115–166, 2018.

Finlay, C., Calder, J., Abbasi, B., and Oberman, A. Lipschitz regularized deep neural networks generalize and are adversarially robust. *arXiv preprint arXiv:1808.09540*, 2018.

Gao, R. Finite-sample guarantees for wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *arXiv preprint arXiv:2009.04382*, 2020.

Gao, R. and Kleywegt, A. J. Distributionally robust stochastic optimization with wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

Gao, R. and Kleywegt, A. J. Distributionally robust stochastic optimization with dependence structure. *arXiv preprint arXiv:1701.04200*, 2017.

Gao, R., Chen, X., and Kleywegt, A. J. Wasserstein distributional robustness and regularization in statistical learning. *arXiv*, pp. arXiv–1712, 2017.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Gozlan, N. and Léonard, C. Transport inequalities. a survey. *Markov Processes and Related Fields*, 16:635–736, 2010.

Khim, J. and Loh, P.-L. Adversarial risk bounds via function transformation. *arXiv preprint arXiv:1810.09519*, 2018.

Kuhn, D., Esfahani, P. M., Nguyen, V. A., and Shafieezadeh-Abadeh, S. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pp. 130–166. INFORMS, 2019.

Lee, J. and Raginsky, M. Minimax statistical learning with wasserstein distances. *Advances in Neural Information Processing Systems*, 31:2687–2696, 2018.

Marchal, O., Arbel, J., et al. On the sub-gaussianity of the beta and dirichlet distributions. *Electronic Communications in Probability*, 22, 2017.

Mohri, M. and Rostamizadeh, A. Rademacher complexity bounds for non-iid processes. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, pp. 1097–1104, 2008.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 5019–5031, 2018.

Shafieezadeh-Abadeh, S., Kuhn, D., and Esfahani, P. M. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.

Si, N., Blanchet, J., Ghosh, S., and Squillante, M. Quantifying the empirical wasserstein distance to a set of measures: Beating the curse of dimensionality. *Advances in Neural Information Processing Systems*, 33, 2020.

Sinha, A., Namkoong, H., and Duchi, J. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2018.

Xu, H. and Mannor, S. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012.

Xu, H., Caramanis, C., and Mannor, S. Robustness and regularization of support vector machines. *Journal of machine learning research*, 10(7), 2009.

Xu, H., Caramanis, C., and Mannor, S. Robust regression and lasso. *IEEE Transactions on Information Theory*, 56(7):3561–3574, 2010.

Yin, D., Kannan, R., and Bartlett, P. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, pp. 7085–7094. PMLR, 2019.

Zahavy, T., Kang, B., Sivak, A., Feng, J., Xu, H., and Mannor, S. Ensemble robustness and generalization of stochastic deep learning algorithms. *arXiv preprint arXiv:1602.02389*, 2016.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pp. 7472–7482. PMLR, 2019.

Check List

For all authors...

(a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Yes.

(b) Have you read the ethics review guidelines and ensured that your paper conforms to them?

Yes.

(c) Did you discuss any potential negative societal impacts of your work? Not Applied.

(d) Did you describe the limitations of your work?

Yes.

If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results?

Yes.

(b) Did you include complete proofs of all theoretical results?

Yes.

If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Not Applied.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Not Applied.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Not Applied.

(d) Did you include the amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Not Applied.

If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? Not Applied.

(b) Did you mention the license of the assets? Not Applied.

(c) Did you include any new assets either in the supplemental material or as a URL? Not Applied.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Not Applied.

If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? Not Applied.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? Not Applied.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Not Applied.