Principled Learning Method for Wasserstein Distributionally Robust Optimization with Local Perturbations

Yongchan Kwon [1] **Wonyoung Kim** [2] **Joong-Ho Won** [2] **Myunghee Cho Paik** [2]

Abstract

Wasserstein distributionally robust optimization (WDRO) attempts to learn a model that minimizes the local worst-case risk in the vicinity of the empirical data distribution defined by Wasserstein ball. While WDRO has received attention as a promising tool for inference since its introduction, its theoretical understanding has not been fully matured. Gao et al. (2017) proposed a minimizer based on a tractable approximation of the local worst-case risk, but without showing risk consistency. In this paper, we propose a minimizer based on a novel approximation theorem and provide the corresponding risk consistency results. Furthermore, we develop WDRO inference for locally perturbed data that include the Mixup (Zhang et al., 2017) as a special case. Numerical experiments demonstrate robustness of the proposed method using image classification datasets. Our results show that the proposed method achieves significantly higher accuracy than baseline models on contaminated datasets.

1. Introduction

Statistical learning problems can be generally formulated as an optimization problem of the form

$$\inf_{h\in\mathcal{H}} R(\mathbb{P}_{\mathrm{data}}, h), \tag{1}$$

where $\mathbb{P}_{\mathrm{data}}$ is the true data distribution, $\mathcal{H}$ is a set of losses, and $R(\mathbb{Q}, h) := \int h(\zeta) d\mathbb{Q}(\zeta)$ is the risk, or the expected value of a loss h with respect to a probability measure $\mathbb{Q}$. In real-world applications, the $\mathbb{P}_{\mathrm{data}}$ is usually unknown, so the computation of the risk in (1) is impossible. We instead observe a set $\mathcal{Z}_n = \{z_1, \ldots, z_n\}$ of independent and identically distributed samples from $\mathbb{P}_{\mathrm{data}}$. Using the dataset $\mathcal{Z}_n$, we solve the empirical risk minimization (ERM) problem

$$\inf_{h\in\mathcal{H}} R(\mathbb{P}_n, h) = \inf_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} h(z_i), \tag{2}$$

where $\mathbb{P}_n := n^{-1}\sum_{i=1}^{n}\delta_{z_i}$ is the empirical data distribution and δ_z is the Dirac delta distribution concentrating unit mass at z.

ERM provides a practical framework for learning models by replacing $\mathbb{P}_{\mathrm{data}}$ in (1) with $\mathbb{P}_n$ (Vapnik, 1999). However, this replacement often yields poor risk estimation, and thus a solution to (2) can have a small training error but a large test error. This phenomenon is well known as overfitting, and to avoid this, a great number of regularization methods have been proposed: penalty-based methods (Tibshirani, 1996; Fan & Li, 2001; Bühlmann & van de Geer, 2011), data augmentations (Zhang et al., 2017; Cubuk et al., 2019), dropout (Wager et al., 2013; Srivastava et al., 2014), and early stopping (Yao et al., 2007), to name a few.

As an alternative approach to prevent overfitting, we consider Wasserstein distributionally robust optimization (WDRO) (Shafieezadeh-Abadeh et al., 2015; Sinha et al., 2017; Blanchet et al., 2019). The goal of WDRO is to learn a model that minimizes the local worst-case risk in the vicinity of the empirical data distribution defined by a Wasserstein ball. To be specific, let $\mathfrak{M}_{\alpha_n,p}(\mathbb{P}_n)$ be a set of probability measures whose p-Wasserstein metric from $\mathbb{P}_n$ is less than $\alpha_n > 0$. The local worst-case risk is defined to be the supremum of the risk over the p-Wasserstein ball. Then WDRO is formulated as follows.

$$\inf_{h\in\mathcal{H}} \sup_{\mathbb{Q}\in\mathfrak{M}_{\alpha_n,p}(\mathbb{P}_n)} R(\mathbb{Q}, h).$$

Detailed definitions are available in Section 2.

By the design of the local worst-case risk, a solution to WDRO can avoid overfitting to $\mathbb{P}_n$ and learn a robust model against local perturbations. However, exact computation of the local worst-case risk is intractable except for few

This work was done while Yongchan Kwon was at Seoul National University. [1]Department of Biomedical Data Science, Stanford University [2]Department of Statistics, Seoul National University. Correspondence to: Myunghee Cho Paik <myungheechopaik@snu.ac.kr>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

Table 1. A summary of contributions of Gao et al. (2017), Lee & Raginsky (2018) and ours. Mark '✓' indicates that the corresponding work has a contribution to the item; mark '✗' means otherwise. We denote a sample space by $\mathcal{Z} \subseteq \mathbb{R}^d$.

	APPROXIMATION	RISK CONSISTENCY	PERTURBATION
GAO ET AL. (2017)	✓($\mathcal{Z} = \mathbb{R}^d$)	✗	✗
LEE & RAGINSKY (2018)	✗	✓($\mathcal{Z}$ IS BOUNDED)	✗
OURS	✓($\mathcal{Z}$ IS BOUNDED)	✓($\mathcal{Z}$ IS BOUNDED; SECTION 3)	✓(SECTION 4)

simple settings because it is difficult (i) to evaluate an exact risk with respect to a probability measure in the Wasserstein ball and (ii) to find the supremum of the risk among infinitely many probability distributions. Gao et al. (2017) obtained an approximation formula for the local worst-case risk and proposed to minimize this surrogate objective, but did not study risk consistency of the minimizer. Under a different assumption on the sample space, Lee & Raginsky (2018) proved that a minimizer of the exact local worst-case risk possesses risk consistency. However, finding such a minimizer is difficult due to the intractability of the local worst-case risk. To the best of our knowledge, there is no known risk consistency result for tractable approximate optimizers.

In this paper, we propose a minimizer based on a novel approximation theorem and provide corresponding risk consistency results. In Section 3, we present a new approximation to the local worst-case risk using gradient penalty assuming that a loss is differentiable and its gradient has a Hölder continuous (Theorem 1). We show that a minimizer of the approximate worst-case risk is consistent in that the risk (*resp.* the worst-case risk) converges to the optimal risk (*resp.* the optimal worst-case risk) (Theorems 2 and 3). Our results show that the proposed minimizer can have the same risk optimality as a minimizer of the exact local worst-case risk attains. In Section 4, we study WDRO inference when data are locally perturbed. We define locally perturbed data distributions and describe examples such as the Mixup (Zhang et al., 2017) and the adversarial training (Goodfellow et al., 2014). We show that our approximation and risk consistency results naturally extend to the cases when data are locally perturbed (Theorems 4, 5, and 6). Such theoretical results provide principled ways to use a group of data augmentation including the Mixup. Numerical experiments demonstrate robustness of the proposed method using image classification datasets. Our experiment results show that the proposed method produces a robust model that achieves significantly higher accuracy than baseline models on contaminated datasets.

A summary of our contributions in relation to Gao et al. (2017) and Lee & Raginsky (2018) is shown in Table 1. Proofs are available in the Supplementary Material.

1.1. Related works

Distributionally robust optimization (DRO) provides a general learning framework of the local worst-case risk minimization. Here, the local worst-case risk is defined as the supremum of the risk in the vicinity of the empirical data distribution, called the ambiguity set. The ambiguity set is often designed as a neighborhood of $\mathbb{P}_n$ and the closeness of two measures is evaluated by ϕ-divergences or probability metrics. Note that WDRO is a special case of DRO when the ambiguity set is designed via the Wasserstein metric. Other examples incorporate the ϕ-divergence (Ben-Tal et al., 2013; Hu et al., 2018; Namkoong & Duchi, 2017; Ghosh & Lam, 2019) and the maximum mean discrepancy (Staib & Jegelka, 2019). We refer to Rahimian & Mehrotra (2019) for a complementary literature review of DRO.

Another related field of this work is data augmentation. Data augmentation has recently emerged as a key technique to improve empirical performance in the field of machine learning (Cubuk et al., 2019; Lim et al., 2019). For example, Mixup and its variants have led remarkable generalization ability in supervised and semi-supervised learning tasks (Zhang et al., 2017; Verma et al., 2019; Berthelot et al., 2019). However, most data augmentations are based on heuristics, and their theoretical bases are limited to account for current successes. In this work, we develop WDRO inference for a group of data augmentations that generate a new data distribution near the original data distribution.

1.2. Notation

For a sequence (a_n) of positive constants and a sequence (b_n) of real numbers, $b_n = O(a_n)$ indicates that there exists constants $C, n_0 \in \mathbb{N}$ such that $|b_n| \leq C a_n$ for all $n \geq n_0$. For a random sequence (B_n), $B_n = O_p(a_n)$ indicates that for any $\varepsilon > 0$, there exists constants $C, n_0 \in \mathbb{N}$ such that $P(|B_n| > C a_n) < \varepsilon$ for all $n \geq n_0$. For a $p \in [1, \infty]$, we denote its Hölder conjugate by $p^* := (1 - 1/p)^{-1}$. Here, we use the conventions $1/\infty = 0$ and $1/0 = \infty$. For $a, b \in \mathbb{R}$, we use $a \vee b$ to denote the maximum between a and b. For $n \in \mathbb{N}$, we use $[n]$ to denote a set of integers $\{1, \ldots, n\}$. A set of all Borel probability measures defined on a set $\mathcal{S}$ is denoted by $\mathcal{P}(\mathcal{S})$. We denote a sample space by $\mathcal{Z} \subseteq \mathbb{R}^d$ and a norm on $\mathcal{Z}$ by $\|\cdot\|$ and the true data distribution by $\mathbb{P}_{\text{data}} \in \mathcal{P}(\mathcal{Z})$.

2. Preliminaries

The goal of this section is to review existing works on WDRO. As discussed in Section 1, the main objective of WDRO is to learn a model that minimizes the local worst-case risk over some Wasserstein ball. Formally, for sets S and $\tilde{S}$, we denote the push-forward measure of $\mu \in \mathcal{P}(S)$ through a map $T : S \to \tilde{S}$ by $T\#\mu \in \mathcal{P}(\tilde{S})$. The definitions of the p-Wasserstein metric and the p-Wasserstein ball are as follows.

Definition 1 (p-Wasserstein metric and p-Wasserstein ball)**.** *For $p \in [1, \infty)$ and $\nu, \mu \in \mathcal{P}(\mathcal{Z})$, the p-Wasserstein metric between ν and μ is defined as*

$$\mathcal{W}_p(\nu, \mu) := \Big(\inf_{\rho \in J(\nu,\mu)} \Big\{ \int_{\mathcal{Z}\times\mathcal{Z}} \left\| \zeta - \tilde{\zeta} \right\|^p d\rho(\zeta, \tilde{\zeta}) \Big\} \Big)^{1/p},$$

where $J(\nu, \mu) := \{\rho \in \mathcal{P}(\mathcal{Z} \times \mathcal{Z}) \mid \pi_1\#\rho = \nu, \pi_2\#\rho = \mu\}$, $\pi_i : \mathcal{Z} \times \mathcal{Z} \to \mathcal{Z}$ is the canonical projection defined by $\pi_i(\zeta_1, \zeta_2) = \zeta_i$ for $i = 1, 2$. For $\alpha > 0$, the p-Wasserstein ball centered at $\mathbb{P} \in \mathcal{P}(\mathcal{Z})$ with radius α is defined as

$$\mathfrak{M}_{\alpha,p}(\mathbb{P}) := \{\mathbb{Q} \in \mathcal{P}(\mathcal{Z}) : \mathcal{W}_p(\mathbb{Q}, \mathbb{P}) \le \alpha\}.$$

Throughout this paper, we denote the radius of the Wasserstein ball by α_n when the sample size is n. With above definitions, the WDRO problem is to minimize the local worst-case risk

$$R^{\text{worst}}_{\alpha_n,p}(\mathbb{P}_n, h) := \sup_{\mathbb{Q}\in\mathfrak{M}_{\alpha_n,p}(\mathbb{P}_n)} R(\mathbb{Q}, h). \tag{3}$$

The local worst-case risk (3) involves the supremum operator over the Wasserstein ball $\mathfrak{M}_{\alpha_n,p}(\mathbb{P}_n)$, which is a set of infinitely many probability distributions. Therefore, the exact computation of (3) is intractable in many cases.

A standard method to handle the intractability is to reformulate (3) by using either a primal-dual pair of infinite-dimensional linear programs (Esfahani & Kuhn, 2018) or first-order optimality conditions of the dual (Gao & Kleywegt, 2016). For the latter, let $\kappa_h = \limsup_{\|\zeta-\tilde{\zeta}\|\to\infty} (h(\zeta) - h(\tilde{\zeta}))/\left\|\zeta - \tilde{\zeta}\right\|^p$ if $\mathcal{Z}$ is unbounded, and zero otherwise. Gao & Kleywegt (2016, Corollary 2) showed that if $\kappa_h < \infty$, then

$$\begin{aligned} &R^{\text{worst}}_{\alpha_n,p}(\mathbb{P}_n, h) \\ &= \min_{\lambda\ge 0} \Big\{ \lambda\alpha_n^p + \frac{1}{n}\sum_{i=1}^n \sup_{z\in\mathcal{Z}} \{h(z) - \lambda\|z - z_i\|^p\} \Big\}. \end{aligned} \tag{4}$$

Similar results are obtained in the literature (Gao & Kleywegt, 2017; Blanchet & Murthy, 2019).

Based on the reformulation (4), relationships between WDRO and penalty-based methods have been investigated in supervised learning settings. For example, Shafieezadeh-Abadeh et al. (2015) and Blanchet et al. (2019) studied classification settings and Chen & Paschalidis (2018) considered regression settings. Although the relationships provide a way to understand WDRO, most existing results focus on linear hypotheses only. Recently, WDRO with nonlinear hypotheses has been developed. Shafieezadeh-Abadeh et al. (2019) showed that (3) has the form of a penalized empirical risk when a loss is Lipschitz continuous and a hypothesis is an element of a reproducing kernel Hilbert space.

In general statistical learning problems, Gao et al. (2017) established a relationship between WDRO and penalty-based methods. They obtained a penalized empirical risk and showed that it approximates (3) when a loss is smooth and $\mathcal{Z} = \mathbb{R}^d$. Although a minimizer of the suggested approximation gives a practical solution for WDRO, its risk consistency has not been studied.

As for the risk consistency, Lee & Raginsky (2018) showed that a minimizer of (3) has a vanishing excess worst-case risk bound when $\mathcal{H}$ is a set of Lipschitz continuous losses. More specifically, let $\hat{h}^{\text{worst}}_{\alpha_n,p} = \operatorname{argmin}_{h\in\mathcal{H}} R^{\text{worst}}_{\alpha_n,p}(\mathbb{P}_n, h)$ and $R^{\text{worst}}_{\alpha_n,p}(\mathbb{P}_{\text{data}}, h) = \sup_{\mathbb{Q}\in\mathfrak{M}_{\alpha_n,p}(\mathbb{P}_{\text{data}})} R(\mathbb{Q}, h)$. For bounded $\mathcal{Z}$, Lee & Raginsky (2018, Theorem 2) showed the following risk consistency result.

$$\mathcal{E}^{\text{worst}}_{\alpha_n,p}(\hat{h}^{\text{worst}}_{\alpha_n,p}) = O_p(n^{-1/2}(\mathfrak{C}(\mathcal{H}) \vee \alpha_n^{1-p})), \tag{5}$$

where

$$\mathcal{E}^{\text{worst}}_{\alpha_n,p}(g) := R^{\text{worst}}_{\alpha_n,p}(\mathbb{P}_{\text{data}}, g) - \inf_{h\in\mathcal{H}} R^{\text{worst}}_{\alpha_n,p}(\mathbb{P}_{\text{data}}, h)$$

is the excess worst-case risk of $g \in \mathcal{H}$, $\mathfrak{C}(\mathcal{S}) := \int_0^\infty \sqrt{\log \mathcal{N}(u, \mathcal{S}, \|\cdot\|_\infty)}du$ is the entropy integral of a set $\mathcal{S}$, and $\mathcal{N}(u, \mathcal{S}, \|\cdot\|_\infty)$ denotes the u-covering number of a set $\mathcal{S}$ with respect to the uniform norm $\|\cdot\|_\infty$ (Györfi et al., 2006, Definition 9.2). The result (5) explains asymptotic behaviors of the WDRO solution $\hat{h}^{\text{worst}}_{\alpha_n,p}$, but as mentioned before, exact computation of (3) is intractable except for few simple cases.

It is noteworthy that Gao et al. (2017) and Lee & Raginsky (2018) have conflicting assumptions on $\mathcal{Z}$, so the results of Lee & Raginsky (2018) cannot be used to show risk consistency of the minimizer by Gao et al. (2017).

3. Tractable WDRO and risk consistency

In this section, we build a principled and tractable learning method for WDRO. In Section 3.1, we propose an approximation of the local worst-case risk that can be easily evaluated by off-the-shelf gradient methods and software. In Section 3.2, we provide asymptotic results: a minimizer of the approximate risk is consistent in that the risk (*resp.* the worst-case risk) converges to the optimal risk (*resp.* the optimal worst-case risk).

3.1. Approximation to the local worst-case risk

For a Lipschitz continuous loss $h : \mathcal{Z} \to \mathbb{R}$, Lee & Raginsky (2018, Proposition 1) showed that

$$\left| R(\mathbb{P}_n, h) - R^{\text{worst}}_{\alpha_n,p}(\mathbb{P}_n, h) \right| = O(\alpha_n). \tag{6}$$

An equivalent result is obtained by Kuhn et al. (2019, Theorem 5). We show that a faster approximation is possible if a loss h is differentiable and its gradient is Hölder continuous. To begin, we define some notations. For $r \in [1, \infty)$, a probability measure $\mathbb{Q} \in \mathcal{P}(\mathbb{R}^d)$, and a function $g : \mathcal{Z} \to \mathbb{R}^d$, we denote a function norm by

$$\|g\|_{\mathbb{Q},r} := (\int \|g(z)\|_*^r \, d\mathbb{Q}(z))^{1/r},$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$. For a constant $C_{\mathrm{H}} > 0$ and $k \in (0, 1]$, a function $g : \mathcal{Z} \to \mathbb{R}^d$ is said to be (C_{H}, k)-Hölder continuous if

$$\|g(z) - g(\tilde{z})\|_* \leq C_{\mathrm{H}} \|z - \tilde{z}\|^k, \quad \forall z, \tilde{z} \in \mathcal{Z}.$$

Let $\mathrm{Conv}(\mathcal{Z})$ be the convex hull of $\mathcal{Z}$ and $\mathbb{E}_{\text{data}}(g) := \int_{\mathcal{Z}} g(z) d\mathbb{P}_{\text{data}}(z)$ for a function $g : \mathcal{Z} \to \mathbb{R}$.

Theorem 1 (Approximation to local worst-case risk)**.** *Let (α_n) be a sequence of positive numbers converging to zero and $\mathcal{Z}$ be an open and bounded subset of $\mathbb{R}^d$. For constants $C_{\mathrm{H}}, C_\nabla > 0$ and $k \in (0, 1]$, assume that a loss $h : \mathrm{Conv}(\mathcal{Z}) \to \mathbb{R}$ is differentiable, its gradient $\nabla_z h(z)$ is (C_{H}, k)-Hölder continuous, and $\mathbb{E}_{\text{data}}(\|\nabla_z h\|_*) \geq C_\nabla$. Then, for $p \in (1 + k, \infty)$, the following holds.*

$$\begin{aligned}&\left| R(\mathbb{P}_n, h) + \alpha_n \|\nabla_z h\|_{\mathbb{P}_n,p^*} - R^{\text{worst}}_{\alpha_n,p}(\mathbb{P}_n, h) \right| \\ &= O_p(\alpha_n^{1+k}).\end{aligned}$$

Remark 1. *Theorem 1 establishes an asymptotic equivalence between WDRO and penalty-based methods. Compared to (6), Theorem 1 provides a sharper approximation to the local worst-case risk. Gao et al. (2017, Theorem 2) obtained a similar result when $\mathcal{Z} = \mathbb{R}^d$, yet our boundedness assumption on $\mathcal{Z}$ is reasonable in a sense that real computers store data in a finite number of states. For example, a d-dimensional gray scale image datum is stored as a d-dimensional vector having integer values range from 0 to 255.*

Remark 2. *The assumption $\mathbb{E}_{\text{data}}(\|\nabla_z h\|_*) \geq C_\nabla$ in Theorem 1 holds as long as there exist positive constants $C_{\nabla,1}$ and $C_{\nabla,2}$ such that $P(\|\nabla_z h\|_* \geq C_{\nabla,1}) \geq C_{\nabla,2}$. Note that by the Markov's inequality, $\mathbb{E}_{\text{data}}(\|\nabla_z h\|_*) \geq C_{\nabla,1} C_{\nabla,2}$. Hence, unless h is a constant function, $\|\nabla_z h\|_*$ is strictly greater than zero and existence of C_∇ is guaranteed.*

Based on Theorem 1, for a vanishing sequence (α_n), we propose to minimize the following surrogate objective:

$$R^{\text{prop}}_{\alpha_n,p}(\mathbb{P}_n, h) := R(\mathbb{P}_n, h) + \alpha_n \|\nabla_z h\|_{\mathbb{P}_n,p^*}. \tag{7}$$

In the sequel, we denote a minimizer of the objective function (7) by $\hat{h}^{\text{prop}}_{\alpha_n,p}$, *i.e.*, $\hat{h}^{\text{prop}}_{\alpha_n,p} = \mathrm{argmin}_{h \in \mathcal{H}} R^{\text{prop}}_{\alpha_n,p}(\mathbb{P}_n, h)$. In contrast to the intractability of (3), the approximate risk (7) can be easily minimized by off-the-shelf gradient methods and software.

3.2. Risk consistency of the proposed estimator

We then study the excess worst-case risk bound of $\hat{h}^{\text{prop}}_{\alpha_n,p}$. To begin with, for a Lipschitz continuous function h, we denote the smallest Lipschitz constant of h by $\mathrm{Lip}(h)$.

Theorem 2 (Excess worst-case risk bound)**.** *Let (α_n) be a sequence of positive numbers converging to zero and $\mathcal{Z}$ be an open and bounded subset of $\mathbb{R}^d$. For constants $C_{\mathrm{H}}, C_\nabla, L > 0$ and $k \in (0, 1]$, assume that $\mathcal{H}$ is a uniformly bounded set of differentiable functions $h : \mathrm{Conv}(\mathcal{Z}) \to \mathbb{R}$ such that its gradient $\nabla_z h$ is (C_{H}, k)-Hölder continuous, $\mathbb{E}_{\text{data}}(\|\nabla_z h\|_*) \geq C_\nabla$, and $\mathrm{Lip}(h) \leq L$. Then, for $p \in (1 + k, \infty)$, the following holds.*

$$\mathcal{E}^{\text{worst}}_{\alpha_n,p}(\hat{h}^{\text{prop}}_{\alpha_n,p}) = O_p\left(\frac{\mathfrak{C}(\mathcal{H}) \vee \alpha_n^{1-p}}{\sqrt{n}} \vee \log(n) \alpha_n^{1+k} \right).$$

Compared to the bound (5) by Lee & Raginsky (2018), the risk bound of the proposed method in Theorem 2 has the additional term $\log(n)\alpha_n^{1+k}$. This additional error is a payoff for the approximation (7), and it is asymptotically negligible when $\alpha_n^{-(p+k)} \geq O(n^{1/2} \log(n))$. Thus the proposed minimizer can have the same risk optimality as $\hat{h}^{\text{worst}}_{\alpha_n,p}$ achieves.

Next, we analyze the excess risk bound of $\hat{h}^{\text{prop}}_{\alpha_n,p}$. Recall that the Rademacher complexity of a set $\mathcal{S}$ is defined as $\mathfrak{R}_n(\mathcal{S}) := \mathbb{E}_{\text{data}} \mathbb{E}_\sigma (\frac{1}{n} \sup_{s \in \mathcal{S}} | \sum_{i=1}^n \sigma_i s(Z_i)|)$ where $\{\sigma_i\}_{i=1}^n$ is a set of independent Rademacher random variables taking 1 or -1 with probability 0.5 each, and $\mathbb{E}_\sigma(\cdot)$ is the expectation operator over the Rademacher random variables (Bartlett & Mendelson, 2002). We denote the excess risk of $g \in \mathcal{H}$ by $\mathcal{E}(g) := R(\mathbb{P}_{\text{data}}, g) - \inf_{h \in \mathcal{H}} R(\mathbb{P}_{\text{data}}, h)$.

Theorem 3 (Excess risk bound)**.** *Under the same assumptions as Theorem 2, the following holds.*

$$\mathcal{E}(\hat{h}^{\text{prop}}_{\alpha_n,p}) = O_p(\mathfrak{R}_n(\mathcal{H}) \vee n^{-1/2} \vee \alpha_n \vee \log(n) \alpha_n^{1+k}).$$

Suppose $\alpha_n = n^{-\epsilon}$ for some $\epsilon > 0$. Then, for a large enough n, we have $\alpha_n \vee \log(n)\alpha_n^{1+k} = \alpha_n$ and the excess risk bound is $O_p(\mathfrak{R}_n(\mathcal{H}) \vee n^{-1/2} \vee \alpha_n)$. Considering the fact that the excess risk bound of the ERM solution is $O_p(\mathfrak{R}_n(\mathcal{H}) \vee n^{-1/2})$ (Mohri et al., 2018, Theorem 11.3), the result of Theorem 3 sounds pessimistic, especially when $\epsilon \le 1/2$. However, Theorem 3 is in fact sensible in that $\hat{h}^{\rm prop}_{\alpha_n,p}$ optimizes the local worst-case risk $R^{\rm worst}_{\alpha_n,p}(\mathbb{P}_n, h)$, not the risk $R(\mathbb{P}_n, h)$. That means, gaining robustness necessarily leads to losing the accuracy of the prediction model. Interested readers in the trade-off between accuracy and robustness are referred to Zhang et al. (2019).

3.3. Example bounds

We now provide an example showing the use of Theorems 2 and 3 in binary classification settings. To begin, we denote a solution of (2) by $\hat{h}^{\rm ERM}_n$. Let $\mathcal{X} \subseteq [-1, 1]^{d-1}$ and $\mathcal{Y} = \{\pm 1\}$ be open sets with respect to the ℓ_2-norm and the discrete norm $I(\cdot \neq 0)$, respectively. We set $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\|(x, y)\| = \|x\|_2 + 4I(y \neq 0)$.

Corollary 1 (Informal)**.** *Let $\mathcal{F}$ be a set of sparse deep neural networks and $\mathcal{H} = \{h(x, y) \mid h(x, y) = \log(1 + \exp(-yf(x)))$ for $f \in \mathcal{F}\}$. Then the excess worst-case risks of $\hat{h}^{\rm prop}_{\alpha_n,p}$ and $\hat{h}^{\rm ERM}_n$ are*

$$\mathcal{E}^{\rm worst}_{\alpha_n,p}(\hat{h}^{\rm prop}_{\alpha_n,p}) = O_p(n^{-1/2}\alpha_n^{1-p} \vee \log(n)\alpha_n^{1+k}),$$
$$\mathcal{E}^{\rm worst}_{\alpha_n,p}(\hat{h}^{\rm ERM}_n) = O_p(n^{-1/2} \vee \alpha_n).$$

Furthermore, the excess risks of $\hat{h}^{\rm prop}_{\alpha_n,p}$ and $\hat{h}^{\rm ERM}_n$ are

$$\mathcal{E}(\hat{h}^{\rm prop}_{\alpha_n,p}) = O_p(n^{-1/2} \vee \alpha_n \vee \log(n)\alpha_n^{1+k}),$$
$$\mathcal{E}(\hat{h}^{\rm ERM}_n) = O_p(n^{-1/2}).$$

Corollary 1 shows that the excess worst-case risk bound of $\hat{h}^{\rm prop}_{\alpha_n,p}$ is sharper than that of $\hat{h}^{\rm ERM}_n$ if $\alpha_n \ge O(n^{-1/(2p)})$. A typical choice of α_n is $O(n^{-p/d})$ to guarantee $\mathbb{P}_{\rm data} \in \mathfrak{M}_{\alpha_n,p}(\mathbb{P}_n)$ with high probability (Shafieezadeh-Abadeh et al., 2019; Kuhn et al., 2019). In such cases the proposed excess worst-case risk bound is sharper if $d > 2p^2$, but slower for the excess risk. This shows the benefit and drawback of $\hat{h}^{\rm prop}_{\alpha_n,p}$. A formal statement for Corollary 1 and other remarks are available in the Supplementary Material.

4. WDRO with locally perturbed data

Recently, the Mixup and its variants have led outstanding performance in many machine learning problems (Zhang et al., 2017; Berthelot et al., 2019). Despite of its empirical successes, theoretical justifications of the Mixup, such as its asymptotic properties, have not been considered much in the literature. In Section 4.1, we define locally perturbed data distributions and describe examples that include the Mixup as a special case. Lastly, we generalize the approximation and risk consistency results presented in Section 3 to the cases when data are locally perturbed in Sections 4.2 and 4.3.

4.1. Locally perturbed data distribution

Definition 2. *For a dataset $\mathcal{Z}_n = \{z_1, \ldots, z_n\}$ and $\beta \ge 0$, we say $\mathbb{P}'_n$ is a β-locally perturbed data distribution if there exists a set $\{z'_1, \ldots, z'_n\}$ such that $\mathbb{P}'_n = \frac{1}{n}\sum_{i=1}^n \delta_{z'_i}$ and z'_i can be expressed as*

$$z'_i = z_i + e_i,$$

for $\|e_i\| \le \beta$ and $i \in [n]$.

Note that $\mathbb{P}_n$ is β-locally perturbed data distribution for any $\beta \ge 0$. The idea of locally perturbed data distribution has been widely applied in machine learning. In the following, we provide three well known examples.

Example 1 (Denoising autoencoder)**.** *Vincent et al. (2010) considered a set $\{z'_1, \ldots, z'_n\}$ of corrupted data defined as follows.*

$$z'_i = z_i D_i,$$

where D_i is a random diagonal matrix with diagonal elements are either one or zero. Let $D_{(n,n)} := \max_{i\in[n]} \sup_{\|z\|\le 1} \|(I - D_i)z\|$ and $\sup_{z\in\mathcal{Z}} \|z\| \le C_{\mathcal{Z}}$. Then, $\|(I - D_i)z_i\| \le \sup_{\|z\|\le 1} \|(I - D_i)z\| C_{\mathcal{Z}} \le D_{(n,n)}C_{\mathcal{Z}}$, and thus training a denoising autoencoder is equivalent to training the autoencoder using a $D_{(n,n)}C_{\mathcal{Z}}$-locally perturbed data distribution.

The next two examples deal with supervised learning settings. For sets $\mathcal{X}$ and $\mathcal{Y}$, suppose $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\|(x, y) - (\tilde{x}, \tilde{y})\| = \|x - \tilde{x}\|_{\mathcal{X}} + \|y - \tilde{y}\|_{\mathcal{Y}}$ for some metrics $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}}$ defined on $\mathcal{X}$ and $\mathcal{Y}$, respectively.

Example 2 (Mixup)**.** *Given a dataset $\mathcal{Z}_n$, we generate a Mixup dataset $\{(x'_i, y'_i)\}_{i=1}^n$ as follows.*

$$x'_i = \gamma_i x_i + (1 - \gamma_i)\tilde{x}_i, \quad y'_i = \gamma_i y_i + (1 - \gamma_i)\tilde{y}_i,$$

for some $(\tilde{x}_i, \tilde{y}_i) \in \mathcal{Z}_n$ and mixing rates $0 \le \gamma_i \le 1$ for all $i \in [n]$. Let $\gamma_{(n,1)} := \min_{i\in[n]} \gamma_i$ and $\sup_{(x,y)\in\mathcal{Z}} \|(x, y)\| \le C_{\mathcal{Z}}$. Then, the Mixup dataset generates a $2(1 - \gamma_{(n,1)})C_{\mathcal{Z}}$-locally perturbed data distribution, since $\|(1 - \gamma_i)((\tilde{x}_i, \tilde{y}_i) - (x_i, y_i))\| \le 2(1 - \gamma_i)C_{\mathcal{Z}} \le 2(1 - \gamma_{(n,1)})C_{\mathcal{Z}}$ for all $i \in [n]$.

Example 3 (Adversarial training)**.** *For a given dataset $\mathcal{Z}_n$, Goodfellow et al. (2014) proposed to minimize a loss with adversarially augmented dataset $\{(x'_i, y_i)\}_{i=1}^n$. Here, each $x'_i = x_i + r_i$ is newly generated data point with perturbation*

$$r_i := \mathrm{argmin}_{\|r\|_{\mathcal{X}} \le \beta_n} \log p_\theta(y_i \mid x_i + r),$$

for some constant $\beta_n > 0$ and $p_\theta(y \mid x)$ is a probability model parametrized by θ. From its construction, it is clear *that adversarial training minimizes the risk under β_n-locally perturbed data distribution. Similar arguments apply to virtual adversarial training (Miyato et al., 2018).*

Remark 3. *The support of a β-locally perturbed data distribution may not be a subset of $\mathcal{Z}$. Instead, it is a subset of $\mathcal{Z} + \mathcal{B}(\beta) := \{z + r \mid z \in \mathcal{Z} \text{ and } \|r\| \leq \beta\}$. As a result, the support of a loss should be larger than $\mathcal{Z}$.*

In the following sections, we present a rigorous analysis of WDRO with a locally augmented data distribution.

4.2. Approximation of the local worst-case risk

We first show that the local worst-case risk can be approximated well by the risk under a locally perturbed data distribution when a loss is Lipschitz continuous:

Proposition 1. *Let (α_n) and (β_n) be sequences of positive numbers converging to zero and $\mathbb{P}'_n$ be a β_n-locally perturbed data distribution. For a constant $M \geq \sup_{n\in\mathbb{N}} \beta_n$, assume that a loss $h : \mathcal{Z} + \mathcal{B}(M) \to \mathbb{R}$ is Lipschitz continuous. Then, for any $p \in [1, \infty)$, the following holds.*

$$\left| R(\mathbb{P}'_n, h) - R^{\text{worst}}_{\alpha_n,p}(\mathbb{P}_n, h) \right| = O(\alpha_n \vee \beta_n).$$

Compared to (6), Proposition 1 reveals that β_n-perturbation causes an additional error $O(\beta_n)$. This error becomes negligible when $\beta_n \leq O(\alpha_n)$. In the following theorem, we obtain a sharper approximation result if a loss has Hölder continuous gradient (cf. Theorem 1).

Theorem 4 (Approximation to the local worst-case risk when data are perturbed)**.** *Let (α_n) and (β_n) be sequences of positive numbers converging to zero and $\mathbb{P}'_n$ be a β_n-locally perturbed data distribution. Let $\mathcal{Z}$ be an open and bounded subset of $\mathbb{R}^d$. For constants $C_{\mathrm{H}}, C_\nabla > 0$, $k \in (0, 1]$, and $M \geq \sup_{n\in\mathbb{N}} \beta_n$, assume that a loss $h : \mathrm{Conv}(\mathcal{Z}) + \mathcal{B}(M) \to \mathbb{R}$ is differentiable, its gradient $\nabla_z h(z)$ is (C_{H}, k)-Hölder continuous, and $\mathbb{E}_{\text{data}}(\|\nabla_z h\|_*) \geq C_\nabla$. Then, for $p \in (1 + k, \infty)$, the following holds.*

$$\begin{aligned}&\left| R(\mathbb{P}'_n, h) + \alpha_n \|\nabla_z h\|_{\mathbb{P}'_n,p^*} - R^{\text{worst}}_{\alpha_n,p}(\mathbb{P}_n, h) \right| \\ &= O_p(\alpha_n^{1+k} \vee \beta_n).\end{aligned}$$

Remark 4. *Theorem 4 extends Theorem 1 to the cases when data are locally perturbed. The cost of perturbation is an additional error $O(\beta_n)$, which is negligible when $\beta_n \leq O(\alpha_n^{1+k})$. Thus Theorem 4 also suggests an appropriate size of perturbation.*

Based on Theorem 4, for vanishing sequences (α_n) and (β_n), and a β_n-locally perturbed data distribution $\mathbb{P}'_n$, we propose to minimize the following objective function.

$$R^{\text{prop}}_{(\alpha_n,\beta_n),p}(\mathbb{P}_n, h) := R(\mathbb{P}'_n, h) + \alpha_n \|\nabla_z h\|_{\mathbb{P}'_n,p^*}, \quad (8)$$

and denote its minimizer by $\hat{h}^{\text{prop}}_{(\alpha_n,\beta_n),p}$, *i.e.*, $\hat{h}^{\text{prop}}_{(\alpha_n,\beta_n),p} = \mathrm{argmin}_{h\in\mathcal{H}} R^{\text{prop}}_{(\alpha_n,\beta_n),p}(\mathbb{P}_n, h)$.

4.3. Risk consistency of the proposed estimator

Now that we study risk consistency when data are locally perturbed. The following two theorems provide risk consistency of the minimizer $\hat{h}^{\text{prop}}_{(\alpha_n,\beta_n),p}$:

Theorem 5 (Excess worst-case risk bound when data are perturbed)**.** *Let (α_n) and (β_n) be sequences of positive numbers converging to zero and $\mathbb{P}'_n$ be a β_n-locally perturbed data distribution. Let $\mathcal{Z}$ be an open and bounded subset of $\mathbb{R}^d$. For constants $C_{\mathrm{H}}, C_\nabla, L > 0$, $k \in (0, 1]$, and $M \geq \sup_{n\in\mathbb{N}} \beta_n$, assume that $\mathcal{H}$ is a uniformly bounded set of differentiable functions $h : \mathrm{Conv}(\mathcal{Z}) + \mathcal{B}(M) \to \mathbb{R}$ such that its gradient $\nabla_z h$ is (C_{H}, k)-Hölder continuous, $\mathbb{E}_{\text{data}}(\|\nabla_z h\|_*) \geq C_\nabla$, and $\mathrm{Lip}(h) \leq L$. Then, for $p \in (1 + k, \infty)$, the following holds.*

$$\begin{aligned}&\mathcal{E}^{\text{worst}}_{\alpha_n,p}(\hat{h}^{\text{prop}}_{(\alpha_n,\beta_n),p}) \\ &= O_p\left(\frac{\mathfrak{C}(\mathcal{H}) \vee \alpha_n^{1-p}}{\sqrt{n}} \vee \log(n)(\alpha_n^{1+k} \vee \beta_n) \right).\end{aligned}$$

Theorem 6 (Excess risk bound when data are perturbed)**.** *Under the same assumptions as Theorem 5, the following holds.*

$$\begin{aligned}&\mathcal{E}(\hat{h}^{\text{prop}}_{(\alpha_n,\beta_n),p}) \\ &= O_p(\mathfrak{R}_n(\mathcal{H}) \vee n^{-1/2} \vee \alpha_n \vee \log(n)(\alpha_n^{1+k} \vee \beta_n)).\end{aligned}$$

Similar to Remark 4, the errors due to the local perturbation both in the order of $\log(n)\beta_n$ are negligible when $\beta_n \leq O(\alpha_n^{1+k})$. In such settings, Theorems 5 and 6 yield the same bound as Theorems 2 and 3, respectively.

Remark 5. *By setting $\beta_n = 2(1 - \gamma_{(n,1)})C_{\mathcal{Z}}$, all the theorems presented in Section 4 apply to the Mixup (see Example 2). To make sure $\lim_{n\to\infty} \beta_n = 0$, we need $\lim_{n\to\infty} \gamma_{(n,1)} = 1$ and it can be satisfied as long as we do not perturb the original data too much as the sample size increases. Similar arguments are applicable to Examples 1 and 3.*

5. Numerical experiments

In this section, we conduct numerical experiments to demonstrate robustness of the proposed method using image classification datasets.

Table 2. Accuracy comparison of the four methods using the clean and noisy test datasets with various training sample sizes. Average and standard deviation are denoted by 'average±standard deviation'. All the results are based on five independent trials. Boldface numbers denote the best and equivalent methods with respect to a t-test with a significance level of 5%.

SAMPLE	CLEAN				1% SALT AND PEPPER NOISE			
SIZE	ERM	WDRO	MIXUP	WDRO+MIX	ERM	WDRO	MIXUP	WDRO+MIX
CIFAR-10								
2500	77.3 ± 0.8	77.1 ± 0.7	$\mathbf{81.4 \pm 0.5}$	$\mathbf{80.8 \pm 0.7}$	69.8 ± 1.8	71.9 ± 0.9	72.7 ± 1.6	$\mathbf{74.8 \pm 0.9}$
5000	83.3 ± 0.4	83.0 ± 0.3	$\mathbf{86.7 \pm 0.2}$	$\mathbf{85.6 \pm 0.3}$	75.2 ± 1.4	77.4 ± 0.5	76.4 ± 1.7	$\mathbf{79.6 \pm 0.9}$
25000	92.2 ± 0.2	91.4 ± 0.1	$\mathbf{93.3 \pm 0.1}$	92.4 ± 0.1	83.3 ± 0.8	$\mathbf{85.8 \pm 0.5}$	82.1 ± 1.7	$\mathbf{86.2 \pm 0.3}$
50000	94.1 ± 0.1	93.1 ± 0.1	$\mathbf{94.8 \pm 0.2}$	93.5 ± 0.2	84.1 ± 1.0	$\mathbf{87.4 \pm 0.5}$	82.5 ± 1.3	$\mathbf{87.3 \pm 0.5}$
CIFAR-100								
2500	33.8 ± 1.0	34.6 ± 1.7	$\mathbf{38.9 \pm 0.6}$	$\mathbf{39.4 \pm 0.2}$	29.2 ± 0.2	30.4 ± 1.2	33.2 ± 1.1	$\mathbf{35.0 \pm 0.5}$
5000	45.2 ± 0.9	43.7 ± 0.7	$\mathbf{49.9 \pm 0.2}$	$\mathbf{49.5 \pm 0.4}$	37.0 ± 0.8	38.1 ± 1.1	39.4 ± 1.3	$\mathbf{42.3 \pm 0.7}$
25000	67.8 ± 0.2	66.6 ± 0.3	$\mathbf{69.3 \pm 0.3}$	68.2 ± 0.3	51.0 ± 1.9	$\mathbf{56.5 \pm 0.8}$	49.6 ± 1.0	$\mathbf{55.8 \pm 0.4}$
50000	74.4 ± 0.2	73.5 ± 0.3	$\mathbf{75.2 \pm 0.2}$	73.8 ± 0.3	51.9 ± 1.3	$\mathbf{62.1 \pm 0.5}$	50.0 ± 3.0	60.6 ± 0.7

Methods We consider the four methods: (i) the empirical risk minimization, denoted by ERM, (ii) the proposed method based on (7), denoted by WDRO, (iii) the empirical risk minimization with the Mixup, denoted by MIXUP, and (iv) the proposed method with the Mixup based on (8), denoted by WDRO+MIX.

Datasets We use the two image classification datasets: CIFAR-10 and CIFAR-100 (Krizhevsky, 2009). For the training, we randomly select 2500, 5000, 25000, or 50000 images from the original datasets, keeping the number of images per class equal. For the testing, we use the original test datasets.

Further implementation details are available in the Supplementary Material and Tensorflow (Abadi et al., 2016)-based scripts are available at `https://github.com/ykwon0407/wdro_local_perturbation`.

5.1. Accuracy comparison

To evaluate robustness of the methods, we compute accuracy on both clean and contaminated datasets. For the latter, we apply the salt and pepper noise to the clean images (Hwang & Haddad, 1995). Figure 1 displays an example of the clean and contaminated images used in our experiments.

Figure 1. An example of clean and contaminated images. We apply the salt and pepper noise to each pixel. The probabilities of noisy pixels are 0%, 1%, 2%, and 4% from left to right, respectively.

Experiment 1 In this experiment, we compare the accuracy of the four methods using the clean and contaminated datasets. For the contaminated datasets, we apply the salt and pepper noise to 1% of pixels. The training sample sizes vary as 2500, 5000, 25000, and 50000. We repeatedly select samples and train models five times[1].

Table 2 compares accuracy of the four methods. For the clean datasets, WDRO+MIX performs comparably with MIXUP and achieves significantly higher accuracy than ERM and WDRO when the sample sizes are 2500 and 5000. When the sample sizes are 25000 and 50000, either WDRO or WDRO+MIX shows lower accuracy than MIXUP and ERM. For the contaminated datasets, either WDRO or WDRO+MIX achieves significantly higher accuracy than ERM and MIXUP in all settings. This shows that the proposed method is robust to contamination of data.

Experiment 2 In this experiment, we compare the reduction of the accuracy from using the clean datasets to the contaminated datasets. For the noise intensity, the probabilities of noisy pixels are set to 1%, 2%, and 4%. We repeatedly train models five times using the original 50000 images.

Table 3 shows accuracy reduction of the four methods. The WDRO, with or without the Mixup, achieves a significantly lower reduction than ERM and MIXUP in every noise level and dataset. For example, on CIFAR-10, the accuracy reduction in WDRO+MIX is 12.7% on average, compared to 24.3% in MIXUP when the probability of noisy pixels is 2%. With the same noise level, on CIFAR-100, the accuracy reduction in WDRO+MIX is 29.7% on average, compared to 45.9% in MIXUP. This result shows that the proposed

[1]Note that when the sample size is 50000, the training set is fixed but trained models can vary due to the randomness of algorithms.

Table 3. The comparison of the accuracy reduction on various salt and pepper noise intensities. Other details are given in Table 2.

PROBABILITY OF NOISY PIXELS	ERM	WDRO	MIXUP	WDRO+MIX
CIFAR-10				
1%	10.1 ± 0.9	$\mathbf{5.7 \pm 0.4}$	12.4 ± 1.2	$\mathbf{6.2 \pm 0.4}$
2%	21.1 ± 1.9	$\mathbf{13.2 \pm 0.5}$	24.3 ± 1.4	$\mathbf{12.7 \pm 0.8}$
4%	39.7 ± 2.9	$\mathbf{32.9 \pm 2.5}$	43.5 ± 1.8	$\mathbf{30.9 \pm 2.0}$
CIFAR-100				
1%	22.5 ± 1.3	$\mathbf{11.4 \pm 0.4}$	25.2 ± 2.5	13.2 ± 0.7
2%	42.8 ± 2.3	$\mathbf{26.5 \pm 1.0}$	45.9 ± 3.4	29.7 ± 0.7
4%	61.7 ± 1.4	$\mathbf{50.0 \pm 0.9}$	63.9 ± 2.0	53.5 ± 0.9

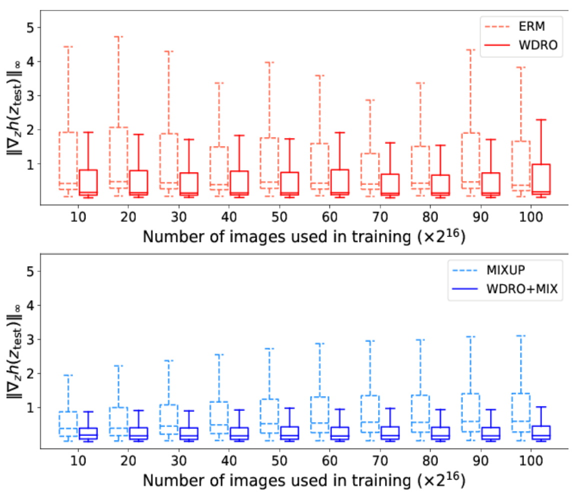

Figure 2. The box plots of the ℓ_∞-norm of the gradients when the number of images used in training increases from 10×2^{16} to 100×2^{16}. We use the original CIFAR-10 test images. The box plots on the top represent the gradient distribution of (dashed) ERM and (solid) WDRO, respectively, and the box plots on the bottom represent that of (dashed) MIXUP and (solid) WDRO+MIX, respectively.

method produces robust models against various noise levels.

5.2. Analysis of the gradient

In this subsection, we demonstrate robustness of our method by analyzing the distribution of the gradients of the loss. We train models with randomly selected 5000 CIFAR-10 images and evaluate the gradients with the CIFAR-10 test images. We consider the ℓ_∞-norm of the gradients $\|\nabla_z h(z_{\text{test}})\|_\infty$ for each test image z_{test}.

Experiment 3 In this experiment, we compare the gradients of ERM vs. WDRO and MIXUP vs. WDRO+MIX as the number of images used in training increases, respectively. Figure 2 shows the box plots of the ℓ_∞-norm of the gradients. Over the entire training phases, the first and third quartiles of the gradients of WDRO (*resp.* WDRO+MIX) are smaller than those of the gradients of ERM (*resp.* MIXUP). This

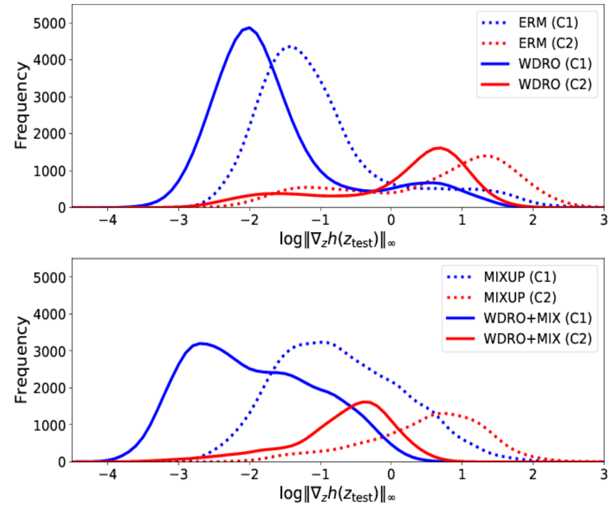

Figure 3. The smoothed histograms of gradients characterizing the two categories: (C1) the images that are correctly classified on both clean and contaminated state and (C2) the other images, respectively. The top panel shows the histograms of ERM (dotted) vs. WDRO (solid) and the bottom panel shows those of MIXUP (dotted) vs. WDRO+MIX (solid), respectively.

result empirically validates that the gradient penalties in (7) and (8) lead to small gradients and robustness of WDRO and WDRO+MIX.

Experiment 4 We visualize smoothed histograms of the gradients for the four methods. We divide the test datasets into the following two categories: (C1) images that are correctly classified on both clean and contaminated state and (C2) images that are incorrectly classified on either clean or contaminated state. In this experiment, the level of noise for the contaminated dataset is 1%.

Figure 3 shows the smoothed histograms of the gradients for (C1) and (C2). The gradients for (C1) depicted in blue are smaller than those for (C2) depicted in red. In both categories (C1) and (C2), the WDRO (*resp.* WDRO+MIX), depicted in the solid line, has smaller gradients than ERM (*resp.* MIXUP) depicted in the dotted line. Thus the proposed method tends to reduce the sizes of gradients in both categories (C1) and (C2), which leads to the robustness of WDRO and WDRO+MIX.

6. Conclusion

In this work, we have developed a principled and tractable statistical inference method for WDRO. In addition, we formally present a locally perturbed data distribution (e.g., Mixup) and develop WDRO inference when data are locally perturbed. Numerical experiments demonstrate robustness of the proposed method.

Acknowledgments

Yongchan Kwon was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT, No.2017R1A2B4008956). Wonyoung Kim and Myunghee Cho Paik were supported by the NRF grant (MSIT, No.2020R1A2C1A0101195011) and Joong-Ho Won was supported by the NRF grant (MSIT, No.2019R1A2C1007126). Wonyoung Kim was also supported by Hyundai Chung Mong-koo foundation.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

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

Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., and Rennen, G. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, pp. 5050–5060, 2019.

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, 2019.

Bühlmann, P. and van de Geer, S. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Chen, R. and Paschalidis, I. C. A robust learning approach for regression models based on distributionally robust optimization. *The Journal of Machine Learning Research*, 19(1):517–564, 2018.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 113–123, 2019.

Esfahani, P. M. and Kuhn, D. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

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 preprint arXiv:1712.06050*, 2017.

Ghosh, S. and Lam, H. Robust analysis in stochastic simulation: Computation and performance guarantees. *Operations Research*, 67(1):232–249, 2019.

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

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

Hu, W., Niu, G., Sato, I., and Sugiyama, M. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pp. 2029–2037, 2018.

Hwang, H. and Haddad, R. A. Adaptive median filters: new algorithms and results. *IEEE Transactions on Image Processing*, 4(4):499–502, April 1995. ISSN 1941-0042. doi: 10.1109/83.370679.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, 2009.

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. In *Advances in Neural Information Processing Systems*, pp. 2687–2696, 2018.

Lim, S., Kim, I., Kim, T., Kim, C., and Kim, S. Fast autoaugment. In *Advances in Neural Information Processing Systems*, pp. 6662–6672, 2019.

Miyato, T., Maeda, S.-i., Koyama, M., and Ishii, S. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41 (8):1979–1993, 2018.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Namkoong, H. and Duchi, J. C. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems*, pp. 2971–2980, 2017.

Rahimian, H. and Mehrotra, S. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

Shafieezadeh-Abadeh, S., Esfahani, P. M. M., and Kuhn, D. Distributionally robust logistic regression. In *Advances in Neural Information Processing Systems*, pp. 1576–1584, 2015.

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

Sinha, A., Namkoong, H., and Duchi, J. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Staib, M. and Jegelka, S. Distributionally robust optimization and generalization in kernel methods. In *Advances in Neural Information Processing Systems*, pp. 9131–9141, 2019.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Vapnik, V. N. An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999, 1999.

Verma, V., Lamb, A., Beckham, C., Najafi, A., Mitliagkas, I., Lopez-Paz, D., and Bengio, Y. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pp. 6438–6447, 2019.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(Dec):3371–3408, 2010.

Wager, S., Wang, S., and Liang, P. S. Dropout training as adaptive regularization. In *Advances in neural information processing systems*, pp. 351–359, 2013.

Yao, Y., Rosasco, L., and Caponnetto, A. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

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, 2019.