PROVABLE ROBUST OVERFITTING MITIGATION IN WASSERSTEIN DISTRIBUTIONALLY ROBUST OPTIMIZATION

Shuang Liu, Yihan Wang, Yifan Zhu, Yibo Miao, Xiao-Shan Gao*
[1] State Key Laboratory of Mathematical Sciences
Academy of Mathematics and Systems Science, Chinese Academy of Sciences
Beijing 100190, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China

ABSTRACT

Wasserstein distributionally robust optimization (WDRO) optimizes against worst-case distributional shifts within a specified uncertainty set, leading to enhanced generalization on unseen adversarial examples, compared to standard adversarial training which focuses on pointwise adversarial perturbations. However, WDRO still suffers fundamentally from the robust overfitting problem, as it does not consider statistical error. We address this gap by proposing a novel robust optimization framework under a new uncertainty set for adversarial noise via Wasserstein distance and statistical error via Kullback-Leibler divergence, called the Statistically Robust WDRO. We establish a robust generalization bound for the new optimization framework, implying that out-of-distribution adversarial performance is at least as good as the statistically robust training loss with high probability. Furthermore, we derive conditions under which Stackelberg and Nash equilibria exist between the learner and the adversary, giving an optimal robust model in certain sense. Finally, through extensive experiments, we demonstrate that our method significantly mitigates robust overfitting and enhances robustness within the framework of WDRO.

1 INTRODUCTION

Optimization in machine learning is often challenged by data ambiguity caused by natural noise, adversarial attacks (Goodfellow et al., 2014), or other distributional shifts (Malinin et al., 2021; 2022). To develop robust models against data ambiguity, Wasserstein Distributionally Robust Optimization (WDRO) (Kuhn et al., 2019) has emerged as a powerful modeling framework, which has recently attracted significant attention attributed to its connections to generalization and robustness (Lee & Raginsky, 2018; Mohajerin Esfahani & Kuhn, 2018; An & Gao, 2021).

Specifically, WDRO optimizes performance under the most adversarial data distribution within a certain Wasserstein distance from the observed empirical distribution $\mathcal{D}_n$, as shown below.

$$\inf_{\theta\in\Theta} \sup_{\mathcal{D}':\mathcal{D}'\in\mathcal{U}(\mathcal{D}_n)} \mathbb{E}_{z\sim\mathcal{D}'}[L(\theta, z)]. \tag{1}$$

Here $\mathcal{U}(\mathcal{D}_n) = \{\mathcal{D}' : \mathrm{W}_p(\mathcal{D}_n, \mathcal{D}') \leq \varepsilon\}$ is the ambiguity set consisting of all possible distributions of interest, and $L(\cdot, \cdot)$ is the loss function. From both intuitive and theoretical perspectives, WDRO can be considered as a more comprehensive framework that subsumes and extends adversarial training (Staib & Jegelka, 2017; Sinha et al., 2017; Bui et al., 2022; Phan et al., 2023). In contrast to standard adversarial training (Madry et al., 2017), WDRO operates on a broader scale by perturbing the entire data distribution rather than pointwise adversarial samples. This approach inherently promotes generalization to unseen adversarial examples.

However, we find that WDRO still fundamentally suffers from the same robust overfitting phenomenon as standard adversarial training (Rice et al., 2020). This phenomenon observed in our

*Corresponding author.

experiments as a degradation in adversarial robustness on test data after a certain point in the training process, which typically occurs shortly after the first learning rate decay (refer to Figure 2).

A primary factor contributing to overfitting behavior is the inherent statistical error arising from finite sampling of train data (Van Parys et al., 2021; Bennouna & Van Parys, 2022). This error, defined as the discrepancy between the empirical distribution of the training data and the true underlying data distribution, manifests as a gap between empirical and population risk. Consequently, models optimized via traditional empirical risk minimization (ERM) may perform well on training data but generalize poorly to unseen samples. Although the recent literature Lam (2019); Van Parys et al. (2021); Bennouna & Van Parys (2022) has acknowledged the significance of statistic error in robust optimization frameworks, its incorporation into WDRO remains an unexplored avenue.

To address these issues, we propose a novel ambiguity set obtained by combining Kullback-Leibler divergence and Wasserstein distance, which is called **Statistically Robust WDRO (SR-WDRO)**. We prove that the SR-WDRO formulation guarantee an upper bound on the adversarial test loss for any desired robustness and any loss function. Furthermore, we formalize the SR-WDRO as a zero-sum game between the learner (decision maker), who chooses a model parameter $\theta \in \Theta$, and the adversary who chooses a distribution from an ambiguity set, and we derive conditions under which Stackelberg and Nash equilibria exist between the learner and the adversary.

Finally, building upon our theoretical framework, we introduce a novel and practical statistically robust loss training algorithm for neural networks, which incurs negligible additional computational burden compared to standard adversarial training or distributionally robust training. We conduct extensive experiments on benchmark datasets, which show that our proposed approach demonstrates superior efficacy in mitigating overfitting compared to existing training methodologies, in that our approach maintains performance on in-distribution data while substantially improving robustness to out-of-distribution samples, thus mitigating the critical challenge in generalization.

The theoretical and practical contributions of this paper are summarized as follows.

- We propose SR-WDRO based on our novel ambiguity set to mitigate robust overfitting in WDRO.
- Theoretically, we prove that the adversarial test loss can be upper bounded by the statistically robust training loss with high probability, which can be considered as generalization bound for SR-WDRO. We also establish the conditions necessary for the existence of Stackelberg and Nash equilibria between the learner and the adversary.
- We introduce a practically robust training algorithm based on SR-WDRO which maintains computational efficiency.
- We conduct extensive evaluations on benchmark datasets and our results show that the SR-WDRO approach effectively mitigates robust overfitting and outperforms other existing robust methods in terms of adversarial robustness.

2 Related Works

Robust Overfitting. The phenomenon of robust overfitting has been a focal point in the field of adversarial machine learning. The seminal work by Rice et al. (2020) brought this issue to the forefront, demonstrating that after a certain point in standard adversarial training, i.e., shortly after the first learning rate decay, the robust performance on test data will continue to degrade with further training. Notably, conventional overfitting remedies such as explicit regularization and data augmentation do not improve performance beyond early stopping. Schmidt et al. (2018) attributed robust overfitting to sample complexity theory and suggested that more training data are required for adversarial robust generalization. This assertion is supported by empirical results in subsequent studies (Schmidt et al., 2018; Zhai et al., 2019). Recent works also proposed various strategies to mitigate robust overfitting without relying on additional training data by sample re-weighting (Zhang et al., 2020; Wang et al., 2019), adversarial weight perturbation (Wu et al., 2020), weight smoothing (Chen et al., 2020), favoring large loss data (Yu et al., 2022), and considering the memorization effect (Dong et al., 2022; Wang et al., 2024). In our work, we will introduce the statistically robustness into WDRO to mitigate overfitting theoretically and practically.

WDRO. WDRO has recently received significant attention in the context of adversarial training (Staib & Jegelka, 2017; Wong et al., 2019; Sinha et al., 2017; Bui et al., 2022; Phan et al., 2023). Unlike standard adversarial training (Madry et al., 2017), which defends against attacks by bounding the perturbation to each individual data point, WDRO provides protection by imposing a bound on the average perturbation constrained by Wasserstein distance applied on the entire data distribution. This distinction positions WDRO as a more holistic approach to handling adversarial perturbations and generalizes better than adversarial training on unseen data samples (Bui et al., 2022). However, despite this advantage, WDRO remains vulnerable to the same robust overfitting phenomenon observed in standard adversarial training, primarily due to its lack of consideration for statistical error. Recently, Bennouna & Van Parys (2022); Bennouna et al. (2023) proposed a DRO formulation using an ambiguity set combining Kullback-Leibler divergence and Levy-Prokhorov metric in an attempt to protect against corruption and statistical errors. In the absence of data poisoning, their framework effectively reduces to improving standard adversarial training with statistical robustness. However, our approach diverges significantly. We leverage the inherent advantages of WDRO as our foundation and then augment it with statistical robustness to mitigate robust overfitting, which enables our framework to achieve superior performance against adversarial perturbations.

3 PRELIMINARIES

Notations. Let $\mathcal{Z} \subset \mathbb{R}^m$ be a compact nonempty set equipped with metric $\mathrm{d}(\cdot\ ,\ \cdot)$ and $\mathrm{diam}(\mathcal{Z}) = \sup\{\mathrm{d}(z, z') : z, z' \in \mathcal{Z}\}$ the diameter of $\mathcal{Z}$ which is finite. The class of Borel probability measures on $\mathcal{Z}$ is denoted by $\mathcal{P}(\mathcal{Z})$. For simplicity, we denote $\mathbb{E}_\mu[f(z)]$ as the expectation of $f(z)$ with $z \sim \mu$.

Definition 1 (Wasserstein metric)**.** *For $p \geq 1$, the p-th Wasserstein metric between $\mu, \nu \in \mathcal{P}(\mathcal{Z})$ is*

$$\mathrm{W}_p(\mu, \nu) := \inf_{\pi \in \Pi(\mu,\nu)} \left\{ \mathbb{E}_{(z,z')\sim\pi} \left[\mathrm{d}^p(z, z') \right] \right\}^{\frac{1}{p}}$$

where the infimum is taken over all coupling of μ and ν, i.e. probability measure π on the product space $\mathcal{Z} \times \mathcal{Z}$ with given marginals μ and ν.

Definition 2 (KL divergence)**.** *Kullback-Leibler (KL) divergence between $\mu, \nu \in \mathcal{P}(\mathcal{Z})$ is*

$$\mathrm{KL}(\mu, \nu) := \int_{\mathcal{Z}} p(z) \log \frac{p(z)}{q(z)} dz$$

where $p(z), q(z)$ are probability density functions of μ, ν, respectively.

Wasserstein metric is a function defined between two probability distributions, which represents the cost of an optimal mass transportation plan. Kullback-Leibler divergence measures the difference between two probability distributions, quantifying how one distribution diverges from a reference distribution. We additionally consider the Levy-Prokhorov (LP) metric and the total variation metric (TV). The Levy-Prokhorov metric is theoretically important because weak convergence of probability distribution is equivalent to the convergence in the Levy-Prokhorov metric. In Appendix A.1, we establish the relationship among these probability discrepancies.

4 STATISTICALLY ROBUST WDRO

In Section 4.1, we first present the definition and game-theoretic description of SR-WDRO. Subsequently, in Section 4.2, we demonstrate that models trained with statistically robust loss achieve out-of-distribution generalization. Finally, in Section 4.3 we examine the existence of Stackelberg and Nash equilibria under some assumptions.

4.1 STATISTICALLY ROBUST WDRO AND GAME DESCRIPTION

We perform stochastic optimization with respect to an unknown data distribution $\mathcal{D}$, given access only to an empirical distribution $\mathcal{D}_n$, where n is the observed finite sample size. The goal of distributionally robust optimization is to learn a robust model from $\mathcal{D}_n$ that will perform well on unseen test distribution $\mathcal{D}_{\text{test}}$, which may either be the true data distribution $\mathcal{D}$ or more likely distributions shifted from $\mathcal{D}$. Our work focuses primarily on distribution shifts caused by adversarial attacks.

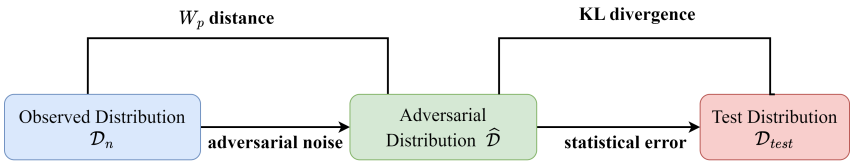

Figure 1: Illustration of our SR-WDRO. The adversarial perturbations are quantified using Wasserstein distance between $\mathcal{D}_n$ and $\widehat{\mathcal{D}}$. The adversarial distribution is then compared with the test distribution $\mathcal{D}_{\text{test}}$ using Kullback-Leibler divergence, which accounts for the statistical error.

The WDRO constructs an ambiguity set based on the empirical distribution $\mathcal{D}_n$, and optimizes performance under the most adversarial distribution within a certain Wasserstein distance from $\mathcal{D}_n$ as shown in Eq (1). However, previous WDRO methods still suffer fundamentally from the robust overfitting phenomenon. To mitigate this issue, we incorporate the Kullback-Leibler (KL) divergence in WDRO, specifically aiming to reduce statistical error caused by training on finite samples, a brief illustration is shown in Figure 1.

Our SR-WDRO can be interpreted as a novel combination of Kullback-Leibler divergence and Wasserstein distributional robust optimizations. We consider the ambiguity set:

$$\mathcal{U}(\mathcal{D}_n) := \{\mathcal{D}' \in \mathcal{P}(\mathcal{Z}) : \exists \mathcal{D}'' \in \mathcal{P}(\mathcal{Z}) \text{ s.t. } \mathrm{W}_p(\mathcal{D}_n, \mathcal{D}'') \leq \varepsilon,\ \mathrm{KL}(\mathcal{D}'', \mathcal{D}') \leq \gamma\} \tag{2}$$

where the desired protection against adversarial noise is controlled by the *adversarial budget* ε and statistical error is accounted by the *statistical budget* $\gamma (\gamma > 0)$.

Now we can formulate this robust optimization problem as a *zero-sum game* between the learner who chooses a decision model $\theta \in \Theta$ and an adversary who chooses a distribution $\mathcal{D}'$ from the ambiguity set $\mathcal{U}(\mathcal{D}_n)$. Let $L(\theta, z)$ be the loss function (e.g. cross-entropy loss). Our game model is as follows:

- The adversary chooses a distribution $\mathcal{D}' \in \mathcal{U}(\mathcal{D}_n)$ to maximize the loss $\mathbb{E}_{\mathcal{D}'}[L(\theta, z)]$.
- The learner selects a decision model $\theta \in \Theta$, aiming to minimize the loss $\mathbb{E}_{\mathcal{D}'}[L(\theta, z)]$.

We introduce the *statistically robust WDRO (SR-WDRO) problem*:

$$\inf_{\theta \in \Theta} \sup_{\mathcal{D}' \in \mathcal{U}(\mathcal{D}_n)} \mathbb{E}_{\mathcal{D}'}[L(\theta, z)] \tag{3}$$

and denote the *statistically robust loss* as:

$$\mathcal{L}_{\varepsilon,\gamma}(\theta, \mathcal{D}_n) := \sup_{\mathcal{D}' \in \mathcal{U}(\mathcal{D}_n)} \mathbb{E}_{\mathcal{D}'}[L(\theta, z)]. \tag{4}$$

4.2 Certified generalization

In this subsection, we rigorously establish that the model trained with statistically robust loss $\mathcal{L}_{\varepsilon,r}(\theta, \mathcal{D}_n)$ enjoys the asymptotic out of distribution generalization guarantee and safeguards against the statistical error and adversarial noises simultaneously. Let $\mathbb{B}(z; \epsilon) := \{z' \in \mathcal{Z} \mid \mathrm{d}(z, z') \leq \varepsilon\}$ be the ϵ-ball around z under metric d. Proofs are provided in Appendix A.2.

Theorem 3 (Generalization certificate)**.** *Let $\mathcal{D}$ be the true data distribution, and $\mathcal{D}_n$ be the observed empirical distribution sampled i.i.d. from $\mathcal{D}$. Then for all $\varepsilon > 0$, let $\delta = (\frac{\varepsilon}{\mathrm{diam}(\mathcal{Z})+1})^p$, we have*

$$\mathbb{P}\Big(\forall \theta \in \Theta, \mathbb{E}_{\mathcal{D}}[L(\theta, z)] \leq \mathcal{L}_{\varepsilon,\gamma}(\theta, \mathcal{D}_n)\Big) \geq 1 - e^{-\gamma n} \left(\frac{4}{\delta}\right)^{m(\mathcal{Z},\delta)} \tag{5}$$

where $m(\mathcal{Z}, \delta) := \min\{k \geq 0 : \exists \xi_1, \cdots, \xi_k \in \mathcal{Z},\ \text{s.t. } \cup_{i=1}^{k} \mathbb{B}(\xi_i, \delta) \supseteq \mathcal{Z}\}$ is the internal covering number *of the support set $\mathcal{Z}$.*

In the proof of Theorem 3, we in fact prove that the true distribution $\mathcal{D}$ can be in the ambiguity set $\mathcal{U}(\mathcal{D}_n)$ with high probability. Compared to the results in (Bennouna & Van Parys, 2022), the bound we derived elucidates the dependency on parameters ε and γ, and the internal covering number is directly related to ε. Unlike classical statistical learning, our bounds do not depend on the dimension of the parameter space Θ but rather on the internal covering number of sample space $\mathcal{Z}$. The result is uniform in that the probability approaches 1 for any θ when the sample size n is sufficiently large.

Remark 4. *To ensure that the bound in equation 5 is meaningful, the condition $\gamma > m(\mathcal{Z}, \delta) \cdot \log(4/\delta)/n$ must hold. Although this condition may seem strong due to the exponential dependency on the dimension of the sample space $\mathcal{Z}$, the cover number is not excessively large in practice because the intrinsic dimensionality of the sample space is often much lower than its ambient dimension. For example, the intrinsic dimensionality of MNIST is around 13, CIFAR-10 approximately 26, and ImageNet between 26 and 43 (Facco et al., 2017; Pope et al., 2021). Our theory can leverage this intrinsic dimensionality instead of the ambient dimension, making the condition more practical.*

We extend our analysis to consider adversarial robustness, wherein the test distribution $\mathcal{D}$ is subject to adversarial perturbations. We formalize this scenario by introducing an adversarial transformation:

$$\mathcal{M} : z \to \arg\max_{z' \in \mathbb{B}(z;\epsilon)} L(\theta, z').$$

This transformation maps each test input to its worst-case adversarial example within the ϵ-neighborhood. Let $\mathcal{M}_{\#}\mathcal{D}$ denote the pushforward measure of $\mathcal{D}$ under $\mathcal{M}$, representing the distribution of adversarial examples generated from $\mathcal{D}$. And the expected loss under this adversarial distribution can be expressed as: $\mathbb{E}_{\mathcal{M}_{\#}\mathcal{D}} L(\theta, z) = \mathbb{E}_{\mathcal{D}}[\max_{z' \in \mathbb{B}(z;\epsilon)} L(\theta, z')]$. Our main result establishes a high-probability bound on the test adversarial loss:

Theorem 5 (Robustness certificate)**.** *Let $\mathcal{D}$ be the true data distribution, $\mathcal{D}_n$ the empirical distribution sampled i.i.d. from $\mathcal{D}$, and $\delta = (\frac{\sigma}{\mathrm{diam}(\mathcal{Z})+1})^p$ for any $\sigma > 0$. Then*

$$\mathbb{P}\Big(\forall \theta \in \Theta, \mathbb{E}_{\mathcal{D}}[\max_{z' \in \mathbb{B}(z;\epsilon)} L(\theta, z')] \leq \mathcal{L}_{\varepsilon+\sigma,\gamma}(\theta, \mathcal{D}_n)\Big) \geq 1 - e^{-n\gamma} \left(\frac{4}{\delta}\right)^{m(\mathcal{Z},\delta)} \tag{6}$$

where $m(\mathcal{Z}, \delta) := \min\{k \geq 0 : \exists \xi_1, \cdots, \xi_k \in \mathcal{Z},\ s.t.\ \cup_{i=1}^{k} \mathbb{B}(\xi_i, \delta) \supseteq \mathcal{Z}\}$ denote the internal covering number of the support set $\mathcal{Z}$.

From Theorem 5, we observe that ensuring ε-robustness on the test set typically needs a larger attack budget $\varepsilon + \sigma$ during training. However, according to Theorem 3, an excessively large training budget may impede generalization for benign data. These findings align with previous experimental results on adversarial training (Tsipras et al., 2018; Andriushchenko & Flammarion, 2020).

Remark 6 (Comparison with Standard WDRO)**.** *From a theoretical perspective, our framework differs from standard WDRO in both the assumptions required and the form of the generalization bounds. Specifically, the generalization bounds in (Azizian et al., 2024) rely on additional assumptions about the loss function, such as Lipschitz continuity and boundedness, which are not required in our framework. Furthermore, the form of the generalization bounds in our framework is fundamentally different from those of WDRO. Our bound provides a stronger guarantee, where the probability of failure (i.e., the complement of the confidence level) decays exponentially with the sample size n. Importantly, this rate of decay is directly influenced by γ.*

Remark 7. *The result is also applicable to general out-of-distribution shifts such as domain generalization. Details are given in Appendix A.3.*

4.3 Nash Equilibrium and Stackelberg Equilibrium

As mentioned in Section 4.1, the SR-WDRO problem (3) constitutes a zero-sum game. In this subsection, we show that Stackelberg equilibrium exists under natural assumptions and Nash equilibrium exists under more assumptions. The detailed definitions of the game equilibrium, along with the proofs for this section, are provided in the Appendix A.4.

We need the following assumptions on the loss function.

Assumption 8 (Loss function)**.**

i) For any $\theta \in \Theta$, the loss function $L(\theta, z)$ is non-negative and upper semi-continuous in z.

ii) For any $z \in \mathcal{Z}$, the loss function $L(\theta, z)$ is lower semi-continuous in θ.

iii) For any $\theta \in \Theta$, there exist $c > 0$, $\widehat{z}_0 \in \mathcal{Z}$ and $k \in (0, p)$ such that $L(\theta, z) \leq c[1 + \mathrm{d}^k(z, \widehat{z}_0)]$ for all $z \in \mathcal{Z}$.

Assumptions (i) and (ii) regarding the continuity of the loss function are common and are satisfied by neural network models, and assumption (iii) is necessary for the existence of the maximum in (4). The following proposition establishes that the inner maximization problem is solvable, which is the key ingredient required to prove the existence of Stackelberg and Nash equilibrium.

Proposition 9. *The ambiguity set $\mathcal{U}(\mathcal{D}_n) := \{\mathcal{D}' \in \mathcal{P}(\mathcal{Z}) : \exists \mathcal{D}'' \in \mathcal{P}(\mathcal{Z}) \text{ s.t. } \mathrm{W}_p(\mathcal{D}_n, \mathcal{D}'') \leq \varepsilon,\ \mathrm{KL}(\mathcal{D}'', \mathcal{D}') \leq \gamma\}$ is compact in $\mathcal{P}(\mathcal{Z})$ with respect to weak topology. If Assumption 8 is satisfied, then the supremum in (4) is finite and can be attained for any $\theta \in \Theta$.*

Stackelberg game: We also consider a threat model in which the adversary has full knowledge of the learner's actions before determining its own strategy. We can model our SR-WDRO problem (3) as a zero-sum Stackelberg game between the leader (decision-maker) who chooses a decision model $\theta \in \Theta$ at first, and the follower (adversary) who chooses a distribution $\mathcal{D}'$ from the ambiguity set $\mathcal{U}(\mathcal{D}_n)$ after observing the leader's action. As a corollary of Proposition 9, we can give the existence of Stackelberg equilibrium.

Theorem 10 (Stackelberg Equilibrium)**.** *If Assumption 8 holds and Θ is compact, then the game has a Stackelberg equilibrium. i.e. there exists a $(\theta^*, \mathcal{D}'^*(\theta^*))$ such that*

$$\theta^* \in \arg\min_{\theta \in \Theta} \max_{\mathcal{D}' \in \mathrm{BR}(\mathcal{D}_\mathrm{n})} \mathbb{E}_{\mathcal{D}'}[L(\theta, z)], \quad \mathcal{D}'^*(\theta^*) \in \mathrm{BR}(\mathcal{D}_\mathrm{n}) = \arg\max_{\mathcal{D}' \in \mathcal{U}(\mathcal{D}_\mathrm{n})} \mathbb{E}_{\mathcal{D}'}[\mathrm{L}(\theta^*, \mathrm{z})].$$

In addition, with the convexity of $L(\theta, z)$ in θ, we can prove the existence of Nash equilibrium.

Theorem 11 (Minimax Theorem)**.** *If Assumption 8 holds, Θ is convex, and $L(\theta, z)$ is convex in θ for any $z \in \mathcal{Z}$, then*

$$\min_{\theta \in \Theta} \max_{\mathcal{D}' \in \mathcal{U}(\mathcal{D}_n)} \mathbb{E}_{\mathcal{D}'}[L(\theta, z)] = \max_{\mathcal{D}' \in \mathcal{U}(\mathcal{D}_n)} \min_{\theta \in \Theta} \mathbb{E}_{\mathcal{D}'}[L(\theta, z)]. \tag{7}$$

We thus have established the existence for the Stackelberg and Nash equilibria of the SR-WDRO problem (3), viewed as a zero-sum game between the learner and the adversary. The existence of Stackelberg equilibrium is considered for the first time for WDRO approaches. It can be observed that the Stackelberg equilibrium requires weaker conditions to exist than the Nash equilibrium, which gives the *smallest statistically robust loss* among all decision models can be considered optimal robust in certain sense. Our result significantly extends the scope of WDRO theory and tackles a substantially more complex and general formulation compared with previous work (Shafieezadeh-Abadeh et al., 2023).

5 Computationally Tractable Reformulation of SR-WDRO

Having established the robustness certificate and equilibrium existence of SR-WDRO, we now turn our attention to practical methods for solving the optimization problem defined in Equation (3). We first show that the inner maximization problem (4) can be simplified to a minimization problem. Next, we apply our previous results to classification tasks and provide an approximate algorithm to solve the robust loss effectively.

Proposition 12 (Strong duality)**.** *$\mathcal{L}_{\varepsilon,\gamma}(\theta, \mathcal{D}_n)$ in (4) admits the dual formulation for all $\gamma > 0$:*

$$\mathcal{L}_{\varepsilon,\gamma}(\theta, \mathcal{D}_n) = \inf_{\substack{\lambda, \beta \geq 0 \\ \eta \geq \max_{z \in \mathcal{Z}} L(\theta, z)}} \{\lambda \varepsilon^p + \mathbb{E}_{\mathcal{D}_n}[\varphi(\lambda, \beta, \eta, z)]\} \tag{8}$$

where $\varphi(\lambda, \beta, \eta, z) = \sup_{\xi \in \mathcal{Z}} \{\beta \log \frac{\beta}{\eta - L(\theta, \xi)} + (\gamma - 1)\beta + \eta - \lambda \mathrm{d}^p(z, \xi)\}$.

In contrast, the dual reformulation for classical WDRO only involves λ and takes the expectation of the implicit function $\sup_{\xi \in \mathcal{Z}} \{L(\theta, \xi) - \lambda \mathrm{d}^p(z, \xi)\}$ with respect to $\mathcal{D}_n$. The additional variables β, η are introduced to account for the KL divergence.

Algorithm 1 Statistically Robust WDRO Training

Input: Training set $\mathcal{S}_n$, number of iterations T, batch size N, learning rate $\eta_\theta,\ \eta_\lambda$, adversary parameters: attack budget ε, steps K, step size η.
Output: Robust model θ.

for $t = 1$ to T **do**
  Sample a mini-batch $\{(x_i, y_i)\}_{i=1}^N \in \mathcal{S}_n$
  Find UDR adversarial examples $\{x_i^a\}_{i=1}^N$:
    (1) $x_i^0 = x_i + \delta$ where $\delta \sim \text{Uniform}(-\varepsilon, \varepsilon)$
    (2) for $k = 1$ to K:
      (a) $x_i^{inter} = x_i^{k-1} + \eta \text{sign}(\nabla_x L(\theta, (x_i^{k-1}, y_i)))$
      (b) $x_i^k = x_i^{inter} - \eta\lambda\nabla_x \widehat{d}(x_i^{inter}, x_i)$
    (3) Clip to valid range: $x_i^a = \text{clip}(x_i^K, 0, 1)$
  Update parameter λ: $\lambda \leftarrow \lambda - \eta_\lambda \left(\varepsilon - \frac{1}{N}\sum_{i=1}^N \widehat{d}_{\mathcal{X}}(x_i^a, x_i)\right)$
  Compute optimal weights $\{p_i\}_{i=1}^N$ in problem (9)
  Update model parameter: $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \mathcal{L}_{\varepsilon,r}(\theta; \{(x_i, y_i)\}_{i=1}^n)$
end for

However, directly solving the statistically robust loss through either (4) or (8) is intractable in practice. In the remainder of this section, we will primarily focus on applying results to classification tasks and provide a finite reformulation to solve the statistically robust loss approximately.

In classification tasks, it is often intuitive to restrict adversarial perturbations solely to input samples (feature vectors). We present an adaptation of existing results to such scenarios. Let $\mathcal{Z} = (X, Y) \in \mathcal{X} \times \mathbb{R}$, where $X \in \mathcal{X} \subset \mathbb{R}^{(m-1)}$ represents input data, and $Y \in \mathbb{R}_+$ denotes a label. In the classification settings, $Y \in \{1, ..., K\}$. We consider an adversary capable of perturbing only the input samples X. This constraint can be elegantly incorporated into our robust formulation by defining the Wasserstein cost function $d : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}+$ as follows. For $z = (x, y)$ and $z' = (x', y')$, we introduce the *sample-shift cost function*:

$$\mathrm{d}(z, z') = \mathrm{d}_{\mathcal{X}}(x, x') + M \cdot \mathbf{1}\{y \neq y'\}$$

where $\mathrm{d}_{\mathcal{X}}$ is the distance (like L_q norm) on $\mathcal{X}$ and we assume that M is a very large positive number to prevent label conversion.

Now we revisit the statistically robust loss:

$$\begin{aligned}\mathcal{L}_{\varepsilon,\gamma}(\theta, \mathcal{D}_n) &= \sup_{\mathcal{D}' \in \mathcal{U}(\mathcal{D}_n)} \mathbb{E}_{\mathcal{D}'}[L(\theta, z)] \\ &= \sup\left\{\sup\left\{\mathbb{E}_{\mathcal{D}'}\left[L(\theta, z)\right] : \mathcal{D}', \mathrm{KL}\left(\mathcal{Q}, \mathcal{D}'\right) \leq \gamma\right\} : \mathcal{Q}, \mathrm{W}_p(\mathcal{D}_n, \mathcal{Q}) \leq \varepsilon\right\}.\end{aligned}$$

Intuitively, we start with the empirical distribution $\mathcal{D}_n$ and identify the most adversarial sample distribution that satisfies the Wasserstein distance constraint. Then, within the KL divergence constraint, similarly to (Bennouna et al., 2023), we adjust the sample weights to find the sample distribution that maximizes the final loss. To identify the most adversarial sample distribution, we employ the UDR method (Bui et al., 2022), which leverages the dual formulation of WDRO. In order to match this method, we set $p = 1$ in our framework. This approach constructs a new Wasserstein cost function to search for adversarial samples that incorporate both local and global information, in contrast to Adversarial Training that considers only local information for each sample.

For any given training data (sub)set $\{z_i = (x_i, y_i)\}_{i=1}^n$, we can compute the statistically robust loss $\mathcal{L}_{\varepsilon,\gamma}(\theta)$ approximately as

$$\mathcal{L}_{\varepsilon,\gamma}(\theta, \{(x_i, y_i)\}_{i=1}^n) = \begin{cases} \max \sum_{i=1}^n p_i L\left(\theta, (x_i', y_i)\right) \\ \qquad x_i' \in \arg\max_{x_i'} L(\theta, (x_i', y_i)) - \lambda \widehat{d}_{\mathcal{X}}\left(x_i', x_i\right) \\ \text{s.t.} \quad \boldsymbol{p} \in \mathbb{R}_+^n, \sum_{i=1}^n p_i = 1, \\ \qquad \boldsymbol{q} \in \mathbb{R}_+^n, q_i = \frac{1}{n}\ \forall i = 1, \cdots, n, \\ \qquad \mathrm{KL}(\boldsymbol{q}||\boldsymbol{p}) = \sum_{i=1}^n q_i \log\left(\frac{q_i}{p_i}\right) \leq \gamma. \end{cases} \tag{9}$$

The λ comes from the dual formulation of WDRO and is a learnable parameter in UDR, and

$$\widehat{d}_{\mathcal{X}}(x, x') := \mathbf{1}\{\mathrm{d}_{\mathcal{X}}(x, x') < \varepsilon\}\, \mathrm{d}_{\mathcal{X}}(x, x') + \mathbf{1}\{\mathrm{d}_{\mathcal{X}}(x, x') \geq \varepsilon\} \left(\varepsilon + \frac{\mathrm{d}_{\mathcal{X}}(x, x') - \varepsilon}{\tau}\right)$$

where $\tau > 0$ is the temperature to control the growing rate of the cost function when x' is out of the perturbation ball.

The statistically robust loss is therefore in essence simply a re-weighting of adversarial loss. Determining the statistically robust loss exactly requires evaluating the adversarial loss $L(\theta, z')$ where the adversarial example is computed through the WDRO method (Bui et al., 2022), subsequently solving the exponential cone problem with n variables and constraints. This optimization can be efficiently executed using standard solvers. More details are shown in Algorithm 1.

6 Experiments

In this section, we investigate the efficacy of our SR-WDRO training through extensive experiments on the CIFAR-10 and CIFAR-100 datasets. We will show that our approach largely circumvents the robust overfitting phenomenon experienced by WDRO and achieves better adversarial robustness than other robust methods. The implementation of our approach is publicly available at the following GitHub repository: https://github.com/hong-xian/SR-WDRO.

We compare our method with PGD-AT (Madry et al., 2017) and distributional robust methods: UDR-AT (Bui et al., 2022), HR training (Bennouna et al., 2023). We train ResNet-18 (He et al., 2016) with 200 epochs, and use SGD as the optimizer with learning rate decay by 0.1 at the epoch 100 and 150. For all methods, we implement adversarial training with $\{k = 10, \varepsilon = 8/255, \eta = 2/255\}$ where k is the iteration number, ε is the attack budget and η is the step size. We use different attacks to evaluate the defense methods, including: 1) PGD-10 with $\{k = 10, \varepsilon = 8/255, \eta = \varepsilon/4\}$, 2) PGD-200 with $\{k = 200, \varepsilon = 8/255, \eta = \varepsilon/4\}$, 3) Auto-Attack (AA) (Croce & Hein, 2020) with $\varepsilon = 8/255$, which is an ensemble method of four different attacks. The l_∞-norm is used for all measures. Unless otherwise specified, we set $\gamma = 0.1$ to its default value. The ablation study on γ is provided in latter part. We repeated all the experiments three times for different seeds. More training details are provided in Appendix A.6.

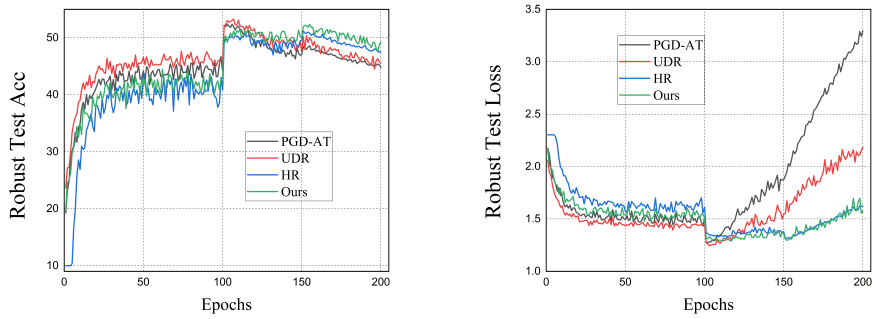

Figure 2: Comparison of SR-WDRO against other robust training methods on CIFAR10 ($\varepsilon = 8/255$). Left: Robust test accuracy. Right: Robust test loss. Our method (green) demonstrates competitive performance in both metrics, particularly in mitigating robust overfitting and higher robust test accuracy.

Mitigating robust overfitting. First, we present the adversarial loss and adversarial accuracy on the test set over the 200 training epochs for different methods in Figure 2. It can be observed that the UDR method, similar to standard AT, suffers from severe robust overfitting. This is particularly evident in the adversarial loss on the test set, which increases sharply when robust overfitting occurs. Both HR and our SR-WDRO effectively mitigate this issue, and our method demonstrates superior robust test accuracy. Furthermore, in Table 1, we report the best robust test accuracy and the final

Table 1: Performance of different methods on CIFAR-10 and CIFAR-100 using a ResNet-18 for l_∞ with budget 8/255. We use PGD-10 as the attack to evaluate robust performance. Nat is the natural test accuracy. The "Best" Robust Test Acc is the highest robust test accuracy achieved during training whereas the "Final" Robust Test Acc is last epoch's robust accuracy. Best scores are highlighted in boldface.

Robust Methods	Nat	Robust Test Acc (%)		
		Final	Best	Diff
		CIFAR-10		
PGD-AT	84.80 ± 0.14	45.16 ± 0.19	52.91 ± 0.11	7.75 ± 0.17
UDR-AT	83.87 ± 0.26	46.60 ± 0.27	$\mathbf{53.23 \pm 0.30}$	6.63 ± 0.57
HR	83.95 ± 0.32	47.32 ± 0.59	51.23 ± 0.25	3.90 ± 0.47
Ours	83.34 ± 0.16	$\mathbf{48.58 \pm 0.21}$	51.95 ± 0.19	$\mathbf{3.36 \pm 0.06}$
		CIFAR-100		
PGD-AT	$\mathbf{57.42 \pm 0.28}$	21.87 ± 0.20	$\mathbf{29.37 \pm 0.20}$	7.50 ± 0.35
UDR-AT	56.20 ± 0.54	22.07 ± 0.12	29.35 ± 0.02	7.28 ± 0.10
HR	56.69 ± 0.43	21.15 ± 0.20	28.35 ± 0.30	7.20 ± 0.49
Ours	56.71 ± 0.08	$\mathbf{23.09 \pm 0.20}$	28.95 ± 0.29	$\mathbf{5.86 \pm 0.50}$

robust test accuracy throughout the training process for different methods. The gap between these two metrics is the smallest in our method, indicating that our approach mitigates robust overfitting more effectively. Furthermore, our method remains consistently effective across different attack budgets as shown in Table 4 in the Appendix A.6.

As another popular adversarial defense strategy, TRADES (Zhang et al., 2019) also exhibits overfitting tendencies, albeit to a lesser extent compared to PGD-AT. Our proposed approach effectively mitigates overfitting when applied to UDR-TRADES, as illustrated in Table 5 in Appendix A.6.

Results on WRN28-10. We would like to provide further experimental results on the CIFAR-10 dataset with WRN28-10 as shown in Table 2. It can be observed that our method achieves the best performance in mitigating adversarial overfitting, as evidenced by the lowest robustness gap (Diff). Furthermore, it achieves the highest robust accuracy, demonstrating the effectiveness of our approach in enhancing robustness while maintaining strong performance on large models like WideResNet28-10.

Table 2: Robust performance of different robust methods using WideResNet28-10 on CIFAR-10 for l_∞ with budget 8/255.

Robust Methods	Nat	Robust Test Acc (%)		
		Final	Best	Diff
		CIFAR-10		
PGD-AT	86.51 ± 0.16	48.81 ± 0.49	55.64 ± 0.07	6.83 ± 0.55
UDR-AT	85.99 ± 0.15	49.01 ± 0.13	$\mathbf{55.82 \pm 0.19}$	6.81 ± 0.31
HR	84.89 ± 0.16	48.45 ± 0.31	52.45 ± 0.27	4.00 ± 0.34
Ours	84.52 ± 0.27	$\mathbf{51.26 \pm 0.42}$	53.87 ± 0.06	$\mathbf{2.61 \pm 0.47}$

Robustness for smaller ε. Theorem 5 shows that to ensure robustness on the test set, it is generally necessary to employ a larger attack budget during training compared to testing. In this experiment, we examine the robustness generalization by attacking different robust methods with smaller attack budget than the training attack budget while keeping other parameters of PGD attack the same. The results of this experiment are shown in Table 3. Our method consistently improves robustness for a smaller test adversarial budget. This empirical observation corroborates the findings of Theorem 5, indicating that our approach provides superior generalization guarantees for test set robustness when the training attack budget marginally exceeds that used during testing.

Table 3: Robustness evaluation on CIFAR-10 using ResNet-18 for l_∞ under different attack budget ε. The adversarial training budget is set to be 8/255 consistently. We use stronger attacks such as PGD-200 and Auto-Attack.

ε	8/255		6/255		4/255	
	PGD-200	AA	PGD-200	AA	PGD-200	AA
PGD-AT	42.86 ± 0.27	41.62 ± 0.25	54.14 ± 0.18	53.30 ± 0.19	65.76 ± 0.03	65.18 ± 0.13
UDR-AT	44.59 ± 0.27	42.81 ± 0.24	55.29 ± 0.18	53.80 ± 0.06	66.04 ± 0.34	64.92 ± 0.24
HR	45.27 ± 0.52	41.99 ± 0.38	56.66 ± 0.30	53.70 ± 0.24	67.38 ± 0.16	65.09 ± 0.24
Ours	$\mathbf{46.79 \pm 0.11}$	$\mathbf{44.06 \pm 0.32}$	$\mathbf{57.57 \pm 0.53}$	$\mathbf{55.09 \pm 0.46}$	$\mathbf{67.56 \pm 0.34}$	$\mathbf{65.76 \pm 0.31}$

Different choice of γ. Figure 3 illustrates the effect of varying values of statistical error γ in our framework on mitigating robust overfitting. We observe that larger γ values demonstrate reduced overfitting tendencies, particularly evident in the loss curves. However, excessively large γ leads to a decrease in natural test accuracy, which drops from $84.8\%(\gamma = 0)$ to $80.4\%(\gamma = 0.2)$. Consequently, we default to $\gamma = 0.1$ as a trade-off between robust overfitting mitigation and natural test accuracy preservation.

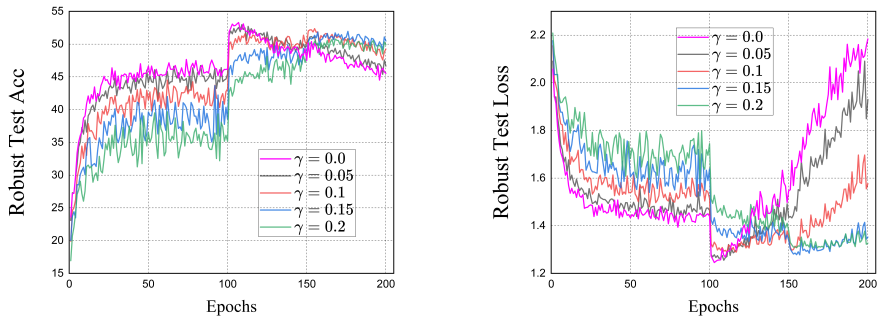

Figure 3: Impact of statistical error γ on mitigating overfitting in our method. Experiments conducted on CIFAR10 with $\varepsilon = 8/255$. Left: Robust test accuracy. Right: Robust test loss.

7 Conclusions

In this paper, we introduce SR-WDRO, a novel approach to address the challenge of robust overfitting in WDRO. Our method combines Kullback-Leibler divergence and Wasserstein distance to create a new ambiguity set, providing theoretical guarantees that adversarial test loss can be upper bounded by the statistically robust training loss with high probability and establishing conditions for Stackelberg and Nash equilibria between the learner and adversary. We developed a practical training algorithm based on this framework, which maintains computational efficiency comparable to standard adversarial training methods. Extensive experiments on benchmark datasets demonstrated that SR-WDRO effectively mitigates robust overfitting and achieves superior adversarial robustness compared to existing approaches. Our work contributes both theoretical insights and practical advancements to the field of robust machine learning. The proposed method offers a promising direction for developing more reliable and robust models in the face of distributional shifts.

Limitations and future works. Our SR-WDRO mitigates robust overfitting and improves robust accuracy compared to existing methods, although compromises with some natural accuracy drop and slight computational overhead. Due to the intractability of Equations (4) and (8), a better approximation is welcomed to solve SR-WDRO to mitigate the compromise of accuracy and computational cost. Furthermore, we mainly focus on supervised learning in this paper, expanding SR-WDRO to broader tasks including unsupervised learning, regressive tasks is a promising direction.

ACKNOWLEDGMENTS

This work is supported by NSFC grant No.12288201, grant GJ0090202, and NKRDP grant No.2018YFA0704705. The authors thank anonymous referees for their valuable comments.

REFERENCES

Yang An and Rui Gao. Generalization bounds for (wasserstein) robust optimization. *Advances in Neural Information Processing Systems*, 34:10382–10392, 2021.

Maksym Andriushchenko and Nicolas Flammarion. Understanding and improving fast adversarial training. *Advances in Neural Information Processing Systems*, 33:16048–16059, 2020.

Waïss Azizian, Franck Iutzeler, and Jérôme Malick. Exact generalization guarantees for (regularized) wasserstein distributionally robust models. *Advances in Neural Information Processing Systems*, 36, 2024.

Amine Bennouna and Bart Van Parys. Holistic robust data-driven decisions. *arXiv preprint arXiv:2207.09560*, 2022.

Amine Bennouna, Ryan Lucas, and Bart Van Parys. Certified robust neural networks: Generalization and corruption resistance. In *International Conference on Machine Learning*, pp. 2092–2112. PMLR, 2023.

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

Anh Tuan Bui, Trung Le, Quan Hung Tran, He Zhao, and Dinh Phung. A unified wasserstein distributional robustness framework for adversarial training. In *International Conference on Learning Representations*, 2022.

Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations*, 2020.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, pp. 2206–2216. PMLR, 2020.

Amir Dembo. *Large deviations techniques and applications*. Springer, 2009.

Yinpeng Dong, Ke Xu, Xiao Yang, Tianyu Pang, Zhijie Deng, Hang Su, and Jun Zhu. Exploring memorization in adversarial training. In *International Conference on Learning Representations*, 2022.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):12140, 2017.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Peter J. Huber. *Robust Statistics*. John Wiley & Sons, New York, 1981.

Daniel Kuhn, Peyman Mohajerin Esfahani, Viet Anh Nguyen, and Soroosh Shafieezadeh-Abadeh. 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.

Henry Lam. Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization. *Operations Research*, 67(4):1090–1105, 2019.

Jaeho Lee and Maxim Raginsky. Minimax statistical learning with wasserstein distances. *Advances in Neural Information Processing Systems*, 31, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Andrey Malinin, Neil Band, German Chesnokov, Yarin Gal, Mark JF Gales, Alexey Noskov, Andrey Ploskonosov, Liudmila Prokhorenkova, Ivan Provilkov, Vatsal Raina, et al. Shifts: A dataset of real distributional shift across multiple large-scale tasks. *arXiv preprint arXiv:2107.07455*, 2021.

Andrey Malinin, Andreas Athanasopoulos, Muhamed Barakovic, Meritxell Bach Cuadra, Mark JF Gales, Cristina Granziera, Mara Graziani, Nikolay Kartashev, Konstantinos Kyriakopoulos, Po-Jui Lu, et al. Shifts 2.0: Extending the dataset of real distributional shifts. *arXiv preprint arXiv:2206.15407*, 2022.

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

Hoang Phan, Trung Le, Trung Phung, Anh Tuan Bui, Nhat Ho, and Dinh Phung. Global-local regularization via distributional robustness. In *International Conference on Artificial Intelligence and Statistics*, pp. 7644–7664. PMLR, 2023.

Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021.

Yu V Prokhorov. Convergence of random processes and limit theorems in probability theory. *Theory of Probability & Its Applications*, 1(2):157–214, 1956.

Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pp. 8093–8104. PMLR, 2020.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *Advances in Neural Information Processing Systems*, 31, 2018.

Soroosh Shafieezadeh-Abadeh, Liviu Aolaritei, Florian Dörfler, and Daniel Kuhn. New perspectives on regularization and computation in optimal transport-based distributionally robust optimization. *arXiv preprint arXiv:2303.03900*, 2023.

Aman Sinha, Hongseok Namkoong, Riccardo Volpi, and John Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

Maurice Sion. On general minimax theorems. *Pacific J. Math.*, 8(1):171–176, 1958.

Matthew Staib and Stefanie Jegelka. Distributionally robust deep learning as a generalization of adversarial training. In *NIPS workshop on Machine Learning and Computer Security*, volume 3, pp. 4, 2017.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Bart PG Van Parys, Peyman Mohajerin Esfahani, and Daniel Kuhn. From data to decisions: Distributionally robust optimization is optimal. *Management Science*, 67(6):3387–3402, 2021.

Yifei Wang, Liangchen Li, Jiansheng Yang, Zhouchen Lin, and Yisen Wang. Balance, imbalance, and rebalance: Understanding robust overfitting from a minimax game perspective. *Advances in Neural Information Processing Systems*, 36, 2024.

Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2019.

Eric Wong, Frank Schmidt, and Zico Kolter. Wasserstein adversarial examples via projected Sinkhorn iterations. In *International Conference on Machine Learning*, volume 97, pp. 6808–6817. PMLR, 2019.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2958–2969, 2020.

Chaojian Yu, Bo Han, Li Shen, Jun Yu, Chen Gong, Mingming Gong, and Tongliang Liu. Understanding robust overfitting of adversarial training and beyond. In *International Conference on Machine Learning*, pp. 25595–25610. PMLR, 2022.

Man-Chung Yue, Daniel Kuhn, and Wolfram Wiesemann. On linear optimization over wasserstein balls. *Mathematical Programming*, 195(1):1107–1122, 2022.

Runtian Zhai, Tianle Cai, Di He, Chen Dan, Kun He, John Hopcroft, and Liwei Wang. Adversarially robust generalization just requires more unlabeled data. *arXiv preprint arXiv:1906.00555*, 2019.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pp. 7472–7482. PMLR, 2019.

Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan Kankanhalli. Geometry-aware instance-reweighted adversarial training. *arXiv preprint arXiv:2010.01736*, 2020.

A APPENDIX

A.1 RELATIONSHIP BETWEEN DIFFERENT DISCREPANCIES

We first introduce Levy-Prokhorov and total variation metrics. Then we will establish relationships among the four metrics: the Wasserstein metric, Levy-Prokhorov metric, total variation metric, and KL divergence.

Definition 13. *The Levy-Prokhorov metric between $\mu, \nu \in \mathcal{P}(\mathcal{Z})$ is:*

$$\mathrm{LP}(\mu, \nu) := \inf \{\tau > 0 \mid \mu(A) \leq \nu(A^\tau) + \tau \quad \forall A \in \mathcal{B}(\mathcal{Z})\}$$

where $A^\tau = \{z \in \mathcal{Z} : \inf_{z' \in A} \mathrm{d}(z, z') \leq \tau\}$.

Definition 14. *The total variation distance between any $\mu, \nu \in \mathcal{P}(\mathcal{Z})$ is:*

$$\mathrm{TV}(\mu, \nu) = \inf \{\tau > 0 \mid \mu(A) \leq \nu(A) + \tau \quad \forall A \in \mathcal{B}(\mathcal{Z})\}.$$

The KL divergence (relative entropy) has some basic properties: (1) $\mathrm{KL}(\mu, \nu) \geq 0$, which equal to 0 if and only if $\mu = \nu$; (2) KL is not symmetric and does not satisfy the triangle inequality.

Here, we give some relationships between these discrepancies.

Proposition 15. *For any $\mu, \nu \in \mathcal{P}(\mathcal{Z})$, the Wasserstein metric and Levy-Prokhorov metric satisfy:*

$$\mathrm{LP}(\mu, \nu)^{\frac{p+1}{p}} \leq \mathrm{W}_p(\mu, \nu) \leq (\mathrm{diam}(\mathcal{Z}) + 1)\mathrm{LP}(\mu, \nu)^{\frac{1}{p}}.$$

Proof. For any joint distribution π on random variables X, Y,

$$\begin{aligned}\mathbb{E}_\pi[\mathrm{d}^p(X, Y)] &\leq \tau^p \cdot \mathbb{P}(\mathrm{d}(X, Y) \leq \tau) + \mathrm{diam}(\mathcal{Z})^p \cdot \mathbb{P}(\mathrm{d}(X, Y) > \tau) \\ &\leq \tau^p + (\mathrm{diam}(\mathcal{Z})^p - \tau^p) \cdot \mathbb{P}(\mathrm{d}(X, Y) > \tau).\end{aligned}$$

If $\mathrm{LP}(\mu, \nu) \leq \tau (\tau \in [0, 1])$, we can choose a coupling of μ and ν so that $\mathbb{P}(\mathrm{d}(X, Y) > \tau)$ is upper bounded by τ (Huber, 1981). Then we have:

$$\mathbb{E}_\pi[\mathrm{d}^p(X, Y)] \leq \tau^p + (\mathrm{diam}(\mathcal{Z})^p - \tau^p)\tau \leq \tau(\mathrm{diam}(\mathcal{Z})^p + 1).$$

Take the infimum of both sides over coupling, we have:

$$\mathrm{W}_p(\mu, \nu) = \inf_\pi (\mathbb{E}_\pi(\mathrm{d}^p(X, Y))^{\frac{1}{p}} \leq (\tau(\mathrm{diam}(\mathcal{Z})^p + 1))^{\frac{1}{p}} \leq \tau^{\frac{1}{p}}(\mathrm{diam}(\mathcal{Z}) + 1).$$

In order to bound Levy-Prokhorov metric by Wasserstein metric, choose τ such that $\mathrm{W}_p(\mu, \nu) = \tau^{\frac{p+1}{p}}$, and use Markov's inequality. We have

$$\mathbb{P}(\mathrm{d}(X, Y) > \tau) = \mathbb{P}(\mathrm{d}^p(X, Y) > \tau^p) \leq \frac{1}{\tau^p}\mathbb{E}_\pi[\mathrm{d}^p(X, Y)] \leq \tau$$

where π is any joint distribution on $X \times Y$. Then by Strassen's theorem (Huber, 1981, Theorem 3.7), $\mathbb{P}(\mathrm{d}(X, Y) > \tau) \leq \tau$ is equivalent to $\mu(A) \leq \nu(A^\tau) + \tau$ for all Borel set $A \in \mathbb{B}(\mathcal{Z})$, which means $\mathrm{LP}(\mu, \nu) \leq \tau$, thus

$$\mathrm{LP}(\mu, \nu)^{\frac{p+1}{p}} \leq \mathrm{W}_p(\mu, \nu).$$

□

Proposition 16. *For any $\mu, \nu \in \mathcal{P}(\mathcal{Z})$, the Wasserstein metric, total variation metric and KL divergence satisfy:*

$$\mathrm{W}_p(\mu, \nu) \leq \mathrm{diam}(\mathcal{Z}) \cdot \mathrm{TV}(\mu, \nu)^{\frac{1}{p}} \leq \mathrm{diam}(\mathcal{Z}) \cdot \left(\frac{\mathrm{KL}(\mu, \nu)}{2}\right)^{\frac{1}{2p}}.$$

Proof. The total variation distance is equivalent to $W_1(\mu, \nu)$ with the optimal transport cost $\mathrm{d}(z, z') = \mathbb{I}(z \neq z')$ by definition, where $\mathbb{I}$ is the indicator function. As $\mathrm{d}(z, z') \leq \mathrm{diam}(\mathcal{Z}) \cdot \mathbb{I}(z \neq$

z'), we have:

$$\begin{aligned}
\mathrm{W}_p(\mu,\nu) &= \inf_{\pi\in\Pi(\mu,\nu)} \left\{\mathbb{E}_{(z,z')\sim\pi}\left[\mathrm{d}^p(z,z')\right]\right\}^{\frac{1}{p}} \\
&\leq \inf_{\pi\in\Pi(\mu,\nu)} \left\{\mathbb{E}_{(z,z')\sim\pi}\left[\mathrm{diam}(\mathcal{Z})^p\cdot\mathbb{I}(z\neq z')\right]\right\}^{\frac{1}{p}} \\
&= \mathrm{diam}(\mathcal{Z}) \inf_{\pi\in\Pi\mu,\nu)} \left\{\mathbb{E}_{(z,z')\sim\pi}\left[\mathbb{I}(z,z')\right]\right\}^{\frac{1}{p}} \\
&= \mathrm{diam}(\mathcal{Z})\cdot \mathrm{TV}(\mu,\nu)^{\frac{1}{p}}.
\end{aligned}$$

And the second part is by Pinsker's inequality that $\mathrm{TV}(\mu,\nu)\leq\sqrt{\frac{1}{2}\mathrm{KL}(\mu,\nu)}$. □

A.2 Proofs of Theorems 3 and 5

In this subsection, we prove formally Theorem 3 and Theorem 5. The goal is to show that our SR-WDRO training loss is an upper bound on the test loss with high probability.

Lemma 17. *Let $\mathcal{D}_n$ be the observed empirical distribution of n independent samples with true distribution $\mathcal{D}$ on a compact space $\mathcal{Z}$. Then for all $\varepsilon>0$, let $\delta=(\frac{\varepsilon}{\mathrm{diam}(\mathcal{Z})+1})^p$. We have*

$$\mathbb{P}\Big(\exists\mathcal{D}'\in\mathcal{P}(\mathcal{Z}),\ \mathrm{W}_p(\mathcal{D}_n,\mathcal{D}')\leq\varepsilon,\ \mathrm{KL}(\mathcal{D}',\mathcal{D})\leq\gamma\Big)\geq 1-e^{-\gamma n}\left(\frac{4}{\delta}\right)^{m(\mathcal{Z},\delta)}$$

where $m(\mathcal{Z},\delta):=\min\{k\geq 0:\exists\xi_1,\cdots,\xi_k\in\mathcal{Z},\ \text{s.t.}\ \cup_{i=1}^k\mathbb{B}(\xi_i,\delta)\supseteq\mathcal{Z}\}$ denote the internal covering number of the support set $\mathcal{Z}$.

Proof. We have:

$$\begin{aligned}
&\mathbb{P}\Big(\exists D'\in\mathcal{P}(\mathcal{Z}),\ \mathrm{W}_p\left(\mathcal{D}_n,\mathcal{D}'\right)\leq\varepsilon,\ \mathrm{KL}\left(\mathcal{D}',\mathcal{D}\right)\leq\gamma\Big) \\
&=\mathbb{P}\left(\mathcal{D}_n\in\left\{\hat{\mathcal{D}}\in\mathcal{P}(\mathcal{Z}):\exists\mathcal{D}'\in\mathcal{P}(\mathcal{Z})\text{ s.t. }\mathrm{W}_p(\hat{\mathcal{D}},\mathcal{D}')\leq\varepsilon,\ \mathrm{KL}(\mathcal{D}',\mathcal{D})\leq\gamma\right\}\right) \\
&=1-\mathbb{P}\left(\mathcal{D}_n\in\mathcal{A}\right)
\end{aligned}$$

where $\mathcal{A}$ is defined as: $\mathcal{A}^c=\left\{\hat{\mathcal{D}}\in\mathcal{P}(\mathcal{Z}):\exists\mathcal{D}'\in\mathcal{P}(\mathcal{Z})\text{ s.t. }\mathrm{W}_p(\hat{\mathcal{D}},\mathcal{D}')\leq\varepsilon,\ \mathrm{KL}\left(\mathcal{D}',\mathcal{D}\right)\leq\gamma\right\}$.

Let $\mathcal{A}^\delta=\{\mathcal{D}'\in\mathcal{P}(\mathcal{Z}):\mathrm{LP}\left(D',D''\right)\leq\delta,D''\in\mathcal{A}\}$ is the δ-inflation of the set $\mathcal{A}$ with respect to the LP metric. Denote $\mathbb{B}_{\mathrm{LP}}\left(\mathcal{D}'\right)=\{D''\in\mathcal{P}(\mathcal{Z}):\mathrm{LP}\left(D',D''\right)\leq\delta\}$ the LP ball, which is compact as LP is continuous in the weak topology and $\mathcal{Z}$ is compact (Prokhorov, 1956).

Dembo (2009) have established that for any set $\mathcal{A}\subset\mathcal{P}(\mathcal{Z})$ and $\delta>0$ we have for all $n\geq 1$ the upper bound:

$$\mathbb{P}\left(\mathcal{D}_n\in\mathcal{A}\right)\leq m_{\mathrm{LP}}(\mathcal{A},\delta)\exp\left(-n\inf_{\mathcal{D}'\in\mathcal{A}^\delta}\mathrm{KL}\left(\mathcal{D}',\mathcal{D}\right)\right)$$

where $m_{\mathrm{LP}}(\mathcal{A},\delta)=\min\{k\geq 0:\exists\mathcal{D}_1\cdots\mathcal{D}_k\in\mathcal{A}\text{ s.t.}\bigsqcup_{i=1}^k\mathbb{B}_{\mathrm{LP}}\left(\mathcal{D}_i,\delta\right)\supset\mathcal{A}\}$ is the internal covering number of set $\mathcal{A}$ with LP balls of radius δ. Furthermore, we can upper bound the internal covering number of any set $\mathcal{A}$ in terms of the internal covering number of the compact event $\mathcal{Z}$ as $m_{\mathrm{LP}}(\mathcal{A},\delta)\leq m_{\mathrm{LP}}(\mathcal{P}(\mathcal{Z}),\delta)\leq(4/\delta)^{m(\mathcal{Z},\delta)}$ for all $\delta>0$ (Dembo, 2009). Hence, we have:

$$\mathbb{P}\left(\mathcal{D}_n\in\mathcal{A}\right)\leq(4/\delta)^{m(\mathcal{Z},\delta)}\cdot\exp\left(-n\min_{\mathcal{D}'\in\mathcal{A}^\delta}\mathrm{KL}\left(\mathcal{D}',\mathcal{D}\right)\right).$$

The final inequality immediately leads to conclusion by remarking that $\mathcal{D}'\in\mathcal{A}^\delta\Rightarrow\mathrm{KL}\left(\mathcal{D}',\mathcal{D}\right)>\gamma$. Otherwise, suppose that there exists $\mathcal{D}''\in\mathcal{A}^\delta$ with $\mathrm{KL}\left(\mathcal{D}'',\mathcal{D}\right)\leq\gamma$. By definition of $\mathcal{A}^\delta$, there exists $\mathcal{D}'\in\mathcal{A}$ such that $\mathrm{LP}(\mathcal{D}'',\mathcal{D}')=\mathrm{LP}(\mathcal{D}',\mathcal{D}'')\leq\delta$. And by Proposition 15, we have $W_p\left(\mathcal{D}',\mathcal{D}''\right)\leq(\mathrm{diam}(\mathcal{Z})+1)\mathrm{LP}(\mathcal{D}',\mathcal{D}'')^{\frac{1}{p}}=\varepsilon$, so we have both $W_p\left(\mathcal{D}',\mathcal{D}''\right)\leq\varepsilon$ and $\mathrm{KL}\left(\mathcal{D}'',\mathcal{D}\right)\leq\gamma$, which implies that $\mathcal{D}'\in\mathcal{A}^c$, it is a contradiction. □

Proof of Theorem 3 We restate the theorem below. Let $\mathcal{D}_n$ be the observed empirical distribution which may be corrupted, and $\mathcal{D}$ be the true data distribution. Then for all $\varepsilon > 0$, let $\delta = (\frac{\varepsilon}{\mathrm{diam}(\mathcal{Z})+1})^p$, we have

$$\mathbb{P}\Big(\forall\theta \in \Theta, \mathcal{L}_{\varepsilon,\gamma}(\theta) \geq \mathbb{E}_{\mathcal{D}}[L(\theta, z)]\Big) \geq 1 - e^{-\gamma n}\left(\frac{4}{\delta}\right)^{m(\mathcal{Z},\delta)} \tag{10}$$

where $m(\mathcal{Z}, \delta) := \min\{k \geq 0 : \exists\xi_1, \cdots, \xi_k \in \mathcal{Z}, \text{ s.t. } \cup_{i=1}^k \mathbb{B}(\xi_i, \delta) \supseteq \mathcal{Z}\}$ denote the internal covering number of the support set $\mathcal{Z}$.

Proof. Recall that the ambiguity set is $\mathcal{U}(\mathcal{D}_n) := \{\mathcal{D}' : \exists\mathcal{D}'' \in \mathcal{P}(\mathcal{Z}) \text{ s.t. } \mathrm{W}_p(\mathcal{D}_n, \mathcal{D}'') \leq \varepsilon,\ \mathrm{KL}(\mathcal{D}'', \mathcal{D}') \leq \gamma\}$. By Lemma 17 we can ensure the ambiguity set $\mathcal{U}(\mathcal{D}_n)$ contains the true distribution $\mathcal{D}$ with high probability:

$$\mathbb{P}\left(\mathcal{D} \in \mathcal{U}(\mathcal{D}_n)\right) \geq 1 - e^{-\gamma n}\left(\frac{4}{\delta}\right)^{m(\mathcal{Z},\delta)}.$$

It follows that

$$\mathbb{P}\left(\sup_{\mathcal{D}' \in \mathcal{U}(\mathcal{D}_n)} \mathbb{E}_{\mathcal{D}'}[L(\theta, z)] \geq \mathbb{E}_{\mathcal{D}}[L(\theta, z)], \forall\theta \in \Theta\right) \geq 1 - e^{-\gamma n}\left(\frac{4}{\delta}\right)^{m(\mathcal{Z},\delta)}.$$

□

To facilitate our analysis on test adversarial distribution, we introduce the Levy-Prokhorov (LP) metric given by Bennouna et al. (2023), LP_ε is defined as:

$$\mathrm{LP}_\varepsilon\left(\mathcal{D}, \mathcal{D}'\right) := \inf\left\{\int 1\left(\mathrm{d}(z, z') > \varepsilon\right) \mathrm{d}\pi\left(z, z'\right) : \pi \in \Pi\left(\mathcal{D}, \mathcal{D}'\right)\right\}. \tag{11}$$

Proof of the Theorem 5 We restate the theorem below. Let $\mathcal{D}$ be the true data distribution, $\mathcal{D}_n$ the empirical distribution sampled i.i.d. from $\mathcal{D}$, and $\delta = (\frac{\sigma}{\mathrm{diam}(\mathcal{Z})+1})^p$ for any $\sigma > 0$. Then

$$\mathbb{P}\Big(\forall\theta \in \Theta, \mathbb{E}_{\mathcal{D}}[\max_{z' \in \mathbb{B}(z;\epsilon)} L(\theta, z')] \leq \mathcal{L}_{\varepsilon+\sigma,\gamma}(\theta, \mathcal{D}_n)\Big) \geq 1 - e^{-n\gamma}\left(\frac{4}{\delta}\right)^{m(\mathcal{Z},\delta)} \tag{12}$$

where $m(\mathcal{Z}, \delta) := \min\{k \geq 0 : \exists\xi_1, \cdots, \xi_k \in \mathcal{Z}, \text{ s.t. } \cup_{i=1}^k \mathbb{B}(\xi_i, \delta) \supseteq \mathcal{Z}\}$ denote the internal covering number of the support set $\mathcal{Z}$.

Proof. We consider the adversarial test distribution $\mathcal{M}_\#\mathcal{D}$, where $\mathcal{M}$ maps every input z to $\arg\max_{z' \in \mathbb{B}(z;\epsilon)} L(\theta, z')$. It is clear that $\mathcal{M}_\#\mathcal{D} \in \{\mathcal{D}' \in \mathcal{P}(\mathcal{Z}) : \mathrm{LP}_\varepsilon(\mathcal{D}, \mathcal{D}') \leq 0\}$. According to Lemma 17, we have that for $\delta = (\frac{\sigma}{\mathrm{diam}(\mathcal{Z})+1})^p$:

$$\mathbb{P}\Big(\exists\mathcal{D}' \in \mathcal{P}(\mathcal{Z}),\ \mathrm{W}_p(\mathcal{D}_n, \mathcal{D}') \leq \sigma,\ \mathrm{KL}(\mathcal{D}', \mathcal{D}) \leq \gamma\Big) \geq 1 - e^{-\gamma n}\left(\frac{4}{\delta}\right)^{m(\mathcal{Z},\delta)}.$$

Let us denote here $\mathcal{U}'_{\sigma,\gamma,\varepsilon}(\mathcal{D}_n) := \{\mathcal{D}'_{\text{test}} : \exists\mathcal{D}', \mathcal{D}'' \in \mathcal{P}(\mathcal{Z}) \text{ s.t. } \mathrm{W}_p(\mathcal{D}_n, \mathcal{D}') \leq \sigma,\ \mathrm{KL}(\mathcal{D}', \mathcal{D}'') \leq \gamma,\ \mathrm{LP}_\varepsilon(\mathcal{D}'', \mathcal{D}'_{\text{test}}) \leq 0\}$, by the definition of $\mathcal{M}_\#\mathcal{D}$, we have

$$\mathbb{P}\left(\mathcal{M}_\#\mathcal{D} \in \mathcal{U}'_{\sigma,r,\varepsilon}(\mathcal{D}_n)\right) \geq 1 - e^{-\gamma n}\left(\frac{4}{\delta}\right)^{m(\mathcal{Z},\delta)}.$$

It follows immediately that:

$$\mathbb{P}\left(\sup_{\mathcal{D}'_{\text{test}} \in \mathcal{U}'_{\sigma,\gamma,\varepsilon}(\mathcal{D}_n)} \mathbb{E}_{\mathcal{D}'_{\text{test}}}[L(\theta, z)] \geq \mathbb{E}_{\mathcal{M}_\#\mathcal{D}}[L(\theta, z)], \forall\theta \in \Theta\right) \geq 1 - e^{-\gamma n}\left(\frac{4}{\delta}\right)^{m(\mathcal{Z},\delta)}.$$

And we have

$$\mathbb{E}_{\mathcal{M}_\#\mathcal{D}} L(\theta, z) = \mathbb{E}_{\mathcal{D}}[\max_{z' \in \mathbb{B}(z;\epsilon)} L(\theta, z')].$$

It remains to show that $\sup_{\mathcal{D}'_{\text{test}} \in \mathcal{U}'_{\sigma,\gamma,\varepsilon}(\mathcal{D}_n)} \mathbb{E}_{\mathcal{D}'_{\text{test}}}[L(\theta, z)] \leq \mathcal{L}_{\varepsilon+\sigma,\gamma}(\theta, \mathcal{D}_n)$ for all θ. We have

$$
\begin{aligned}
&\sup_{\mathcal{D}'_{\text{test}} \in \mathcal{U}'_{\sigma,\gamma,\varepsilon}(\mathcal{D}_n)} \mathbb{E}_{\mathcal{D}'_{\text{test}}}[L(\theta, z)] \\
&= \sup_{\mathcal{D}' : \mathrm{W}_p(\mathcal{D}_n, \mathcal{D}') \leq \sigma} \sup\{\mathbb{E}_{\mathcal{D}'_{\text{test}}}[L(\theta, z)] : \exists \mathcal{D}'', \mathcal{D}'_{\text{test}} \textbf{ s.t. } \mathrm{KL}(\mathcal{D}', \mathcal{D}'') \leq \gamma,\ \mathrm{LP}_\varepsilon(\mathcal{D}'', \mathcal{D}'_{\text{test}}) \leq 0\} \\
&\overset{(1)}{=} \sup_{\mathcal{D}' : \mathrm{W}_p(\mathcal{D}_n, \mathcal{D}') \leq \sigma} \sup\{\mathbb{E}_{\mathcal{D}'_{\text{test}}}[L(\theta, z)] : \exists \mathcal{D}'', \mathcal{D}'_{\text{test}} \textbf{ s.t. } \mathrm{LP}_\varepsilon(\mathcal{D}', \mathcal{D}'') \leq 0,\ \mathrm{KL}(\mathcal{D}'';, \mathcal{D}'_{\text{test}}) \leq \gamma\} \\
&= \sup\{\mathbb{E}_{\mathcal{D}'_{\text{test}}}[L(\theta, z)] : \exists \mathcal{D}', \mathcal{D}'' \textbf{ s.t. } \mathrm{W}_p(\mathcal{D}_n, \mathcal{D}') \leq \sigma, \mathrm{LP}_\varepsilon(\mathcal{D}', \mathcal{D}'') \leq 0,\ \mathrm{KL}(\mathcal{D}'';, \mathcal{D}'_{\text{test}}) \leq \gamma\} \\
&\overset{(2)}{\leq} \sup\{\mathbb{E}_{\mathcal{D}'_{\text{test}}}[L(\theta, z)] : \exists \mathcal{D}', \mathcal{D}'' \textbf{ s.t. } \mathrm{W}_p(\mathcal{D}_n, \mathcal{D}') \leq \sigma, \mathrm{W}_p(\mathcal{D}', \mathcal{D}'') \leq \varepsilon,\ \mathrm{KL}(\mathcal{D}'';, \mathcal{D}'_{\text{test}}) \leq \gamma\} \\
&\overset{(3)}{\leq} \sup\{\mathbb{E}_{\mathcal{D}'_{\text{test}}}[L(\theta, z)] : \exists \mathcal{D}' \textbf{ s.t. } \mathrm{W}_p(\mathcal{D}_n, \mathcal{D}') \leq \varepsilon + \sigma,\ \mathrm{KL}(\mathcal{D}', \mathcal{D}'_{\text{test}}) \leq \gamma\} \\
&\leq \mathcal{L}_{\varepsilon+\sigma,\gamma}(\theta, \mathcal{D}_n).
\end{aligned}
$$

The first equality follows from Lemma C.3 in Bennouna et al. (2023), the second inequality is clear as under LP_ε constraint we have $\mathrm{d}(z, z') \leq \varepsilon$ for any z, z' almost surely. The third inequality follows from the triangle inequality. □

A.3 General Distributional Shifts

Our framework is applicable to a variety of distributional shifts, not limited to the adversarial example distributions that we focus on. This includes other shifts such as those encountered in domain generalization. If the distance between the test distribution and the original data distribution remains within certain bounds, we can still provide generalization guarantees. Here we assume that

$$\mathcal{D}_{\text{test}} \in \{\mathcal{D}' \in \mathcal{P}(\mathcal{Z}) : \mathrm{LP}_\varepsilon(\mathcal{D}, \mathcal{D}') \leq 0\},$$

where $\varepsilon > 0$ is the adversarial budget and LP_ε is defined in equation 11. Here, we consider a more general scenario where potential shifts might include corruptions or other variations, resulting in discrepancies between test samples and true samples. This formulation allows us to address a broader range of distributional shifts beyond just adversarial examples. Similar to Theorem 5, we have:

Theorem 18. *Let $\mathcal{D}$ be the true data distribution, $\mathcal{D}_n$ the empirical distribution sampled i.i.d. from $\mathcal{D}$, $\mathcal{D}_{\text{test}}$ the test distribution as defined above, and $\delta = (\frac{\sigma}{\mathrm{diam}(\mathcal{Z})+1})^p$ for $\sigma > 0$. Then*

$$\mathbb{P}\Big(\forall \theta \in \Theta, \mathbb{E}_{\mathcal{D}_{\text{test}}}[L(\theta, z)] \leq \mathcal{L}_{\varepsilon+\sigma,\gamma}(\theta, \mathcal{D}_n)\Big) \geq 1 - e^{-n\gamma} \left(\frac{4}{\delta}\right)^{m(\mathcal{Z},\delta)} \tag{13}$$

where $m(\mathcal{Z}, \delta) := \min\{k \geq 0 : \exists \xi_1, \cdots, \xi_k \in \mathcal{Z},\ s.t.\ \cup_{i=1}^k \mathbb{B}(\xi_i, \delta) \supseteq \mathcal{Z}\}$ denote the internal covering number of the support set $\mathcal{Z}$.

A.4 Proofs for Section 4.3

In this subsection, we first give the definition of Nash equilibrium and Stackelberg equilibrium, then we provide the detailed proof of the existence of two game equilibrium.

Here we consider two player zero sum game, let the strategy spaces of player 1 and player 2 be X and Y, respectively. Their utility functions are denoted by $u_1(x, y)$ and $u_2(x, y)$, where $x \in X$ is the strategy of player 1, and $y \in Y$ is the strategy of player 2. In a zero-sum game, $u_1(x, y) = -u_2(x, y)$.

Definition 19 (Nash Equilibrium)**.** *A Nash equilibrium is a pair of strategies (x^*, y^*) such that neither player can improve their payoff by unilaterally changing their strategy, given the strategy of the other player.*

$$u_1(x^*, y^*) \geq u_1(x, y^*), \quad \forall x \in X; \quad u_1(x^*, y^*) \leq u_1(x^*, y), \quad \forall y \in Y.$$

In a Stackelberg game, one player (the leader) moves first, and the other player (the follower) observes the leader's action before selecting their own strategy. Let player 1 be the leader and player 2 be the follower.

Definition 20 (Stackelberg Equilibrium)**.** *A Stackelberg equilibrium is a pair of strategies $(x^*, y^*(x^*))$, where x^* is the leader's optimal strategy and $y^*(x^*)$ is the follower's best response given the leader's strategy.*

1. Follower's best response: $y^*(x) = \arg\max_{y \in Y} u_2(x, y)$

2. Leader's optimal strategy: $x^* = \arg\max_{x \in X} u_1(x, y^*(x))$.

Thus, the Stackelberg equilibrium consists of the leader's strategy x^*, which maximizes the leader's payoff, and the follower's response $y^*(x^*)$, which is the best response to x^*.

Proof of Proposition 9: (Restate) The ambiguity set $\mathcal{U}(\mathcal{D}_n) := \{\mathcal{D}' \in \mathcal{P}(\mathcal{Z}) : \exists \mathcal{D}'' \in \mathcal{P}(\mathcal{Z}) \text{ s.t. } \mathrm{W}_p(\mathcal{D}_n, \mathcal{D}'') \leq \varepsilon,\ \mathrm{KL}(\mathcal{D}'', \mathcal{D}') \leq \gamma\}$ is compact in $\mathcal{P}(\mathcal{Z})$ with respect to weak topology. And if the loss function satisfies Assumption 8, then the supremum in (4) is finite and can be attained for any $\theta \in \Theta$.

Proof. Firstly, we prove $\mathcal{U}(\mathcal{D}_n)$ is closed. Assume that any sequence $\{\mathcal{D}^i\}_{i=1}^{\infty} \subset \mathcal{U}(\mathcal{D}_n)$ weakly converges to $\mathcal{D}^0$. We will show that $\mathcal{D}^0 \in \mathcal{U}(\mathcal{D}_n)$, which is equivalent to $\mathcal{U}(\mathcal{D}_n)$ is closed. By definition of $\mathcal{U}(\mathcal{D}_n)$, for each i, there exists $\widehat{\mathcal{D}^i}$ such that $\mathrm{W}_p(\mathcal{D}_n, \widehat{\mathcal{D}^i}) \leq \varepsilon,\ \mathrm{KL}(\widehat{\mathcal{D}^i}, \mathcal{D}^i) \leq \gamma$. We aim to show that such $\widehat{\mathcal{D}^0}$ also exist for the limit $\mathcal{D}^0$, preserving the Wasserstein and KL constraints. The set $\{\mathcal{D}'' \in \mathcal{P}(\mathcal{Z}) : \mathrm{W}_p(\mathcal{D}_n, \mathcal{D}'') \leq \varepsilon\}$ is compact in the weak topology of $\mathcal{P}(\mathcal{Z})$, as the Wasserstein ball is compact with respect to weak convergence. Therefore, from the sequence $\widehat{\mathcal{D}^i}$, we can extract a subsequence $\{\widehat{\mathcal{D}^{i_j}}\}_{j=1}^{\infty}$ that converges weakly to some distribution $\widehat{\mathcal{D}^0}$, and by the compactness of the Wasserstein ball, $\widehat{\mathcal{D}^0}$ satisfies: $\mathrm{W}_p(\mathcal{D}_n, \widehat{\mathcal{D}^0}) \leq \varepsilon$. Here, we abuse the notation and still denote the subsequence as $\{\widehat{\mathcal{D}^i}\}_{i=1}^{\infty}$.

Next, we use the lower semicontinuity property of the Kullback-Leibler (KL) divergence with respect to weak convergence. Since $\mathrm{KL}(\widehat{\mathcal{D}^i}, \mathcal{D}^i) \leq \gamma$ for all i, and $\widehat{\mathcal{D}^i} \to \widehat{\mathcal{D}^0}$ weakly and $\mathcal{D}^i \to \mathcal{D}^0$ weakly, we apply the lower semicontinuity of the KL divergence to conclude:

$$\mathrm{KL}(\widehat{\mathcal{D}}^0, \mathcal{D}^0) \leq \liminf_{i \to \infty} \mathrm{KL}(\widehat{\mathcal{D}}^i, \mathcal{D}^i) \leq \gamma.$$

Thus, the limit point $\mathcal{D}^0$ of the sequence $\{\mathcal{D}^i\}_{i=1}^{\infty}$ satisfies the conditions:

$$\exists \widehat{\mathcal{D}^0} \in \mathcal{P}(\mathcal{Z}) \text{ s.t. } \mathrm{W}_p(\mathcal{D}_n, \widehat{\mathcal{D}}^0) \leq \varepsilon \quad \text{and} \quad \mathrm{KL}(\widehat{\mathcal{D}}^0, \mathcal{D}^0) \leq \gamma,$$

which implies that $\mathcal{D}^0 \in \mathcal{U}(\mathcal{D}_n)$. Therefore, $\mathcal{U}(\mathcal{D}_n)$ is closed with respect to the weak topology.

Then we prove that $\mathcal{U}(\mathcal{D}_n)$ is subset of a larger Wasserstein ball, as any weak closed subset of a weakly compact set is weakly compact, we can prove $\mathcal{U}(\mathcal{D}_n)$ is weak compact. For any $\mathcal{D}' \in \mathcal{U}(\mathcal{D}_n)$, there exists $\mathcal{D}'' \in \mathcal{P}(\mathcal{Z})$ such that $\mathrm{W}_p(\mathcal{D}_n, \mathcal{D}'') \leq \varepsilon$ and $\mathrm{KL}(\mathcal{D}'', \mathcal{D}') \leq \gamma$. By Proposition 16, we have $\mathrm{W}_p(\mathcal{D}'', \mathcal{D}') \leq \mathrm{diam}(\mathcal{Z})(\gamma/2)^{\frac{1}{2p}}$, then we have that $\mathrm{W}_p(\mathcal{D}_n, \mathcal{D}'') \leq \varepsilon$ and $\mathrm{W}_p(\mathcal{D}'', \mathcal{D}') \leq \mathrm{diam}(\mathcal{Z})(\gamma/2)^{\frac{1}{2p}}$, thus $\mathrm{W}_p(\mathcal{D}_n, \mathcal{D}') \leq \varepsilon + \mathrm{diam}(\mathcal{Z})(\gamma/2)^{\frac{1}{2p}}$, which means that $\mathcal{U}(\mathcal{D}_n) \subset \mathbb{B}_{W_p}\left(\mathcal{D}_n,\ \varepsilon + \gamma \cdot \mathrm{diam}(\mathcal{Z})(\gamma/2)^{\frac{1}{2p}}\right)$.

Next, we show the supremum in (4) is finite. As $\mathcal{U}(\mathcal{D}_n) \subset \mathbb{B}_{W_p}\left(\mathcal{D}_n,\ \varepsilon + \gamma a \cdot d_{\min}^p(\mathcal{Z})\right)$, it is clear that:

$$\sup_{\mathcal{D}' \in \mathcal{U}(\mathcal{D}_n)} \mathbb{E}_{\mathcal{D}'}[L(\theta, z)] \leq \sup_{\mathcal{D}' \in \mathbb{B}_{W_p}\left(\mathcal{D}_n,\ \varepsilon + \gamma a \cdot d_{\min}^p(\mathcal{Z})\right)} \mathbb{E}_{\mathcal{D}'}[L(\theta, z)].$$

The supremum on the right side of the above inequality is finite by Yue et al. (2022, Theorem 2), which applies Assumption 8 (iii) and $\mathcal{D}_n$ has a finite p-th moment. Thus the supremum have a finite upper bound.

It remains to be shown that supremum in (4) can be attained. We know that the objective function $\sup_{\mathcal{D}' \in \mathcal{U}(\mathcal{D}_n)} \mathbb{E}_{\mathcal{D}'}[L(\theta, z)]$ is weak upper semi-continuous in $\mathcal{D}'$ over the ambiguity set $\mathcal{U}(\mathcal{D}_n)$. This follows immediately from the proof of Yue et al. (2022, Theorem 3), which reveals that $\sup_{\mathcal{D}' \in \mathcal{U}(\mathcal{D}_n)} \mathbb{E}_{\mathcal{D}'}[L(\theta, z)]$ is weak upper semi-continuous over $\mathbb{B}_{W_p}\left(\mathcal{D}_n,\ \varepsilon + \gamma a \cdot d_{\min}^p(\mathcal{Z})\right)$. As $\mathcal{U}(\mathcal{D}_n)$ is weak compact, Weirestrass's theorem the guarantees the supremum in (4) is indeed attained for any θ. $\square$

Proof of the Stackelberg Equilibrium (Theorem 10) (Restate) If the loss function satisfies Assumption 8 and we assume Θ is compact, the game exists a Stackelberg equilibrium. i.e. there exists

$$\theta^* \in \arg\inf_{\theta\in\Theta} \sup_{\mathcal{D}'\in\mathcal{U}(\mathcal{D}_n)} \mathbb{E}_{\mathcal{D}'}[L(\theta, z)], \quad \mathcal{D}'^*(\theta^*) \in \arg\sup_{\mathcal{D}'\in\mathcal{U}(\mathcal{D}_n)} \mathbb{E}_{\mathcal{D}'}[L(\theta^*, z)].$$

Proof. By Proposition 9, for any θ, the supremum in 4 can be attained, that is, there exists a best response $\mathcal{D}'^*(\theta) \in \arg\max_{\mathcal{D}'\in\mathcal{U}(\mathcal{D}_n)} \mathbb{E}_{\mathcal{D}'}[L(\theta^*, z)]$ for any θ, for simplicity we can assume $\mathcal{D}'^*(\theta)$ is unique since the supremum is unique. We consider the minimization problem: $\inf_{\theta\in\Theta} \mathbb{E}_{\mathcal{D}'^*(\theta)}[L(\theta, z)]$. Furthermore, the expected loss objective $\mathbb{E}_{Z\sim\mathcal{D}'}[L(\theta, z)]$ inherits lower-semicontinuity in θ from the loss function $L(\theta, z)$. Specifically, lower semi-continuity holds because

$$\liminf_{\theta_n\to\theta} \mathbb{E}_{Z\sim\mathcal{D}'}\left[L\left(\theta_n, Z\right)\right] \geq \mathbb{E}_{Z\sim\mathcal{D}'} \liminf_{\theta_n\to\theta}[L(\theta, z)] \geq \mathbb{E}_{Z\sim\mathcal{D}'}[L(\theta, z)],$$

then we have

$$\liminf_{\theta_n\to\theta} \mathbb{E}_{\mathcal{D}'^*(\theta)}\left[L\left(\theta_n, Z\right)\right] \geq \mathbb{E}_{\mathcal{D}'^*(\theta)}[L(\theta, z)]$$

and by definition of best response:

$$\liminf_{\theta_n\to\theta} \mathbb{E}_{\mathcal{D}'^*(\theta_n)}\left[L\left(\theta_n, Z\right)\right] \geq \liminf_{\theta_n\to\theta} \mathbb{E}_{\mathcal{D}'^*(\theta)}\left[L\left(\theta_n, Z\right)\right],$$

thus we have

$$\liminf_{\theta_n\to\theta} \mathbb{E}_{\mathcal{D}'^*(\theta_n)}\left[L\left(\theta_n, Z\right)\right] \geq \mathbb{E}_{\mathcal{D}'^*(\theta)}[L(\theta, z)],$$

which means the function $\inf_{\theta\in\Theta} \mathbb{E}_{\mathcal{D}'^*(\theta)}[L(\theta, z)]$ is lower-semi continuous on θ. Finally, since Θ is compact, the minimization problem $\inf_{\theta\in\Theta} \mathbb{E}_{\mathcal{D}'^*(\theta)}[L(\theta, z)]$ has a solution θ^*. □

Proof of the Minimax Theorem 11 (Restate) If the loss function satisfies Assumption 8 holds, Θ is convex, and $L(\theta, z)$ is convex in θ for any $z \in \mathcal{Z}$, then we have

$$\min_{\theta\in\Theta} \max_{\mathcal{D}'\in\mathcal{U}(\mathcal{D}_n)} \mathbb{E}_{\mathcal{D}'}[L(\theta, z)] = \max_{\mathcal{D}'\in\mathcal{U}(\mathcal{D}_n)} \min_{\theta\in\Theta} \mathbb{E}_{\mathcal{D}'}[L(\theta, z)]. \tag{14}$$

Proof. We will verify the conditions of Sion's minimax theorem (Sion, 1958). First, we show that the ambiguity set $\mathcal{U}(\mathcal{D}_n)$ is convex. Assume that $\mathcal{D}^1, \mathcal{D}^2 \in \mathcal{U}(\mathcal{D}_n)$. Then there exist $\widehat{\mathcal{D}}^i (i = 1, 2) \in \mathcal{P}(\mathcal{Z})$ such that $\mathrm{W}_p(\mathcal{D}_n, \widehat{\mathcal{D}^i}) \leq \varepsilon$, $\mathrm{KL}(\widehat{\mathcal{D}^i}, \mathcal{D}^i) \leq \gamma$. As KL-divergence is convex, for any $\lambda \in [0, 1]$, we have

$$\mathrm{KL}(\lambda\widehat{\mathcal{D}^1} + (1-\lambda)\widehat{\mathcal{D}^2}, \lambda\mathcal{D}^1 + (1-\lambda)\mathcal{D}^2) \leq \lambda\mathrm{KL}(\widehat{\mathcal{D}^1}, \mathcal{D}^1) + (1-\lambda)\mathrm{KL}(\widehat{\mathcal{D}^2}, \mathcal{D}^2) \leq \gamma.$$

And for Wasserstein we have

$$\mathrm{W}_p(\mathcal{D}_n, \lambda\widehat{\mathcal{D}^1} + (1-\lambda)\widehat{\mathcal{D}^2}) \leq \lambda\mathrm{W}_p(\mathcal{D}_n, \widehat{\mathcal{D}^1}) + (1-\lambda)\mathrm{W}_p(\mathcal{D}_n, \widehat{\mathcal{D}^2}) \leq \varepsilon.$$

Let $\mathcal{D}'' = \lambda\widehat{\mathcal{D}^1} + (1-\lambda)\widehat{\mathcal{D}^2}$, which means $\lambda\mathcal{D}^1 + (1-\lambda)\mathcal{D}^2) \in \mathcal{U}(\mathcal{D}_n)$. Now we have Θ is convex and $\mathcal{U}(\mathcal{D}_n)$ is convex and weak compact by proposition 9.

Furthermore, the expected loss objective $\mathbb{E}_{Z\sim\mathcal{D}'}[L(\theta, z)]$ inherits convexity and lower-semicontinuity in θ from the loss function $L(\theta, z)$. Specifically, lower semi-continuity holds because

$$\liminf_{\theta_n\to\theta} \mathbb{E}_{Z\sim\mathcal{D}'}\left[L\left(\theta_n, Z\right)\right] \geq \mathbb{E}_{Z\sim\mathcal{D}'} \liminf_{\theta_n\to\theta}[L(\theta, z)] \geq \mathbb{E}_{Z\sim\mathcal{D}'}[L(\theta, z)].$$

These inequalities leverage two key properties. The first inequality follows from Fatou's lemma for random variables with uniformly integrable negative parts, a property guaranteed by Assumption 8 (i) that loss function is non-negative. The second inequality exploits the lower semi-continuity of the loss function $L(\theta, \cdot)$ in θ, as established in Assumption 8 (ii).

Finally, it is readily verified that the objective function $\mathbb{E}_{Z\sim\mathcal{D}'}[L(\theta, z)]$ is concave (in fact, linear) and weakly upper semi-continuous in $\mathcal{D}'$. This follows directly from the proof of Proposition 9. This analysis establishes that all the conditions of Sion's minimax theorem are satisfied. Consequently, the infimum and supremum in Equation (7) can be interchanged.

It remains to show that both maxima in (7) are reached. However, this follows immediately from the weak compactness of $\mathcal{U}(\mathcal{D}_n)$ and the weak upper semi-continuity in $\mathcal{D}'$ of both the expected loss $\mathbb{E}_{Z\sim\mathcal{D}'}[L(\theta, z)]$ and the optimal expected loss $\inf_{\theta\in\Theta} \mathbb{E}_{Z\sim\mathcal{D}'}[L(\theta, z)]$. □

A.5 Proof of Proposition 12

In this subsection, we prove the dual formulation of the robust maximization problem.

Proof of Strong Duality (Proposition 12) (Restate) $\mathcal{L}_{\varepsilon,r}(\theta,\mathcal{D})$ admits the dual formulation for all $\gamma > 0$:

$$\mathcal{L}_{\varepsilon,\gamma}(\theta,\mathcal{D}) = \inf_{\substack{\lambda,\beta\geq 0\\ \eta\geq \max_{z\in\mathcal{Z}} L(\theta,z)}} \{\lambda\varepsilon^p + \mathbb{E}_{\mathcal{D}}[\varphi(\lambda,\beta,\eta,z)]\}$$

where $\varphi(\lambda,\beta,\eta,z) = \sup_{\xi\in\mathcal{Z}}\{\beta\log\frac{\beta}{\eta-L(\theta,\xi)} + (\gamma-1)\beta + \eta - \lambda \mathrm{d}^p(z,\xi)\}$.

Proof.

$$\begin{aligned}
&L_{\varepsilon,\gamma}(\theta,\mathcal{D})\\
&=: \sup\{\mathbb{E}_{\mathcal{D}'}[L(\theta,z)] : D', \mathcal{Q}\in\mathcal{P}(\mathcal{Z}), W_p(\mathcal{D},\mathcal{Q})\leq\varepsilon, \mathrm{KL}(\mathcal{Q},\mathcal{D}')\leq\gamma\}\\
&= \sup\{\sup\{\mathbb{E}_{\mathcal{D}'}[L(\theta,z)] : \mathcal{D}'\in\mathcal{P}(\mathcal{Z}), \mathrm{KL}(\mathcal{Q},\mathcal{D}')\leq\gamma\} : \mathcal{Q}\in\mathcal{P}(\mathcal{Z}), \mathrm{W}_p(\mathcal{D},\mathcal{Q})\leq\varepsilon\}\\
&\overset{(1)}{=} \sup_{\substack{\mathcal{Q}\in\mathcal{P}(\mathcal{Z}),\\ \mathrm{W}_p(\mathcal{D},\mathcal{Q})\leq\varepsilon}} \left\{ \inf_{\substack{\beta\geq 0\\ \eta\geq\max_{z\in\mathcal{Z}} L(\theta,z)}} \left\{ \int \beta\cdot\log\frac{\beta}{\eta-L(\theta,z)} d\mathcal{Q}(z) + (\gamma-1)\beta+\eta \right\}\right\}\\
&\overset{(2)}{=} \lim_{\tau\to 0_+} \sup_{\substack{\mathcal{Q}\in\mathcal{P}(\mathcal{Z}),\\ \mathrm{W}_p(\mathcal{D},\mathcal{Q})\leq\varepsilon}} \left\{ \inf_{\substack{\beta\geq 0\\ \eta\geq\max_{z\in\mathcal{Z}} L(\theta,z)+\tau}} \left\{ \int \beta\cdot\log\frac{\beta}{\eta-L(\theta,z)} d\mathcal{Q}(z) + (\gamma-1)\beta+\eta \right\}\right\}\\
&\overset{(3)}{=} \lim_{\tau\to 0_+} \left\{ \inf_{\substack{\beta\geq 0\\ \eta\geq\max_{z\in\mathcal{Z}} L(\theta,z)+\tau}} \left\{ \sup_{\substack{\mathcal{Q}\in\mathcal{P}(\mathcal{Z}),\\ \mathrm{W}_p(\mathcal{D},\mathcal{Q})\leq\varepsilon}} \int \beta\cdot\log\frac{\beta}{\eta-L(\theta,z)} d\mathcal{Q}(z) + (\gamma-1)\beta+\eta \right\}\right\}\\
&= \lim_{\tau\to 0_+} \left\{ \inf_{\substack{\beta\geq 0\\ \eta\geq\max_{z\in\mathcal{Z}} L(\theta,z)+\tau}} \left\{ \sup_{\substack{\mathcal{Q}\in\mathcal{P}(\mathcal{Z}),\\ \mathrm{W}_p(\mathcal{D},\mathcal{Q})\leq\varepsilon}} \mathbb{E}_{\mathcal{Q}}\left[\beta\cdot\log\frac{\beta}{\eta-L(\theta,z)} + (\gamma-1)\beta+\eta\right] \right\}\right\}\\
&\overset{(4)}{=} \lim_{\tau\to 0_+} \left\{ \inf_{\substack{\beta\geq 0\\ \eta\geq\max_{z\in\mathcal{Z}} L(\theta,z)+\tau}} \left\{ \inf_{\lambda\geq 0} \lambda\varepsilon^p + \mathbb{E}_{\mathcal{D}}[\varphi(\lambda,\beta,\eta,z)] \right\}\right\}\\
&= \inf_{\substack{\beta\geq 0,\lambda\geq 0\\ \eta\geq\max_{z\in\mathcal{Z}} L(\theta,z)}} \lambda\varepsilon^p + \mathbb{E}_{\mathcal{D}}[\varphi(\lambda,\beta,\eta,z)].
\end{aligned}$$

Here, the equality (1) follows from Lemma 21. The equality (2) follows from the fact that

$$\begin{aligned}
&\inf\left\{\int \beta\log\left(\frac{\beta}{\eta-L(\theta,z)}\right)\mathrm{d}\mathcal{Q}(z) + (\gamma-1)\beta+\eta : \beta\geq 0, \eta\geq\max_{z\in\mathcal{Z}} L(\theta,z)\right\}\\
&\leq \inf\left\{\int \beta\log\left(\frac{\beta}{\eta-L(\theta,z)}\right)\mathrm{d}\mathcal{Q}(z) + (\gamma-1)\beta+\eta : \beta\geq 0, \eta\geq\max_{z\in\mathcal{Z}} L(\theta,z)+\tau\right\}\\
&\leq \inf\left\{\int \beta\log\left(\frac{\beta}{\eta-L(\theta,z)}\right)\mathrm{d}\mathcal{Q}(z) + (\gamma-1)\beta+\eta : \beta\geq 0, \eta\geq\max_{z\in\mathcal{Z}} L(\theta,z)\right\}+\tau.
\end{aligned}$$

for any $\tau > 0$. The equality (3) follows from the minimax theorem of (Sion, 1958), and we refer to (Theorem 3.4's proof (Bennouna & Van Parys, 2022)) for the proof of this part. Finally, equality (4) follows from the strong duality of Wasserstein strong duality. □

Lemma 21. *(Van Parys et al., 2021) We have*

$$\sup_{\substack{\mathcal{Q}\in\mathcal{P}(\mathcal{Z}),\\ \mathrm{KL}(\mathcal{D}',\mathcal{Q})\leq\gamma}} \mathbb{E}_{\mathcal{D}'}\left[L(\theta,z)\right] = \inf_{\substack{\beta\geq 0,\\ \eta\geq\max_{\xi\in\mathcal{Z}} L(\theta,\xi)}} \left\{\int \beta\log\left(\frac{\beta}{\eta - L(\theta,\xi)}\right)\mathrm{d}\mathcal{D}'(\xi) + (\gamma-1)\beta + \eta\right\}$$

for all $\mathcal{D}' \in \mathcal{P}(\mathcal{Z})$ and $\gamma > 0$.

Lemma 22. *(Blanchet & Murthy, 2019) For all $\mathcal{D}'$ we have:*

$$\sup_{\substack{\mathcal{Q}\in\mathcal{P}(\mathcal{Z}),\\ \mathrm{W}_p(\mathcal{Q},\mathcal{D}')\leq\varepsilon}} \mathbb{E}_{\mathcal{Q}}\left[L(\theta,z)\right] = \inf_{\lambda\geq 0}\left\{\lambda\varepsilon^p + \mathbb{E}_{\mathcal{D}'}[\varphi(\lambda,z)]\right\}$$

where $\varphi(\lambda,z) := \sup_{\xi\in\mathcal{Z}}\{L(\theta,\xi) - \lambda \mathrm{d}^p(z,\xi)\}$.

A.6 Further details on experiments

We use the ResNet-18 for the CIFAR-10 and CIFAR-100 dataset. We use the SGD optimizer with momentum 0.9, weight decay 5e-4. The starting learning rate is 0.1 and reduce the learning rate (×0.1) at epoch $\{100, 150\}$. We train with 200 epochs.

Mitigating overfitting on more attack budgets. We would like to provide further experimental results on the CIFAR-10 dataset on more attack budgets (training and test attack budget is same) as shown in Table 4. The gap between these final robust accuracy and best robust accuracy is the smallest in our method, indicating that our approach most effectively mitigates overfitting even for different attack budgets.

Table 4: Robust performance of different methods on CIFAR-10 and CIFAR-100 using a ResNet-18 for l_∞ with budget 6/255, 10/255 and 12/255. We use PGD-10 as the attack to evaluate robust performance. Nat is the natural test accuracy. The "Best" robust test acc is the highest robust test accuracy achieved during training whereas the "Final" robust test acc is last epoch's robust accuracy.

Robust Methods	Nat	Robust Test Acc (%)		
		Final	Best	Diff
		$\varepsilon = 6/255$		
PGD-AT	**87.46 ± 0.17**	54.33 ± 0.12	60.41 ± 0.08	6.08 ± 0.18
UDR-AT	86.59 ± 0.04	54.35 ± 0.55	**60.85 ± 0.16**	6.51 ± 0.41
HR	87.02 ± 0.12	55.23 ± 0.35	58.67 ± 0.12	3.44 ± 0.42
Ours	86.36 ± 0.23	**56.52 ± 0.54**	59.57 ± 0.13	**3.06 ± 0.66**
		$\varepsilon = 8/255$		
PGD-AT	**84.80 ± 0.14**	45.16 ± 0.19	52.91 ± 0.11	7.75 ± 0.17
UDR-AT	83.87 ± 0.26	46.60 ± 0.27	**53.23 ± 0.30**	6.63 ± 0.57
HR	83.95 ± 0.32	47.32 ± 0.59	51.23 ± 0.25	3.90 ± 0.47
Ours	83.34 ± 0.16	**48.58 ± 0.21**	51.95 ± 0.19	**3.36 ± 0.06**
		$\varepsilon = 10/255$		
PGD-AT	**82.32 ± 0.33**	38.51 ± 0.06	46.45 ± 0.22	7.94 ± 0.26
UDR-AT	81.49 ± 0.54	39.54 ± 0.64	**47.33 ± 0.55**	7.79 ± 0.27
HR	80.90 ± 0.40	41.78 ± 0.43	45.70 ± 0.02	3.93 ± 0.41
Ours	80.34 ± 0.97	**43.52 ± 0.59**	46.47 ± 0.13	**3.11 ± 0.64**
		$\varepsilon = 12/255$		
PGD-AT	**79.55 ± 0.06**	34.16 ± 0.30	41.79 ± 0.12	7.63 ± 0.41
UDR-AT	78.67 ± 0.61	35.93 ± 1.06	**42.69 ± 0.10**	6.76 ± 1.01
HR	77.86 ± 0.98	37.69 ± 0.61	41.50 ± 0.26	3.81 ± 0.76
Ours	76.33 ± 1.18	**39.45 ± 0.58**	42.24 ± 0.31	**2.79 ± 0.53**

Table 5: Robust performance of different robust methods based on TRADES on CIFAR-10 using a ResNet-18 for l_∞ with budget 8/255. We use PGD-10 as the attack to evaluate robust performance. The "Best" robust test acc is the highest robust test accuracy achieved during training whereas the "Final" robust test acc is last epochs's robust accuracy.

Robust Methods	Nat	Robust Test Acc (%)		
		Final	Best	Diff
TRADES	82.69 ± 0.32	51.07 ± 0.18	53.54 ± 0.17	2.47 ± 0.21
UDR-TRADES	**82.74** $\pm$ **0.59**	51.16 ± 0.53	**53.64** $\pm$ **0.15**	2.48 ± 0.69
Ours-TRADES	81.79 ± 0.55	**52.22** $\pm$ **0.42**	53.58 ± 0.19	**1.36** $\pm$ **0.62**

Table 6: Robustness evaluation on CIFAR-10 using ResNet-18 for l_∞ under attack budget $8/255$, where models are trained with a larger attack budget $10/255$.

Methods($10/255$)	Nat	PGD-10	PGD-200	AA
PGD-AT	**82.32** $\pm$ **0.33**	46.98 ± 0.06	45.03 ± 0.12	43.17 ± 0.08
UDR-AT	81.49 ± 0.54	48.12 ± 0.41	46.45 ± 0.63	43.85 ± 0.57
HR	80.90 ± 0.40	50.04 ± 0.45	48.69 ± 0.41	44.05 ± 0.42
Ours	80.58 ± 0.98	**50.86** $\pm$ **0.72**	**49.71** $\pm$ **0.88**	**45.86** $\pm$ **0.88**

Table 7: Robustness evaluation on CIFAR-10 using ResNet-18 for l_∞ under attack budget $8/255$, where models are trained with attack budget $12/255$.

Methods($12/255$)	Nat	PGD-10	PGD-200	AA
PGD-AT	**79.55** $\pm$ **0.06**	48.99 ± 0.38	47.20 ± 0.26	44.48 ± 0.07
UDR-AT	78.67 ± 0.61	50.38 ± 0.63	49.10 ± 0.75	45.47 ± 0.72
HR	76.65 ± 0.65	51.96 ± 0.05	51.03 ± 0.12	45.45 ± 0.44
Ours	76.33 ± 1.17	**52.32** $\pm$ **0.23**	**51.52** $\pm$ **0.26**	**46.87** $\pm$ **0.17**

Mitigating overfitting on TRADES. We would like to provide further experimental results on the CIFAR-10 dataset on TRADES and its counterpart as shown in Table 5. It can be observed that our method also can mitigates the robust overfitting, and enhances robustness while maintaining comparable accuracy.

Training with larger budgets. Theorem 5 shows that to ensure robustness on the test set, it is generally necessary to employ a larger attack budget during training compared to testing. In this experiment, we examine the robustness generalization by attacking different robust methods with fixed attack budget $8/255$ with respect to larger training budgets $10/255, 12/255$ while keeping other parameters of PGD attack the same. The results shown in Table 6 and Table 7 demonstrate that our proposed SR-WDRO achieves the best robustness under various adversarial attacks.

Computation cost We evaluate the computational efficiency of our proposed SR-WDRO method against baseline approaches. Table presents the average training time per epoch and total training time for CIFAR-10 on a single NVIDIA A800 GPU. Our SR-WDRO incurs a modest increase in per-epoch time (approximately 12.7% longer than standard adversarial training).

Table 8: Training time per epoch and total training time for CIFAR-10 on a single NVIDIA A800 GPU.

Time Cost (s)	PGD-AT	UDR-AT	HR	Ours
200 epoch	12614.59	14104.05	13705.56	14216.43
per epoch	63.07	70.52	68.53	71.08