# Statistical Properties of Robust Satisficing

**Zhiyi Li** [1] **Yunbei Xu** [2] **Ruohan Zhan** [3 4]

## Abstract

The Robust Satisficing (RS) model is an emerging approach to robust optimization, offering streamlined procedures and robust generalization across various applications. However, the statistical theory of RS remains unexplored in the literature. This paper fills in the gap by comprehensively analyzing the theoretical properties of the RS model. Notably, the RS structure offers a more straightforward path to deriving statistical guarantees compared to the seminal Distributionally Robust Optimization (DRO), resulting in a richer set of results. In particular, we establish two-sided confidence intervals for the optimal loss without the need to solve a minimax optimization problem explicitly. We further provide finite-sample generalization error bounds for the RS optimizer. Importantly, our results extend to scenarios involving distribution shifts, where discrepancies exist between the sampling and target distributions. Our numerical experiments show that the RS model consistently outperforms the baseline empirical risk minimization in small-sample regimes and under distribution shifts. Furthermore, compared to the DRO model, the RS model exhibits lower sensitivity to hyperparameter tuning, highlighting its practicability for robustness considerations.

## 1. Introduction

Robust methods are optimization techniques that guarantee performances even when environments vary slightly (Ronchetti, 2021). These methods are resilient against variations or uncertainties, ensuring consistent and reliable outcomes. Robustness provided by these methods is particularly valuable in scenarios where limited sample sizes may not fully capture the entire distribution, or where the target environment differs from the initial sampling distribution.

The application of robust methods spans across various domains: in machine learning, they are utilized to enhance the robustness of algorithms, ensuring they maintain strong performance even when there are adversarial attacks in the input data (Blanchet et al., 2019; Sim et al., 2021). In energy systems, they are adopted to optimize the operation and planning, including bidding strategies in electricity markets, operation scheduling of power systems, and integration of renewable energy (Li et al., 2023; Huang et al., 2023). In supply chains, they are employed to optimize various aspects such as production planning, inventory management, logistics, and transportation. (Chen & Chen, 2023; Deng et al., 2023; Wang et al., 2023). These examples represent a fraction of the wide-ranging applications of robust methods. In fact, robust methods can be applied to any field that involves optimization problems, making it a vital tool for decision-making under uncertainty.

Among various robust methods, Distributionally Robust Optimization (DRO) is a pivotal approach (Hu & Hong, 2013; Bayraksan & Love, 2015; Esfahani & Kuhn, 2015). DRO's significance lies in its more robust handling of ambiguity compared to conventional stochastic programming models. This is achieved by optimizing the worst-case performance over a set of potential distributions rather than for a single distribution. Specifically, the DRO problem is formulated as follows:

$$\min_{x\in\mathcal{X}} \max_{P\in\mathbb{P}_r} \ \mathbb{E}_P[h(x,\xi)], \tag{1}$$
$$\text{where } \mathbb{P}_r = \{P \in \mathbb{P} : d(P, \hat{P}_N) \le r\}.$$

Above, $x$ represents the decision variable, which is contained in a non-empty decision space $\mathcal{X}$, and $\xi$ is a random variable. The function $h(x,\xi)$ denotes the loss associated with $x$ and $\xi$. $\hat{P}_N$ is the empirical distribution derived from the data[1]. The function $d(\cdot,\cdot)$ is a distance measure to quantify discrepancies between distributions. The hyperparame-

[1] School of Mathematical Sciences, Peking University, Beijing, China [2] Yunbei Xu, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, USA [3] Department of Industrial Engineering and Decision Analytics, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR [4] HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen, China. Correspondence to: Ruohan Zhan <rhzhan@ust.hk>.

*Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1] Other nominal distributions are also viable. For example, when provided with a parametric distributional class, the distribu-

ter $r$, referred to as the "radius", defines the ambiguity set $\mathbb{P}_r$, a subset of $\mathbb{P}$ that encompasses all feasible distributions. It plays a crucial role in controlling robustness—the larger the value of $r$, the greater the robustness demanded.

Despite its strengths, DRO has a few shortcomings. First, DRO can be overconservative in practice, as Esfahani & Kuhn pointed out. This is because the DRO framework optimizes the worst-case scenario in the distribution domain, which may be unnecessarily large to incorporate the target distribution. Second, selecting an appropriate and interpretable radius $r$ is a challenging task in practice, as noted by Sim et al.. This difficulty stems from the abstract nature of the radius, which characterizes the distance within the distributional space and is hard to be intuitively translated into tangible, real-world values. In addition, there is a growing demand for incorporating globalized distributions–as opposed to restricting to an ambiguity set under the DRO framework–to further increase robustness (Liu et al., 2023).

To address these issues, the Robust Satisficing (RS) model has been proposed (Long et al., 2023), as structured below:

$$\begin{aligned} k_\tau = \min \quad & k \\ \text{s.t. } & \mathbb{E}_P[h(x,\xi)] - \tau \leq k d(P, \hat{P}_N), \quad \forall P \in \mathbb{P} \\ & \boldsymbol{x} \in \mathcal{X}, \quad k \geq 0. \end{aligned} \tag{2}$$

Here, the hyperparameter is no longer the radius $r$ of the ambiguity set, but a reference value $\tau$, which can be interpreted as an anticipated cost in practical applications. The constraint (2) then ensures that once the expected loss under a certain distribution exceeds our anticipated cost, the excess part should not be too large: it will be controlled by a multiple of the distribution's distance to the empirical distribution $\hat{P}_N$. Hence, the RS model compromises some training set performance for robustness in the target distribution, as it doesn't aim for minimizing the empirical loss. Unlike DRO that focuses on worst-case optimization, RS follows a satisficing strategy to avoid over-conservatism, thereby providing better generalization performance on the target distribution. Another key aspect of the RS model is its global consideration of probability distributions, unlike DRO's restriction to ambiguity set.

Current research on the RS model primarily centers around forming tractable new optimization models and experimental analysis. Notable examples of tractable RS model encompass Risk-based Linear Optimization and Linear Optimization with Recourse (Sim, 2023; Long et al., 2023), illustrating RS's practical optimization and generalization advantages compared to the DRO models. Ruan et al. proposed Robust Satisficing Markov Decision Processes and demonstrated its superiority over traditional robust MDP through experiments. Saday et al. proposed the Robust Bayesian Satisficing model, established upper bound on regret and outperformed Distributionally Robust Bayesian Optimization in experiments. Despite RS's notable results in the realm of optimization, to the best of our knowledge, there are no existing studies on the statistical properties of the RS model. This gap leads to the central research questions of this paper:

*Are there statistical guarantees for the RS model? What are the statistical merits it holds, potentially surpassing DRO?*

### 1.1. Contributions

Our work delves into the statistical theory of the RS model, with a focus on deriving and analyzing its statistical properties. In particular, we provide a two-sided confidence interval estimate for the optimal loss using the reference value, and present non-asymptotic upper bound of the generalization error. These results fill a crucial gap in the literature, where statistical guarantees for the RS model have been seldom studied. It is noteworthy that our results extend beyond cases where the sampling distribution matches the target distribution of interest, a context where robustness still remains relevant due to potential discrepancies between the empirical and sampling distributions, especially in small-sample regimes; we also consider scenarios involving distribution shifts, where disparities exist between the sampling distribution and the target distribution. We highlight the contributions of this paper as follows:

1) We obtain two-sided, non-asymptotic confidence intervals for the optimal loss $J^*$ in the RS model, where $J^*$ is the minimum expected loss under the true distribution. Notably, this result does not necessitate solving a minimax optimization problem explicitly.

2) We present finite-sample generalization error bounds for the optimizer derived from the RS model, achieved through an insightful and succinct derivation.

3) We demonstrate that, even under distribution shifts, our key findings – confidence intervals and generalization error bounds for the RS model optimizer – remain valid. These results incorporate an additional term, a finite multiple of the distance between the sampling and target distributions. This adaptation highlights the RS model's robust generalization abilities.

4) Our numerical experiments reveal that RS model's advantages over the empirical risk minimization baseline becomes more pronounced in small-sample regimes or with increasing distribution shifts. Furthermore, our analysis reveals the relationship between the RS and DRO models under the Lipschitz loss scenarios, which also

tion estimated using maximum likelihood estimation can serve as a substitute for $\hat{P}_N$.

highlights that the RS model has lower sensitivity to hyperparameter tuning as compared to DRO.

In all these aspects, we perform an extensive comparison with DRO. Our analysis reveals that these advantageous properties are closely associated with the inherent structure of the RS model itself. It becomes evident that obtaining statistical guarantees is more straightforward within the RS framework compared to the DRO framework.

## 2. Set up

We start by describing our learning problem. Let $\xi \in \Xi$ be the $m$-dimensional random variable of observations and $x \in \mathcal{X}$ be the decision variable to be learned. Let $h(x, \xi)$ be the loss function (which can accommodate a wide range of machine learning problems as detailed in Appendix D). We use $J^*$ to denote the minimum expected loss under the optimal decision variable $x^*$:

$$J^* := \inf_{x \in \mathcal{X}} \mathbb{E}_{P^*}[h(x, \xi)] = \mathbb{E}_{P^*}[h(x^*, \xi)]. \tag{3}$$

Given $N$ observations $\{\xi_i\}_{i=1}^N$ sampled from the distribution $P^*$, the decision maker wants to learn a decision variable such that the expected loss is minimized.

Consider the **Robust Satisficing** (RS) model (2), and we focus on the Wasserstein distance for the distance measure between distributions. Here, $\tau$ is the "reference value", which can be interpreted as the anticipated cost in practical applications, and its choice will be further discussed in Section 2.1. $\mathbb{P}$ is the set of all feasible distributions, on which the RS model does not impose any constraint; this allows the RS model to consider probability distributions globally. $\hat{P}_N$ denotes the empirical distribution of samples $\{\xi_i\}_{i=1}^N$, which converges to $P^*$ as sample size goes to infinity. And $d_{\mathrm{W}}$ denotes the type-1 Wasserstein distance between two distributions[2]:

$$d_{\mathrm{W}}(Q_1, Q_2) := \inf_{\Pi} \left\{ \int_{\Xi \times \Xi} c(\xi_1, \xi_2) \Pi(\mathrm{d}\xi_1, \mathrm{d}\xi_2) \right\},$$

where $\Pi$ is a joint distribution over $(\xi_1, \xi_2)$, with its marginal distributions on $\xi_1$ and $\xi_2$ being $Q_1$ and $Q_2$ respectively; the cost function $c(\cdot, \cdot)$ used for Wasserstein distance is chosen as the type-I version, with $c(x, y) = \|x - y\|_2$.

Let $\hat{x}_N$ be the solution derived from the RS model (2), the reformulation of which will be elaborated upon in Section 2.2. Our goal is to provide statistical guarantees on $\hat{x}_N$ and $J^*$.

### 2.1. Reference Value $\tau$

The reference value $\tau$ introduced in the RS model (2) is critical in controlling the robustness of the learned solution $\hat{x}_N$. Conceptually, following the satisficing criterion, the RS model ensures that any excess beyond the reference value $\tau$ under a certain distribution is controlled by a multiple of the distance between this distribution and the empirical distribution of the data. A larger $\tau$ indicates increased robustness considered in the RS model.

Choose $P$ as $\hat{P}_N$ in (2), we easily obtain:

$$\tau \geq \mathbb{E}_{\hat{P}_N}[h(x, \xi)]. \tag{4}$$

Inspired by this, Long et al. suggest choosing $\tau$ as:

$$\tau_\epsilon := (1 + \epsilon) \inf_x \mathbb{E}_{\hat{P}_N}[h(x, \xi)], \tag{5}$$

where $\epsilon$ is referred to as "tolerance rate" that the RS model allows for excess empirical loss. This means that the reference value $\tau$, which we choose or tolerate, is $\epsilon$ more than the smallest cost achievable under the empirical distribution. We adopt this approach, focusing on characterizing the role of $\epsilon$ in the statistical guarantees provided by the RS model. Additionally, $\epsilon$ will be the primary hyperparameter we adjust and analyze in the numerical experiments section.

### 2.2. Reformulation

The original RS optimization (2) requires enumerating over all possible distributions over $\mathbb{P}$, which may not be tractable. We now reformulate the model (2), following the practice by Long et al.. Let $\eta$ and $\xi$ be samples from $P$ and $\hat{P}_N$ respectively, and let $\pi(\eta|\epsilon)$ be the conditional distribution of $\eta$ when conditioning on $\xi$. We have:

$$\begin{aligned} &\sup_P \{\mathbb{E}_P[h(x, \eta)] - k d_W(P, \hat{P}_N)\} \\ =&\sup_\pi \iint [h(x, \eta) - kc(\xi, \eta)] d\hat{P}_N(\xi) \, d\pi(\eta|\xi) \\ =&\mathbb{E}_{\hat{P}_N}[\sup_{z \in \Xi} h(x, z) - kc(\xi, z)], \end{aligned} \tag{6}$$

where the last equation is achieved by choosing the maximizer $\pi$ as the Dirac distribution, which concentrates the mass at the point to maximize $\{h(x, \cdot) - kc(\xi, \cdot)\}$. Then the RS model (2) can be reformulated as:

$$\begin{aligned} \min \quad & k \geq 0 \\ \text{s.t.} \quad & \mathbb{E}_{\hat{P}_N}[\sup_{z \in \Xi} h(x, z) - kc(\xi, z)] \leq \tau. \\ & x \in \mathcal{X} \end{aligned} \tag{7}$$

With that, the optimizer $\hat{x}_N$ of RS model can be obtained in a hierarchical way. First, for a fixed decision variable $x$, let

[2] We follow the literature (Long et al., 2023) and consider the Wasserstein distance instead of f-divergence to avoid the requirement that $P$ is absolute continuous with respect to $\hat{P}_N$, which is impractical for continuous distribution.

$k_\tau(x)$ be the smallest $k$ that satisfies the RS constraint:

$$\begin{aligned} k_\tau(x) &:= \min k(x), \\ \text{s.t. } &\mathbb{E}_{\hat{P}_N}[\sup_{z\in\Xi} h(x,z) - kc(\xi,z)] \leq \tau. \end{aligned}$$

Then $\hat{x}_N$ is the minimizer of $k_\tau(x)$:

$$\hat{x}_N := \text{argmin}_x k_\tau(x). \tag{8}$$

We note that a similar reformulation technique has been employed by Blanchet & Murthy to derive tractable solutions for DRO. However in the DRO framework, the distribution is restricted to an ambiguity set, necessitating the use of a Lagrange multiplier for constraint conditions and the existence of strong duality. These constraints introduce additional assumptions, including the continuity of functions $h(x,\xi)$ and $c(\cdot,\cdot)$. In contrast, robust satisficing, which does not limit the distribution set, avoids these extra assumptions.

## 3. Statistical Properties

This section presents our main results for the statistical properties of the optimizer $\hat{x}_N$ in the RS model (2). We start by describing the assumptions required for our analysis.

**Assumption 1** (Exponential tail decay in random variable)**.** *There exists an $a > 1$, such that $\mathbb{E}_{P^*}[\exp(||\xi||^a)] < \infty$.*

Assumption 1 requires that $\xi$ is relatively light-tailed. It plays a key role in bounding the rate at which the empirical distribution $\hat{P}_N$ approximates the true distribution $P^*$ under the type-1 norm Wasserstein distance (Fournier & Guillin, 2015) (see Proposition 2 for details). This assumption is relatively mild and is applicable to a broad range of distributions including sub-Gaussian random variables.

**Assumption 2** (Lipschitz continuity of loss function)**.** *The loss $h(x,\xi)$ is Lipschitz with a uniform constant $L$ in $\xi$.*

Assumption 2 is essential for deriving the dual expression form of the type-1 norm Wasserstein distance (Esfahani & Kuhn, 2015) (see Proposition 1 for details). This assumption holds true for a wide range of machine learning problems, which we further elaborate in Appendix D. Note that we don't need the Lipschitz continuity assumption of $h(x,\xi)$ with respect to $x$, which we defer the detailed discussion in the Appendix E.

### 3.1. Confidence Intervals of Optimal Loss

This section provides both non-asymptotic and asymptotic confidence intervals for the optimal loss $J^*$, the smallest attainable expected loss as defined in (3), and the true loss of $\hat{x}_N$.

**Theorem 1** (Confidence intervals of optimal loss)**.** *Suppose Assumptions 1 & 2 hold. For any $N$, let $\beta_N$ be the confidence level. We have with probability at least $1-\beta_N$:*

$$-L\cdot r_N + \frac{\tau_\epsilon}{1+\epsilon} \leq J^* \leq \mathbb{E}_{P^*}[h(\hat{x}_N,\xi)] \leq k_{\tau_\epsilon}\cdot r_N + \tau_\epsilon, \tag{9}$$

*where $r_N$, denoted as the "remainder", is solved from the below equation:*

$$\beta_N = \begin{cases} c_1 \exp\big(-c_2 N {r_N}^{\max\{m,2\}}\big) & \text{if } r_N \leq 1, \\ c_1 \exp\big(-c_2 N r_N^a\big) & \text{if } r_N > 1, \end{cases}$$

*with $c_1, c_2$ as positive constants that only depend on exponential decay rate $a$ and dimension $m$.*

*Moreover, when choosing the confidence sequence $\{\beta_N\}$ satisfying $\sum_{N=1}^{\infty}\beta_N < \infty$, we have*

$$\begin{aligned} P\Big\{ -L\cdot r_N + \frac{\tau_\epsilon}{1+\epsilon} \leq J^* \leq \mathbb{E}_{P^*}[h(\hat{x}_N,\xi)] \\ \leq k_{\tau_\epsilon}\cdot r_N + \tau_\epsilon \textit{ for all sufficiently large } N\Big\} = 1. \end{aligned} \tag{10}$$

Table 1 outlines typical selections of $\beta_N$ and their respective rates of decay for the remainder $r_N$. Notably, the last two $\beta_N$ options satisfy $\sum_{N=1}^{\infty}\beta_N < \infty$ and $\lim_{N\to\infty} r_N = 0$, under which (10) suggests asymptotic consistency of $\mathbb{E}_{P^*}[h(\hat{x}_N,\xi)]$(Note that this asymptotic interval applies to $J^*$ directly by the convergence of empirical loss to the true loss as $N$ increases):

$$\begin{aligned} P\Big\{\frac{\tau_\epsilon}{1+\epsilon} \leq \mathbb{E}_{P^*}[h(\hat{x}_N,\xi)] \leq \tau_\epsilon \text{ for all} \\ \text{sufficiently large } N\Big\} = 1. \end{aligned} \tag{11}$$

We recognize the challenge posed by the curse of dimensionality, as indicated by the exponent $m$ in $r_N$, which is a common issue associated with the Wasserstein distance (Esfahani & Kuhn, 2015; Kuhn et al., 2019), and we leave as a promising future research question.

We also note that the upper bound in Eq. (9) includes $k_{\tau_\epsilon}$, which may be difficult to derive analytically. Fortunately, the following lemma provides an upper bound guarantee for $k_{\tau_\epsilon}$.

**Lemma 1** (Fragility Upper Bound)**.** *Under Assumption 2, we have $k_\tau \leq L$, where $k_\tau$ is solved from the RS model (2).*

Lemma 1 that we prove is noteworthy on its own. As pointed out by Long et al., $k_\tau$ characterizes the fragility of the model, with lower values indicating more robustness. Lemma 1 sets an upper bound for $k_\tau$ based on the Lipschitz constant $L$, suggesting the model fragility being controlled.

**Remark 1.** *In the proof detailed in the Appendix B.2, we establish the following relationship:*

$$\begin{aligned} -L\cdot d_W(P^*,\hat{P}_N) + \frac{\tau_\epsilon}{1+\epsilon} \leq J^* \leq \mathbb{E}_{P^*}[h(\hat{x}_N,\xi)] \\ \leq k_{\tau_\epsilon}\cdot d_W(P^*,\hat{P}_N) + \tau_\epsilon. \end{aligned} \tag{12}$$

*Equation (12) illustrates that the true loss of $\hat{x}_N$ (the optimizer obtained from the RS model), under the target distribution $P^*$, also falls within the confidence interval provided by Equation (9). Furthermore, Equation (9) further facilitates the derivation of the upper bound of the generalization error in Theorem 3.*

**Remark 2.** *Equation (9) provides a guideline on determining the sufficient sample size required to achieve a predefined accuracy at a specified confidence level $\beta_N$. This sample size is primarily quantified by the width of the confidence interval and mainly driven by $r_N$. Table 1 illustrates various selections of $\beta_N$ along with their corresponding $r_N$ values, which allows us to explicitly compute the sample size required for specific scenarios.*

By integrating Lemma 1 with Theorem 1, we derive a simpler form of confidence intervals for $J^*$, which depends solely on the Lipschitz constant $L$ and the reference value $\tau_\epsilon$, eliminating the need to compute $k_{\tau_\epsilon}$ from the RS model.

**Corollary 2.** *Suppose Assumptions 1 & 2 hold. For any $N$ and the confidence level $\beta_N$, let $r_N$ be solved as in Theorem 1. With probability at least $1-\beta_N$, we have*

$$-L \cdot r_N + \frac{\tau_\epsilon}{1+\epsilon} \leq J^*, \mathbb{E}_{P^*}[h(\hat{x}_N, \xi)] \leq L \cdot r_N + \tau_\epsilon.$$

The remainder $r_N$ becomes negligible for choices of $\beta_N$ listed in Table 1. Thus Corollary 2 indicates that as $N$ approaches $\infty$, the expected loss $E_{P^*}h(\hat{x}_N, \xi)$ of the optimizer $\hat{x}_N$ will also fall within the interval $[\frac{\tau_\epsilon}{1+\epsilon}, \tau_\epsilon]$. This allows us to characterize the loss value that the optimizer can achieve and also shows that the regret of our optimizer $\hat{x}_N$ (the gap between $E_{P^*}h(\hat{x}_N, \xi)$ and the true loss) will be controlled by the length of the interval asymptotically.

To conclude this section, we offer a brief comparison of our confidence intervals with those derived by Esfahani & Kuhn. In their DRO framework, they define

$$\tilde{J}_N = \inf_{x \in \mathcal{X}} \sup_{P \in B(\hat{P}_N, \epsilon(\beta_N))} \mathbb{E}_P\left[h(x, \xi)\right],$$

where $B(\hat{P}_N, \epsilon(\beta_N))$ represents a Wasserstein ball with its center $\hat{P}_N$ and radius $\epsilon(\beta_N)$. Under similar assumptions, Esfahani & Kuhn show that

$$P\{J^* \leq \tilde{J}_N\} \geq 1 - \beta_N,$$

which provides only an upper bound for the optimal loss $J^*$. Moreover, this upper bound $\tilde{J}_N$ requires to solve the minimax problem in the DRO framework. In contrast, our confidence intervals from Corollary 2 are derived through the relatively easier optimization of the ERM problem than the minimax problem, and our results provide two-sided rather than one-sided confidence intervals.

### 3.2. Finite-Sample Generalization Error Bound

We now focus on characterizing the generalization error of the optimizer $\hat{x}_N$ derived from the RS model. The generalization error, denoted as $R(P^*, \hat{x}_N)$, is defined as follows:

$$\begin{aligned} R(P^*, \hat{x}_N) :=& \mathbb{E}_{P^*}[h(\hat{x}_N, \xi)] - J^* \\ =& \mathbb{E}_{P^*}[h(\hat{x}_N, \xi)] - \mathbb{E}_{P^*}[h(x^*, \xi)]. \end{aligned}$$

**Theorem 3.** *Suppose Assumptions 1 & 2 hold. With probability at least $1-\beta_N$, we have:*

$$R(P^*, \hat{x}_N) \leq \epsilon \cdot J^* + (2+\epsilon) \cdot L \cdot r_N, \tag{13}$$

*where $r_N$ is the reminder solved as in Theorem 1. Taking expectation with respect to data, we have:*

$$\mathbb{E}_{P^*}\left[R(P^*, \hat{x}_N)\right] \leq \epsilon \cdot J^* + O(L \cdot N^{-\min\{\frac{1}{m}, \frac{1}{2}\}}). \tag{14}$$

**Remark 3.** *We further elaborate on the "expectation with respect to data". Recall that we derive the optimizer $\hat{x}_N$ based on sample data, which are random variables that follow the source distribution $P^*$. As a result, the $\hat{x}_N$ and its generalization error upper bound are also random variables. So we take the expectation with respect to the randomness from the sample data to derive our expected version of the generalization error upper bound.*

Theorem 3 explicitly characterizes how the generalization error is influenced by $\epsilon$. By reducing $\epsilon$ as the sample size $N$ increases — indicating less tolerance for empirical loss excess with more data — we can bound the generalization error more succinctly, as outlined in the following result.

**Corollary 4.** *Suppose Assumptions 1 & 2 hold. Choose reference value $\tau_{\epsilon_N}$ with $\epsilon_N = N^{-\min\{\frac{1}{m}, \frac{1}{2}\}}$. Then*

$$\mathbb{E}_{P^*}\left[R(P^*, \hat{x}_N)\right] = O(L \cdot N^{-\min\{\frac{1}{m}, \frac{1}{2}\}}). \tag{15}$$

## 4. Guarantees under Distribution Shift

As discussed, the distribution selection under the RS framework is globalized, eliminating the need to pre-select a radius to restrict the distribution domain. We take this advantage further and integrate it into the earlier derivation process, allowing us to straightforwardly derive the confidence intervals and the finite-sample generalization error bound under distribution shifts.

Consider that samples are drawn from the source distribution $P^*$, and the empirical distribution is denoted as $\hat{P}_N$. The decision variable learned from the RS model (2) is $\hat{x}_N$. Under distribution shifts, we evaluate the performance when applying $\hat{x}_N$ to another distribution $\tilde{P}$, which may shift from $P^*$, resulting in a certain degree of discrepancy.

Define the optimal loss under the new distribution $\tilde{P}$ as $\tilde{J}$:

$$\tilde{J} := \inf_{x \in \mathcal{X}} \mathbb{E}_{\tilde{P}}[h(x, \xi)] = \mathbb{E}_{\tilde{P}}[h(\tilde{x}, \xi)],$$

*Table 1.* Choices of Confidence Level $\beta_N$

| Choice of $\beta_N$ | Corresponding $r_N$ |
|---|---|
| $\beta_N \equiv \beta$ | $r_N = \begin{cases} \left(\frac{\log(c_1\beta^{-1})}{c_2N}\right)^{1/\max\{m,2\}} & if\ N \geq \frac{\log(c_1\beta^{-1})}{c_2}, \\ \left(\frac{\log(c_1\beta^{-1})}{c_2N}\right)^{1/a} & if\ N < \frac{\log(c_1\beta^{-1})}{c_2}. \end{cases}$ |
| $\beta_N = \exp(-\gamma\sqrt{N}), \gamma > 0$ | $r_N = \begin{cases} \left(\frac{\log c_1}{c_2N} + \frac{\gamma}{c_2\sqrt{N}}\right)^{1/\max\{m,2\}} & if\ c_2N - \gamma\sqrt{N} \geq \log c_1, \\ \left(\frac{\log c_1}{c_2N} + \frac{\gamma}{c_2\sqrt{N}}\right)^{1/a} & if\ c_2N - \gamma\sqrt{N} < \log c_1. \end{cases}$ |
| $\beta_N = N^{-\alpha}, \alpha > 0$ | $r_N = \begin{cases} \left(\frac{\log c_1}{c_2N} + \alpha\frac{\log N}{c_2N}\right)^{1/\max\{m,2\}} & if\ c_2N - \alpha\log N \geq \log c_1, \\ \left(\frac{\log c_1}{c_2N} + \alpha\frac{\log N}{c_2N}\right)^{1/a} & if\ c_2N - \alpha\log N < \log c_1. \end{cases}$ |

where $\tilde{x} = \text{argmin}_x \mathbb{E}_{\tilde{P}}[h(x,\xi)]$. For the learned decision variable $\hat{x}_N$, denote the corresponding generalization error as

$$\begin{aligned} R(\tilde{P}, \hat{x}_N) :=\ & \mathbb{E}_{\tilde{P}}[h(\hat{x}_N,\xi)] - \tilde{J} \\ =\ & \mathbb{E}_{\tilde{P}}[h(\hat{x}_N,\xi)] - \mathbb{E}_{\tilde{P}}[h(\tilde{x},\xi)]. \end{aligned}$$

Our goal is to derive confidence intervals for $\tilde{J}$ and generalization error bound for $R(\tilde{P}, \hat{x}_N)$.

**Theorem 5** (Distribution Shift)**.** *Suppose Assumptions 1 & 2 hold. For any $N$, let $\beta_N$ be some nominal confidence level. We have with probability at least $1-\beta_N$:*

$$\begin{aligned} -L\cdot r_N - L\cdot d_W(P^*,\tilde{P}) + \frac{\tau_\epsilon}{1+\epsilon} \leq \tilde{J} \leq \mathbb{E}_{\tilde{P}} h(\hat{x}_N,\xi) \\ \leq k_{\tau_\epsilon}\cdot r_N + k_\tau \cdot d_W(P^*,\tilde{P}) + \tau_\epsilon, \end{aligned}$$

*and*

$$\begin{aligned} R(\tilde{P},\hat{x}_N) \leq \epsilon\cdot\tilde{J} + (2+\epsilon)\cdot L\cdot d_W(P^*,\tilde{P}) \\ +(2+\epsilon)\cdot L\cdot r_N, \end{aligned}$$

*where the reminder $r_N$ is solved as Theorem 1.*

*Taking the expectation on data, we have:*

$$\begin{aligned} \mathbb{E}_{P^*}\left[R(\tilde{P},\hat{x}_N)\right] \leq \epsilon\cdot\tilde{J} + (2+\epsilon)\cdot L\cdot d_W(P^*,\tilde{P}) \\ +O\left(L\cdot N^{-\min\{\frac{1}{m},\frac{1}{2}\}}\right). \end{aligned}$$

This theorem shows that results under distribution shifts merely require adding a multiple of the shift distance.

**Remark 4.** *While our results face the common curse of dimensionality issue associated with the Wasserstein distance, they embody a trade-off. In higher dimensions, despite the slow decay of the remainder term $r_N$, a greater degree of distribution shift is tolerable. Specifically, when the distribution shift decays at the rate of $N^{-\min\{\frac{1}{m},\frac{1}{2}\}}$, this rate can be integrated with the remainder term to yield the following guarantee:*

$$\mathbb{E}_{P^*}\left[R(\tilde{P},\hat{x}_N)\right] \leq \epsilon\cdot\tilde{J} + O(L\cdot N^{-\min\{\frac{1}{m},\frac{1}{2}\}}).$$

*This implies that the RS model can accommodate a distribution shift up to $N^{-\min\{\frac{1}{m},\frac{1}{2}\}}$ while still maintaining performance comparable to scenarios with no shift.*

Finally, we compare our results with DRO. Under the DRO framework, if the distribution shifts, we must require the radius to reach a certain magnitude so that the ambiguity set can contain the distribution after the shift. However, as this ball expands, the worst-case expected value within DRO's conservative minimax framework deteriorates. In contrast, the RS framework, benefiting from its globalized distribution selection, only requires the inclusion of a linear multiple of the shift distance to address the same situation.

## 5. Numerical Experiments

In this section, we conduct numerical evaluations of the RS model under both the original sampling distribution and distributional shifts. We compare RS with the baseline method, which is empirical risk minimization (ERM). Additionally, we establish connections with DRO and demonstrate that RS exhibits lower sensitivity to hyperparameter tuning.

All experiments are based on a data generating process detailed below. We define the random variable $\xi$ as $\xi = (u,y)$, with $u \in \mathbb{R}^{m_u}$ representing the feature variable and $y \in \mathbb{R}$ as the label variable. The sampling distribution $P^*$ is specified as follows: the feature variable $u$ is drawn from a normal distribution:

$$u \sim \mathcal{N}\left([0.5, 0.5, ..., 0.5]^T, 0.5I_{m_u}\right); \qquad (16)$$

and the label variable $y$ is generated via a linear model:

$$y = u\cdot x^* + e,$$

where $\cdot$ means the inner product, $x^*$ is the true model parameter, and $e$ is the exogenous noise sampled from $\mathcal{N}(0, 0.1)$. $P^*$ satisfies Assumption 1 because Gaussian distribution is light-tailed. Let the training data $\{(u_i, y_i)\}_{i=1}^N$ be i.i.d. samples from the distribution $P^*$.

We use $\ell_1$ loss for model parameter $x$: $h(u, y, x) = |y - u \cdot x|$, which satisfies the Lipschitz condition in Assumption 2. For the cost function used in the type-I Wasserstein distribution, we follow (Blanchet et al., 2019) and slightly modify its original definition of the $l_2$ norm as follows[3]:

$$c(\xi_1, \xi_2) = c((u_1, y_1), (u_2, y_2)) = \begin{cases} ||u_1 - u_2||_2 & \text{if } y_1 = y_2, \\ +\infty & \text{otherwise.} \end{cases}$$

The learned parameter $\hat{x}_N$ is evaluated on the target distribution $\tilde{P}$. The marginal ditribution of $u$ under $\tilde{P}$ is identical to that under $P^*$, following (16). However, the label variable $y$ is generated under a potentially different parameter $\tilde{x}$:

$$y = u \cdot \tilde{x} + e.$$

In the following sections, we will evaluate the performances of RS under two scenarios: when $P^* = \tilde{P}$ (i.e., $x^* = \tilde{x}$), representing settings without distribution shift, and when $P^* \neq \tilde{P}$ (i.e., $x^* \neq \tilde{x}$), indicating settings with distribution shift. We focus on the mean square error (MSE) in the target distribution as the performance metric. We will conclude this section by drawing connections between RS and DRO.

### 5.1. RS Performance in the Sampling Distribution

In this section, we evaluate the RS optimizer $\hat{x}_N$ under the sampling distribution. Although the target distribution aligns with the sampling distribution, discrepancies between the empirical and sampling distributions may arise, particularly in small-sample regimes for high dimensional random variables. For this purpose, we consider a relatively high-dimensional setting with the dimension $m_u = 10$ and the true model parameter $\tilde{x} = x^* = [2.0, -1.0, ..., 2.0, -1.0]^T$. We investigate the generalization performance of $\hat{x}_N$ across various sample sizes.

Figure 1 demonstrates how the RS model's performance varies with different settings of the tolerate rate $\epsilon$ and across various sample sizes. For smaller sample sizes, the RS model outperforms the baseline that does not incorporate robustness; among those RS models, those configured with a larger $\epsilon$ (indicating a greater emphasis on robustness) perform better. This result notes the importance of accounting for robustness, particularly when there is a notable gap between the sampling and empirical distributions in small-sample regimes. As the sample size increases, the relative benefit of the RS model decreases. This trend is expected since a larger dataset allows to better picture the sampling distribution, making the baseline approach of empirical risk minimization increasingly effective.

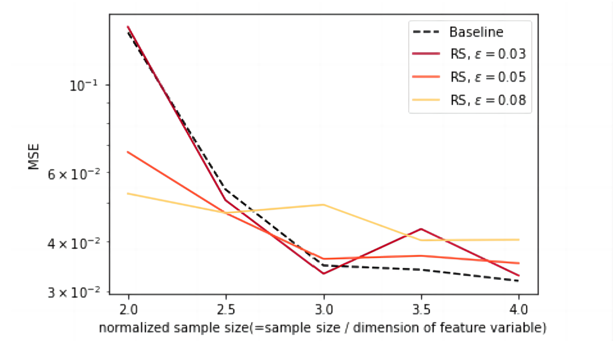

*Figure 1.* Performances across various sample sizes. RS outperforms the ERM baseline in small-sample regimes.

### 5.2. RS Performance under Distribution Shift

We now evaluate the performance of RS under distribution shift. Let the model parameter in the sampling distribution be $x^* = [2.0, -1.0]$, and the model parameter in the target distribution be:

$$\tilde{x} = [2.00 - 0.05 \times \text{DEGREE}, -1.00 + 0.025 \times \text{DEGREE}],$$

where DEGREE is a positive number characterizing the degree of distribution shift[4]. As DEGREE increases, the discrepancy between $x^*$ and $\tilde{x}$ increases, leading to a larger distribution shift between the sampling distribution $P^*$ and the target distribution $\tilde{P}$.

Figure 2 shows how the RS model performs when configured with different tolerance rates $\epsilon$ and under various distribution shifts. For minor distribution shifts (DEGREE less than 2), the RS model's performance is comparable to the baseline, deteriorating slightly at models of larger $\epsilon$ for stronger robustness control. However, with more substantial distribution shifts (DEGREE greater than 5), the RS model almost consistently outperforms the baseline, presenting stronger robustness under larger distribution shifts. This result highlights the potential of RS framework for strong generalization guarantee in uncertain environments.

### 5.3. Connection to DRO under Lipschitz loss

We proceed to compare RS and DRO. We establish explicit correspondence between the hyperparameters of RS and

[3] The purpose of this adjustment is to make the subsequent exposition more concise. We leave the results under the original $l_2$ norm in Appendix C.2

[4] Here our focus is on evaluating robustness in scenarios with smoothed parameters, rather than under arbitrary perturbations of $x^*$. This setup is based on the observation that both DRO and RS, known for their robustness, typically yield smoother parameters than those derived from direct empirical risk minimization.

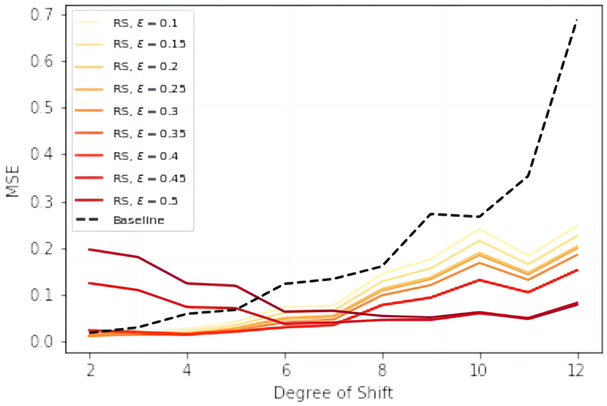

*Figure 2.* Performances across various degree of distribution shifts. RS outperforms the ERM baseline under distribution shifts.

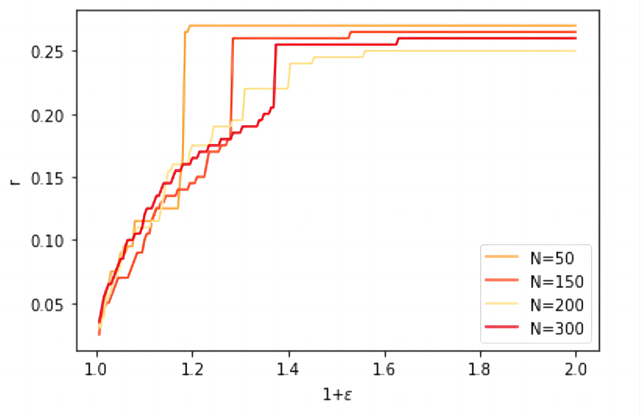

*Figure 3.* Correspondence between RS torelance rate parameter $\epsilon$ and DRO radius parameter $r$.

DRO in this type of problem. And we conduct experiments to compare their sensitivities to hyperparameter tuning.

### 5.3.1. Hyperparameter correspondence

Consider a Lipschitz loss function $L(\cdot)$. We reformulate the general Lipschitz-loss learning problem following the DRO literature (Blanchet et al., 2019; Shafieezadeh-Abadeh et al., 2019):

$$\min_{x} \frac{1}{N}\sum_{i=1}^{N} L(y_i - u_i \cdot x) + r \cdot ||x||_2. \tag{17}$$

Building on the reformulation presented in Section 2.2, this Lipschitz-loss learning problem under the RS framework is equivalent to:

$$\begin{aligned} &\min_{x} \; ||x||_2, \\ &\text{s.t.} \; \frac{1}{N}\sum_{i=1}^{N} L(y_i - u_i \cdot x) \le \tau. \end{aligned} \tag{18}$$

By applying the Lagrangian method to solve Equation (18) and with the strong duality held, we further deduce that Equation (18) simplifies to the following expression (see Appendix C.1 for the proof):

$$\sup_{\lambda>0} \inf_{x} \frac{\frac{1}{N}\sum_{i=1}^{N} L(y_i - u_i \cdot x) + \lambda \cdot ||x||_2 - \tau}{\lambda}. \tag{19}$$

We immediately observe a clear link between RS and DRO for the general Lipschitz-loss learning problem: given a reference value $\tau$, solving the RS model (19) yields the optimizer $(\hat{\lambda}_N, \hat{x}_N)$. Then, by setting the radius $r$ in the DRO model (17) to $\hat{\lambda}_N$, the DRO model (17) generates the same optimizer for the model parameter $x$. Recall that the reference value is set to be $\tau_\epsilon = (1+\epsilon)\inf_x \mathbb{E}_{\hat{P}_N}[h(x,\xi)]$ throughout this paper, where $\epsilon$ is the tolerate rate that controls the robustness of RS. Thus each hyperparameter $\epsilon$ in the RS model is associated with a specific radius $r(\epsilon)$, the hyperparameter in the DRO model.

Figure 3 illustrates the relationship between the robustness-controlling hyperparameters of the two models: the DRO radius $r$ and the RS torelance rate $\epsilon$. Notably, the function $r(\epsilon)$ is concave with respect to $\epsilon$, flatting as $1+\epsilon$ nears $1.6$. This indicates that, to achieve comparable performance, the RS model can accommodate larger variations in $\epsilon$ compared to variations in $r$ for the DRO model, implying that RS is less sensitive to hyperparameter tuning. This observation will be further supported by the following experiments.

### 5.3.2. Numerical sensitivity analysis

We now conduct experiments to evaluate the sensitivity of RS and DRO to hyperparameter tuning. Specifically, we vary the tolerance rate $\epsilon$ in the RS model and the radius $r$ in the DRO model. We set the model parameter in the sampling distribution to $x^* = [2.0, -1.0]^T$, as in Section 5.2; and set the target environment to be $\tilde{x} = [1.80, -0.90]^T$.

Figure 4 shows that DRO outperforms the baseline for radius smaller than $0.24$, with the optimal $r$ around $0.215$. Figure 5 shows that RS exceeds the baseline when $1+\epsilon$ is below $1.4$, with the optimal $1+\epsilon$ around $1.285$. The smallest MSEs from both DRO and RS models are comparable.

However, there is a drastic MSE surge in response to changes of $r$ for DRO in Figure 4, in contrast to the milder variation of MSE to $\epsilon$ for RS showed in Figure 5. This difference in hyperparameter sensitivity aligns with the nuanced relationship between $r$ and $\epsilon$ depicted in Figure 3. In particular, within a 15% relative error range around the optimal hyperparameters, the MSE for DRO may spike to

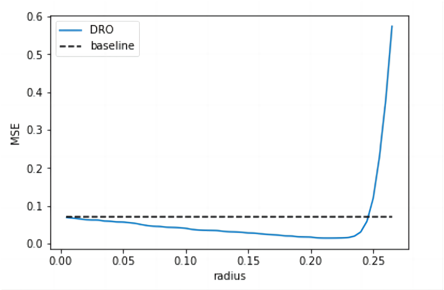

*Figure 4.* DRO performance under distribution shifts. DRO model shows higher sensitivity to haperparameter $r$.

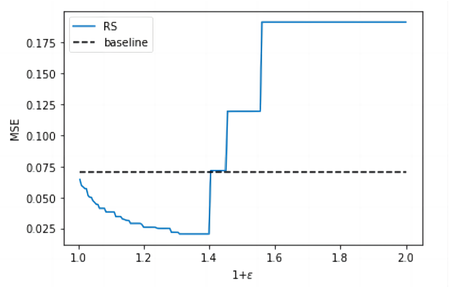

*Figure 5.* RS performance under distribution shifts. RS model shows lower sensitivity to hyperparameter $\epsilon$.

$0.30$, whereas RS maintains a more stable MSE of $0.125$. This suggests that RS offers greater flexibility in setting the tolerate rate hyperparameter $\epsilon$, unlike DRO, which requires more precise tuning for the radius $r$.

## 6. Conclusions

This paper focuses on exploring the statistical properties of the RS model, a recent robust optimization framework introduced in (Long et al., 2023). We provide theoretical guarantees for the RS model, including two-sided confidence intervals for the optimal loss and finite-sample generalization error bounds. These guarantees extend to scenarios involving distribution shifts, highlighting the RS model's robust generalization performance. Our numerical experiments reveal the superiority of the RS model compared to the baseline empirical risk minimization method, particularly in small-sample regimes and under distribution shifts. We establish explicit connections between the RS and DRO frameworks within specific models, showcasing that RS exhibits lower sensitivity to hyperparameter tuning than DRO, making it a more practical and interpretable choice. Future research directions include extending our analysis of RS to other distribution distances such as $f$-divergence, proving lower bounds for generalization error, and applying RS to various practical applications.

## Acknowledgement

We would like to thank the reviewers for their constructive feedback, which has greatly improved the current version of our work. Yunbei Xu acknowledges support from ARO through award W911NF-21-1-0328 and from the Simons Foundation and NSF through award DMS-2031883. Ruohan Zhan is partly supported by the Guangdong Provincial Key Laboratory of Mathematical Foundations for Artificial Intelligence (2023B1212010001) and the Project of Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone (HZQB-KCZYB-2020083).

## Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

## References

Bayraksan, G. and Love, D. K. Data-driven stochastic programming using phi-divergences. In *The operations research revolution*, pp. 1–19. INFORMS, 2015.

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.

Chen, S. and Chen, Y. Designing a resilient supply chain network under ambiguous information and disruption risk. *Computers & Chemical Engineering*, 179:108428, 2023.

Deng, M., Bian, B., Zhou, Y., and Ding, J. Distributionally robust production and replenishment problem for hydrogen supply chains. *Transportation Research Part E: Logistics and Transportation Review*, 179:103293, 2023.

Esfahani, P. M. and Kuhn, D. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *arXiv preprint arXiv:1505.05116*, 2015.

Fournier, N. and Guillin, A. On the rate of convergence in wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3-4):707–738, 2015.

Hu, Z. and Hong, L. J. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 1(2):9, 2013.

Huang, H., Li, Z., Gooi, H. B., Qiu, H., Zhang, X., Lv, C., Liang, R., and Gong, D. Distributionally robust energy-transportation coordination in coal mine integrated energy systems. *Applied Energy*, 333:120577, 2023.

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.

Li, Y., Han, M., Shahidehpour, M., Li, J., and Long, C. Data-driven distributionally robust scheduling of community integrated energy systems with uncertain renewable generations considering integrated demand response. *Applied Energy*, 335:120749, 2023.

Liu, F., Chen, Z., and Wang, S. Globalized distributionally robust counterpart. *INFORMS Journal on Computing*, 2023.

Long, D. Z., Sim, M., and Zhou, M. Robust satisficing. *Operations Research*, 71(1):61–82, 2023.

Ronchetti, E. The main contributions of robust statistics to statistical science and a new challenge. *Metron*, 79(2): 127–135, 2021.

Ruan, H., Zhou, S., Chen, Z., and Ho, C. P. Robust satisficing mdps. In *International Conference on Machine Learning*, pp. 29232–29258. PMLR, 2023.

Saday, A., Yıldırım, Y. C., and Tekin, C. Robust bayesian satisficing. *arXiv preprint arXiv:2308.08291*, 2023.

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

Sim, M. The analytics of robust satisficing: predict, optimize, satisfice, then fortify. In *Book of Abstracts*, pp. 42, 2023.

Sim, M., Zhao, L., and Zhou, M. Tractable robust supervised learning models. *Available at SSRN 3981205*, 2021.

Wang, D., Yang, K., and Yang, L. Risk-averse two-stage distributionally robust optimisation for logistics planning in disaster relief management. *International Journal of Production Research*, 61(2):668–691, 2023.

## A. Key Propositions

**Proposition 1.** *For any distributions $\mathbb{Q}_1$, $\mathbb{Q}_2 \in \mathcal{M}(\Xi)$, we have*

$$d_{\mathrm{W}}\big(\mathbb{Q}_1, \mathbb{Q}_2\big) = \sup_{f \in \mathcal{L}} \Big\{ \int_{\Xi} f(\xi)\, \mathbb{Q}_1(\mathrm{d}\xi) - \int_{\Xi} f(\xi)\, \mathbb{Q}_2(\mathrm{d}\xi) \Big\}, \tag{20}$$

*where $\mathcal{L}$ denotes the space of all Lipschitz functions with $|f(\xi) - f(\xi')| \leq \|\xi - \xi'\|$ for all $\xi, \xi' \in \Xi$. and $||\cdot||$ is a norm*

This is the dual representation of Wasserstein distance, and it needs the Assumption 2.

We will also utilize the bound of Wasserstein distance between $P^*$ and $\hat{P}_N$, which is presented below.

**Proposition 2.** *(Fournier & Guillin, 2015) If Assumption1 holds, we have*

$$P^N\Big\{d_W\big(P^*, \hat{P}_N\big) \geq r\Big\} \leq \begin{cases} c_1 \exp\big(-c_2 N r^{\max\{m,2\}}\big) & if\ r \leq 1, \\ c_1 \exp\big(-c_2 N r^{a}\big) & if\ r > 1, \end{cases}$$

*for all $N \geq 1$, the dimension of $\xi$ : $m \neq 2$ and $r > 0$, where $c_1, c_2$ are positive constants that only depend on $a$ and $m$.*

## B. Proof of Lemmas and Theorems

### B.1. Proof of Lemma 1.

Choose $x = \hat{x}_N$, we have

$$\mathbb{E}_P[h(\hat{x}_N, \xi)] - \tau \leq k_\tau d_W(P, \hat{P}_N) \quad \forall P \in \mathbb{P}. \tag{21}$$

Moreover, by the definition of $k_\tau$, for any $\delta > 0$, we can choose one distribution $P_1$, which satisfies:

$$\mathbb{E}_{P_1}[h(\hat{x}_N, \xi)] - \tau \geq (k_\tau - \delta) d_W(P_1, \hat{P}_N). \tag{22}$$

Then (22) minus (21), we have:

$$\begin{aligned} (k_\tau - \delta) d_W(P_1, \hat{P}_N) - k_\tau d_W(P, \hat{P}_N) & \qquad (23)\\ \leq \mathbb{E}_{P_1}[h(\hat{x}_N, \xi)] - \mathbb{E}_P[h(\hat{x}_N, \xi)] & \qquad (24)\\ \leq L \cdot d_W(P_1, P), & \qquad (25) \end{aligned}$$

for all $\delta > 0$ and $P \in \mathbb{P}$, Where the second inequality utilizes (20). Due to the arbitrariness of $\delta$, we can ignore the terms of $\delta$ and choose $\hat{P}_N$ as $P$ in (23), we get:

$$k_\tau d_W(P_1, \hat{P}_N) \leq L \cdot d_W(P_1, \hat{P}_N). \tag{26}$$

If $d_W(P_1, \hat{P}_N) > 0$, we will complete the proof.

Now we explain why $d_W(P_1, \hat{P}_N) > 0$ holds. Actually, $d_W(P_1, \hat{P}_N) = 0$ if and only if $P_1 = \hat{P}_N$, a.s. But if $P_1 = \hat{P}_N$, then (22) leads to that $\mathbb{E}_{\hat{P}_N}[h(\hat{x}_N, \xi)] \geq \tau$. Meanwhile, we can also choose $P = \hat{P}_N$ in (21) and we find $\mathbb{E}_{\hat{P}_N}[h(\hat{x}_N, \xi)] \leq \tau$. So $\mathbb{E}_{\hat{P}_N}[h(\hat{x}_N, \xi)] = \tau$. But it is impossible to hold because we can change $\epsilon$ in $\tau = \tau_\epsilon$ randomly. ■

### B.2. Proof of Theorem 1.

For the right side , $J^* = \mathbb{E}_{P^*}[h(x^*, \xi)] \leq \mathbb{E}_{P^*}[h(\hat{x}_N, \xi)] \leq k_{\tau_\epsilon} \cdot d_W(P^*, \hat{P}_N) + \tau_\epsilon$, where the second inequality is derived from the model (2) itself and the fact that $k_\tau(\hat{x}_N) = k_\tau$.
For the other side, by the definition of $J^*$, for any $\eta > 0$, choose $x_\eta$ satisfies:

$$\mathbb{E}_{P^*}[h(x^\eta, \xi)] \leq J^* + \eta. \tag{27}$$

Then

$$
\begin{aligned}
\tau_\epsilon &= (1+\epsilon)\inf_x \mathbb{E}_{\hat{P}_N}[h(x,\xi)] && (28)\\
&\le (1+\epsilon)\mathbb{E}_{\hat{P}_N}[h(x_\eta,\xi)] && (29)\\
&\le (1+\epsilon)\Big[L\cdot d_W(P^*,\hat{P}_N) + \mathbb{E}_{P^*}[h(x_\eta,\xi)]\Big] && (30)\\
&\le (1+\epsilon)L\cdot d_W(P^*,\hat{P}_N) + (1+\epsilon)(J^*+\eta), && (31)
\end{aligned}
$$

for all $\eta > 0$. The second inequality is derived from (20) and the third inequality uses (27). Due to the arbitrariness of $\eta$, we have $\tau_\epsilon \le (1+\epsilon)L\cdot d_W(P^*,\hat{P}_N) + (1+\epsilon)J^*$, which can be solved:

$$
-L\cdot d_W(P^*,\hat{P}_N) + \frac{\tau_\epsilon}{1+\epsilon} \le J^*.
$$

For (9), we can simply utilize (2) :

$$
P\Big\{-L\cdot r_N + \frac{\tau_\epsilon}{1+\epsilon} \le J^* \le k_{\tau_\epsilon}\cdot r_N + \tau_\epsilon\Big\} \ge P\Big\{d_W\big(P^*,\hat{P}_N\big) \le r_N\Big\} \ge 1-\beta_N, \qquad (32)
$$

where $\beta_N$ is given in (9).

Then we denote events $A_N = \Big\{-L\cdot r_N + \frac{\tau_\epsilon}{1+\epsilon} \le J^* \le k_{\tau_\epsilon}\cdot r_N + \tau_\epsilon\Big\}$. Then $P(A_N^c) \le \beta_N$. Because $\sum_{N=1}^\infty \beta_N < \infty$, we use Borel-Cantelli Lemma and the limit supremum of the sequence of events satisfies:

$$
P(\lim_{N\to\infty}\sup A_N^c) = 0. \qquad (33)
$$

So $P(\lim_{N\to\infty}\inf A_N) = 1$, which implies the consistency.

Finally, if $\lim_{N\to\infty} r_N(\beta_N) = 0$, we can take $N\to\infty$ and obtain (11). ■

### B.3. Proof of Theorem 3.

Utilize Theorem 1 and Remark 1, we can obtain that:

$$
\begin{aligned}
J^* \le A_N &\le k_{\tau_\epsilon}\cdot d_W(P^*,\hat{P}_N) + \tau_\epsilon\\
&\le L\cdot d_W(P^*,\hat{P}_N) + \tau_\epsilon\\
&\le (2+\epsilon)\cdot L\cdot d_W(P^*,\hat{P}_N) + (1+\epsilon)\cdot J^*,
\end{aligned}
$$

where the second inequality utilizes Lemma 1.

Then we have the similar step to (32):

$$
P\Big\{J^* \le A_N \le (1+\epsilon)\cdot J^* + (2+\epsilon)\cdot L\cdot r_N\Big\} \ge P\Big\{d_W\big(P,\hat{P}_N\big) \le r_N\Big\} \ge 1-\beta_N.
$$

Then we take the expectation on data and we have:

$$
\begin{aligned}
\mathbb{E}_{\sim P_{data}}[d_W(P^*,\hat{P}_N)] &= \int_0^\infty P_{data}\Big\{d_W\big(P,\hat{P}_N\big) \ge r\Big\}dr\\
&\le \int_0^1 c_1\exp\big(-c_2 N r^{\max\{m,2\}}\big)dr + \int_1^\infty c_1\exp\big(-c_2 N r^a\big)dr\\
&= N^{-\frac{1}{\max\{m,2\}}}\int_0^{N^{\frac{1}{\max\{m,2\}}}} c_1\exp\big(-c_2 t^{\max\{m,2\}}\big)dt + N^{-\frac{1}{a}}\int_{N^{\frac{1}{a}}}^\infty c_1\exp\big(-c_2 t^a\big)dt\\
&\le N^{-\frac{1}{\max\{m,2\}}}\int_0^\infty c_1\exp\big(-c_2 t^{\max\{m,2\}}\big)dt + N^{-\frac{1}{a}}\int_0^\infty c_1\exp\big(-c_2 t^a\big)dt\\
&= O(N^{-\min\{\frac{1}{m},\frac{1}{a},\frac{1}{2}\}})).
\end{aligned}
$$

The first inequality is derived from (2) and we utilize the convergence of two exponential integrals. Finally, we have:Here $a$ can be removed: When $a > 2$ satisfies Assumption 1, it can be weaken for $a = 2$; when $a < 2$ in Assumption 1, here we have $\frac{1}{a} > \frac{1}{2}$, so $\frac{1}{a}$ can be omitted due to minimum. The result should be $O(N^{-\min\{\frac{1}{m},\frac{1}{2}\}})$.

$$\begin{aligned} J^* \leq \mathbb{E}_{\sim P_{data}} A_N &\leq (2+\epsilon)\cdot \mathbb{E}_{\sim P_{data}}[d_W(P^*, \hat{P}_N)] + (1+\epsilon)\cdot J^* \\ &= (1+\epsilon)\cdot J^* + O(\frac{1}{N^\eta}). \end{aligned}$$

■

### B.4. Proof of Theorem 5.

The proof here is very straightforward following the proof of Theorem 1 and 3. Combine with the formula (12) that was proven earlier, we can easily obtain similar result:

$$-L\cdot d_W(\tilde{P}, \hat{P}_N) + \frac{\tau_\epsilon}{1+\epsilon} \leq \tilde{J} \leq \mathbb{E}_{\tilde{P}}[h(\hat{x}_N, \xi)] \leq k_{\tau_\epsilon}\cdot d_W(\tilde{P}, \hat{P}_N) + \tau_\epsilon. \tag{34}$$

Then we use the triangle inequality of distance: $d_W(\tilde{P}, \hat{P}_N) \leq d_W(P^*, \hat{P}_N) + d_W(\tilde{P}, P^*)$ and we get the extra term $d_W(\tilde{P}, P^*)$. And for the term $d_W(P^*, \hat{P}_N)$, continue to use the tail probability (2) to get guarantees for various probabilities and expected values. ■

## C. Supplementary results for Section 5

### C.1. Derivation of Equivalent Models in Section 5.1

**Proof of (17).** For convenience, let $x_a$ denote the augmented parameter vector $(-x, 1)^T$. Consider Lipschitz loss $h(x, \xi) = L(x_a \cdot \xi)$. For DRO, under mild assumptions, Esfahani & Kuhn have shown that:

$$\max_{P\in\{P:d(P,\hat{P}_N)\leq r\}} E_P[h(x,\xi)] = \inf_{\lambda\geq 0} \lambda r + \frac{1}{N}\sum_{i=1}^{N}\sup_{\xi}(h(x,\xi) - \lambda c(\xi,\xi_i)). \tag{35}$$

Next, denote $\Delta = u - u_i$, Utilizing the proof given by Shafieezadeh-Abadeh et al., we can obtain:

$$\begin{aligned} &\sup_{\xi}(h(x,\xi) - \lambda c(\xi,\xi_i)) = \sup_{\xi}(L(x_a\cdot\xi) - \lambda c(\xi,\xi_i)) \; = \sup_{u}(L(x\cdot u - y_i) - \lambda||u-u_i||_2) \\ &= \sup_{\Delta}(L((u_i+\Delta)\cdot x - y_i) - \lambda||\Delta||_2) = \left\{ \begin{array}{ll} L(u_i\cdot x - y_i) & if \lambda \geq ||x||_2, \\ +\infty & otherwise. \end{array} \right. \end{aligned}$$

The second equality here utilizes the definition of our fine-tuned cost function: if $y$ in $\xi$ and $y_i$ in $\xi_i$ are not equal, the distance will become $\infty$, thereby making the entire expression $-\infty$. Therefore, only the distance of the feature variable $u$ is retained.

Back to (35), we have:

$$\inf_{\lambda\geq 0}\lambda r + \frac{1}{N}\sum_{i=1}^{N}\sup_{\xi}(h(x,\xi) - \lambda c(\xi,\xi_i)) = \inf_{\lambda\geq||x||_2}\lambda r + \frac{1}{N}\sum_{i=1}^{N}L(u_i\cdot x - y_i) = r||x||_2 + \frac{1}{N}\sum_{i=1}^{N}L(u_i\cdot x - y_i).$$

Then we have

$$\min_{x\in\mathcal{X}}\max_{P\in\{P:d(P,\hat{P}_N)\leq r\}} E_P[h(x,\xi)] = \min_{x}\frac{1}{N}\sum_{i=1}^{N}L(y_i - u_i\cdot x) + r\cdot||x||_2. \tag{36}$$

■

**Proof of (18).** We have already given the reformulation of the RS model(7) in Section 2.2. The constraint condition is:

$$\frac{1}{N}\sum_{i=1}^{N}[\sup_{z\in\Xi} h(x,z) - kc(\xi_i, z)] \leq \tau. \tag{37}$$

The proof will follow the proof of (17):

$$\frac{1}{N}\sum_{i=1}^{N}[\sup_{z\in\Xi} h(x,z) - kc(\xi_i,z)] = \begin{cases} \frac{1}{N}\sum_{i=1}^{N} L(u_i\cdot x - y_i) & if k \geq ||x||_2, \\ +\infty \quad otherwise. \end{cases}$$

Since $\tau$ serves as the upper bound in (37), to ensure that the left side of (37) is not infinite, it is necessary to satisfy $k \geq ||x||_2$. Therefore, in (7), taking the minimum value of $k$ is equivalent to minimizing $||x||_2$ and let $k = \min_x ||x||_2$, while satisfying the constraint condition $\frac{1}{N}\sum_{i=1}^{N} L(u_i\cdot x - y_i) \leq \tau$. ■

**Proof of (19).** Following (18), we can express it in its dual form:

$$\inf_x \sup_{\lambda>0} ||x||_2 + \lambda\cdot(\frac{1}{N}\sum_{i=1}^{N} L(y_i - u_i\cdot x) - \tau).$$

Notably, this equation represents a convex problem. As long as $\tau > \min_x \frac{1}{N}\sum_{i=1}^{N} L(y_i - u_i\cdot x)$ which is mentioned in (4), there exists a point in the relative interior, hence the Slater's strong duality condition holds. Subsequently, to facilitate comparative analysis with DRO, we replace $\lambda$ with $\frac{1}{\lambda}$ to obtain:

$$\inf_x \sup_{\lambda>0} \frac{\frac{1}{N}\sum_{i=1}^{N} L(y_i - u_i\cdot x) + \lambda\cdot||x||_2 - \tau}{\lambda}.$$

Finally, note that the above equation is convex with respect to $x$ and concave with respect to $\lambda$. According to the Mini-max theorem, we can interchange the order of sup and inf to obtain the desired formula. ■

### C.2. Other Equivalent Conclusions

This section answers the question mentioned in the previous footnote. If we still use the $l_2$ norm of the entire vector as the cost function i.e. $c(\xi_1,\xi_2) = ||\xi_1 - \xi_2||_2$, then the DRO model will be equivalent to:

$$\min_{x\in\mathcal{X}} \max_{P\in\{P:d(P,\hat{P}_N)\leq r\}} E_P[h(x,\xi)] = \min_x \frac{1}{N}\sum_{i=1}^{N} L(y_i - u_i\cdot x) + r\cdot||x_a||_2,$$

where $x_a$ is the augmented vector $(x,-1)^T$. Similarly, RS model is equilvalent to:

$$\begin{aligned} &\min_x \ ||x_a||_2, \\ &\text{s.t. } \frac{1}{N}\sum_{i=1}^{N} L(y_i - u_i\cdot x) \leq \tau. \end{aligned}$$

And we can also write its dual equivalent form as:

$$\sup_{\lambda>0}\inf_x \frac{\frac{1}{N}\sum_{i=1}^{N} L(y_i - u_i\cdot x) + \lambda\cdot||x_a||_2 - \tau}{\lambda}.$$

Therefore, after modifying the definition of the cost function, the only difference is whether one term in the model is the $l_2$ norm of the parameter $x$ itself or the $l_2$ norm of the augmented vector $(x,-1)^T$ obtained by adding an element 1.

### C.3. Function $r(\epsilon)$ in Ten Dimensions

In Section 5.1, we set the feature variable to be ten-dimensional. As a supplement to Section 5.2 , we also plot the $\tau - \epsilon$ relationship graph under the ten-dimensional situation of the feature variable.

Figure 6 shows similar relationship between the robustness-controlling hyperparameters of the two models: the DRO radius $r$ and the RS tolerance rate $\epsilon$. The overall trend of the graph is concave. It tends to flatten when $1+\epsilon = 1.4$. Moreover, when the function tends to be flat, the corresponding radius value $r$ is smaller than that in the two-dimensional case in Figure 3.

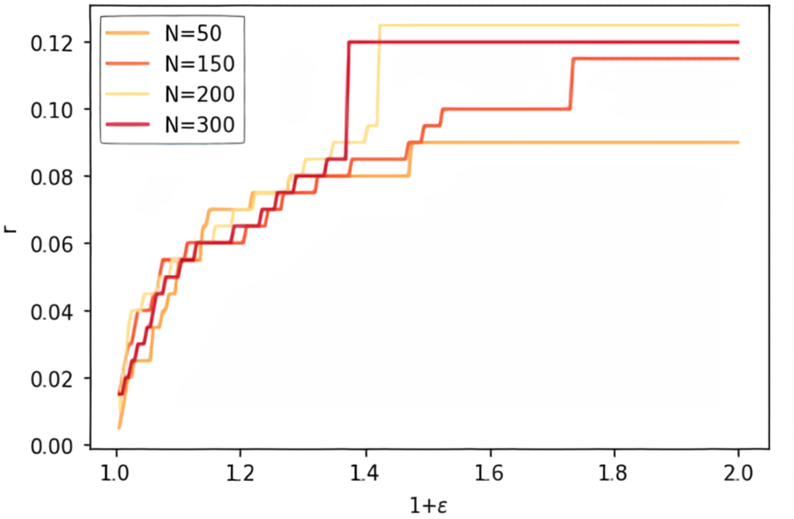

*Figure 6.* function $r$-$\epsilon(m_u = 10)$

## D. Some Classical Loss Functions

Here we present a few common loss functions.

| | L(z) | Classification(C) or Regression(R) | Learning Model |
|---|---|---|---|
| **Hinge Loss** | $\max\{0, 1-z\}$ | C | SVM |
| **Smooth Hinge Loss** | $\begin{cases} \frac{1}{2} - z & if z \leq 0 \\ \frac{1}{2}(1-z)^2 & if 0 < z < 1 \\ 0 & z \geq 1 \end{cases}$ | C | smooth SVM |
| **Logloss** | $\log(1+e^{-z})$ | C | Logistic Regression |
| **Squared Loss** | $z^2$ | R | MSE |
| $L_1$ **Loss** | $\|z\|$ | R | MAE |
| **Huber Loss** | $\begin{cases} \frac{1}{2}z^2 & if \|z\| \leq \delta \\ \delta(\|z\| - \frac{1}{2}\delta) & otherwise \end{cases}$ | R | Huber regression |
| $\delta$**-insensitive Loss** | $\max\{0, \|z\| - \delta\}$ | R | support vector regression |
| **Pinball Loss** | $\max\{-\delta z, (1-\delta)z\}$ | R | Quantile regression |

*Table 2.* Some Classical Loss Functions

Here we consider $h(x, \xi)$ as a loss function for machine learning applications, where $x$ denotes the parameters in the classification or regression model, and $\xi = (\xi^f, \xi^l)^T$ represents the data with $\xi^f$ as the feature variable and $\xi^l$ as the label variable. For binary classification problems, the loss function can be defined as follows:

$$h(x, \xi) = L(\xi^l \cdot x^T \xi^f). \tag{38}$$

For regression problems, the loss function can be defined as follows:

$$h(x, \xi) = L(\xi^l - x^T \xi^f). \tag{39}$$

Here we present a few common loss functions (See Table 2). Apart from the squared loss, all other loss functions in Table 2 are Lipschitz, so our Assumption 2 is relatively weak and reasonable. Furthermore, in practical applications, $x$ and $\xi$ are often bounded, so even if we use squared loss, it is Lipschitz in the case of a bounded domain.

## E. Discussion on Lipschitz continuity Assumption

In our paper, we do not need to assume the Lipschitz continuity of the loss function with respect to $x$, but with respect to $\xi$. This assumption follows that of (Esfahani & Kuhn, 2015), with the aim of using the inequality in Proposition 1. We however understand that it is a common condition to assume the Lipschitz continuity of parameter $x$. In response to this, we provide a conservative answer: at least for the regression problem in Appendix D where $h(x, \xi) = L(\xi^l - x^T \cdot \xi^f)$, if both the random variable space $\Xi$ and the parameter space $\mathcal{X}$ are bounded, then as long as we assume that $L(\cdot)$ is a Lipschitz function, it can be simultaneously derived that $h(x, \xi)$ is Lipschitz with respect to both $x$ and $\xi = (\xi^f, \xi^l)$. In such scenarios, the Lipschitz assumptions for $x$ and $\xi$ hold simultaneously.