Statistical Properties of Robust Optimization under Distribution Shifts

Zhiyi Li[1] **Xiaojie Mao**[2] **Yunbei Xu**[1] **Ruohan Zhan**[3]
[1]National University of Singapore [2]Tsinghua University [3]University College London
e1632488@u.nus.edu maoxj@sem.tsinghua.edu.cn yunbei@nus.edu.sg ruohan.zhan@ucl.ac.uk

Abstract

Distributional shifts commonly arise in practice when the target environment differs from the source environment that provides training data. Robust learning frameworks such as Distributionally Robust Optimization (DRO) and Robust Satisficing (RS) have been developed to address this challenge, yet their theoretical behavior under such shifts remains insufficiently understood. This paper analyzes their performance under distributional shifts measured by the Wasserstein distance, focusing on the generalization error defined as the excess loss in the target environment. We derive the first generalization error bounds that explicitly characterize how DRO and RS balance improved robustness in the target environment with the regularization cost of robustness, while avoiding the curse of dimensionality. When partial shift information such as magnitude or direction is available, we conduct a systematic comparison of both methods and provide theory-based guidelines for selecting between them, supported by simulation results. Finally, we demonstrate the practical relevance of our framework through an application to a network lot-sizing problem. This work fills theoretical gaps in robust learning under distributional shifts and provides practical guidance for algorithm design.

1. Research Problem Machine learning systems often face distribution shifts between training and deployment, arising from covariate or label shifts, non-stationary environments, or domain-specific variations. These shifts can severely degrade models trained via Empirical Risk Minimization (ERM), the widely adopted baseline that minimizes empirical loss under the source distribution [Sinha et al., 2018, Ben-David et al., 2010].

To address the problem caused by distribution shifts, one widely studied paradigm is Distributionally Robust Optimization (DRO). Formally, DRO solves

$$\min_{f\in\mathcal{F}} \max_{P\in\mathbb{P}_r} \mathbb{E}_P[\ell(f,z)], \quad \text{where } \mathbb{P}_r = \{P : d_W(P,\hat{P}_n) \le r\}, \tag{1}$$

where $\ell(f,z)$ denotes the loss function with f as the objective function and z a random variable. The radius r defines the Wasserstein ball around the empirical distribution $\hat{P}_n$ and reflects a fundamental trade-off: enlarging r increases robustness to shifts but simultaneously induces conservatism and higher optimization cost. Existing theoretical results analyze vanishing-radius asymptotics or adversarial settings [Esfahani and Kuhn, 2015, Blanchet et al., 2019], but they do not characterize this r-induced trade-off or provide guarantees under genuine distribution shifts.

A parallel robust approach is Robust Satisficing (RS) proposed by Long et al. [2023]. Formally, RS solves

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

$$
\begin{aligned}
k_\tau = \min \quad & k \\
\text{s.t. } & \mathbb{E}_P[\ell(f,z)] - \tau \le k d(P, \hat{P}_n), \quad \forall P \in \mathbb{P} \\
& f \in \mathcal{F}, \quad k \ge 0.
\end{aligned} \tag{2}
$$

where the hyperparameter τ represents a reference value corresponding to tolerable performance. RS achieves robustness via a satisficing strategy: loss deviations beyond τ are constrained to scale proportionally with the distributional distance. Most prior work on RS has focused on optimization and reformulation aspects. From a statistical perspective, Li et al. [2024] established the first theoretical characterization under distribution shifts, but their results suffer from the curse of dimensionality and fail to capture the τ-trade-off mechanism, namely the balance between capturing target loss and sacrificing empirical performance for robustness.

In this work we aim to theoretically characterize the generalization error bounds of DRO and RS under distributional shifts and to highlight their distinct mechanisms for handling such shifts. These bounds make explicit the trade-off inherent in setting robustness hyperparameters: increasing robustness improves protection against shifts but also raises optimization costs, either through additional regularization in DRO or satisficing penalties in RS. Building on these results, we further analyze how partial knowledge of the shift (e.g., its direction or magnitude) influences the relative performance of DRO and RS, providing actionable guidance for method selection in practice.

2. Key Methodology and Assumptions One of the key methodologies of this work is the distribution shift setting. We assume that models are trained on a source distribution P^* and evaluated on a different target distribution $\tilde{P}$. Let $f \in \mathcal{F}$ denote the predictive function, and let $z \in \mathcal{Z} \subseteq \mathbb{R}^d$ be a random variable. The loss function $\ell(f,z)$, following Vapnik's general learning setting [Vapnik and Vapnik, 1998], measures the prediction quality of f on z.

We denote the minimal achievable risks under the two distributions as

$$
J^* := \inf_{f \in \mathcal{F}} \mathbb{E}_{P^*}[\ell(f,z)], \quad \tilde{J} := \inf_{f \in \mathcal{F}} \mathbb{E}_{\tilde{P}}[\ell(f,z)]. \tag{3}
$$

For an estimator $\hat{f} \in \mathcal{F}$ trained on n i.i.d. samples from P^*, we evaluate its performance on the target distribution by the excess risk

$$
\mathcal{R}_{\tilde{P}}(\hat{f}) := \mathbb{E}_{\tilde{P}}[\ell(\hat{f},z)] - \tilde{J}. \tag{4}
$$

This definition generalizes the classical notion of generalization error, which typically assumes $P^* = \tilde{P}$ and focuses only on overfitting within the same environment [Esfahani and Kuhn, 2015, Shafieezadeh-Abadeh et al., 2019, Gao, 2023]. It also differs fundamentally from adversarial formulations [Lee and Raginsky, 2018, An and Gao, 2021], which measure error under the worst-case distribution in a neighborhood of P^*. By focusing on the loss of source-trained optimizers under a shifted target, our setting directly captures the challenges posed by distributional shifts.

Below are mild assumptions required for our results.

Assumption 1 (Regularity)**.** *We assume:*

- *(a)* ***Bounded instance space.*** *The instance space $\mathcal{Z}$ is bounded:* $\operatorname{diam}(\mathcal{Z}) = \sup_{z,z' \in \mathcal{Z}} d(z,z') < \infty$.
- *(b)* ***Bounded functions.*** *The loss functions of $f \in \mathcal{F}$ are upper semicontinuous and uniformly bounded:* $0 \le l(f,z) \le M < \infty,\ \forall f \in \mathcal{F},\ z \in \mathcal{Z}$.
- *(c)* ***Lipschitz continuity of loss.*** *The functions in $\mathcal{F}$ are Lipschitz:*

$$
\sup_{z \neq z'} \frac{|\, l(f,z) - l(f,z') \,|}{d(z,z')} \le L, \ \ \forall f \in \mathcal{F}.
$$

3. Generalization Error Bounds We define the induced loss function class $\mathcal{A} := \{ z \mapsto \ell(f,z) \mid f \in \mathcal{F} \}$. To measure its complexity, we use the entropy integral $\mathcal{C}(\mathcal{A}) := \int_0^\infty \sqrt{\log \mathcal{N}(\mathcal{A}, \|\cdot\|_\infty, u)}\, du$, where $\mathcal{N}(\mathcal{A}, \|\cdot\|_\infty, u)$ denotes the ℓ_∞-norm covering number of $\mathcal{A}$ at radius u. Then for ERM, we have:

Proposition 1 (ERM, Generalization Upper Bound)**.** *If Assumptions 1 hold, then with probability at least* $1-\delta$*, we have*

$$\mathcal{R}_{\tilde{P}}(\hat{f}_{ERM}) \leq \underbrace{[J^* - \tilde{J}]}_{\textit{Common term}} + \underbrace{L \cdot d_W(P^*, \tilde{P})}_{\textit{Shift term}} + \underbrace{\tfrac{24}{\sqrt{n}}\mathcal{C}(\mathcal{A}) + 3M\sqrt{\tfrac{\log(2/\delta)}{2n}}}_{\textit{Statistical error}}, \quad \forall \tilde{P}. \tag{5}$$

Here the common term will exist in every generalization upper bound for robust methods below.

Proposition 2 (ERM, Lower Bound)**.** *Choose* $\ell(\theta, z) = |1 - z\theta|$ *with* $\theta \in [0,1]$*,* $d = 1$ *and* $P^* = \delta(1)$*. For any target distribution* $\tilde{P}$*, we have*

$$\mathcal{R}_{\tilde{P}}(\hat{f}_{ERM}) \geq L \cdot d_W(P^*, \tilde{P}).$$

These results together imply that, under distribution shifts, the excess risk of ERM inevitably requires shift term $L \cdot d_W(P^*, \tilde{P})$.

We adopt the notation of Gao [2023] and define the DRO regularizer as $\Lambda_r(Q, f) := \sup_{P \in B(Q,r)} \mathbb{E}_P[\ell(f,z)] - \mathbb{E}_Q[\ell(f,z)]$, where $B(Q,r) := \{P \in \mathcal{P}(\mathcal{Z}) : d_W(P,Q) \leq r\}$ is the Wasserstein ball of radius r centered at Q. Then for DRO, we have:

Theorem 1 (Generalization error bound of DRO)**.** *If Assumptions 1 hold, then with probability at least* $1-\delta$*, we have*

$$\begin{aligned}\mathcal{R}_{\tilde{P}}(\hat{f}_{DRO}) \leq \underbrace{[J^* - \tilde{J}]}_{\textit{Common term}} + \underbrace{L \cdot \inf_{P \in B(P^*,r)} d_W(\tilde{P}, P)}_{\textit{Shift term}} + \underbrace{\Lambda_r(P^*, f^*)}_{\textit{Regularization term}} \\ + \underbrace{\tfrac{48}{\sqrt{n}}\mathcal{C}(\mathcal{A}) + \tfrac{48L \cdot \mathrm{diam}(\mathcal{Z})}{\sqrt{n}} + 3M\sqrt{\tfrac{\log(2/\delta)}{2n}}}_{\textit{Statistical error}}, \quad \forall \tilde{P}.\end{aligned} \tag{6}$$

The bound contains a shift term and a regularization term, which together reflect the trade-off induced by the hyperparameter r: small r may fail to capture the target distribution, while large r incurs regularization-driven degradation. It is important to note that such trade-off has not appeared in existing results. Moreover, when $r = 0$, DRO bound 1 reduces to ERM bound 1.

To ensure feasibility of RS, the reference value τ must be no smaller than the empirical loss. Motivated by this, we parameterize $\tau_\epsilon := \inf_{f \in \mathcal{F}} \mathbb{E}_{\hat{P}_n}[\ell(f,z)] + \epsilon$, with $\epsilon \geq 0$ controlling the satisficing margin. Next is one Lemma which shows an upper bound for optimizer k_{τ_ϵ}:

Lemma 1 (Li et al. [2024])**.** *If Assumption 1(c) holds, then*

$$k_{\tau_\epsilon} \leq L.$$

Theorem 2 (Generalization error bound of RS)**.** *If Assumptions 1(a), 1(b), and 1(c) hold, then with probability at least* $1-\delta$*, we have*

$$\begin{aligned}\mathcal{R}_{\tilde{P}}(\hat{f}_{RS}) \leq \underbrace{[J^* - \tilde{J}]}_{\textit{Common term}} + \underbrace{k_{\tau_\epsilon} \cdot d_W(P^*, \tilde{P})}_{\textit{Shift term}} + \underbrace{\epsilon}_{\textit{Satisficing term}} \\ + \underbrace{\tfrac{48}{\sqrt{n}}\mathcal{C}(\mathcal{A}) + \tfrac{48L \cdot \mathrm{diam}(\mathcal{Z})}{\sqrt{n}} + 3M\sqrt{\tfrac{\log(2/\delta)}{2n}}}_{\textit{Statistical error}}, \quad \forall \tilde{P}.\end{aligned} \tag{7}$$

The bound reveals a trade-off governed by ϵ: larger τ_ϵ decreases coefficient k_{τ_ϵ} and hence suppresses the shift term, but incurs a higher satisficing penalty ϵ. Notably, our bound avoids the curse of dimensionality, and when $\epsilon = 0$, the RS bound is no worse than that of ERM.

Comparing the two bounds 1 2, we reveal the different mechanisms of two robust models in handing distributional shifts: DRO mitigates distribution shifts by directly shrinking the distance, whereas RS reduces the coefficient of the shift distance.

4. Comparative Statics under Partial Shift Information We specify distributional shift information through its magnitude or direction. We quantify shift magnitude by the shift distance $d_W(P^*, \tilde{P})$, and capture shift direction by a distribution family $\{P_t\}_{t\geq 0}$. To ensure that t can be interpreted as the shift magnitude, we impose Assumption 2, which holds for many translation- or scaling-type distribution families:

Assumption 2 (Monotonicity)**.** $d_W(P_0, P_s) \leq d_W(P_0, P_t)$ *for all* $0 \leq s \leq t$*, where* $P_0 \equiv P^*$ *denotes the source distribution.*

We next demonstrate the hyperparameter selection procedure according to available shift information.

4.1. Scenario I: Known Magnitude, Unknown Direction Suppose the shift magnitude $d_W(P^*, \tilde{P})$ is known, while its direction is unspecified. For DRO, the natural choice is to set $r = d_W(P^*, \tilde{P})$. For RS, we set $\tau_r = \sup_{P\in\mathbb{P}_r} \mathbb{E}_P[\ell(\hat{f}_{ERM}, z)]$, reflecting anticipated degradation of the ERM optimizer under shifts within $\mathbb{P}_r$.

4.2. Scenario II: Known Direction, Unknown Magnitude Suppose the shift direction is known but the true magnitude t_2 is unknown. We pre-specify t (may be underspecified or overspecified) and set the DRO radius $r_t = d_W(P^*, P_t)$ and RS reference $\tau_t = \mathbb{E}_{P_t}[\ell(\hat{f}_{ERM}, z)]$. In this scenario, we specifically define the *adversarial direction* by selecting $P_t \in \arg\max_{P\in Arg(r_t)} d_W(P, P^*)$ as the adversarial scenario where $Arg(r_t) := \arg\max_{P\in B(P^*, r_t)} \mathbb{E}_P\big[\ell(f^*, z)\big]$.

Figure 1 summarizes the key takeaways and offers practical guidance for method adoption under partial shift information.

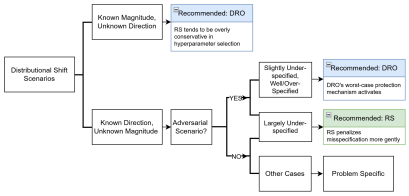
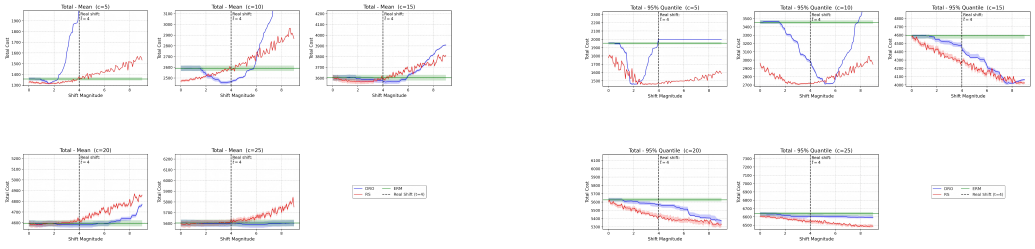


Figure 1: Comparison between DRO and RS under Different Distributional Shift Scenarios

5. Application to Network Lot-sizing We consider the network lot-sizing problem under distributional shifts in demand z. The optimization formulation and the robust counterparts for DRO and RS are provided in Appendix D.



(a) Mean of total costs across shift magnitude t

(b) 95% quantile of total costs across shift magnitude t

Figure 2: Mean and 95% quantile of total cost under distribution shift

We highlight the empirical performance in Figure 2. Regarding the mean in Figure 2a, RS exhibits superior performance under the underspecified setting and when the unit cost of initial ordering c is low. In contrast, under the overspecified setting, DRO gradually outperforms RS as c becomes higher, consistent with our theoretical analysis. As shown in Figure 2b, both RS and DRO yield improved results and RS yield more stable performance than DRO.

References

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. *ICLR*, 2018. URL `https://arxiv.org/abs/1710.10571`.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

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

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

Daniel Zhuoyu Long, Melvyn Sim, and Minglong Zhou. Robust satisficing. *Operations Research*, 71(1):61–82, 2023.

Zhiyi Li, Yunbei Xu, and Ruohan Zhan. Statistical properties of robust satisficing. *arXiv preprint arXiv:2405.20451*, 2024.

Vladimir Vapnik and Vlamimir Vapnik. Statistical learning theory wiley. *New York*, 1(624):2, 1998.

Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Peyman Mohajerin Esfahani. Regularization via mass transportation. *Journal of Machine Learning Research*, 20(103):1–68, 2019.

Rui Gao. Finite-sample guarantees for wasserstein distributionally robust optimization: Breaking the curse of dimensionality. *Operations Research*, 71(6):2291–2306, 2023.

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

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

Leonid Vasilevich Kantorovich and SG Rubinshtein. On a space of totally additive functions. *Vestnik of the St. Petersburg University: Mathematics*, 13(7):52–59, 1958.

Tong Zhang. *Mathematical analysis of machine learning algorithms*. Cambridge University Press, 2023.

Zhiyuan Wang, Lun Ran, Minglong Zhou, and Long He. On the equivalence and performance of distributionally robust optimization and robust satisficing models in om applications. *Available at SSRN 4455065*, 2023.

A Proofs of Main Results

A.1 Proof of Proposition 1

Proof. Let $\mathcal{A} := \{z \mapsto \ell(f,z) : f \in \mathcal{F}\}$ and recall Assumptions 1(b)–1(c). By the Kantorovich–Rubinstein duality (Proposition 3), for any L-Lipschitz function g we have

$$\left|\mathbb{E}_{\tilde{P}}[g(z)] - \mathbb{E}_{P^*}[g(z)]\right| \le L\, d_W(P^*, \tilde{P}).$$

Applying this to $g(\cdot) = \ell(\hat{f}_{ERM}, \cdot)$ yields

$$\mathbb{E}_{\tilde{P}}\big[\ell(\hat{f}_{ERM}, z)\big] \le \mathbb{E}_{P^*}\big[\ell(\hat{f}_{ERM}, z)\big] + L\, d_W(P^*, \tilde{P}). \tag{8}$$

We now control $\mathbb{E}_{P^*}[\ell(\hat{f}_{ERM}, z)]$ via a standard decomposition. Add and subtract $\mathbb{E}_{\hat{P}_n}[\cdot]$ terms and insert $f^* \in \arg\min_{f\in\mathcal{F}} \mathbb{E}_{P^*}[\ell(f,z)]$:

$$\begin{aligned}\mathbb{E}_{P^*}\big[\ell(\hat{f}_{ERM}, z)\big] = &\underbrace{\big(\mathbb{E}_{P^*}\ell(\hat{f}_{ERM}, z) - \mathbb{E}_{\hat{P}_n}\ell(\hat{f}_{ERM}, z)\big)}_{(I)} + \underbrace{\big(\mathbb{E}_{\hat{P}_n}\ell(\hat{f}_{ERM}, z) - \mathbb{E}_{\hat{P}_n}\ell(f^*, z)\big)}_{\le 0} \\ &+ \underbrace{\big(\mathbb{E}_{\hat{P}_n}\ell(f^*, z) - \mathbb{E}_{P^*}\ell(f^*, z)\big)}_{(II)} + \mathbb{E}_{P^*}\ell(f^*, z).\end{aligned}$$

The middle term is non-positive because $\hat{f}_{ERM}$ minimizes the empirical loss. Hence

$$\mathbb{E}_{P^*}\big[\ell(\hat{f}_{ERM}, z)\big] \le \sup_{f\in\mathcal{F}} \big(\mathbb{E}_{P^*}\ell(f,z) - \mathbb{E}_{\hat{P}_n}\ell(f,z)\big) + \big(\mathbb{E}_{\hat{P}_n}\ell(f^*,z) - \mathbb{E}_{P^*}\ell(f^*,z)\big) + J^*, \quad (9)$$

where $J^* = \inf_{f\in\mathcal{F}} \mathbb{E}_{P^*}\ell(f,z)$.

Define the (one-sided) empirical process

$$G_n := \sup_{f\in\mathcal{F}} \Big\{\mathbb{E}_{P^*}\ell(f,z) - \mathbb{E}_{\hat{P}_n}\ell(f,z)\Big\}.$$

By symmetrization and standard concentration, with probability at least $1-\delta/2$,

$$G_n \le 2\,\mathfrak{R}_n(\mathcal{A}) + M\sqrt{\frac{2\log(2/\delta)}{n}}, \quad (10)$$

where $\mathfrak{R}_n(\mathcal{A})$ is the empirical Rademacher complexity of $\mathcal{A}$. For the fixed function f^*, Hoeffding's inequality (using $0 \le \ell \le M$) implies that with probability at least $1-\delta/2$,

$$\big|\mathbb{E}_{\hat{P}_n}\ell(f^*,z) - \mathbb{E}_{P^*}\ell(f^*,z)\big| \le M\sqrt{\frac{\log(2/\delta)}{2n}}. \quad (11)$$

Combining (10) and (11) via a union bound, and noting that

$$\sup_{f\in\mathcal{F}} \big(\mathbb{E}_{P^*}\ell(f,z) - \mathbb{E}_{\hat{P}_n}\ell(f,z)\big) \le G_n,$$

we obtain, with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}} \big(\mathbb{E}_{P^*}\ell(f,z) - \mathbb{E}_{\hat{P}_n}\ell(f,z)\big) + \big(\mathbb{E}_{\hat{P}_n}\ell(f^*,z) - \mathbb{E}_{P^*}\ell(f^*,z)\big) \le 2\,\mathfrak{R}_n(\mathcal{A}) + 3M\sqrt{\frac{\log(2/\delta)}{2n}}. \quad (12)$$

Substitute (12) into (9), and then plug the result into (8):

$$\mathbb{E}_{\tilde{P}}\big[\ell(\hat{f}_{ERM}, z)\big] \le J^* + L\,d_W(P^*, \tilde{P}) + 2\,\mathfrak{R}_n(\mathcal{A}) + 3M\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Finally, by Dudley's entropy integral bound,

$$\mathfrak{R}_n(\mathcal{A}) \le \frac{12}{\sqrt{n}} \int_0^\infty \sqrt{\log \mathcal{N}\big(\mathcal{A}, \|\cdot\|_\infty, u\big)}\, du = \frac{12}{\sqrt{n}}\,\mathcal{C}(\mathcal{A}),$$

which yields, with probability at least $1-\delta$,

$$\mathbb{E}_{\tilde{P}}\big[\ell(\hat{f}_{ERM}, z)\big] \le J^* + L\,d_W(P^*, \tilde{P}) + \frac{24}{\sqrt{n}}\mathcal{C}(\mathcal{A}) + 3M\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Subtracting $\tilde{J} := \inf_{f\in\mathcal{F}} \mathbb{E}_{\tilde{P}}\ell(f,z)$ from both sides completes the claim:

$$\mathcal{R}_{\tilde{P}}(\hat{f}_{ERM}) := \mathbb{E}_{\tilde{P}}\big[\ell(\hat{f}_{ERM}, z)\big] - \tilde{J} \le \underbrace{(J^* - \tilde{J})}_{\text{Common term}} + \underbrace{L\,d_W(P^*, \tilde{P})}_{\text{Shift term}} + \underbrace{\frac{24}{\sqrt{n}}\mathcal{C}(\mathcal{A}) + 3M\sqrt{\frac{\log(2/\delta)}{2n}}}_{\text{Statistical error}}.$$

Proposition 3 (Kantorovich and Rubinshtein [1958])**.** *For any distributions $\mathbb{Q}_1, \mathbb{Q}_2 \in \mathcal{M}(\Xi)$, we have*

$$d_{\mathrm{W}}\big(\mathbb{Q}_1, \mathbb{Q}_2\big) = \sup_{f\in\mathcal{L}} \Big\{ \int_\Xi f(\xi)\,\mathbb{Q}_1(\mathrm{d}\xi) - \int_\Xi f(\xi)\,\mathbb{Q}_2(\mathrm{d}\xi)\Big\}, \quad (13)$$

where $\mathcal{L}$ denotes the space of all Lipschitz functions with $|f(\xi) - f(\xi')| \le \|\xi - \xi'\|$ for all $\xi, \xi' \in \Xi$.

A.2 Proof of Proposition 2

Proof. Consider the ℓ_1 loss $\ell(\theta, z) = |1 - z\theta|$ with $\Theta = [0, 1]$ and source distribution $P^* = \delta(1)$ ($\hat{P}_n = \delta(1)^{\otimes n}$). Then Lipschitz constant $L = 1$ and

$$\hat{\theta}_{ERM} \in \arg\min_{\theta} \mathbb{E}_{\hat{P}_n}|1 - Z\theta| = \arg\min_{\theta} |1 - \theta| = \{1\},$$

hence $\hat{\theta}_{ERM} = 1$. For any target distribution $\tilde{P}$ on $\mathbb{R}$,

$$\mathbb{E}_{\tilde{P}}\big[\,|1 - Z\hat{\theta}_{ERM}|\,\big] = \mathbb{E}_{\tilde{P}}\big[\,|Z - 1|\,\big].$$

For the 1-Wasserstein distance with ground cost $c(z, z') = |z - z'|$, any coupling $\pi \in \Pi(\tilde{P}, \delta(1))$ must satisfy $\pi(\mathrm{d}z, \mathrm{d}z') = \tilde{P}(\mathrm{d}z)\delta_1(\mathrm{d}z')$, whence

$$d_W(\tilde{P}, \delta(1)) = \inf_{\pi \in \Pi(\tilde{P}, \delta(1))} \int |z - z'| \,\mathrm{d}\pi(z, z') = \int |z - 1| \,\mathrm{d}\tilde{P}(z) = \mathbb{E}_{\tilde{P}}\big[\,|Z - 1|\,\big].$$

Combining the displays yields

$$\mathbb{E}_{\tilde{P}}\big[\,|1 - Z\hat{\theta}_{ERM}|\,\big] = d_W(\tilde{P}, \delta(1)) = d_W(\tilde{P}, P^*),$$

which establishes the claimed lower bound without requiring any additional condition on $d_W(\tilde{P}, P^*)$.

A.3 Proof of Theorem 1

Proof. Fix the Wasserstein radius $r > 0$ and define the population and empirical balls $B_*(r) := \{P : d_W(P, P^*) \le r\}$ and $\widehat{B}_n(r) := \{P : d_W(P, \widehat{P}_n) \le r\}$. Let

$$\hat{f}_{\mathrm{DRO}} \in \arg\min_{f \in \mathcal{F}} \sup_{P \in \widehat{B}_n(r)} \mathbb{E}_P[\ell(f, z)], \qquad f^* \in \arg\min_{f \in \mathcal{F}} \mathbb{E}_{P^*}[\ell(f, z)], \qquad \tilde{f} \in \arg\min_{f \in \mathcal{F}} \mathbb{E}_{\tilde{P}}[\ell(f, z)].$$

Introduce the finite-sample discrepancy

$$\tilde{\Delta}(f) := \Big|\, \sup_{P \in \widehat{B}_n(r)} \mathbb{E}_P[\ell(f, z)] \;-\; \sup_{P \in B_*(r)} \mathbb{E}_P[\ell(f, z)] \,\Big|.$$

A three-term decomposition yields, for any $f \in \mathcal{F}$,

$$\sup_{P \in B_*(r)} \mathbb{E}_P[\ell(\hat{f}_{\mathrm{DRO}}, z)] - \sup_{P \in B_*(r)} \mathbb{E}_P[\ell(f, z)] = \underbrace{\sup_{P \in B_*(r)} \mathbb{E}_P[\ell(\hat{f}_{\mathrm{DRO}}, z)] - \sup_{P \in \widehat{B}_n(r)} \mathbb{E}_P[\ell(\hat{f}_{\mathrm{DRO}}, z)]}_{\le \sup_{f' \in \mathcal{F}} \tilde{\Delta}(f')}$$

$$+ \underbrace{\sup_{P \in \widehat{B}_n(r)} \mathbb{E}_P[\ell(\hat{f}_{\mathrm{DRO}}, z)] - \sup_{P \in \widehat{B}_n(r)} \mathbb{E}_P[\ell(f, z)]}_{\le 0} \quad + \underbrace{\sup_{P \in \widehat{B}_n(r)} \mathbb{E}_P[\ell(f, z)] - \sup_{P \in B_*(r)} \mathbb{E}_P[\ell(f, z)]}_{\le \tilde{\Delta}(f)}$$

$$\le \sup_{f' \in \mathcal{F}} \tilde{\Delta}(f') + \tilde{\Delta}(f).$$

Taking $f = f^*$ and adding/subtracting suitable terms gives

$$\mathbb{E}_{\tilde{P}}[\ell(\hat{f}_{\mathrm{DRO}}, z)] - \mathbb{E}_{\tilde{P}}[\ell(\tilde{f}, z)] \le \underbrace{\mathbb{E}_{\tilde{P}}[\ell(\hat{f}_{\mathrm{DRO}}, z)] - \sup_{P \in B_*(r)} \mathbb{E}_P[\ell(\hat{f}_{\mathrm{DRO}}, z)]}_{(A)} + \underbrace{\sup_{P \in B_*(r)} \mathbb{E}_P[\ell(f^*, z)] - \mathbb{E}_{P^*}[\ell(f^*, z)]}_{(B)}$$

$$+ \mathbb{E}_{P^*}[\ell(f^*, z)] - \mathbb{E}_{\tilde{P}}[\ell(\tilde{f}, z)] + \sup_{f' \in \mathcal{F}} \tilde{\Delta}(f') + \tilde{\Delta}(f^*). \tag{14}$$

Then by definition $(B) = \Lambda_r(P^*, f^*)$.

For (A), if $\tilde{P} \in B_*(r)$ then $(A) \le 0$; otherwise let $P_0 \in B_*(r)$ attain $P_0 \in \arg\min_{P \in B_*(r)} d_W(\tilde{P}, P)$, whence

$$(A) \;\le\; \mathbb{E}_{\tilde{P}}[\ell(\hat{f}_{\mathrm{DRO}}, z)] - \mathbb{E}_{P_0}[\ell(\hat{f}_{\mathrm{DRO}}, z)] \;\le\; L\, d_W(\tilde{P}, P_0) \;=\; L \inf_{P \in B_*(r)} d_W(\tilde{P}, P).$$

Plugging these bounds into (14) and replacing $\tilde{\Delta}(\tilde{f})$ by the looser $\tilde{\Delta}(f^*)$ yields

$$\mathbb{E}_{\tilde{P}}[\ell(\hat{f}_{\mathrm{DRO}}, z)] - \mathbb{E}_{\tilde{P}}[\ell(\tilde{f}, z)] \;\le\; L \inf_{P \in B_*(r)} d_W(\tilde{P}, P) + \Lambda_r(P^*, f^*) + \big(\mathbb{E}_{P^*}[\ell(f^*, z)] - \mathbb{E}_{\tilde{P}}[\ell(\tilde{f}, z)]\big) + \sup_{f' \in \mathcal{F}} \tilde{\Delta}(f') + \tilde{\Delta}(f^*).$$

Control of the uniform term via the ϕ-envelope class. Below is similar to the proof given in Lee and Raginsky [2018]. Define, for $f \in \mathcal{F}$ and $0 \le k \le L$,

$$\phi_{f,k}(z) \;:=\; \sup_{z' \in \mathcal{Z}} \Big\{ \ell(f, z') - k\, d(z, z') \Big\}, \quad \text{and} \quad \Phi \;:=\; \big\{ \phi_{f,k} : f \in \mathcal{F},\; 0 \le k \le L \big\}.$$

Let $\Delta(f, k) \;:=\; \mathbb{E}_{P^*}[\phi_{f,k}(z)] \;-\; \mathbb{E}_{\hat{P}_n}[\phi_{f,k}(z)]$. By the Kantorovich dual formulation of the Wasserstein-robust risk,

$$\sup_{P \in B_*(\rho)} \mathbb{E}_P[\ell(f, z)] = \inf_{0 \le k \le L} \Big\{ k\rho + \mathbb{E}_{P^*}[\phi_{f,k}(z)] \Big\},$$
$$\sup_{P \in \widehat{B}_n(\rho)} \mathbb{E}_P[\ell(f, z)] = \inf_{0 \le k \le L} \Big\{ k\rho + \mathbb{E}_{\hat{P}_n}[\phi_{f,k}(z)] \Big\}.$$

Hence, for every $f \in \mathcal{F}$,

$$\tilde{\Delta}(f) = \Big| \sup_{P \in \widehat{B}_n(\rho)} \mathbb{E}_P[\ell(f, z)] \;-\; \sup_{P \in B_*(\rho)} \mathbb{E}_P[\ell(f, z)] \Big| \;\le\; \sup_{0 \le k \le L} \Big| \mathbb{E}_{P^*}[\phi_{f,k}(z)] - \mathbb{E}_{\hat{P}_n}[\phi_{f,k}(z)] \Big|.$$

Taking sup over f and applying symmetrization yields, with probability at least $1 - \delta_1$,

$$\sup_{f \in \mathcal{F},\, 0 \le k \le L} \Delta(f, k) \;\le\; 2\, \mathcal{R}_n(\Phi) \;+\; M \sqrt{\frac{2 \log(2/\delta_1)}{n}}. \tag{15}$$

By Lee and Raginsky [2018], the Rademacher complexity of Φ is bounded as

$$\mathcal{R}_n(\Phi) \;\le\; \frac{24}{\sqrt{n}}\, \mathcal{C}(\mathcal{A}) \;+\; \frac{24\, L\, \mathrm{diam}(\mathcal{Z})}{\sqrt{n}}, \qquad \mathcal{C}(\mathcal{A}) := \int_0^\infty \sqrt{\log \mathcal{N}\big(\mathcal{A}, \|\cdot\|_\infty, u\big)}\, du. \tag{16}$$

Combining the dual representation with (15)–(16) and a union bound for the two occurrences of $\tilde{\Delta}(\cdot)$ in (14) yields the desired statistical term (w.p. $\ge 1 - \delta_1$):

$$\sup_{f \in \mathcal{F}} \tilde{\Delta}(f) \;\le\; \frac{48}{\sqrt{n}}\, \mathcal{C}(\mathcal{A}) \;+\; \frac{48\, L\, \mathrm{diam}(\mathcal{Z})}{\sqrt{n}} \;+\; M \sqrt{\frac{2 \log(2/\delta_1)}{n}}. \tag{17}$$

Pointwise control for $\tilde{\Delta}(f^*)$. Fix $f = f^*$ and define the class $\Phi_{f^*} := \{\phi_{f^*,k} : 0 \le k \le L\}$ with $\phi_{f^*,k}(z) := \sup_{z'} \{\ell(f^*, z') - k\, d(z, z')\}$. Since $0 \le \ell \le M$, we have $0 \le \phi_{f^*,k} \le M$ and, for any $k, k' \in [0, L]$ and $z \in \mathcal{Z}$,

$$\phi_{f^*,k}(z) - \phi_{f^*,k'}(z) \le \sup_{z'} \{(k' - k)\, d(z, z')\} \le |k - k'|\, \mathrm{diam}(\mathcal{Z}),$$

and the reverse inequality by swapping k, k', hence

$$\big| \phi_{f^*,k}(z) - \phi_{f^*,k'}(z) \big| \;\le\; \mathrm{diam}(\mathcal{Z})\, |k - k'|, \qquad 0 \le \phi_{f^*,k} \le M. \tag{18}$$

Let the pseudometric $d_\Phi(k, k') := \mathrm{diam}(\mathcal{Z})\, |k - k'|$ and the Rademacher process

$$X_k \;:=\; \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i\, \phi_{f^*,k}(Z_i), \qquad \varepsilon_i \overset{i.i.d.}{\sim} \mathrm{Unif}\{\pm 1\}.$$

By Hoeffding's lemma and (18),

$$\mathbb{E} \exp\big(t(X_k - X_{k'})\big) = \prod_{i=1}^n \mathbb{E} \exp\Big(\tfrac{t}{\sqrt{n}} \varepsilon_i \big(\phi_{f^*,k}(Z_i) - \phi_{f^*,k'}(Z_i)\big) \Big) \;\le\; \exp\Big(\tfrac{t^2}{2}\, d_\Phi(k, k')^2 \Big),$$

so (X_k) has subgaussian increments w.r.t. d_Φ. Dudley's entropy integral therefore gives

$$\mathcal{R}_n(\Phi_{f^*}) \;\le\; \frac{12}{\sqrt{n}} \int_0^\infty \sqrt{\log \mathcal{N}(\Phi_{f^*}, d_\Phi, u)}\, du. \tag{19}$$

Since Φ_{f^*} is parameterized by the interval $[0, L]$ and d_Φ scales linearly with $|k - k'|$, we have

$$\mathcal{N}(\Phi_{f^*}, d_\Phi, u) = \mathcal{N}\Big([0, L], |\cdot|, \tfrac{u}{\mathrm{diam}(\mathcal{Z})} \Big) \;\le\; 1 + \frac{L\, \mathrm{diam}(\mathcal{Z})}{u}.$$

Truncating the integral at $u = L\,\mathrm{diam}(\mathcal{Z})$ (above which the covering number is 1) and evaluating the elementary integral in 19, we obtain

$$\mathcal{R}_n(\Phi_{f^*}) \;\le\; \frac{12\,L\,\mathrm{diam}(\mathcal{Z})}{\sqrt{n}}. \tag{20}$$

By symmetrization and McDiarmid's inequality,

$$\tilde{\Delta}(f^*) \le \sup_{0\le k\le L} \left|\mathbb{E}_{P^*}\phi_{f^*,k} - \mathbb{E}_{\hat{P}_n}\phi_{f^*,k}\right| \;\le\; 2\,\mathcal{R}_n(\Phi_{f^*}) \;+\; M\sqrt{\frac{2\log(2/\delta_2)}{n}}$$

with probability at least $1-\delta_2$. Combining with (20),

$$\tilde{\Delta}(f^*) \;\le\; \frac{24\,L\,\mathrm{diam}(\mathcal{Z})}{\sqrt{n}} \;+\; M\sqrt{\frac{2\log(2/\delta_2)}{n}} \qquad \text{w.p. } \ge 1-\delta_2. \tag{21}$$

Finally, choose $\delta_1 = \delta_2 = \delta/2$ and apply a union bound to (17) and (21) to get, with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}} \tilde{\Delta}(f) \;+\; \tilde{\Delta}(f^*) \;\le\; \frac{48}{\sqrt{n}}\,\mathcal{C}(\mathcal{A}) \;+\; \frac{72\,L\,\mathrm{diam}(\mathcal{Z})}{\sqrt{n}} \;+\; 2M\sqrt{\frac{2\log(4/\delta)}{n}}. \tag{22}$$

Substituting (22) back into (14) completes the proof.

A.4 Proof of Theorem 2

Proof. By feasibility of the RS solution with parameter $\tau_\epsilon = \inf_{f\in\mathcal{F}} \mathbb{E}_{\hat{P}_n}[\ell(f,z)] + \epsilon$ and its sensitivity $k_{\tau_\epsilon} \in [0, L]$, we have for all $P \in \mathcal{P}$:

$$\mathbb{E}_P\big[\ell(\hat{f}_{RS}, z)\big] \;\le\; \tau_\epsilon \;+\; k_{\tau_\epsilon}\, d_W(P, \hat{P}_n). \tag{23}$$

Using the dual representation of the 1-Wasserstein ball, (23) is equivalent to

$$\mathbb{E}_{\hat{P}_n}\Big[\sup_{y\in\mathcal{Z}} \big\{\ell(\hat{f}_{RS}, y) - k_{\tau_\epsilon}\, c(z,y)\big\}\Big] \;\le\; \tau_\epsilon. \tag{24}$$

Replacing $\hat{P}_n$ by P^* and adding a uniform deviation term yields

$$\mathbb{E}_{P^*}\Big[\sup_{y\in\mathcal{Z}} \big\{\ell(\hat{f}_{RS}, y) - k_{\tau_\epsilon}\, c(z,y)\big\}\Big] \;\le\; \tau_\epsilon \;+\; \sup_{f\in\mathcal{F},\, 0\le k\le L} \tilde{\Delta}(f,k), \tag{25}$$

where

$$\tilde{\Delta}(f,k) \;:=\; \mathbb{E}_{P^*}\big[\sup_y \{\ell(f,y) - k\,c(z,y)\}\big] \;-\; \mathbb{E}_{\hat{P}_n}\big[\sup_y \{\ell(f,y) - k\,c(z,y)\}\big]\,.$$

Applying again the dual representation to (25) gives, for every P,

$$\mathbb{E}_P\big[\ell(\hat{f}_{RS}, z)\big] \;\le\; \tau_\epsilon \;+\; k_{\tau_\epsilon}\, d_W(P, P^*) \;+\; \sup_{f\in\mathcal{F},\, 0\le k\le L} \Delta(f,k). \tag{26}$$

Specializing (26) to $P = \tilde{P}$ and subtracting $\tilde{J} := \inf_{f\in\mathcal{F}} \mathbb{E}_{\tilde{P}}[\ell(f,z)]$,

$$\begin{aligned}\mathcal{R}_{\tilde{P}}(\hat{f}_{RS}) &:= \mathbb{E}_{\tilde{P}}\big[\ell(\hat{f}_{RS}, z)\big] - \tilde{J} \\ &\le (J^* - \tilde{J}) \;+\; k_{\tau_\epsilon}\, d_W(P^*, \tilde{P}) \;+\; \epsilon \;+\; \big(\mathbb{E}_{\hat{P}_n}[\ell(f^*,z)] - \mathbb{E}_{P^*}[\ell(f^*,z)]\big) \;+\; \sup_{f\in\mathcal{F},\, 0\le k\le L} \tilde{\Delta}(f,k),\end{aligned} \tag{27}$$

where $f^* \in \arg\min_f \mathbb{E}_{P^*}[\ell(f,z)]$ and $J^* = \mathbb{E}_{P^*}[\ell(f^*,z)]$.

The last two terms in (27) are the same uniform deviations that appear in the DRO proof's "control of the uniform term" via the ϕ-envelope class but we adopt a one-sided control scheme rather than a two-sided one. We therefore *directly reuse* that result without rederiving it: with probability at least $1-\delta/2$,

$$\sup_{f\in\mathcal{F},\,0\le k\le L}\tilde{\Delta}(f,k)\;\le\;\frac{48}{\sqrt{n}}\mathcal{C}(\mathcal{A})\;+\;\frac{48L\,\mathrm{diam}(\mathcal{Z})}{\sqrt{n}}+M\sqrt{\frac{2\log(2/\delta)}{n}}.$$

Also with probability at least $1-\delta/2$,

$$\mathbb{E}_{\hat{P}_n}[\ell(f^*,z)]-\mathbb{E}_{P^*}[\ell(f^*,z)]\;\le\;M\sqrt{\frac{\log(2/\delta)}{2n}},$$

Substituting these bounds into (27) yields, with probability at least $1-\delta$,

$$\mathcal{R}_{\tilde{P}}(\hat{f}_{RS})\;\le\;(J^*-\tilde{J})\;+\;k_{\tau_\epsilon}\,d_W(P^*,\tilde{P})\;+\;\epsilon\;+\;\frac{48}{\sqrt{n}}\mathcal{C}(\mathcal{A})\;+\;\frac{48L\,\mathrm{diam}(\mathcal{Z})}{\sqrt{n}}\;+\;3M\sqrt{\frac{\log(2/\delta)}{2n}},$$

which is the stated RS generalization bound.

B Supplementary Theoretical Results

B.1 RS Generalization Error Bound without Distributional Shift

Remark 1 (A subtle improvement without distributional shift)**.** *When there is no distributional shift ($P^*=\tilde{P}$), $\hat{f}_{RS}$ reduces to the ϵ-approximate ERM introduced in Equation 3.3 of Zhang [2023]. By Proposition 4.20 in Zhang [2023], under Assumption 1(b), with probability at least $1-\delta$, we have*

$$\mathcal{R}_{\tilde{P}}(\hat{f}_{RS})\le\epsilon+\frac{24}{\sqrt{n}}\mathcal{C}(\mathcal{A})+2M\sqrt{\frac{\log(2/\delta)}{2n}}.\tag{28}$$

Comparing (28) with the general RS bound in Theorem 2, we see that in the absence of distributional shifts, the term involving the instance diameter $\mathrm{diam}(\mathcal{Z})$ disappears. Consequently, Assumptions 1(a) and 1(c) are no longer required.

Our interpretation is that under genuine distributional shifts, these regularity conditions are necessary to control optimizer migration across distributions, thereby justifying the form of Theorem 2. In the special case without shifts, however, the analysis reduces to the classical ERM setting [Zhang, 2023], where only a uniform upper bound M on the loss function is needed.

B.2 Supplement on ERM Lower Bound

Building on the ERM lower bound setting 2, we now consider a special case where the target distribution is $\tilde{P}=\delta(x)$, i.e. a Dirac measure concentrated at x. Distributional shift occurs whenever $x\neq 1$. We evaluate the optimizers $\hat{\theta}_{ERM},\hat{\theta}_{DRO},\hat{\theta}_{RS}$ (learned from the source distribution) under such target distributions.

	ERM	DRO	RS
Optimizer	$\hat{\theta}_{\mathrm{ERM}}=1$	$\hat{\theta}_{\mathrm{DRO}}=\mathbb{I}[r\le 1]$	$\hat{\theta}_{\mathrm{RS}}=(1-\epsilon)\,\mathbb{I}[\epsilon\le 1](=k_{\tau_\epsilon})$

Table 1: Solutions under the ℓ_1-loss example.

Figure 3 plots target losses of DRO and RS relative to ERM under varying hyperparameters, with the x-axis representing the point where $\delta(x)$ is concentrated. Moving away from $x=1$ quantifies both direction and magnitude of distributional shift.

The plot shows that for nontrivial DRO ($r\le 1$), DRO coincides with ERM: both exhibit identical target loss growth rates, with slope equal to $L=1$. In contrast, RS with $\epsilon\in(0,1)$ demonstrates two distinctive properties. First, its target loss grows at a rate strictly less than 1, reflecting improved robustness against shifts. Second, an ϵ-dependent trade-off emerges: larger ϵ suppresses growth under larger shifts but harms performance near $x=1$.

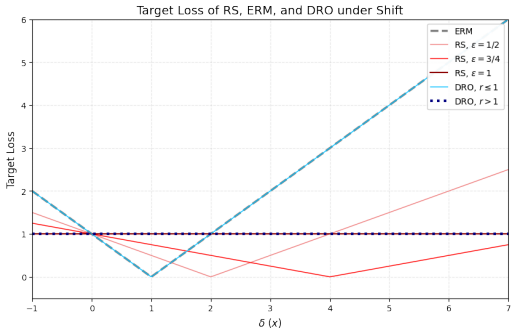


Figure 3: Target losses of RS, ERM, and DRO under ℓ_1-loss.

B.3 Additional Theorems for Comparative Statics under Paartial Shift Information

We provide a quantitative comparison of the generalization error bounds of DRO and RS presented in Theorems 1 and 2. We begin by introducing the following assumptions, which require the loss function to be Lipschitz continuous and strongly convex in the functional space. These conditions are standard in the literature and hold for many common loss functions, including logistic loss and mean squared error.

Assumption 3. *We assume:*

(a) ***Parameter-Lipschitz loss:*** *There exists a constant $L' > 0$ and a norm $\|\cdot\|$ on the parameter space such that*

$$\big|\ell(f,z) - \ell(g,z)\big| \;\le\; L'\,\|f-g\|, \qquad \forall f,g \in \mathcal{F},\ \forall z \in \mathcal{Z}.$$

(b) ***Strong convexity of population risk:*** *Let $\mathcal{E}(f) := \mathbb{E}_{P^*}[\ell(f,z)]$. There exists $\alpha > 0$ such that for all $f, g \in \mathcal{F}$,*

$$\mathcal{E}(g) \;\ge\; \mathcal{E}(f) + \langle \nabla\mathcal{E}(f),\, g-f\rangle + \frac{\alpha}{2}\,\|g-f\|^2.$$

B.3.1 Scenario I: Known Magnitude, Unknown Direction

Proposition 4. *Under Assumption 1 and 3, with probability at least $1-\delta$, we have:*

$$UB_{DRO}(r) \le UB_{RS}(\tau_r) - k_{\tau_\epsilon} r + \rho_n(\delta),$$

where $UB_{DRO}(r)$ and $UB_{RS}(\tau)$ are the generalization upper bounds of DRO and RS in (6) *and* (7) *respectively, $\rho_n(\delta) = L'\sqrt{\frac{8}{\alpha}\,\mathfrak{G}_n(\delta)} + \frac{24\,L\,\mathrm{diam}(\mathcal{Z})}{\sqrt{n}} + 3M\sqrt{\frac{\log(8/\delta)}{2n}}$ is the statistical error with $\mathfrak{G}_n(\delta) := \frac{24}{\sqrt{n}}\,\mathcal{C}(\mathcal{A}) + M\sqrt{\frac{\log(4/\delta)}{2n}}$.*

Proposition 4 shows that with known shift information, DRO theoretically outperforms RS by a term $k_{\tau_\epsilon} r$ in the asymptotic regime (as the sample size grows to infinity). This advantage stems from DRO's ability to naturally incorporate the shift magnitude into its robustness framework and capture the worst-case scenario.

B.3.2 Scenario II: Known Direction, Unknown Magnitude

Proposition 5. *If $\frac{d_W(P_t,P^*)}{d_W(\bar{P},P^*)} \le 1 - \frac{k_{\tau_t}}{L}$, under Assumption 1, 2 and 3, with probability at least $1-\delta$, we have:*

$$UB_{RS}(\tau_t) \le UB_{DRO}(r_t) + \bar{\rho}_n(\delta),$$

where $UB_{DRO}(r)$ and $UB_{RS}(\tau)$ are the generalization upper bounds of DRO and RS in (6) *and* (7) *respectively, $\bar{\rho}_n(\delta) = 2L'\sqrt{\frac{8}{\alpha}\bar{\mathfrak{G}}_n(\delta)} + \bar{\mathfrak{G}}_n(\delta)$ is the statistical error with $\bar{\mathfrak{G}}_n(\delta) = \frac{24}{\sqrt{n}}\mathcal{C}(\mathcal{A}) + M\sqrt{\frac{\log(2/\delta)}{2n}}$ (the same below).*

This proposition implies that when the shift magnitude is substantially under-specified, RS provides a tighter generalization guarantee than DRO. However, when the shift magnitude is only slightly under-specified, or when it is well-specified ($t = \tilde{t}$) or over-specified ($t > \tilde{t}$), the shift term in the DRO bound (6) becomes smaller than that in the RS bound (7), and no clear theoretical comparison can be made. When the shift direction aligns with the worst-case distributional direction, referred to as the adversarial scenario, additional results can be derived where DRO tends to perform better because it is designed to handle worst-case distributions.

Specially for adversarial scenario, we have:

Proposition 6. *Under Assumption 1, 2 and3, under the conditions either: (i) the underspecified case $t < \tilde{t}$ and $\inf_{0\le s\le t} d_W(\tilde{P}, P_s) \le \frac{k_{\tau_t}}{L} \cdot d_W(P^*, \tilde{P})$, or (ii) the wellspecified or overspecified case $t \ge \tilde{t}$, with probability at least $1-\delta$, we have:*

$$UB_{DRO}(r_t) \le UB_{RS}(\tau_t) + \bar{\rho}_n(\delta),$$

where the statistical error is the same as Proposition 5.

In such adversarial scenarios, the RS satisficing term is defined with respect to the worst-case distribution and is therefore asymptotically equivalent to the DRO regularization term. When the shift magnitude is mildly under-specified, well-specified, or over-specified, the DRO shift term becomes small or even vanishes, but the RS bound retains a non-negligible shift term. This results in a tighter bound for DRO, as characterized by the following proposition.

C Simulation Studies

We present linear regression experiments to evaluate DRO and RS under two distribution shift scenarios: known shift magnitude and known shift direction. ERM serves as the non-robust baseline.

C.1 Simulation Setup

We generate data (x, y) where $x \sim \mathcal{N}(0, I_3)$ and $y = \langle x, \beta\rangle + \epsilon$ with $\beta = [2, 3, -1]^T$ and Gaussian noise ϵ. Training samples come from source distribution P^*. Distribution shift is introduced via parameter drift: the conditional law changes to $y|x = \langle x, \tilde{\beta}\rangle$, while x's distribution remains unchanged.

We adopt the Huber loss (satisfying Assumption 1(c)), and use a type-1 Wasserstein cost function following Blanchet et al. [2019]. The optimization problems after dual reformulations are:

1. ERM: $\hat{\beta}_{ERM} = \arg\min_\beta \frac{1}{n}\sum_i \ell(\beta, x_i, y_i)$.
2. DRO: $\hat{\beta}_{DRO} = \arg\min_\beta \frac{1}{n}\sum_i \ell(\beta, x_i, y_i) + r\|\beta\|_2$.
3. RS: $\hat{\beta}_{RS} = \arg\min_\beta \|\beta\|_2$ s.t. $\frac{1}{n}\sum_i \ell(\beta, x_i, y_i) \le \tau$.

Both DRO and RS are regularized versions of ERM, shrinking β toward the origin. We evaluate $\mathbb{E}_{\tilde{P}}[\ell(\hat{\beta}, z)]$ on target data.

C.2 Scenario I: Known Shift Magnitude

We assume $d_W(P^*, \tilde{P}) \le r$. DRO sets its radius to r, while RS uses $\tau_r = \frac{1}{n}\sum_i \ell(\hat{\beta}_{ERM}, x_i, y_i) + r\|\hat{\beta}_{ERM}\|_2$. We simulate 500 random shift directions $\tilde{\beta} = \beta + td$ with $\|d\| = 1$ and adjust t so that $d_W(P^*, \tilde{P}) = 0.15$. Figure 4 summarizes target losses.

DRO and ERM exhibit similar performance, with loss ratios concentrated near 1.0. In contrast, RS performs significantly worse, with average target loss nearly twice as large. Moreover, DRO's loss

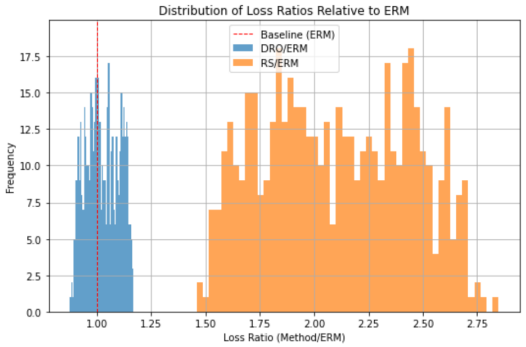


Figure 4: Distribution of loss ratios under known shift magnitude.

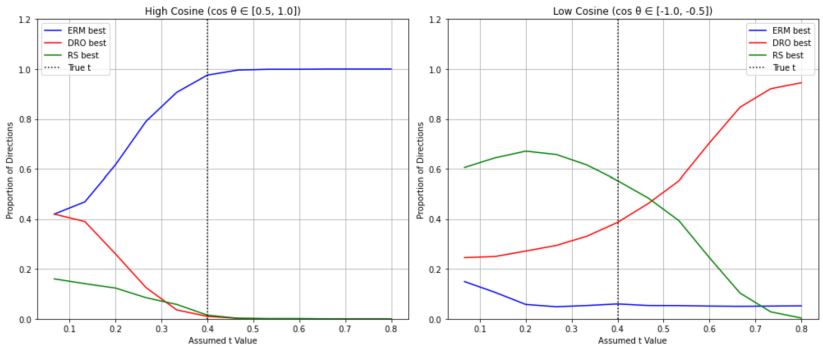


Figure 5: Pairwise method comparisons under known shift direction.

distribution shows tighter concentration and smaller spread than RS's, indicating greater stability against unknown shift directions. These findings are consistent with theoretical results that DRO dominates RS when the shift magnitude is known but the direction is unknown.

C.3 Scenario II: Known Shift Direction

We fix a direction d and vary magnitude t in $\tilde{\beta} = \beta^* + td$. True target is at $t = 0.4$, but hypothesized magnitudes may be under/over specified. Let $\cos\theta$ denote the cosine similarity between the shift direction d and the negative direction of source parameter vector $-\beta^*$.

We categorize directions into two regimes: *high-cosine* ($\cos\theta \in [0.5, 1]$) where d is largely aligned with β, and *low-cosine* ($\cos\theta \in [-1, -0.5]$) where d is strongly opposed to β. Figure 5 shows method comparisons across 500 sampled directions.

Results: in the high-cosine region, ERM is often competitive with the robust methods, while DRO consistently outperforms RS due to its worst-case protection. In the low-cosine region, when the shift magnitude is under- or well-specified, RS demonstrates clear advantages over both DRO and ERM. However, once the magnitude becomes overspecified, DRO gradually regains superiority and RS performance degrades. These findings confirm that the relative performance of DRO and RS depends critically on the alignment between the shift direction and the parameter shrinkage direction.

D Supplementary Material on the Network Lot-sizing Problem

D.1 Problem Formulation Details

We provide the complete formulation of the network lot-sizing problem. The decision maker determines the initial stock allocation x, which represents the amount of goods pre-stocked at each store before demand is realized. After demand z is observed, the system may adjust through second-stage decisions (y, w): y_{ij} denotes the amount transported from store i to store j, and w_i denotes

the amount obtained via emergency order at store i. The objective is to minimize the total cost, which includes the pre-stocking cost $c^\top x$, the transportation cost $\sum_{i\in[N]} d_i^\top y_i$, and the emergency replenishment cost $l^\top w$. Formally, the two-stage optimization model is

$$f(x,z) = c^\top x + \min_{(y,w)\in\mathcal{Y}} \Big\{ \sum_{i\in[N]} d_i^\top y_i + l^\top w \Big\},$$

where the feasible set $\mathcal{Y}$ is defined by the balance constraints

$$x_i + w_i + \sum_{j\in[N]} y_{ji} - \sum_{j\in[N]} y_{ij} - z_i \geq 0, \quad \forall i \in [N].$$

Here c_i is the unit ordering cost for initial stock, l_i is the emergency order cost, d_{ij} is the transportation cost between stores, and z_i is the random demand at store i.

We define several key quantities that decompose the total cost, which quantities will be used in the subsequent analysis. For a given initial allocation $\hat{x}$ and a realized random demand $\bar{z}$, the *total cost* is denoted by $f(\hat{x}, \bar{z})$. It consists of two parts: the *initial ordering cost* $c^T\hat{x}$ and the *operational cost*, which corresponds to the optimal value of the following second-stage problem:

$$\min_{(y,w)\in\mathcal{Y}_{(\hat{x},\bar{z})}} \Big\{ \sum_{i\in[N]} d_i^T y_i + l^T w \Big\}.$$

D.2 Robust Counterparts of DRO and RS

To handle distributional shifts, we introduce robust counterparts of the model. Following Long et al. [2023], Wang et al. [2023], both DRO and RS admit tractable dual formulations. We present them below.

DRO counterpart:

$$\begin{aligned}
&\min kr + c^\top x + \frac{1}{S}\sum_{s\in[S]} v_s \\
&\text{s.t.} \quad \mathbf{c}^\top x - \tau + \frac{1}{S}\sum_{s\in[S]} v_s \leq 0, \\
&\sum_{i\in[N]} \Big(\mathbf{d}_i^\top y_i^{(s)}(z,u) + l^\top w^{(s)}(z,u) \Big) - ku - v_s \leq 0, \\
&x_i + w_i^{(s)}(z,u) + y_{ji}^{(s)}(z,u) - y_{ij}^{(s)}(z,u) - z_i \geq 0, \quad \forall i \in [N], \\
&y^{(s)}(z,u) \geq 0, \quad w^{(s)}(z,u) \geq 0, \\
&0 \leq x \leq \bar{\delta}, \\
&y^{(s)} \in \mathcal{L}^{N+1, N\times N}, \quad w^{(s)} \in \mathcal{L}^{N+1,N}, \quad \forall s \in [S].
\end{aligned}$$

RS counterpart:

$$\begin{aligned}
&\min k \\
&\text{s.t.} \quad \mathbf{c}^\top x - \tau + \frac{1}{S}\sum_{s\in[S]} v_s \leq 0, \\
&\sum_{i\in[N]} \Big(\mathbf{d}_i^\top y_i^{(s)}(z,u) + l^\top w^{(s)}(z,u) \Big) - ku - v_s \leq 0, \\
&x_i + w_i^{(s)}(z,u) + y_{ji}^{(s)}(z,u) - y_{ij}^{(s)}(z,u) - z_i \geq 0, \quad \forall i \in [N], \\
&y^{(s)}(z,u) \geq 0, \quad w^{(s)}(z,u) \geq 0, \\
&0 \leq x \leq \bar{\delta}, \\
&y^{(s)} \in \mathcal{L}^{N+1, N\times N}, \quad w^{(s)} \in \mathcal{L}^{N+1,N}, \quad \forall s \in [S].
\end{aligned}$$

Both models can be solved efficiently using standard solvers such as `Gurobi`.

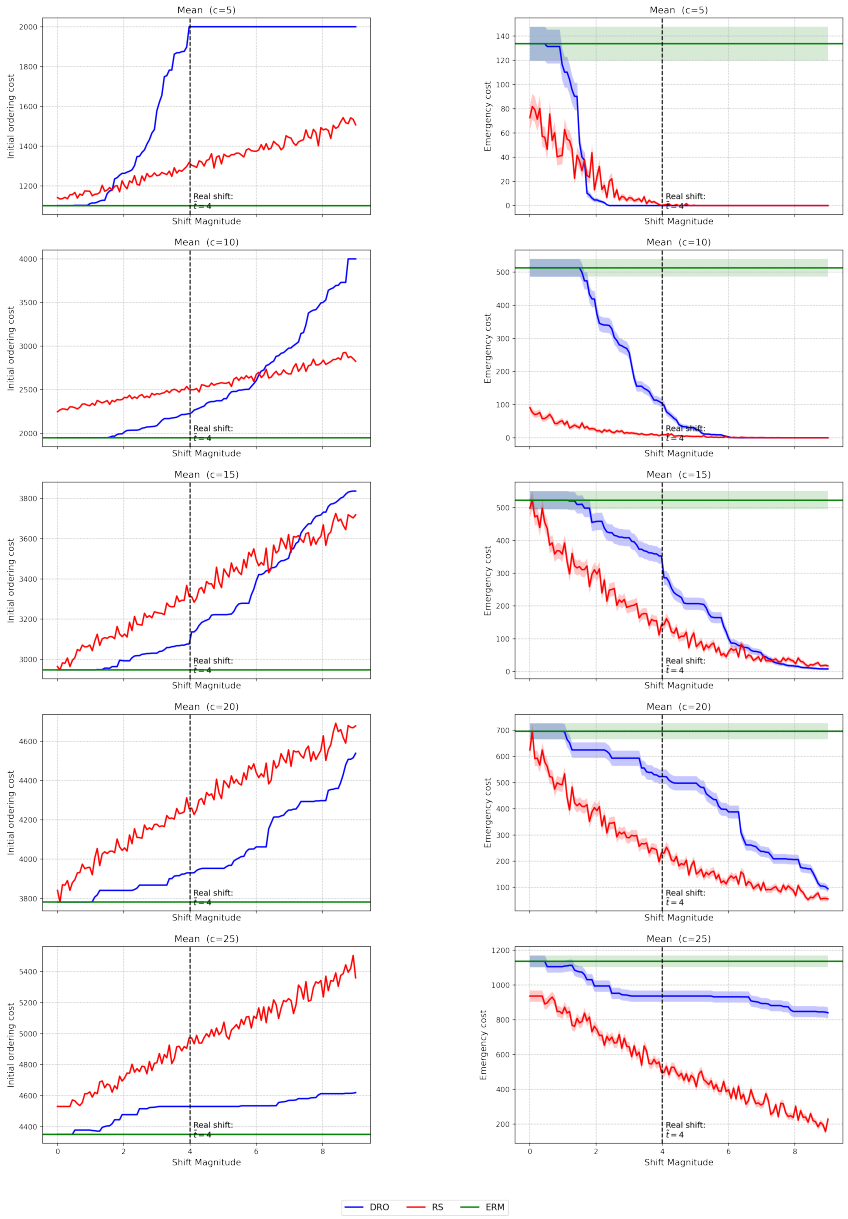


Figure 6: Mean of initial ordering costs and operational costs under distribution shift.

D.3 Cost decomposition: initial ordering vs. operational costs.

We now analyze how the robust methods influence the composition of the total cost, which comprises the initial ordering cost and the operational cost. Figure 6 reports the mean values of these two components.

Both DRO and RS incur higher initial ordering costs than the baseline ERM, while their operational costs are lower. This pattern reflects the preventive behavior of robust methods: by anticipating potential demand increases, they promote higher initial stocking of relatively inexpensive inventory to mitigate the risk of incurring substantially higher transportation or emergency-order costs later.

Moreover, both the initial and operational costs under RS increase approximately linearly with the anticipated demand level, across various choices of unit ordering cost c. In contrast, DRO exhibits greater sensitivity to c. When c is low, DRO's initial ordering cost rises faster than the demand shift, reflecting a heightened urgency to preempt potential shortages. As c increases, this urgency weakens:

DRO becomes more conservative in its upfront allocation, preferring to retain flexibility for later adjustments through transportation or emergency replenishment. These observations help explain the patterns in Figure 2a. In the underspecified regime, RS attains a lower total cost by slightly increasing initial orders while substantially reducing operational expenses. In the overspecified regime, when the unit cost of initial orders is low, DRO becomes overly aggressive, purchasing excessive initial stock without achieving a proportional reduction in operational costs, which leads to a higher total cost. As the unit ordering cost increases, DRO grows more conservative in its upfront purchasing, whereas RS continues to scale its orders roughly linearly with the anticipated demand shift. This gradual shift in behavior causes DRO's total cost to fall below that of RS at higher unit costs.

D.4 Hyperparameter Correspondence

We now adopt the perspective of hyperparameter correspondence to explain why RS performs more stably and achieves lower total cost than DRO in under-specified regimes ($t < \tilde{t}$) but the pattern reverses in over-specified regimes ($t > \tilde{t}$). Following Wang et al. [2023], Li et al. [2024], we derive a correspondence between the DRO radius r and the RS reference value τ. Figure 7 illustrates this relationship for an initial ordering cost of $c = 10$, where each point (r, τ) on the curve represents a pair of hyperparameters yielding identical solutions from DRO and RS. In the under-specified regime, corresponding to small r and τ with small anticipated shift, small changes in τ lead to large changes in r. This indicates that RS is more stable, while DRO behaves more conservatively. In the over-specified regime, corresponding to large r and τ with large anticipated shift, the relationship reverses: small changes in r induce large changes in τ, implying that DRO becomes more stable while RS turns overly conservative.

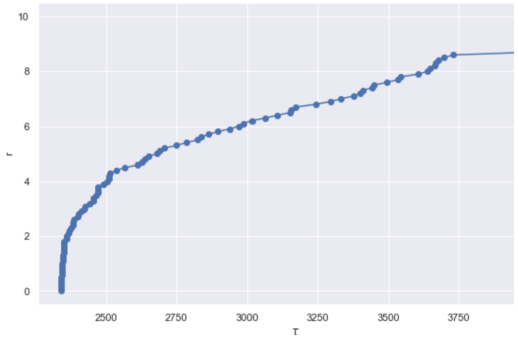


Figure 7: Hyperparameter correspondence between DRO (r) and RS (τ). Each pair on the curve yields the same optimizer.