

---

# Distributionally Robust Performative Optimization

---

Zhuangzhuang Jia<sup>1</sup>, Yijie Wang<sup>2</sup>, Roy Dong<sup>1</sup>, Grani A. Hanusanto<sup>1</sup>

<sup>1</sup>Department of Industrial and Enterprise Systems Engineering  
University of Illinois Urbana-Champaign

<sup>2</sup>School of Economics and Management, Tongji University  
`{zj12,roydong,gah}@illinois.edu, yijiewang@tongji.edu.cn`

## Abstract

In performative stochastic optimization, decisions can influence the distribution of random parameters, rendering the data-generating process itself decision-dependent. In practice, decision-makers rarely have access to the true distribution map and must instead rely on imperfect surrogate models, which can lead to severely suboptimal solutions under misspecification. Data scarcity or costly collection further exacerbates these challenges in real-world settings. To address these challenges, we propose a distributionally robust framework for performative optimization that explicitly accounts for ambiguity in the decision-dependent distribution. Our framework introduces three modeling paradigms that capture a broad range of applications in machine learning and decision-making under uncertainty. This latter setting has not previously been explored in the performative optimization literature. To tackle the intractability of the resulting nonconvex objectives, we develop an iterative algorithm named repeated robust risk minimization, which alternates between solving a decision-independent distributionally robust optimization problem and updating the ambiguity set based on the previous decision. This decoupling ensures computational tractability at each iteration while enhancing robustness to model uncertainty. We provide reformulations compatible with off-the-shelf solvers and establish theoretical guarantees on convergence and suboptimality. Extensive numerical experiments in strategic classification, revenue management, and portfolio optimization demonstrate significant performance gains over state-of-the-art baselines, highlighting the practical value of our approach.

## 1 Introduction

Decision-makers' actions often have a ripple effect on the external environment, which can lead to changes in the distribution of uncertain parameters. For instance, in the realm of portfolio management, institutional investors' decisions can have a profound effect on stock prices. This is partly due to their substantial capital, which can propel stock prices to rise (or fall) when they are buying (or selling) [57], and partly because their actions shape market sentiment towards those particular stocks [2]. Similarly, in revenue management, airlines often make forecasts and design pricing strategies based on the historical patterns of passenger behavior. However, passengers are not static; they often modify their behaviors in reaction to the new pricing strategies of airlines, which in turn shifts the overall pattern of consumer behavior [11].

When the solution of a stochastic optimization problem affects the distribution of the uncertain parameters, we call such a problem *performative* [40, 15]. The primary goal of the decision maker within such a dynamic environment is to find a decision that minimizes the expected risk after the environment has reacted to its deployment. A natural way of introducing this decision-dependent uncertainty is to construct a distributional map from the set of decisions to the space of distributions.

However, the true underlying map is typically unknown in reality. Consequently, practitioners usually rely on a nominal or reference distribution map constructed from historical observations or domain expertise. While methods grounded in the reference distribution map might perform satisfactorily on the observed samples, they often fail to achieve acceptable performance in out-of-sample circumstances.

This paper aims to tackle this core deficiency by leveraging the ideas of *Distributionally Robust Optimization*. Unlike traditional approaches that assume a single distribution map, distributionally robust performative optimization (DRPO) adopts a more flexible strategy: it establishes an ambiguity set of plausible distributions centered at the reference distribution. Next, the objective of the decision maker is to derive an optimal decision that minimizes the worst-case expected risk, where the worst case is taken over all distribution maps from within this ambiguity set. By optimizing against this adversarial perspective within a neighborhood of the reference map, it mitigates the overfitting issue and improves the out-of-sample performance.

The difficulty in solving DRPO stems from the dependence of the distribution map on the decision, which prohibits the direct use of existing solution schemes from the robust and distributionally robust optimization literature. For instance, even under the simplest setting where the loss function is linear and the distribution map is linear in decisions, the resultant problem is a nonconvex bilinear program. We address this challenge by designing a repeated robust risk minimization algorithm. Specifically, the decision maker repeatedly obtains an optimal decision that minimizes DRPR risk using the reference distribution from the previous iteration. Finally, we conduct a theoretical analysis of the algorithm, providing convergence and sub-optimality guarantees. Our main contributions can be summarized as follows:

1. **General DRPO Framework:** We propose a distributionally robust framework based on Wasserstein distance for performative optimization. This approach enables safe decision-making when we lack full information about the underlying decision-dependent distributions and only have access to some reference distributions. It encompasses a wide class of loss functions and can be applied to numerous machine learning and decision-making problems. As a byproduct of our reformulations, we identify a relatively general setting where the distributionally robust model is equivalent to Tikhonov regularization.
2. **Repeated Robust Risk Minimization Algorithm:** We develop a repeated robust risk minimization algorithm for the problem that effectively mitigates the intractability of decision-dependent uncertainty. Our approach decouples the decisions associated with the ambiguity set from the expected loss, optimizing the latter while fixing the former to the previous decision in each iteration. This transforms the challenging DRPO problem into a sequence of tractable conic programs, rendering the framework computationally feasible and amenable to solutions via off-the-shelf solvers.
3. **Rigorous Theoretical Guarantees:** We provide convergence results of the repeated robust risk minimization algorithm to the stable solutions for our proposed models. To our knowledge, the convergence of such an algorithm has not previously been established in the distributionally robust settings. Our results show that, in general, the distributionally robust models lead to faster convergence than non-robust schemes. We further prove that our stable solutions are near to the robust performatively optimal ones.

## 1.1 Related Work

**Stochastic Optimization:** Our study closely relates to a body of research delving into endogenous uncertainty in stochastic optimization. Early examples include production planning problems with decision-dependent production costs [24] and oil planning problem with decision-dependent information discovery [18]. Their solution entails constructing a scenario-tree-based stochastic programming model and implementing a decomposition algorithm for resolution. Building on this, a subsequent study [19] extends the methodology to the multi-stage setting, focusing on production scheduling that minimizes costs while fulfilling diverse product demands. More broadly, [54] explore general multi-stage stochastic optimization problems beset with endogenous uncertainty, devising a conservative solution framework by leveraging piecewise linear decision rule approximations.

**Robust Optimization:** In robust optimization, decision-dependent uncertainty is usually imposed directly on the uncertainty set. Specific applications include customized endogenous uncertainty

sets for software partitioning [50] and robust scheduling, where the uncertainty set is constructed as a decision-dependent combination of simpler sets [56]. Decision-dependent sets have also been designed for primitive uncertainties in control systems [63]. The complexity analysis of the robust selection problem with decision-dependent information discovery is studied by [33], where they present polynomial complexity results for two special cases. For general settings, [37] show the problem is NP-complete and demonstrate the benefit of using endogenous uncertainty sets via a shortest-path problem. Algorithmic approaches have been developed, including exact nested decomposition schemes for two-stage problems [38] and approximation methods based on decision rules for the multistage setting [62].

**Distributionally Robust Optimization:** Our research expands upon the conventional framework of distributionally robust optimization by incorporating the influence of decision-making on the probability distribution. However, due to the problem difficulty, decision-dependent ambiguity sets generally lead to intractable reformulations, which could be computationally intensive for large instances. [61] study a broad class of distributionally robust optimization problems with decision-dependent moment ambiguity sets and conduct stability analysis. For the two stage setting, [31] analyze a wide range of decision-dependent ambiguity sets and establish non-convex semi-infinite reformulations and [23] explore decision-dependent information discovery using the K-adaptability approximation scheme. In multistage settings, [59] focus on moment-based ambiguity sets and derive a mixed-integer semidefinite programming reformulation. This decision-dependent DRO framework has been applied to a wide range of operations management problems, including nurse staffing [47], facility location [5], and retrofitting planning [14]. In contrast to the existing approaches that require an explicit specification of the ambiguity set, our scheme relies only on reference distributions at finitely many decisions.

**Performative Supervised Learning:** Our study also belongs to an emerging class of research on performative supervised learning, where algorithmic predictions can actively mold the surrounding environment and alter the underlying distributions of uncertain parameters [15, 20, 34, 40, 32]. The origins of performative learning can be traced back to studies on supervised learning under distribution drifts [3, 4, 16]. The foundational framework of performative prediction was introduced by [40], who designed a retraining algorithm and analyze the convergence behaviors to stochastic performatively stable points. Unfortunately, the convergence results do not extend to the robust settings, as taking worst-case expectations introduces non-smoothness into the objective function. Gradient-based methods have been developed in [40, 32, 15] for non-robust formulations, but they are inapplicable in robust settings due to the intractability of computing gradients of worst-case expectations. [28] has extended the scope of performative prediction to decision-making via performative omniprediction. We refer readers who are interested in performative learning to a comprehensive review [21].

Finally, we would like to highlight the difference between our paper and two related works on distributionally robust performative prediction. [39] propose a distributionally robust performative model to promote machine learning fairness. However, their robust smoothness assumption on the objective and robust sensitivity assumption on the worst-case distributions are unrealistic. Additionally, unlike the Wasserstein ambiguity set adopted in our paper, their phi-divergence ambiguity set precludes any continuous distribution, so it cannot ensure true probability distribution coverage guarantee. [58] study distributionally robust performative prediction based on a KL-divergence ambiguity set. Similar to [39], such an ambiguity set may fail to provide sufficient protections against distributions whose scenarios do not coincide with those in the reference distribution. In the paper, the authors develop an alternating minimization (coordinate descent) algorithm to solve the distributionally robust model. Unfortunately, such an algorithm may fail to converge, and the paper also does not provide any convergence guarantee. The algorithm also assumes the existence of a solver that can solve performative risk-sensitive minimization problems; however, to our knowledge, there is no existing literature that studies such problems and provides convergent algorithms.

## 1.2 Notation and Terminology

We use  $\mathbb{R}_+$  and  $\mathbb{R}_{++}$  to denote the sets of nonnegative and strictly positive real numbers, respectively. The identity matrix is denoted by  $\mathbb{I}$ . For any  $n \in \mathbb{N}$ , we define  $[n]$  as the index set  $\{1, \dots, n\}$ . The Dirac measure concentrating unit mass at  $\xi \in \Xi$  is denoted by  $\delta_\xi$ . For any real-valued matrix  $\mathbf{A}$ , its Schatten- $q$  norm is defined as  $\|\mathbf{A}\|_q = (\text{tr}(\mathbf{A}^\top \mathbf{A})^{q/2})^q$ . Random variables are designated with tilde signs (e.g.,  $\tilde{z}$ ), while their realizations are denoted by the same symbols without tildes (e.g.,  $z$ ).

## 2 Preliminaries

In stochastic optimization, the goal of the decision maker is to find a decision  $\theta \in \Theta \subseteq \mathbb{R}^d$  that minimizes the risk

$$R(\theta) = \mathbb{E}_{\mathbb{P}}[\ell(\tilde{z}, \theta)], \quad (1)$$

where the vector  $\tilde{z} \in \mathcal{Z} \subseteq \mathbb{R}^m$  comprises the random parameters, and  $\mathbb{P}$  denotes the underlying distribution. We assume that the loss function is convex in  $\theta$  for any fixed  $z \in \mathcal{Z}$ . When the solution of a stochastic optimization problem affects the distribution of the uncertain parameters, we call such problem *performative* [40, 15]. A natural way of introducing this decision-dependent uncertainty is through a map  $\mathbb{P}(\cdot)$  from the set of decisions to the space of distributions. Hence, in performative stochastic optimization problems, the objective of the decision maker is to obtain a decision  $\theta$  that minimizes the *performative risk*

$$PR(\theta) = \mathbb{E}_{\mathbb{P}(\theta)}[\ell(\tilde{z}, \theta)], \quad (2)$$

where  $\mathbb{P}(\theta)$  is the distribution of uncertain parameters  $\tilde{z}$  given the choice of the decision  $\theta$ . Unfortunately, the true underlying distribution map  $\mathbb{P}(\theta)$  is unknown to the decision makers and usually approximated using the reference distribution  $\hat{\mathbb{P}}(\theta) = \sum_{s \in [S]} \hat{p}_s(\theta) \delta_{\hat{z}_s(\theta)}$ , where  $\{\hat{z}_s(\theta)\}_{s \in [S]}$  represent plausible scenarios and  $\{\hat{p}_s(\theta)\}_{s \in [S]}$  are their respective probabilities. This reference distribution could be derived from observations or expert knowledge elicitation.

Although methods based on a reference distribution may perform well when the true distribution closely aligns with it, their performance often deteriorates when this assumption is violated. *Distributionally robust optimization* has proven effective in addressing such distributional ambiguity in non-performative stochastic optimization settings. Unlike traditional approaches that assume a single distribution map, distributionally robust optimization adopts a more flexible strategy: it establishes an ambiguity set  $\mathbb{B}(\hat{\mathbb{P}}(\theta))$  of distributions that are close to the reference distribution. In this paper, we construct this ambiguity set based on the Wasserstein distance. A formal definition of the Wasserstein distance is as follows.

**Definition 1** (Wasserstein metric). For any  $r \geq 1$ , let  $\mathcal{M}(\mathcal{Z})$  be the space of all probability distributions  $\mathbb{Q}$  supported on  $\mathcal{Z}$  satisfying  $\mathbb{E}_{\mathbb{Q}}[c(\tilde{z}, z_0)^r] = \int_{\Xi} c(z, z_0)^r \mathbb{Q}(dz) < \infty$ , where  $z_0 \in \mathcal{Z}$  is some reference point and  $c(z, z_0)$  is a non-negative, continuous and thus lower semi-continuous [55] reference metric on  $\mathcal{Z}$ . The type- $r$  ( $1 \leq r$ ) Wasserstein distance between two distributions  $\mathbb{Q}_1$  and  $\mathbb{Q}_2$  is defined as

$$W_r(\mathbb{Q}_1, \mathbb{Q}_2) = \inf_{\pi \in \Pi(\mathbb{Q}_1, \mathbb{Q}_2)} \left( \int_{\mathcal{Z} \times \mathcal{Z}} c(z_1, z_2)^r \pi(dz_1, dz_2) \right)^{\frac{1}{r}},$$

where  $\Pi(\mathbb{Q}_1 \times \mathbb{Q}_2)$  is the set of all joint probability distributions of random vectors  $z_1$  and  $z_2$  with marginals  $\mathbb{Q}_1$  and  $\mathbb{Q}_2$ , respectively.

The Wasserstein metric offers a natural way of comparing two distributions when one is derived from the other by small perturbations. The decision variable  $\pi$  can be interpreted as a transportation plan for moving a mass distribution denoted by  $\mathbb{Q}_1$  to another one denoted by  $\mathbb{Q}_2$ , where the transportation cost between two points  $z_1$  and  $z_2$  is given by  $c(z_1, z_2)$ . Therefore, the  $r$ -Wasserstein distance can be viewed as the  $r$ -th root of the minimum transportation cost between  $\mathbb{Q}_1$  and  $\mathbb{Q}_2$ . An important advantage of the Wasserstein distance is its ability to handle distributions with non-overlapping supports, i.e., even when  $\mathbb{Q}_1$  and  $\mathbb{Q}_2$  have different supports, the Wasserstein distance provides a finite, meaningful value. By contrast, measures like KL-divergence become undefined under such scenarios. We now consider the following Wasserstein ambiguity set

$$\mathbb{B}(\hat{\mathbb{P}}(\theta)) = \{ \mathbb{Q} \in \mathcal{M}(\Xi) : W_r(\mathbb{Q}, \hat{\mathbb{P}}(\theta)) \leq \rho \}, \quad (3)$$

which is a neighborhood around the reference distribution  $\hat{\mathbb{P}}(\theta)$ . The Wasserstein ambiguity set contains all the distributions whose  $r$ -Wasserstein distance from  $\hat{\mathbb{P}}(\theta)$  is less than or equal to  $\rho$ .

Equipped with the Wasserstein ambiguity set, the decision maker minimizes the *distributionally robust performative risk*

$$DRPR(\theta) := \sup_{\mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}}(\theta))} \mathbb{E}_{\mathbb{Q}} \ell(\tilde{z}, \theta), \quad (4)$$

where  $\mathbb{Q}$  is a distribution from within the prescribed ambiguity set  $\mathbb{B}(\hat{\mathbb{P}}(\boldsymbol{\theta}))$ . In other words, the model optimizes the expected risk over the worst-case distribution map, thereby mitigating overfitting to the reference distribution and improving generalization performance to other plausible distributions. We now introduce the following concepts regarding the optimality and stability of the solutions.

**Definition 2** (Robust performative optimality). A decision  $\boldsymbol{\theta}_{\text{RPO}}$  is *robust performatively optimal* if the following relationship holds:

$$\boldsymbol{\theta}_{\text{RPO}} \in \arg \min_{\boldsymbol{\theta} \in \Theta} \sup_{\mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}}(\boldsymbol{\theta}))} \mathbb{E}_{\mathbb{Q}} \ell(\tilde{\mathbf{z}}, \boldsymbol{\theta}).$$

**Definition 3** (Robust performative stability). A decision  $\boldsymbol{\theta}_{\text{RPS}}$  is *robust performatively stable* if the following relationship holds:

$$\boldsymbol{\theta}_{\text{RPS}} \in \arg \min_{\boldsymbol{\theta} \in \Theta} \sup_{\mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}}(\boldsymbol{\theta}_{\text{RPS}}))} \mathbb{E}_{\mathbb{Q}} \ell(\tilde{\mathbf{z}}, \boldsymbol{\theta}).$$

While distinct from the robust performatively optimal solution, a robust performatively stable solution constitutes a fixed point of the problem and is optimal with respect to the worst-case expected loss over the ambiguity set it induces.

**Definition 4** (Robust decoupled performative risk). We define

$$\mathbb{J}_{\boldsymbol{\eta}}(\boldsymbol{\theta}) := \sup_{\mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}}(\boldsymbol{\eta}))} \mathbb{E}_{\mathbb{Q}} \ell(\tilde{\mathbf{z}}, \boldsymbol{\theta})$$

as the *robust decoupled performative risk*, separating the decision  $\boldsymbol{\eta}$  associated with the ambiguity set and the decision  $\boldsymbol{\theta}$  associated with the risk; then,  $\boldsymbol{\theta}_{\text{RPS}} \in \arg \min_{\boldsymbol{\theta}} \mathbb{J}_{\boldsymbol{\theta}_{\text{RPS}}}(\boldsymbol{\theta})$ .

### 3 Repeated Robust Risk Minimization Algorithm

In this section, we introduce the *repeated robust risk minimization* algorithm for solving the distributionally robust performative risk minimization problem and investigate its fundamental properties. The algorithm starts with an initial solution  $\boldsymbol{\theta}_0$ , and for every  $t \geq 0$ , the subsequent  $\boldsymbol{\theta}_{t+1}$  can be obtained by solving the following robust risk minimization problem:

$$\min_{\boldsymbol{\theta} \in \Theta} \mathbb{J}_{\boldsymbol{\theta}_t}(\boldsymbol{\theta}). \quad (\text{RRMP})$$

The algorithm addresses the computational challenges posed by decision-dependent distributions by constructing the ambiguity set using the reference distribution based on the optimal decision from the previous iteration. Thus, it effectively decouples the current decision from the ambiguity set, simplifying the optimization process.

#### 3.1 Models and Their Tractable Reformulations

We present three models that cover a broad spectrum of problems in machine learning and decision-making under uncertainty. We first consider the case where the loss function is given by the composition of a Lipschitz continuous function and a quadratic function. This class of problems is particularly relevant in machine learning and robust statistics, as it captures many commonly used models such as linear regression [36], logistic regression [22], and certain types of support vector machines [48].

**Model 1.** Assume that the loss function is defined as

$$\ell(\mathbf{Z}, \boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}^\top \mathbf{Y} \boldsymbol{\theta} + 2\mathbf{z}^\top \boldsymbol{\theta} + z^0), \quad (5)$$

where  $\tilde{\mathbf{Z}} = (\tilde{\mathbf{Y}}, \tilde{\mathbf{z}}, \tilde{z}^0)$  includes random variables  $\tilde{\mathbf{Y}} \in \mathbb{S}^N$ ,  $\tilde{\mathbf{z}} \in \mathbb{R}^d$  and  $\tilde{z}^0 \in \mathbb{R}$ . The function  $\mathcal{L}(\cdot)$  is assumed to be  $L$ -Lipschitz continuous.

We consider the 1-Wasserstein ball, where the ground cost function  $c$  is given by the Schatten- $\infty$  norm. Under this setting, (RRMP) is equivalent to the Tikhonov regularized problem

$$\inf_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_{\hat{\mathbb{P}}(\boldsymbol{\theta}_t)} \left[ \mathcal{L} \left( \boldsymbol{\theta}^\top \tilde{\mathbf{Y}} \boldsymbol{\theta} + 2\tilde{\mathbf{z}}^\top \boldsymbol{\theta} + \tilde{z}^0 \right) \right] + \rho L \|\boldsymbol{\theta}\|_2^2.$$

This result demonstrates that, under mild assumptions, (RRMP) can be reformulated as a regularized risk minimization problem. This substantially generalizes the findings of [30], who established an equivalence to Tikhonov regularization for strongly convex quadratic loss functions and martingale-constrained Wasserstein ambiguity sets. Our result reveals that Tikhonov regularization can be obtained for a broader class of loss functions without requiring complicating martingale constraints, thereby sharpening the theoretical understanding of the connection between distributional robustness and regularization.

The following model provides a general formulation that is typically considered in the distributionally robust optimization literature [43, 44]. This model is pertinent to many applications in decision-making under uncertainty, including in inventory management [29] and energy systems [27]. To the best of our knowledge, such a formulation has not previously been proposed in the performative optimization literature, which mainly focuses on prediction problems. Note that if the  $L$ -lipschitz continuous function  $\mathcal{L}$  in (5) is piecewise linear, then this model constitutes a generalization.

**Model 2.** *Assume that the loss function is defined as*

$$\ell(\mathbf{Z}, \boldsymbol{\theta}) = \max_{j \in [J]} Q_j(\mathbf{Z}, \boldsymbol{\theta}),$$

where  $\tilde{\mathbf{Z}} = (\tilde{\mathbf{Y}}, \tilde{\mathbf{z}}, \tilde{z}^0)$  includes random variables  $\tilde{\mathbf{Y}} \in \mathbb{S}^N$ ,  $\tilde{\mathbf{z}} \in \mathbb{R}^d$  and  $\tilde{z}^0 \in \mathbb{R}$ . Each component  $Q_j(\mathbf{Z}, \boldsymbol{\theta})$  is a quadratic function of the form

$$Q_j(\mathbf{Z}, \boldsymbol{\theta}) = \mathbf{a}_j(\boldsymbol{\theta})^\top \mathbf{Y} \mathbf{a}_j(\boldsymbol{\theta}) + 2\mathbf{b}_j(\boldsymbol{\theta})^\top \mathbf{z} + c_j(\boldsymbol{\theta})z^0,$$

with parameter-dependent coefficients given by affine functions  $\mathbf{a}_j(\boldsymbol{\theta}) = \bar{\mathbf{a}}_j + \bar{\mathbf{A}}_j \boldsymbol{\theta}$ ,  $\mathbf{b}_j(\boldsymbol{\theta}) = \bar{\mathbf{b}}_j + \bar{\mathbf{B}}_j \boldsymbol{\theta}$ , and  $c_j(\boldsymbol{\theta}) = \bar{c}_{j0} + \bar{\mathbf{c}}_j^\top \boldsymbol{\theta}$  for all  $j \in [J]$ , where  $\bar{\mathbf{a}}_j, \bar{\mathbf{b}}_j \in \mathbb{R}^N$ ,  $\bar{c}_{j0} \in \mathbb{R}$ ,  $\bar{\mathbf{A}}_j, \bar{\mathbf{B}}_j \in \mathbb{R}^{N \times d}$ , and  $\bar{\mathbf{c}}_j \in \mathbb{R}^d$ .

Consider the 1-Wasserstein ball with the Schatten- $\infty$  norm ground cost. Then, the optimal value of the following exponential conic program provides an arbitrarily tight conservative approximation for (RRMP).

$$\begin{aligned} \inf_{\boldsymbol{\theta}} \quad & \sum_{s \in [S]} \hat{p}_s(\boldsymbol{\theta}_t) t_s + t_{S+1} \\ \text{s. t.} \quad & \boldsymbol{\theta} \in \Theta, \quad t_s \in \mathbb{R} \quad \forall s \in [S+1] \\ & \zeta_{s,j}, r_{s,j} \in \mathbb{R}, \quad (r_{s,j}, \mu, \zeta_{s,j} - t_s) \in K_{\text{exp}} \quad \forall s \in [S+1] \quad j \in [J] \\ & \boldsymbol{\theta}^\top \bar{\mathbf{A}}_j^\top \hat{\mathbf{Y}}_s \bar{\mathbf{A}}_j \boldsymbol{\theta} + (2\bar{\mathbf{a}}_j^\top \hat{\mathbf{Y}}_s \bar{\mathbf{A}}_j + 2\hat{\mathbf{z}}_s^\top \bar{\mathbf{B}}_j + \hat{z}_s^0 \bar{\mathbf{c}}_j^\top) \boldsymbol{\theta} \\ & \quad + \bar{\mathbf{a}}_j^\top \hat{\mathbf{Y}}_s \bar{\mathbf{a}}_j + 2\bar{\mathbf{b}}_j^\top \hat{\mathbf{z}}_s + \hat{z}_s^0 \bar{c}_{j0} \leq \zeta_{s,j} \quad \forall s \in [S+1] \quad j \in [J] \\ & \sum_{j \in J} r_{s,j} \leq \mu \quad \forall s \in [S+1]. \end{aligned} \tag{6}$$

where  $\hat{\mathbf{Y}}_{S+1} = \rho \mathbb{I}$ ,  $\hat{\mathbf{z}}_{S+1} = \mathbf{0}$ ,  $\hat{z}_{S+1}^0 = \rho$ , and  $\mu \in \mathbb{R}_+$  is the smoothing parameter.

The formulation (6) relies on the exponential smoothing techniques described in [7, Section 2.2]. The primary motivation for this approach lies in the need to handle the nonsmoothness incurred by the inner maximization over quadratic functions, which will hinder the convergence of our proposed algorithm. To address this challenge, we apply a log-sum-exp smoothing approximation with smoothing parameter  $\mu > 0$ , which yields a smoothed robust decoupled performative risk, denoted as  $\mathbb{J}_{\boldsymbol{\theta}_t}^\mu$ . And problem (6) is equivalent to  $\inf_{\boldsymbol{\theta} \in \Theta} \mathbb{J}_{\boldsymbol{\theta}_t}^\mu(\boldsymbol{\theta})$  (see details in Appendix D).

As is standard in exponential smoothing, the smoothed objective serves as a uniform upper bound to the original nonsmooth function:  $\mathbb{J}_{\boldsymbol{\theta}_t}(\boldsymbol{\theta}) \leq \mathbb{J}_{\boldsymbol{\theta}_t}^\mu(\boldsymbol{\theta}) \quad \forall \boldsymbol{\theta} \in \Theta$ , which ensures that optimization over the smooth surrogate does not underestimate the original objective. Importantly,  $\mathbb{J}_{\boldsymbol{\theta}_t}^\mu$  epi-converges to  $\mathbb{J}_{\boldsymbol{\theta}_t}$  as the smoothing parameter  $\mu \downarrow 0$  [46, Theorem 7.17], which ensures convergence of optimal solutions whenever  $\Theta$  is compact [46, Theorem 7.33].

Finally, inspired by minimax [52] and adversarially robust optimization literature [49], we turn to the setting where the loss function is convex-concave. In this setting, we exploit the 2-Wasserstein ambiguity set to ensure the convergence of the repeated risk minimization algorithm. This model further allows one to impose additional structural support information that may reduce the overconservatism of the distributionally robust solutions. The following theorem provides a convex reformulation for the model, leveraging convex conjugate representations and support functions.

**Model 3.** Let  $\mathcal{Z} \subseteq \mathbb{R}^m$  be a nonempty, convex and closed set, and consider the 2-Wasserstein ball, where the ground cost  $c$  is given by the Euclidean norm on  $\mathbb{R}^d$ . Suppose that for every  $\theta$ , the function  $\ell(z, \theta)$  is proper, concave, and upper-semicontinuous in  $z$ . Then, the optimal value of the following finite convex program provides an arbitrarily tight conservative approximation for (RRMP):

$$\begin{aligned} \inf \quad & \sum_{s \in [S]} \hat{p}_s(\theta_t) \left( [-\ell]^*(r_s - \zeta_s, \theta) + \sigma_{\mathcal{Z}}(\zeta_s) - r_s^\top \hat{z}_s + \frac{1}{4\lambda} \|r_s\|_2^2 \right) + \rho^2 \lambda + \tau \lambda^2 \\ \text{s. t.} \quad & \theta \in \Theta, \quad \lambda \in \mathbb{R}_+, \quad r_s \in \mathbb{R}^m \quad \forall s \in [S], \quad \zeta_s \in \mathbb{R}^m \quad \forall s \in [S] \end{aligned} \quad (7)$$

where  $\tau \in \mathbb{R}_{++}$  is a constant,  $[-\ell]^*(\xi, \theta) = \sup_{z \in \mathbb{R}^m} \xi^\top z - [-\ell(z, \theta)]$  denotes the conjugate of  $-\ell$  with respect to  $z$ , and  $\sigma_{\mathcal{Z}}(\xi) = \sup_{z \in \mathcal{Z}} \xi^\top z$  is the support function of  $\mathcal{Z} \in \mathbb{R}^m$ .

Note that in problem (7), the loss function  $\ell(z, \theta)$  and the support set  $\mathcal{Z}$  enter the formulation through the convex conjugate of the negative loss  $[-\ell]^*(\cdot, \theta)$ , and the support function  $\sigma_{\mathcal{Z}}(\cdot)$ , respectively. Both transformations yield convex functions under the assumptions stated in Model 3. Furthermore, the term  $(1/4\lambda) \|r_s\|_2^2$  is jointly convex in  $(\lambda, r_s)$  [9, section 3.2.6]. Consequently, all objective and constraint functions in problem (7) are convex, and the overall optimization problem is manifestly convex.

The reformulation in Model 3 introduces the dual variable  $\lambda$ , which appears alongside the decision variable  $\theta$ . This motivates the use of an augmented vector  $\bar{\theta} = (\theta, \lambda) \in \Theta \times \mathbb{R}_+ = \bar{\Theta}$ . To ensure strong convexity in the joint variable  $\bar{\theta}$  which is important for the convergence guarantees of the R<sup>3</sup>M algorithm, we introduce a regularization term  $\tau \lambda^2$ , where  $\tau > 0$ . This leads to the regularized robust decoupled performative risk, denoted by  $\mathbb{J}_{\theta_t}^\tau(\bar{\theta})$ . And problem (7) is equivalent to  $\inf_{\bar{\theta} \in \bar{\Theta}} \mathbb{J}_{\theta_t}^\tau(\bar{\theta})$ . Notice that for any  $\theta \in \Theta$ ,  $\inf_{\lambda \geq 0} \mathbb{J}_{\theta_t}^\tau(\theta, \lambda)$  converges to  $\mathbb{J}_{\theta_t}^\tau(\theta)$  as  $\tau \downarrow 0$ . If  $\ell(z, \theta)$  is lower semicontinuous in  $\theta$ , then so is  $\mathbb{J}_{\theta_t}^\tau$ . Hence,  $\inf_{\lambda \geq 0} \mathbb{J}_{\theta_t}^\tau(\cdot, \lambda)$ , epi-converges to  $\mathbb{J}_{\theta_t}^\tau(\cdot)$  [46, Theorem 7.17], and the minimizer  $\hat{\theta}$  of (7) converges to a minimizer of (RRMP) whenever  $\Theta$  is compact [46, Theorem 7.33].

### 3.2 Convergence Analysis

We now establish convergence guarantees for the RRRM algorithm when applied to the reformulations introduced in Models 1, 2, and 3. As noted earlier, we employ the smoothed objective  $\mathbb{J}_{\theta_t}^\mu(\theta)$  for (RRMP) in Model 2 and the regularized objective  $\mathbb{J}_{\theta_t}^\tau(\bar{\theta})$  in Model 3. Each model requires specific assumptions to ensure contraction of the risk map and hence convergence:

- (A1) Model 1: The loss function satisfies the  $\gamma$ -strong convexity (B1) and  $\beta$ -smoothness (B2).
- (A2) Model 2: For all  $j \in [J]$ ,  $\bar{A}_j^\top \bar{A}_j \succ 0$  and let  $\alpha = 2 \min_{j \in [J]} \lambda_{\min}(\bar{A}_j^\top \bar{A}_j)$ . In addition, the feasible set  $\Theta$  and the support of  $\hat{P}(\theta_t)$  are bounded.
- (A3) Model 3: The loss function satisfies the  $\gamma$ -strong convexity (B1) and  $\beta$ -jointly smoothness (B2). Additionally the support  $\mathcal{Z}$  has a finite diameter  $D < \infty$ .

Under these respective conditions, the algorithm converges linearly to a stable point.

**Theorem 1.** Suppose the loss functions in Models 1, 2, and 3 satisfy Assumptions (A1), (A2), and (A3) respectively, and that the distribution map  $\hat{P}(\cdot)$  satisfies the  $\epsilon$ -sensitivity condition (B3). Then:

- (a)  $\|\theta_{t+1} - \theta'_{t+1}\|_2 \leq \epsilon \kappa \|\theta_t - \theta'_t\|_2$  for all  $\theta_t, \theta'_t \in \Theta$ .
- (b) if  $\epsilon \kappa < 1$ , the iterate  $\theta_t$  of (RRMP) converges linearly to a unique performatively stable point  $\theta_{\text{RPS}}$ :

$$\|\theta_t - \theta_{\text{RPS}}\|_2 \leq \Delta \text{ for } t \geq (1 - \epsilon \kappa)^{-1} \log \left( \|\theta_0 - \theta_{\text{RPS}}\|_2 / \Delta \right).$$

where  $\Delta > 0$  is a predefined tolerance level. The fixed point  $\theta_{\text{RPS}}$  depends on the model and is denoted  $\theta_{\text{RPS}}^\mu$  under Model 2, and  $\theta_{\text{RPS}}^\tau$  under Model 3.

Here,  $\kappa = \beta / (\gamma + 2\rho L)$  for Model 1,  $\kappa = Jdk_3 \left( \frac{k_1 k_2}{\mu} + 1 \right) / \rho \alpha$  for Model 2, and  $\kappa = (\beta + 4D) / \min\{2\tau, \gamma\}$  for Model 3, where  $k_1, k_2, k_3 < \infty$  are model-dependent constants.

This convergence result highlights several compelling advantages of our DRPO framework, particularly when contrasted with traditional approaches to performative optimization that necessitate strong convexity and smoothness for convergence. First, our DRPO framework establishes convergence guarantees for loss functions that are convex but not necessarily strongly convex. Additionally, our exponential smoothing tricks applied in Model 2 enable the RRRM algorithm to converge for non-smooth loss functions. Finally, our DRPO framework accelerates the convergence rate compared with its non-robust counterpart. These advancements significantly broadens the scope of problems amenable to performative optimization and enhance the computational efficiency in practical problems.

### 3.3 Suboptimality Guarantees

Our next result demonstrates that the robust performatively stable solution  $\theta_{\text{RPS}}$  is close to the robust performatively optimal solution  $\theta_{\text{RPO}}$  whenever the  $\epsilon$ -sensitivity of the distribution map is small, the loss function has a large strong convexity parameter  $\gamma$ , or the distributionally robust model induces a regularization with large strong convexity parameter  $\rho$ .

**Theorem 2.** *Suppose all conditions in Theorem 1 hold. Additionally, assume that the loss function is  $L_z$ -Lipschitz in  $z$  for both models, and  $L_\theta$ -Lipschitz in  $\theta$  for Model 3. Then, the suboptimality gap between the robust performatively stable solution and the optimal solution under the true distribution is bounded as follows:*

- (a) **Model 1:**  $\mathbb{J}_{\theta_{\text{RPS}}}(\theta_{\text{RPS}}) - \mathbb{J}_{\theta_{\text{RPO}}}(\theta_{\text{RPO}}) \leq \frac{2\epsilon^2 L_z^2}{\gamma + 2\rho L}$ .
- (b) **Model 2:**  $\mathbb{J}_{\theta_{\text{RPS}}^\mu}^\mu(\theta_{\text{RPS}}^\mu) - \mathbb{J}_{\theta_{\text{RPO}}}(\theta_{\text{RPO}}) \leq \frac{2(\epsilon L_z + 2\mu' \log J)^2}{\rho \alpha}$ .
- (c) **Model 3:**  $\mathbb{J}_{\theta_{\text{RPS}}^\tau}^\tau(\bar{\theta}_{\text{RPS}}^\tau) - \mathbb{J}_{\theta_{\text{RPO}}}(\bar{\theta}_{\text{RPO}}) \leq \tau \bar{\lambda}^2 + \frac{(2\tau \bar{\lambda} + \rho^2 + D^2 + L_\theta + \epsilon L_z) 2\epsilon L_z}{\min(\gamma, 2\tau)}$ .

Here,  $\mu' \in [0, 1]$  is a constant that satisfies  $\mu \leq \mu' \|\theta_{\text{RPS}}^\mu - \theta_{\text{RPO}}\|_2$ , and  $\bar{\lambda}$  is the upper bound on the optimal  $\lambda$  as given in Lemma F.1.

The suboptimality bound for Model 2 also highlights the advantages of our DRPO framework alongside the smoothing tricks. For non-smooth, convex but not strongly-convex functions, the suboptimality bound under the traditional performative prediction framework can be arbitrarily large (as  $\rho \rightarrow 0, \mu \rightarrow 0$ ) as suggested by our result.

## 4 Experiments

In this section, we present numerical experiments to evaluate the performance of our proposed models across three applications: strategic classification, revenue management, and portfolio optimization. All experiments were conducted on a laptop equipped with a 6-core, 2.3 GHz Intel Core i7 CPU and 16 GB of RAM. The optimization problems were implemented in Python 3.11.

### 4.1 Strategic Classification

We consider a simulated strategic classification problem from [40] using a class-balanced subset of a Kaggle credit scoring dataset [25]. The dataset contains features  $\tilde{x} \in \mathbb{R}^P$  about borrowers, such as their ages and the number of open loans. The outcomes  $\tilde{y} \in \{-1, 1\}$  are equal to 1 if the individual defaulted on a loan and  $-1$  otherwise. The institution's objective is to predict whether an individual will default on their debt.

Under the strategic classification setting, individuals respond to the institution's classifier by altering their features to increase their likelihood of receiving a favorable classification. The institution employs logistic regression for classification, with  $\tilde{z} = \tilde{x}\tilde{y}$ , and the loss function is given by  $\log(1 + \exp(-\mathbf{x}^\top \theta y))$ . This setting aligns with Theorem 1, where  $\mathcal{L}$  represents the logloss function with Lipschitz constant  $L = 1$ , and the quadratic function  $\theta^\top \mathbf{Y}\theta + 2z^\top \theta + z^0$  simplifies to the affine function  $z^\top \theta$ . See Appendix G for additional details.

We compare the performance of our robust models (with type-1 and type-2 Wasserstein ambiguity sets) against the alternating minimization algorithm under KL divergence ambiguity set (AMKL)

from [58]. Additionally, we include a non-robust model as a baseline for comparison. The training set is fixed at 200 samples, while approximately 3,600 data points are used for out-of-sample testing. This setup reflects realistic scenarios where data collection is costly or limited. In credit scoring, for example, obtaining labeled data often requires expensive evaluations, expert assessments, or lengthy observation periods. All models are trained for 40 iterations, with the robust parameter set to 0.1 for all robust variants.

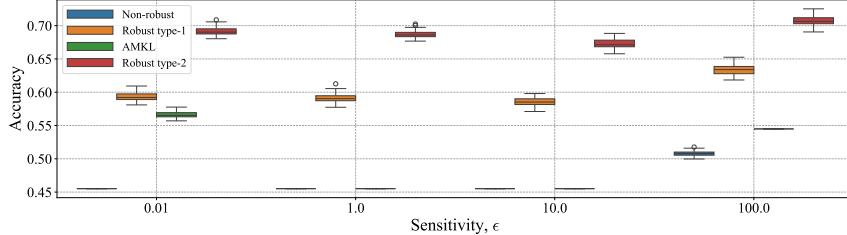


Figure 1: Out of sample performance of different approaches

Figure 1 displays the box plots of 50 independent trials. As shown, our robust models outperform the AMKL algorithm, which performs quite poorly. This may be due to its susceptibility to local optima, as well as the limitations of the KL divergence ambiguity set, which may not sufficiently guard against distributional shifts that are not well represented in the reference distribution. All robust models significantly outperform the non-robust baseline, highlighting the effectiveness of distributional robustness. Finally, we find that the robust model with a type-2 Wasserstein ambiguity set outperforms the one with a type-1 Wasserstein ambiguity set. This arises primarily due to the geometry of the ambiguity set. As discussed in [10], selecting the optimal radius  $\rho$  is challenging, and the 2-Wasserstein ball often provides better performance because it offers a wider range of radius values for which the robust solution outperforms its non-robust counterpart.

## 4.2 Revenue Management

In this experiment, we address the revenue management problem where the decision-maker determines the unit price  $\theta \geq 0$  for a fixed quantity of perishable products  $q \in \mathbb{Z}_{++}$ , such as hotel rooms or airplane seats, under uncertain demand  $\tilde{z} \sim \mathbb{P}(\theta)$ , with higher prices inducing lower demand [1, 13, 42].

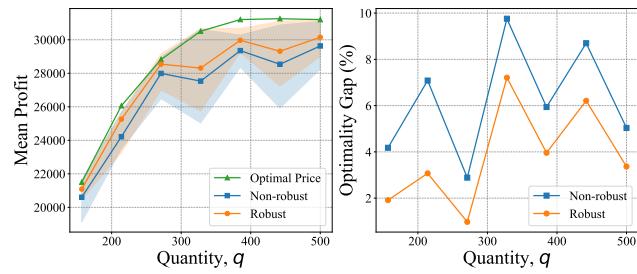


Figure 2: Out-of-sample performance of pricing schemes

Following [41], we model the price-dependent demands using an additive function  $\tilde{z}(\theta) = -a\theta + b + \tilde{\epsilon}$ . Here, unknown parameters  $a > 0$  and  $b > 0$  capture a deterministic linear demand curve, and  $\tilde{\epsilon}$  is a random variable with a bounded support that is characterized by an unknown density function. Instead of the true model, we only have access to a surrogate model  $-\bar{a}\theta + \bar{b}$ , which deviates from the true model as the parameters  $\bar{a}$  and  $\bar{b}$  do not accurately represent their counterparts. Under this additive demand setting, the associated loss function becomes a piecewise quadratic function in  $\theta$ , allowing us to apply Model 2 to formulate a robust version of the revenue management problem, which can be efficiently solved within seconds using an off-the-shelf commercial solver as it is a convex problem.

Figure 2 compares the out-of-sample performance of our robust scheme (orange) with benchmarks. Due to the non-smoothness of the loss function, the AMKL approach is not applicable in this setting. As shown in [41], when the true distribution map is available, the optimal price (green) can be derived in closed form, achieving the highest expected profit for each fixed quantity  $q$  in the left subplot. Similarly, the non-robust price (blue) can be obtained by treating the surrogate model as the true model. The confidence region in the left subplot, representing the 10th-90th percentile range of 100 independent tests, and the smaller optimality gap relative to the mean profit induced by the optimal price in the right subplot show that our robust price consistently outperforms the benchmark.

### 4.3 Demand Response Portfolio Optimization

We evaluate our robust scheme in a power system application, focusing on demand response (DR) portfolio optimization [51, 60]. In electricity markets, consumers capable of lowering electricity consumption during certain periods are called DR resources. In this experiment, we consider a DR aggregator (decision-maker) managing  $n$  DR resources over a planning horizon of  $T$  periods. The goal is to maximize the expected profit by determining commitment level  $\theta_t \in \mathbb{R}_+^n$  to meet a required deterministic demand reduction  $D_t$  at each time  $t \in [T]$ .

The challenge lies in the uncertainty of DR resource's performance, where the scheduled commitment level  $\theta_t$  may significantly differ from the actual reduction level  $\tilde{\theta}_t$  due to random noise  $\tilde{z}_t$ . This noise is decision-dependent, as larger commitments lead to higher variability. In this experiment, we model the actual reduction of each resource  $i$  as  $\tilde{\theta}_{t,i} = \theta_{t,i}\tilde{z}_{t,i}$  for all  $i \in [n]$ . Here, the multiplicative noise follows a beta distribution whose parameters  $\alpha$  and  $\beta$  depend linearly on the commitment level:  $\tilde{z}_{t,i} \sim 2 \cdot \text{Beta}(\alpha = \beta = a_i\theta_{t,i} + b_i)$  with  $a_i < 0$ . This beta distribution has a support of  $[0, 2]$  and a mean of 1, regardless of the value of  $a_i\theta_{t,i} + b_i$ . Therefore, the decision  $\theta_{t,i}$  only influences the distribution shape, with higher commitment levels leading to heavier tails, hence, higher variability of the actual reduction level  $\tilde{\theta}_{t,i}$ .

We consider three DR resources with distinct characteristics: Resource 1 has high revenue but large variability, Resource 2 has low revenue but high predictability, and Resource 3 offers a balanced trade-off between the two. We follow the experiment setup in [12], including the loss function and the values of unit revenue, over-commitment cost, and under-commitment penalty. As their loss function is piecewise-linear, the resulting optimization problem can be formulated using Model 2 and efficiently solved. For further details, we refer the reader to the original paper. We conduct two out-of-sample tests: low and high demand loads over the planning horizon, corresponding to small  $D_t$  and large  $D_t$ , respectively. Figure 3 compares the mean profits of our robust scheme (orange) and the non-robust scheme (blue), showing that the robust approach outperforms the non-robust one. Notably, under high demand, the non-robust scheme often performs significantly poorly, resulting in losses in some cases.

## 5 Concluding Remarks

We have presented the first Wasserstein distributionally robust optimization framework for performative optimization. In contrast to existing approaches, our framework accommodates a broader class of problems in decision-making under uncertainty, thereby extending the original scope of performative prediction. We proposed an efficient algorithm and established its convergence and suboptimality guarantees. To our knowledge, these theoretical results have not been previously established in the literature on robust performative prediction. Our experimental results demonstrate the superiority of our approach over existing methods on the standard strategic classification benchmark, as well as in two decision-making applications: revenue management and demand response portfolio optimization.

Notably, as the ambiguity set radius  $\rho$  approaches zero, the robust objective coincides with the non-robust counterpart, which more directly targets the performative risk. This observation suggests a possible direction: designing algorithms that gradually shrink the ambiguity set over time, potentially trading robustness for improved approximation of the true performative risk as more information becomes available. Another promising direction is contextual performative optimization, where incorporating side information could further improve decision quality by enabling more accurate modeling of uncertainty.

**Broader impacts.** Our framework extends the scope of performative prediction beyond its original focus, enabling its application to a wider range of decision-making problems. In high-stakes settings, adopting a distributionally robust optimization perspective allows our approach to prioritize safe and reliable deployment in the presence of uncertainty and potential adversarial conditions.

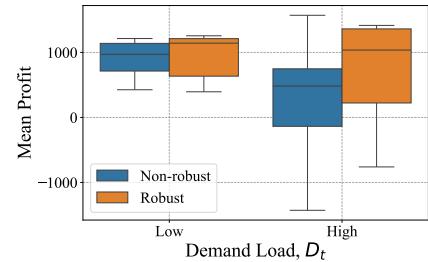


Figure 3: Out-of-sample performance of DR scheduling

## Acknowledgments and Disclosure of Funding

Grani A. Hanususanto is supported in part by the National Science Foundation (NSF) under Grants CCF-2343869 and ECCS-2404413. Roy Dong is supported in part by NSF under Grant CCF-2236484. Yijie Wang is supported in part by the Fundamental Research Funds for the Central Universities. We thank Hyuk Park for assistance with the numerical experiments and the anonymous reviewers for their constructive feedback that helped improve this work.

## References

- [1] J. Alstrup, S.-E. Andersson, S. Boas, O. B. Madsen, and R. V. V. Vidal. Booking control increases profit at scandinavian airlines. *Interfaces*, 19(4):10–19, 1989.
- [2] M. Baker and J. Wurgler. Investor sentiment and the cross-section of stock returns. *The journal of Finance*, 61(4):1645–1680, 2006.
- [3] P. L. Bartlett. Learning with a slowly changing distribution. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 243–252, 1992.
- [4] P. L. Bartlett, S. Ben-David, and S. R. Kulkarni. Learning changing concepts by exploiting the structure of change. In *Proceedings of the ninth annual conference on Computational learning theory*, pages 131–139, 1996.
- [5] B. Basciftci, S. Ahmed, and S. Shen. Distributionally robust facility location problem under decision-dependent stochastic demand. *European Journal of Operational Research*, 292(2):548–561, 2021.
- [6] D. Bertsekas. *Convex optimization theory*, volume 1. Athena Scientific, 2009.
- [7] D. Bertsekas. *Convex optimization algorithms*. Athena Scientific, 2015.
- [8] J. Blanchet and K. Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.
- [9] S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- [10] G. Byeon. Comparative analysis of two-stage distributionally robust optimization over 1-wasserstein and 2-wasserstein balls. *arXiv preprint arXiv:2501.05619*, 2025.
- [11] A. P. Calmon, F. D. Ciocan, and G. Romero. Revenue management with repeated customer interactions. *Management Science*, 67(5):2944–2963, 2021.
- [12] H. Chen, X. A. Sun, and H. Yang. Robust optimization with continuous decision-dependent uncertainty with applications to demand response management. *SIAM Journal on Optimization*, 33(3):2406–2434, 2023.
- [13] G. D. DeYong. The price-setting newsvendor: review and extensions. *International Journal of Production Research*, 58(6):1776–1804, 2020.
- [14] X. V. Doan. Distributionally robust optimization under endogenous uncertainty with an application in retrofitting planning. *European Journal of Operational Research*, 300(1):73–84, 2022.
- [15] D. Drusvyatskiy and L. Xiao. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 48(2):954–998, 2023.
- [16] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.
- [17] B. Gao and L. Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

[18] V. Goel and I. E. Grossmann. A stochastic programming approach to planning of offshore gas field developments under uncertainty in reserves. *Computers & chemical engineering*, 28(8):1409–1429, 2004.

[19] V. Goel and I. E. Grossmann. A class of stochastic programs with decision dependent uncertainty. *Mathematical programming*, 108(2):355–394, 2006.

[20] M. Hardt, M. Jagadeesan, and C. Mendler-Dünner. Performative power. *Advances in Neural Information Processing Systems*, 35:22969–22981, 2022.

[21] M. Hardt and C. Mendler-Dünner. Performative prediction: Past and future. *arXiv preprint arXiv:2310.16608*, 2023.

[22] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. *Applied logistic regression*. John Wiley & Sons, 2013.

[23] Q. Jin, A. Georghiou, P. Vayanos, and G. A. Hanasusanto. Distributionally robust optimization with decision-dependent information discovery. *arXiv preprint arXiv:2404.05900*, 2024.

[24] T. W. Jónsbråten, R. J. Wets, and D. L. Woodruff. A class of stochastic programs with decision dependent random elements. *Annals of Operations Research*, 82(0):83–106, 1998.

[25] Kaggle. Give me some credit. <https://www.kaggle.com/c/GiveMeSomeCredit/data>, 2012.

[26] L. V. Kantorovich and S. Rubinshtein. On a space of totally additive functions. *Vestnik of the St. Petersburg University: Mathematics*, 13(7):52–59, 1958.

[27] J. H. Kim and W. B. Powell. Optimal energy commitments with storage and intermittent supply. *Operations research*, 59(6):1347–1360, 2011.

[28] M. P. Kim and J. C. Perdomo. Making decisions under outcome performativity. In *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2023.

[29] S. Lee, H. Kim, and I. Moon. A data-driven distributionally robust newsvendor model with a wasserstein ambiguity set. *Journal of the Operational Research Society*, 72(8):1879–1897, 2021.

[30] J. Li, S. Lin, J. Blanchet, and V. A. Nguyen. Tikhonov regularization is optimal transport robust under martingale constraints. *Advances in Neural Information Processing Systems*, 35:17677–17689, 2022.

[31] F. Luo and S. Mehrotra. Distributionally robust optimization with decision dependent ambiguity sets. *Optimization Letters*, 14(8):2565–2594, 2020.

[32] C. Mendler-Dünner, J. Perdomo, T. Zrnic, and M. Hardt. Stochastic optimization for performativity prediction. *Advances in Neural Information Processing Systems*, 33:4929–4939, 2020.

[33] G. Michel, J. Omer, and M. Poss. Robust selection problem with decision-dependent information discovery under budgeted uncertainty. In *23ème congrès annuel de la Société Française de Recherche Opérationnelle et d'Aide à la Décision*, 2022.

[34] J. P. Miller, J. C. Perdomo, and T. Zrnic. Outside the echo chamber: Optimizing the performative risk. In *International Conference on Machine Learning*, pages 7710–7720. PMLR, 2021.

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

[36] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

[37] O. Nohadani and K. Sharma. Optimization under decision-dependent uncertainty. *SIAM Journal on Optimization*, 28(2):1773–1795, 2018.

[38] R. Paradiso, A. Georghiou, S. Dabia, and D. Tönissen. Exact and approximate schemes for robust optimization problems with decision dependent information discovery. *arXiv preprint arXiv:2208.04115*, 2022.

[39] L. Peet-Pare, N. Hegde, and A. Fyshe. Long term fairness for minority groups via performative distributionally robust optimization. *arXiv preprint arXiv:2207.05777*, 2022.

[40] J. Perdomo, T. Zrnic, C. Mendler-Dünner, and M. Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

[41] N. C. Petrucci and M. Dada. Pricing and the newsvendor problem: A review with extensions. *Operations research*, 47(2):183–194, 1999.

[42] A. Popescu, P. Keskinocak, E. Johnson, M. LaDue, and R. Kasilingam. Estimating air-cargo overbooking based on a discrete show-up-rate distribution. *Interfaces*, 36(3):248–258, 2006.

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

[44] H. Rahimian and S. Mehrotra. Frameworks and results in distributionally robust optimization. *Open Journal of Mathematical Optimization*, 3:1–85, 2022.

[45] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[46] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[47] M. Ryu and R. Jiang. Nurse staffing under absenteeism: A distributionally robust optimization approach. *arXiv preprint arXiv:1909.09875*, 2019.

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

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

[50] S. A. Spacey, W. Wiesemann, D. Kuhn, and W. Luk. Robust software partitioning with multiple instantiation. *INFORMS Journal on Computing*, 24(3):500–515, 2012.

[51] J.-H. Teng and C.-H. Hsieh. Modeling and investigation of demand response uncertainty on reliability assessment. *Energies*, 14(4):1104, 2021.

[52] K. K. Thekumpampil, P. Jain, P. Netrapalli, and S. Oh. Efficient algorithms for smooth minimax optimization. *Advances in neural information processing systems*, 32, 2019.

[53] J. v. Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

[54] P. Vayanos, D. Kuhn, and B. Rustem. Decision rules for information discovery in multi-stage stochastic programming. In *2011 50th IEEE Conference on Decision and Control and European Control Conference*, pages 7368–7373. IEEE, 2011.

[55] C. Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

[56] R. Vujanic, P. Goulart, and M. Morari. Robust optimization of schedules affected by uncertain events. *Journal of Optimization Theory and Applications*, 171:1033–1054, 2016.

[57] R. Wermers. Mutual fund herding and the impact on stock prices. *the Journal of Finance*, 54(2):581–622, 1999.

[58] S. Xue and Y. Sun. Distributionally robust performative prediction. *arXiv preprint arXiv:2412.04346*, 2024.

- [59] X. Yu and S. Shen. Multistage distributionally robust mixed-integer programming with decision-dependent moment-based ambiguity sets. *Mathematical Programming*, 196(1):1025–1064, 2022.
- [60] J. Zhang and A. D. Domínguez-García. Evaluation of demand response resource aggregation system capacity under uncertainty. *IEEE Transactions on Smart Grid*, 9(5):4577–4586, 2017.
- [61] J. Zhang, H. Xu, and L. Zhang. Quantitative stability analysis for distributionally robust optimization with moment constraints. *SIAM Journal on Optimization*, 26(3):1855–1882, 2016.
- [62] Q. Zhang and W. Feng. A unified framework for adjustable robust optimization with endogenous uncertainty. *AIChE journal*, 66(12):e17047, 2020.
- [63] X. Zhang, M. Kamgarpour, A. Georgiou, P. Goulart, and J. Lygeros. Robust optimal control with adjustable uncertainty sets. *Automatica*, 75:249–259, 2017.
- [64] J. Zhen, D. Kuhn, and W. Wiesemann. Mathematical foundations of robust and distributionally robust optimization. *arXiv preprint arXiv:2105.00760*, 2021.

## Technical Appendices and Supplementary Material

### A Preliminary Definitions

**Definition 5** (Generalized strong convexity). We say that a loss function  $\ell(z, \theta)$  is  $\gamma$ -strongly convex in  $\theta$  if

$$\ell(z, \theta) \geq \ell(z, \theta') + \nabla_{\theta} \ell(z, \theta')^\top (\theta - \theta') + \frac{\gamma}{2} \|\theta - \theta'\|_2^2, \quad (\text{B1})$$

for all  $\theta, \theta' \in \Theta$  and  $z \in \mathcal{Z}$ . If  $\gamma = 0$ , this condition reduces to the standard definition of convexity. We will also use the following equivalent definition of strong convexity. A loss function  $\ell(z, \theta)$  is  $\gamma$ -strongly convex in  $\theta$  if the function

$$f(z, \theta) = \ell(z, \theta) - \frac{\gamma}{2} \|\theta\|_2^2 \quad (\text{B1}')$$

is convex for all  $z \in \mathcal{Z}$ .

**Definition 6** (Smoothness). We say that a loss function  $\ell(z, \theta)$  is  $\beta$ -smooth if the gradient  $\nabla_{\theta} \ell(z, \theta)$  is  $\beta$ -Lipschitz in  $z$ , that is

$$\|\nabla_{\theta} \ell(z, \theta) - \nabla_{\theta} \ell(z', \theta)\|_2 \leq \beta \|z - z'\|_2, \quad (\text{B2})$$

for all  $z, z' \in \mathcal{Z}$ .

**Definition 7** ( $\epsilon$ -sensitivity). We say that a distribution map  $\mathbb{P}(\cdot)$  is  $\epsilon$ -sensitive if for all  $\theta, \theta' \in \Theta$

$$W_1(\mathbb{P}(\theta), \mathbb{P}(\theta')) \leq \epsilon \|\theta - \theta'\|_2, \quad (\text{B3})$$

where  $W_1$  denotes the 1-Wasserstein metric.

### B Background on Exponential Smoothing

Let  $\mathcal{Z} = \{z_1, \dots, z_n\}$  denote a finite support set. Given a loss function  $\ell : \Theta \times \mathcal{Z} \rightarrow \mathbb{R}$  and a distribution  $\hat{p} \in \Delta_n$ , where  $\Delta_n$  denotes the probability simplex. We define the exponentially smoothed objective as:

$$f_{\mu}(\theta) := \mu \cdot \log \left( \sum_{i=1}^n \hat{p}_i \cdot \exp \left( \frac{\ell(\theta, z_i)}{\mu} \right) \right),$$

where  $\mu > 0$  is a smoothing parameter.

This function, also known as the log-sum-exp function, is a widely used smooth approximation to the pointwise maximum function and has well-established properties in convex analysis [Bertsekas, 2015; Section 2.2]. This function provides a smooth approximation to  $\max_i \ell(\theta, z_i)$  whenever  $\hat{p}_i > 0$  for all  $i$ .

**Properties.** The function  $f_{\mu}(\theta)$  satisfies the following:

- Approximation Bounds: The approximation error can be precisely quantified:

$$\max_i \ell(\theta, z_i) \leq f_{\mu}(\theta) \leq \max_i \ell(\theta, z_i) + \mu \cdot \log \left( \frac{1}{\min_i \hat{p}_i} \right),$$

provided  $\hat{p}_i > 0$  for all  $i$ . Thus, as  $\mu \rightarrow 0$ , the smoothed objective  $f_{\mu}(\theta)$  approaches the exact maximum, with the error vanishing linearly in  $\mu$  up to a logarithmic multiplicative factor.

- Convexity and Differentiability: If each function  $\ell(\theta, z_i)$  is convex in  $\theta$ , then  $f_{\mu}(\theta)$  is also convex, as it is a composition of convex functions closed under nonnegative weighted log-sum-exp operations. Moreover,  $f_{\mu}(\theta)$  is continuously differentiable for all  $\mu > 0$ , with gradient:

$$\nabla_{\theta} f_{\mu}(\theta) = \sum_{i=1}^n \pi_{\mu}(z_i; \theta) \cdot \nabla_{\theta} \ell(\theta, z_i),$$

where the weight vector  $\pi_{\mu}(\cdot; \theta) \in \Delta_n$  defines a softmax distribution:

$$\pi_{\mu}(z_i; \theta) := \frac{\hat{p}_i \cdot \exp(\ell(\theta, z_i)/\mu)}{\sum_{j=1}^n \hat{p}_j \cdot \exp(\ell(\theta, z_j)/\mu)}.$$

This smoothing mechanism not only ensures differentiability but also facilitates efficient computation of gradients for robust optimization objectives involving maxima.

## C Auxiliary Lemmas

**Lemma C.1** (First-order optimality condition; Section 4.2.3 in [9]). *Let  $f$  be a convex function and let  $\Omega$  be a closed convex set on which  $f$  is differentiable, then*

$$\mathbf{x}^* \in \arg \min_{\mathbf{x} \in \Omega} f(\mathbf{x})$$

*if and only if*

$$\nabla f(\mathbf{x}^*)^T(\mathbf{y} - \mathbf{x}^*) \geq 0 \quad \forall \mathbf{y} \in \Omega.$$

**Lemma C.2** ([26]). *A distribution map  $\mathbb{P}(\cdot)$  is  $\epsilon$ -sensitive if and only if for all  $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$ , we have*

$$\sup_{g \in \mathcal{L}} \left| \mathbb{E}_{\mathbb{P}(\boldsymbol{\theta})}[g(\tilde{\mathbf{z}})] - \mathbb{E}_{\mathbb{P}(\boldsymbol{\theta}')}[g(\tilde{\mathbf{z}})] \right| \leq \epsilon L \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2,$$

where

$$\mathcal{L} = \{g : \mathbb{R}^p \rightarrow \mathbb{R} \mid |g(\mathbf{z}) - g(\mathbf{z}')| \leq L \|\mathbf{z} - \mathbf{z}'\|_2 \quad \forall \mathbf{z}, \mathbf{z}' \in \mathcal{Z}\}$$

is the space of all  $L$ -Lipschitz continuous functions.

**Lemma C.3.** *If  $\ell(\mathbf{z}, \boldsymbol{\theta})$  is  $\gamma$ -strongly convex, then the worst case expectation*

$$\mathbb{J}_{\boldsymbol{\eta}}(\boldsymbol{\theta}) = \sup_{\mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}}(\boldsymbol{\eta}))} \mathbb{E}_{\mathbb{Q}}[\ell(\tilde{\mathbf{z}}, \boldsymbol{\theta})]$$

is  $\gamma$ -strongly convex in  $\boldsymbol{\theta}$ .

*Proof.* By the equivalent definition of  $\gamma$ -strong convexity in (B1'), we have that  $\ell(\mathbf{z}, \boldsymbol{\theta}) - \frac{\gamma}{2} \|\boldsymbol{\theta}\|^2$  is convex in  $\boldsymbol{\theta}$ . Hence, the worst-case expectation

$$\sup_{\mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}}(\boldsymbol{\eta}))} \mathbb{E}_{\mathbb{Q}} \left[ \ell(\tilde{\mathbf{z}}, \boldsymbol{\theta}) - \frac{\gamma}{2} \|\boldsymbol{\theta}\|_2^2 \right]$$

is convex in  $\boldsymbol{\theta}$  since the expectation and the pointwise supremum operations preserve convexity. Thus, we have

$$\mathbb{J}_{\boldsymbol{\eta}}(\boldsymbol{\theta}) = \sup_{\mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}}(\boldsymbol{\eta}))} \mathbb{E}_{\mathbb{Q}} \left[ \ell(\tilde{\mathbf{z}}, \boldsymbol{\theta}) - \frac{\gamma}{2} \|\boldsymbol{\theta}\|_2^2 \right] + \frac{\gamma}{2} \|\boldsymbol{\theta}\|_2^2$$

is  $\gamma$ -strongly convex in  $\boldsymbol{\theta}$  by the definition (B1'). □

**Lemma C.4.** *Let  $\mathbf{x} \in \mathbb{R}^J$  and define the smooth maximum function*

$$f_{\mu}(\mathbf{x}) = \mu \log \left( \sum_{j \in [J]} e^{x_j/\mu} \right)$$

for any  $\mu > 0$ . Then the following bounds hold:

$$\max_{j \in [J]} x_j \leq f_{\mu}(\mathbf{x}) \leq \max_{j \in [J]} x_j + \mu \log J.$$

*Proof.* Let  $M := \max_{j \in [J]} x_j$ . Then for each  $j \in [J]$ , we have  $x_j \leq M$ , and hence:

$$\sum_{j \in [J]} e^{x_j/\mu} \leq \sum_{j \in [J]} e^{M/\mu} = J e^{M/\mu}.$$

Taking logarithms and multiplying by  $\mu$ , we obtain the upper bound:

$$f_{\mu}(\mathbf{x}) = \mu \log \left( \sum_{j \in [J]} e^{x_j/\mu} \right) \leq \mu \log \left( J e^{M/\mu} \right) = \mu \log J + M. \quad (8)$$

For the lower bound, observe that

$$\sum_{j \in [J]} e^{x_j/\mu} \geq e^{M/\mu}$$

since at least one term in the sum equals  $e^{M/\mu}$ . Thus:

$$f_\mu(\mathbf{x}) = \mu \log \left( \sum_{j \in [J]} e^{x_j/\mu} \right) \geq \mu \log \left( e^{M/\mu} \right) = M \quad (9)$$

Combining both bounds in (8) and (9), we have the desired result:

$$\max_{j \in [J]} x_j \leq f_\mu(\mathbf{x}) \leq \max_{j \in [J]} x_j + \mu \log J.$$

□

## D Deferred Proofs Related to Reformulations

### D.1 Proof of Reformulation for Model 1

*Proof.* We begin by rewriting the random parameters  $\tilde{\mathbf{Z}}$  as

$$\tilde{\mathbf{Z}} = \begin{bmatrix} \tilde{\mathbf{Y}} & \tilde{\mathbf{z}} \\ \tilde{\mathbf{z}}^\top & \tilde{z}^0 \end{bmatrix} \in \mathbb{S}^{N+1}.$$

Next, we introduce the matrix variable

$$\boldsymbol{\Gamma} = \begin{bmatrix} \boldsymbol{\theta} \boldsymbol{\theta}^\top & \boldsymbol{\theta} \\ \boldsymbol{\theta}^\top & 1 \end{bmatrix}, \quad (10)$$

which allows us to rewrite the loss function as  $\ell(\mathbf{Z}, \boldsymbol{\theta}) = \mathcal{L}(\langle \boldsymbol{\Gamma}, \mathbf{Z} \rangle)$ . According to [8, Remark 1], the worst-case expected loss over a 1-Wasserstein ambiguity set  $\mathbb{B}(\hat{\mathbb{P}}(\boldsymbol{\theta}_t))$ , with cost induced by the Schatten- $\infty$  norm, is given by:

$$\sup_{\mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}}(\boldsymbol{\theta}_t))} \mathbb{E}_{\mathbb{Q}}[\ell(\tilde{\mathbf{Z}}, \boldsymbol{\theta})] = \inf_{\lambda \in \mathbb{R}_+} \rho \lambda + \mathbb{E}_{\hat{\mathbb{P}}(\boldsymbol{\theta}_t)} \left[ \sup_{\mathbf{Z} \in \mathbb{S}^{N+1}} \mathcal{L}(\langle \boldsymbol{\Gamma}, \mathbf{Z} \rangle) - \lambda \|\mathbf{Z} - \tilde{\mathbf{Z}}\|_\infty \right].$$

By applying [48, Lemma 47], the inner maximization problem can be simplified as:

$$\sup_{\mathbf{Z} \in \mathbb{S}^{N+1}} \mathcal{L}(\langle \boldsymbol{\Gamma}, \mathbf{Z} \rangle) - \lambda \|\mathbf{Z} - \hat{\mathbf{Z}}\|_\infty = \begin{cases} \mathcal{L}(\langle \boldsymbol{\Gamma}, \hat{\mathbf{Z}} \rangle) & \text{if } L\|\boldsymbol{\Gamma}\|_1 \leq \lambda \\ +\infty & \text{otherwise.} \end{cases}$$

Hence, the worst-case expectation reduces to

$$\sup_{\mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}}(\boldsymbol{\theta}_t))} \mathbb{E}_{\mathbb{Q}}[\ell(\tilde{\mathbf{Z}}, \boldsymbol{\theta})] = \rho L \|\boldsymbol{\Gamma}\|_1 + \mathbb{E}_{\hat{\mathbb{P}}(\boldsymbol{\theta}_t)}[\mathcal{L}(\langle \boldsymbol{\Gamma}, \tilde{\mathbf{Z}} \rangle)].$$

To conclude, we observe that

$$\|\boldsymbol{\Gamma}\|_1 = \left\| \begin{bmatrix} \boldsymbol{\theta} \boldsymbol{\theta}^\top & \boldsymbol{\theta} \\ \boldsymbol{\theta}^\top & 1 \end{bmatrix} \right\|_1 = \text{tr} \left( \begin{bmatrix} \boldsymbol{\theta} \boldsymbol{\theta}^\top & \boldsymbol{\theta} \\ \boldsymbol{\theta}^\top & 1 \end{bmatrix} \right) = \|\langle \boldsymbol{\theta}, 1 \rangle\|_2^2.$$

This completes the proof. □

### D.2 Proof of Reformulation for Model 2

*Proof.* Rewriting the random parameters  $\tilde{\mathbf{Z}}$  as

$$\tilde{\mathbf{Z}} = \begin{bmatrix} \tilde{\mathbf{Y}} & \tilde{\mathbf{z}} \\ \tilde{\mathbf{z}}^\top & \tilde{z}^0 \end{bmatrix} \in \mathbb{S}^{N+1},$$

and introducing the matrix variable

$$\boldsymbol{\Gamma}_j = \begin{bmatrix} \mathbf{a}_j(\boldsymbol{\theta}) \mathbf{a}_j(\boldsymbol{\theta})^\top & \mathbf{b}_j(\boldsymbol{\theta}) \\ \mathbf{b}_j(\boldsymbol{\theta})^\top & c_j(\boldsymbol{\theta}) \end{bmatrix}, \quad (11)$$

allow us to rewrite  $Q_j(\mathbf{Z}, \boldsymbol{\theta})$  as  $\langle \boldsymbol{\Gamma}_j, \mathbf{Z} \rangle$ . By [35, Remark 6.6], the robust decoupled performatice risk can be expressed as

$$\mathbb{J}_{\boldsymbol{\theta}_t}(\boldsymbol{\theta}) = \sup_{\mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}}(\boldsymbol{\theta}_t))} \mathbb{E}_{\mathbb{Q}}[\ell(\tilde{\mathbf{Z}}, \boldsymbol{\theta})] = \sum_{s \in [S]} \hat{p}_s(\boldsymbol{\theta}_t) \left[ \max_{j \in [J]} Q_j(\tilde{\mathbf{Z}}_s, \boldsymbol{\theta}) \right] + \max_{j \in [J]} \rho \|\boldsymbol{\Gamma}_j\|_1,$$

where

$$\|\boldsymbol{\Gamma}_j\|_1 = \left\| \begin{bmatrix} \mathbf{a}_j(\boldsymbol{\theta}) \mathbf{a}_j(\boldsymbol{\theta})^\top & \mathbf{b}_j(\boldsymbol{\theta}) \\ \mathbf{b}_j(\boldsymbol{\theta})^\top & c_j(\boldsymbol{\theta}) \end{bmatrix} \right\|_1 = \text{tr} \left( \begin{bmatrix} \mathbf{a}_j(\boldsymbol{\theta}) \mathbf{a}_j(\boldsymbol{\theta})^\top & \mathbf{b}_j(\boldsymbol{\theta}) \\ \mathbf{b}_j(\boldsymbol{\theta})^\top & c_j(\boldsymbol{\theta}) \end{bmatrix} \right) = \mathbf{a}_j(\boldsymbol{\theta})^\top \mathbf{a}_j(\boldsymbol{\theta}) + c_j(\boldsymbol{\theta}).$$

Applying the exponential smoothing techniques described in [7, Section 2.2], we obtain the following smooth approximation of  $\mathbb{J}_{\boldsymbol{\eta}}(\boldsymbol{\theta})$ ;

$$\mathbb{J}_{\boldsymbol{\theta}_t}^\mu(\boldsymbol{\theta}) = \mu \sum_{s \in [S]} \hat{p}_s(\boldsymbol{\theta}_t) \log \left( \sum_{j \in [J]} e^{Q_j(\tilde{\mathbf{Z}}_s, \boldsymbol{\theta})/\mu} \right) + \mu \log \left( \sum_{j \in [J]} e^{\rho \|\boldsymbol{\Gamma}_j\|_1/\mu} \right), \quad (12)$$

where  $\mu \in \mathbb{R}_{++}$  is smoothing parameter. Introducing epigraphical variables, we reformulate the objective function as the optimal value of the convex program

$$\begin{aligned} \inf & \sum_{s \in [S]} \hat{p}_s(\boldsymbol{\theta}_t) t_s + t_{S+1} \\ \text{s. t.} & t_s \in \mathbb{R} \quad \forall s \in [S+1] \\ & \zeta_{s,j} \in \mathbb{R} \quad \forall s \in [S+1] \quad j \in [J] \\ & \rho \|\boldsymbol{\Gamma}_j\|_1 \leq \zeta_{s+1,j} \quad \forall j \in [J] \\ & Q_j(\tilde{\mathbf{Z}}_s, \boldsymbol{\theta}) \leq \zeta_{s,j} \quad \forall s \in [S+1] \quad j \in [J] \\ & \mu \log \left( \sum_{j \in [J]} e^{\zeta_{s,j}/\mu} \right) \leq t_s \quad \forall s \in [S+1]. \end{aligned} \quad (13)$$

The last constraint of (13) is equivalent to  $\sum_{j \in [J]} \mu e^{\zeta_{s,j}/\mu - t_s/\mu} \leq \mu$  and can be reformulated using the exponential cone:

$$\sum_{j \in [J]} r_{s,j} \leq \mu, \quad (r_{s,j}, \mu, \zeta_{s,j} - t_s) \in K_{\text{exp}} \quad \forall j \in [J]$$

where the exponential cone  $K_{\text{exp}}$  is defined as

$$K_{\text{exp}} = \{(x_1, x_2, x_3) : x_1 \geq x_2 e^{x_3/x_2}, x_2 > 0\} \cup \{(x_1, 0, x_3) : x_1 \geq 0, x_3 \leq 0\}.$$

To complete the proof, we substitute the expressions for  $\|\boldsymbol{\Gamma}_j\|_1 \quad \forall j \in [J]$  with

$$\boldsymbol{\theta}^\top \bar{\mathbf{A}}_j^\top \bar{\mathbf{A}}_j \boldsymbol{\theta} + (2\bar{\mathbf{a}}_j^\top \bar{\mathbf{A}}_j + \bar{\mathbf{c}}_j^\top) \boldsymbol{\theta} + \bar{\mathbf{a}}_j^\top \bar{\mathbf{a}}_j + \bar{c}_{j0},$$

and for  $Q_j(\tilde{\mathbf{Z}}_s, \boldsymbol{\theta}) \quad \forall s \in [S] \quad j \in [J]$  with

$$\boldsymbol{\theta}^\top \bar{\mathbf{A}}_j^\top \hat{\mathbf{Y}}_s \bar{\mathbf{A}}_j \boldsymbol{\theta} + (2\bar{\mathbf{a}}_j^\top \hat{\mathbf{Y}}_s \bar{\mathbf{A}}_j + 2\hat{\mathbf{z}}_s^\top \bar{\mathbf{B}}_j + \hat{\mathbf{z}}_s^0 \bar{\mathbf{c}}_j^\top) \boldsymbol{\theta} + \bar{\mathbf{a}}_j^\top \hat{\mathbf{Y}}_s \bar{\mathbf{a}}_j + 2\bar{\mathbf{b}}_j^\top \hat{\mathbf{z}}_s + \hat{\mathbf{z}}_s^0 \bar{c}_{j0}.$$

This completes the proof.  $\square$

### D.3 Proof of Reformulation for Model 3

*Proof.* By using Definition 1, the robust decoupled performatice risk  $\mathbb{J}_{\boldsymbol{\theta}_t}(\boldsymbol{\theta})$  can be rewritten as

$$\begin{aligned} \mathbb{J}_{\boldsymbol{\theta}_t}(\boldsymbol{\theta}) &= \sup_{\mathbb{Q} \in \mathbb{B}(\hat{\mathbb{P}}(\boldsymbol{\theta}_t))} \mathbb{E}_{\mathbb{Q}}[\ell(\tilde{\mathbf{z}}, \boldsymbol{\theta})] \\ &= \begin{cases} \sup_{\mathbb{Q}_s \in \mathcal{M}(\mathcal{Z})} \sum_{s \in [S]} \hat{p}_s(\boldsymbol{\theta}_t) \int_{\mathcal{Z}} \ell(\mathbf{z}, \boldsymbol{\theta}) \mathbb{Q}_s(d\mathbf{z}) \\ \text{s. t.} \quad \sum_{s \in [S]} \hat{p}_s(\boldsymbol{\theta}_t) \int_{\mathcal{Z}} \|\mathbf{z} - \hat{\mathbf{z}}_s\|_2^2 \mathbb{Q}_s(d\mathbf{z}) \leq \rho^2. \end{cases} \end{aligned}$$

where  $\mathcal{M}(\mathcal{Z})$  denotes the space of all probability distributions  $\mathbb{Q}$  supported on  $\mathcal{Z}$  satisfying  $\mathbb{E}_{\mathbb{Q}}[\|\mathbf{z}\|_2^2] = \int_{\mathcal{Z}} \|\mathbf{z}\|_2^2 \mathbb{Q}(\mathbf{d}z) < \infty$ . This reformulation follows from the law of total probability, where  $\mathbb{Q}_s$  represents the conditional distribution of  $\tilde{\mathbf{z}}$  given that the scenario  $\hat{\mathbf{z}}_s(\boldsymbol{\theta}_t)$  is realized. Using the Lagrangian, we have

$$\begin{aligned} \mathbb{J}_{\boldsymbol{\theta}_t}(\boldsymbol{\theta}) &= \sup_{\mathbb{Q}_s \in \mathcal{M}(\mathcal{Z})} \inf_{\lambda \in \mathbb{R}_+} \sum_{s \in [S]} \hat{p}_s(\boldsymbol{\theta}_t) \int_{\mathcal{Z}} \ell(\mathbf{z}, \boldsymbol{\theta}) \mathbb{Q}_s(\mathbf{d}z) \\ &\quad + \lambda \left( \rho^2 - \sum_{s \in [S]} \hat{p}_s(\boldsymbol{\theta}_t) \int_{\mathcal{Z}} \|\mathbf{z} - \hat{\mathbf{z}}_s\|_2^2 \mathbb{Q}_s(\mathbf{d}z) \right). \end{aligned}$$

By the minimax theorem [53], which is valid under the assumption that  $\ell$  is upper-semicontinuous and concave in  $\mathbf{z}$ , and the support set  $\mathcal{Z}$  is convex, we can exchange the supremum and infimum to obtain:

$$\mathbb{J}_{\boldsymbol{\theta}_t}(\boldsymbol{\theta}) = \inf_{\lambda \in \mathbb{R}_+} \sup_{\mathbb{Q}_s \in \mathcal{M}(\mathcal{Z})} \rho^2 \lambda + \sum_{s \in [S]} \hat{p}_s(\boldsymbol{\theta}_t) \int_{\mathcal{Z}} (\ell(\mathbf{z}, \boldsymbol{\theta}) - \lambda \|\mathbf{z} - \hat{\mathbf{z}}_s\|_2^2) \mathbb{Q}_s(\mathbf{d}z).$$

From the fact that the space  $\mathcal{M}(\mathcal{Z})$  contains all the Dirac distributions supported on  $\mathcal{Z}$ , we have

$$\mathbb{J}_{\boldsymbol{\theta}_t}(\boldsymbol{\theta}) = \inf_{\lambda \in \mathbb{R}_+} \rho^2 \lambda + \sum_{s \in [S]} \hat{p}_s(\boldsymbol{\theta}_t) \sup_{\mathbf{z} \in \mathcal{Z}} (\ell(\mathbf{z}, \boldsymbol{\theta}) - \lambda \|\mathbf{z} - \hat{\mathbf{z}}_s\|_2^2).$$

Adding a regularized term  $\tau \lambda^2$  where  $\tau \in \mathbb{R}_{++}$  is a positive constant, we have

$$\mathbb{J}_{\boldsymbol{\theta}_t}(\boldsymbol{\theta}) \leq \inf_{\lambda \in \mathbb{R}_+} \sum_{s \in [S]} \hat{p}_s(\boldsymbol{\theta}_t) \sup_{\mathbf{z} \in \mathcal{Z}} (\ell(\mathbf{z}, \boldsymbol{\theta}) - \lambda \|\mathbf{z} - \hat{\mathbf{z}}_s\|_2^2 + \rho^2 \lambda + \tau \lambda^2).$$

Therefore, minimizing the right-hand side provides an upper bound on  $\mathbb{J}_{\boldsymbol{\theta}_t}(\boldsymbol{\theta})$ . Next, we introduce auxiliary variables  $t_s \forall s \in [S]$ , which yields the equivalent formulation for the right hand side of the above inequality

$$\begin{aligned} \inf & \sum_{s \in [S]} \hat{p}_s(\boldsymbol{\theta}_t) t_s + \rho^2 \lambda + \tau \lambda^2 \\ \text{s. t.} & \lambda \in \mathbb{R}_+, \quad t_s \in \mathbb{R} \quad \forall s \in [S] \\ & \sup_{\mathbf{z} \in \mathcal{Z}} (\ell(\mathbf{z}, \boldsymbol{\theta}) - \lambda \|\mathbf{z} - \hat{\mathbf{z}}_s\|_2^2) \leq t_s \quad \forall s \in [S]. \end{aligned} \tag{14}$$

By the definition of conjugate functions, we have

$$\sup_{\mathbf{z} \in \mathcal{Z}} (\ell(\mathbf{z}, \boldsymbol{\theta}) - \lambda \|\mathbf{z} - \hat{\mathbf{z}}_s\|_2^2) = [-\ell + \chi_{\mathcal{Z}} + \lambda \|\mathbf{z} - \hat{\mathbf{z}}_s\|_2^2]^*(0),$$

where  $\chi_{\mathcal{Z}}$  denotes the characteristic function of  $\mathcal{Z}$ . Based on results from [35, Theorem 4.2], [46, Theorem 11.23], and [64, Lemma B.8], the conjugate functions of infimal convolutions and 2-norm balls is given by

$$[-\ell + \chi_{\mathcal{Z}} + \lambda \|\mathbf{z} - \hat{\mathbf{z}}_s\|_2^2]^*(0) = \inf_{\mathbf{r}_s, \boldsymbol{\zeta}_s} ([-\ell]^*(\mathbf{r}_s - \boldsymbol{\zeta}_s, \boldsymbol{\theta}) + \sigma_{\mathcal{Z}}(\boldsymbol{\zeta}_s) + [\lambda \|\mathbf{z} - \hat{\mathbf{z}}_s\|_2^2]^*(-\mathbf{r}_s)).$$

with

$$[\lambda \|\mathbf{z} - \hat{\mathbf{z}}_s\|_2^2]^*(-\mathbf{r}_s) = \sup_{\mathbf{v}_s} (-\mathbf{r}_s^\top \mathbf{v}_s - \lambda \|\mathbf{z} - \hat{\mathbf{z}}_s\|_2^2) = -\mathbf{r}_s^\top \hat{\mathbf{z}}_s + \frac{1}{4\lambda} \|\mathbf{r}_s\|_2^2,$$

Substituting this back into the formulation (14), we thus obtain that  $\mathbb{J}_{\boldsymbol{\theta}_t}(\boldsymbol{\theta})$  is upper bounded by the optimal value of the following convex program:

$$\begin{aligned} \inf & \sum_{s \in [S]} \hat{p}_s(\boldsymbol{\theta}_t) \left( [-\ell]^*(\mathbf{r}_s - \boldsymbol{\zeta}_s, \boldsymbol{\theta}) + \sigma_{\mathcal{Z}}(\boldsymbol{\zeta}_s) - \mathbf{r}_s^\top \hat{\mathbf{z}}_s + \frac{1}{4\lambda} \|\mathbf{r}_s\|_2^2 \right) + \rho^2 \lambda + \tau \lambda^2 \\ \text{s. t.} & \lambda \in \mathbb{R}_+, \quad t_s \in \mathbb{R} \quad \forall s \in [S], \quad \mathbf{r}_s \in \mathbb{R}^m \quad \forall s \in [S], \quad \boldsymbol{\zeta}_s \in \mathbb{R}^m \quad \forall s \in [S]. \end{aligned}$$

Combining with outer minimization over  $\boldsymbol{\theta} \in \Theta$  completes the proof.  $\square$

## E Deferred Proofs Related to Convergence

**Lemma E.1.** Consider the loss function defined in Model 2 and assume  $\Theta$  is bounded. We define the smoothed loss function

$$\ell_\mu(\mathbf{Z}, \boldsymbol{\theta}) = \mu \log \left( \sum_{j \in [J]} e^{Q_j(\mathbf{Z}, \boldsymbol{\theta})/\mu} \right), \quad (15)$$

where  $\mathbf{Z} \sim \hat{\mathbb{P}}(\boldsymbol{\theta}_t)$  satisfies and  $\|\mathbf{Z}\|_2 \leq k_1$  for some constant  $k_1 < \infty$ . Then the gradient  $\nabla_{\boldsymbol{\theta}} \ell_\mu(\mathbf{Z}, \boldsymbol{\theta})$  is  $\beta$ -Lipschitz in  $\mathbf{Z}$  for  $\beta = Jdk_3 \left( \frac{k_1 k_2}{\mu} + 1 \right)$  for some constants  $k_2, k_3 < \infty$  defined below.

*Proof.* Since the coefficients  $\mathbf{a}_j(\boldsymbol{\theta}), \mathbf{b}_j(\boldsymbol{\theta})$ , and  $c_j(\boldsymbol{\theta})$  are affine for all  $j \in [J]$  and  $\Theta$  is bounded, there exist some constants  $k_2, k_3 < \infty$  such that

- $\|\boldsymbol{\Gamma}_j\|_2 \leq k_2$  for all  $j \in [J]$ ,
- $\|\nabla_{\theta_i} \boldsymbol{\Gamma}_j\|_2 \leq k_3$  for all  $i \in [d], j \in [J]$ .

By definition, we have the gradient of the smoothed loss (15):

$$\nabla_{\boldsymbol{\theta}} \ell_\mu(\mathbf{Z}, \boldsymbol{\theta}) = \sum_{j \in [J]} w_j(\mathbf{Z}, \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} Q_j(\mathbf{Z}, \boldsymbol{\theta})$$

where the softmax weights are

$$w_j(\mathbf{Z}, \boldsymbol{\theta}) = \frac{e^{Q_j(\mathbf{Z}, \boldsymbol{\theta})/\mu}}{\sum_{i \in [J]} e^{Q_i(\mathbf{Z}, \boldsymbol{\theta})/\mu}},$$

and the gradient of  $Q_j$  with respect to  $\boldsymbol{\theta}$  is given by

$$\nabla_{\boldsymbol{\theta}} Q_j(\mathbf{Z}, \boldsymbol{\theta}) = 2\bar{\mathbf{A}}_j^\top \mathbf{Y} \bar{\mathbf{A}}_j \boldsymbol{\theta} + 2\bar{\mathbf{A}}_j^\top \mathbf{Y}^\top \bar{\mathbf{a}}_j + 2\bar{\mathbf{B}}_j^\top \mathbf{z} + z^0 \bar{\mathbf{c}}_j.$$

To prove that  $\nabla_{\boldsymbol{\theta}} \ell_\mu$  is Lipschitz in  $\mathbf{Z}$ , we examine  $\|\nabla_{\boldsymbol{\theta}} \ell_\mu(\mathbf{Z}_1, \boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} \ell_\mu(\mathbf{Z}_2, \boldsymbol{\theta})\|_2$ . We first decompose the difference and apply the triangle inequality

$$\begin{aligned} & \left\| \sum_{j \in [J]} w_j(\mathbf{Z}_1, \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} Q_j(\mathbf{Z}_1, \boldsymbol{\theta}) - \sum_{j \in [J]} w_j(\mathbf{Z}_2, \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} Q_j(\mathbf{Z}_2, \boldsymbol{\theta}) \right\|_2 \\ & \leq \left\| \sum_{j \in [J]} w_j(\mathbf{Z}_1, \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} Q_j(\mathbf{Z}_1, \boldsymbol{\theta}) - \sum_{j \in [J]} w_j(\mathbf{Z}_2, \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} Q_j(\mathbf{Z}_1, \boldsymbol{\theta}) \right\|_2 \\ & \quad + \left\| \sum_{j \in [J]} w_j(\mathbf{Z}_2, \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} Q_j(\mathbf{Z}_1, \boldsymbol{\theta}) - \sum_{j \in [J]} w_j(\mathbf{Z}_2, \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} Q_j(\mathbf{Z}_2, \boldsymbol{\theta}) \right\|_2 \\ & = \left\| \sum_{j \in [J]} (w_j(\mathbf{Z}_1, \boldsymbol{\theta}) - w_j(\mathbf{Z}_2, \boldsymbol{\theta})) \nabla_{\boldsymbol{\theta}} Q_j(\mathbf{Z}_1, \boldsymbol{\theta}) \right\|_2 + \left\| \sum_{j \in [J]} (\nabla_{\boldsymbol{\theta}} Q_j(\mathbf{Z}_1, \boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} Q_j(\mathbf{Z}_2, \boldsymbol{\theta})) \right\|_2. \end{aligned} \quad (16)$$

Next, we bound the terms involved:

- Weight difference:

$$\begin{aligned}
\|w_j(\mathbf{Z}_1, \boldsymbol{\theta}) - w_j(\mathbf{Z}_2, \boldsymbol{\theta})\|_2 &= \left\| \frac{e^{Q_j(\mathbf{Z}_1, \boldsymbol{\theta})/\mu}}{\sum_{i \in [J]} e^{Q_i(\mathbf{Z}_1, \boldsymbol{\theta})/\mu}} - \frac{e^{Q_j(\mathbf{Z}_2, \boldsymbol{\theta})/\mu}}{\sum_{i \in [J]} e^{Q_i(\mathbf{Z}_2, \boldsymbol{\theta})/\mu}} \right\|_2 \\
&\stackrel{(a)}{\leq} \frac{1}{\mu} \|Q_j(\mathbf{Z}_1, \boldsymbol{\theta}) - Q_j(\mathbf{Z}_2, \boldsymbol{\theta})\|_2 \\
&= \frac{1}{\mu} \|\langle \boldsymbol{\Gamma}_j, \mathbf{Z}_1 \rangle - \langle \boldsymbol{\Gamma}_j, \mathbf{Z}_2 \rangle\|_2 \\
&\stackrel{(b)}{\leq} \frac{1}{\mu} \|\boldsymbol{\Gamma}_j\|_2 \|\mathbf{Z}_1 - \mathbf{Z}_2\|_2 \\
&\leq \frac{k_2}{\mu} \|\mathbf{Z}_1 - \mathbf{Z}_2\|_2,
\end{aligned}$$

where (a) comes from the Lipschitz continuity of the softmax function [17], and (b) uses the Cauchy–Schwarz inequality.

- Gradient:

$$\|\nabla_{\boldsymbol{\theta}} Q_j(\mathbf{Z}, \boldsymbol{\theta})\|_2 = \sum_{i \in [d]} \|\langle \nabla_{\theta_i} \boldsymbol{\Gamma}_j, \mathbf{Z} \rangle\|_2 \stackrel{(a)}{\leq} \sum_{i \in [d]} \|\nabla_{\theta_i} \boldsymbol{\Gamma}_j\|_2 \|\mathbf{Z}\|_2 \stackrel{(b)}{\leq} k_1 k_3 d,$$

where  $\nabla_{\theta_i} \boldsymbol{\Gamma}_j$  is a matrix whose  $(i_1, i_2)$ -th matrix slice is the gradient of the  $(i_1, i_2)$ -th component of the matrix with respect to  $\theta_i$ .

- Gradient difference:

$$\|\nabla_{\boldsymbol{\theta}} Q_j(\mathbf{Z}_1, \boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} Q_j(\mathbf{Z}_2, \boldsymbol{\theta})\|_2 \leq d \|\nabla_{\theta_i} \boldsymbol{\Gamma}_j\|_2 \|\mathbf{Z}_1 - \mathbf{Z}_2\|_2 \leq d k_3 \|\mathbf{Z}_1 - \mathbf{Z}_2\|_2.$$

Substituting the above bounds into (16), we obtain

$$\|\nabla_{\boldsymbol{\theta}} \ell_{\mu}(\mathbf{Z}_1, \boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} \ell_{\mu}(\mathbf{Z}_2, \boldsymbol{\theta})\|_2 \leq J d k_3 \left( \frac{k_1 k_2}{\mu} + 1 \right) \|\mathbf{Z}_1 - \mathbf{Z}_2\|_2.$$

Thus, the claim follows.  $\square$

**Lemma E.2.** Consider the loss function defined in Model 2, and assume that  $\overline{\mathbf{A}}_j^\top \overline{\mathbf{A}}_j \succ \mathbf{0}$  for all  $j \in [J]$ . Define the smoothed loss function

$$\ell_{\mu}^{\text{reg}}(\boldsymbol{\theta}) = \mu \log \left( \sum_{j \in [J]} e^{\rho \|\boldsymbol{\Gamma}_j\|_1 / \mu} \right), \quad (17)$$

where  $\boldsymbol{\Gamma}_j$  is defined in (11). Then  $\ell_{\mu}^{\text{reg}}(\boldsymbol{\theta})$  is  $\rho\alpha$ -strongly convex in  $\boldsymbol{\theta}$  where

$$\alpha = 2 \min_{j \in [J]} \lambda_{\min}(\overline{\mathbf{A}}_j^\top \overline{\mathbf{A}}_j).$$

*Proof.* By definition, the gradient of the smoothed loss (17) is given by

$$\nabla_{\boldsymbol{\theta}} \ell_{\mu}^{\text{reg}}(\boldsymbol{\theta}) = \rho \sum_{j \in [J]} w_j(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \boldsymbol{\Gamma}_j(\boldsymbol{\theta}),$$

where the softmax weights are defined as

$$w_j(\boldsymbol{\theta}) = \frac{e^{\rho \boldsymbol{\Gamma}_j(\boldsymbol{\theta}) / \mu}}{\sum_{i \in [J]} e^{\rho \boldsymbol{\Gamma}_i(\boldsymbol{\theta}) / \mu}} \quad \forall j \in [J],$$

and the gradient of  $\boldsymbol{\Gamma}_j$  with respect to  $\boldsymbol{\theta}$  is given by

$$\nabla_{\boldsymbol{\theta}} \boldsymbol{\Gamma}_j(\boldsymbol{\theta}) = 2 \overline{\mathbf{A}}_j^\top \overline{\mathbf{A}}_j \boldsymbol{\theta} + 2 \overline{\mathbf{A}}_j^\top \overline{\mathbf{a}}_j + \overline{\mathbf{c}}_j.$$

Using the product rule and softmax identity:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}}^2 \ell_{\mu}^{\text{reg}}(\boldsymbol{\theta}) &= \rho \sum_{j \in [J]} w_j(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^2 \boldsymbol{\Gamma}_j(\boldsymbol{\theta}) + \frac{\rho^2}{\mu} \sum_{j \in [J]} w_j(\boldsymbol{\theta}) [\nabla_{\boldsymbol{\theta}} \boldsymbol{\Gamma}_j(\boldsymbol{\theta}) - \bar{g}] [\nabla_{\boldsymbol{\theta}} \boldsymbol{\Gamma}_j(\boldsymbol{\theta}) - \bar{g}]^\top \\ &= 2\rho \sum_{j \in [J]} w_j(\boldsymbol{\theta}) \bar{\mathbf{A}}_j^\top \bar{\mathbf{A}}_j + \frac{\rho^2}{\mu} \sum_{j \in [J]} w_j(\boldsymbol{\theta}) [\nabla_{\boldsymbol{\theta}} \boldsymbol{\Gamma}_j(\boldsymbol{\theta}) - \bar{g}] [\nabla_{\boldsymbol{\theta}} \boldsymbol{\Gamma}_j(\boldsymbol{\theta}) - \bar{g}]^\top,\end{aligned}$$

where

$$\bar{g} = \rho \sum_{j \in [J]} w_j(\boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}} \boldsymbol{\Gamma}_j(\boldsymbol{\theta}).$$

Notice that the first term is a convex combination of positive definite matrices  $\bar{\mathbf{A}}_j^\top \bar{\mathbf{A}}_j$ , so it is positive definite. The second term is a Gram matrix, hence positive semidefinite. Therefore, the minimum eigenvalue of the Hessian is lower bounded by the minimum eigenvalue of the first term and we have

$$\nabla_{\boldsymbol{\theta}}^2 \ell_{\mu}^{\text{reg}}(\boldsymbol{\theta}) \succeq 2\rho \min_{j \in [J]} \lambda_{\min}(\bar{\mathbf{A}}_j^\top \bar{\mathbf{A}}_j) \mathbb{I}.$$

Thus, the claim follows.  $\square$

**Corollary E.1.** *Consider the setting of Model 2, and define the following smoothed loss function*

$$g(\mathbf{Z}, \boldsymbol{\theta}) = \ell_{\mu}(\mathbf{Z}, \boldsymbol{\theta}) + \ell_{\mu}^{\text{reg}}(\boldsymbol{\theta})$$

where  $\ell_{\mu}(\mathbf{Z}, \boldsymbol{\theta})$  is the smoothed loss function defined in (15), and  $\ell_{\mu}^{\text{reg}}(\boldsymbol{\theta})$  is defined in (17). Suppose that  $\bar{\mathbf{A}}_j^\top \bar{\mathbf{A}}_j \succ 0$  for all  $j \in [J]$ . Then  $g(\mathbf{Z}, \boldsymbol{\theta})$  is  $\rho\alpha$ -strongly convex in  $\boldsymbol{\theta}$  with  $\alpha = 2 \min_{j \in [J]} \lambda_{\min}(\bar{\mathbf{A}}_j^\top \bar{\mathbf{A}}_j)$ , and the gradient  $\nabla_{\boldsymbol{\theta}} g(\mathbf{Z}, \boldsymbol{\theta})$  is  $\beta$ -Lipschitz in  $\mathbf{Z}$  for  $\beta = Jdk_3 \left( \frac{k_1 k_2}{\mu} + 1 \right)$  for some constants  $k_1, k_2, k_3 < \infty$ .

*Proof.* By a standard result in convex analysis, the function  $\ell_{\mu}(\mathbf{Z}, \boldsymbol{\theta})$  is convex in  $\boldsymbol{\theta}$ , as the log-sum-exp operator preserves convexity when applied to a collection of convex functions. Furthermore, Lemma E.2 establishes that  $\ell_{\mu}^{\text{reg}}(\boldsymbol{\theta})$  is  $\rho\alpha$ -strongly convex, where

$$\alpha = \min_j \lambda_{\min}(\bar{\mathbf{A}}_j^\top \bar{\mathbf{A}}_j).$$

As the sum of a convex function and a strongly convex function is strongly convex,  $g(\mathbf{Z}, \bar{\boldsymbol{\theta}})$  is  $\rho\alpha$ -strongly convex.

The dependence of  $g$  on  $\mathbf{Z}$  comes only through  $\ell_{\mu}(\mathbf{Z}, \boldsymbol{\theta})$ . Therefore, the Lipschitz continuity of the gradient  $\nabla_{\boldsymbol{\theta}} g(\mathbf{Z}, \boldsymbol{\theta})$  with respect to  $\mathbf{Z}$  follows from Lemma 2, which provides the bound on the Lipschitz constant  $\beta$ . This concludes the proof.  $\square$

**Lemma E.3.** *Assume that the loss function  $\ell(\mathbf{z}, \boldsymbol{\theta})$  is concave in  $\mathbf{z}$ , and that the set  $\mathcal{Z}$  has a finite diameter  $D = \sup_{\mathbf{z}, \mathbf{z}' \in \mathcal{Z}} \|\mathbf{z} - \mathbf{z}'\|_2 < \infty$ . Define the function*

$$g(\mathbf{v}, (\boldsymbol{\theta}, \lambda)) = \sup_{\mathbf{z} \in \mathcal{Z}} \ell(\mathbf{z}, \boldsymbol{\theta}) - \lambda \|\mathbf{z} - \mathbf{v}\|_2^2 + \rho^2 \lambda + \tau \lambda^2.$$

Then the function  $g$  is  $\alpha$ -strongly convex in  $(\boldsymbol{\theta}, \lambda)$  where

$$\alpha = \min(\gamma, 2\tau),$$

and the gradient  $\nabla_{(\boldsymbol{\theta}, \lambda)} g(\mathbf{v}, (\boldsymbol{\theta}, \lambda))$  is  $(\beta + 4D)$ -Lipschitz in  $\mathbf{v}$ .

*Proof.* Define

$$\phi((\boldsymbol{\theta}, \lambda), \mathbf{v}, \mathbf{z}) = \ell(\mathbf{z}, \boldsymbol{\theta}) - \lambda \|\mathbf{z} - \mathbf{v}\|_2^2 + \rho^2 \lambda + \tau \lambda^2$$

Then we have

$$g(\mathbf{v}, (\boldsymbol{\theta}, \lambda)) = \sup_{\mathbf{z} \in \mathcal{Z}} \phi((\boldsymbol{\theta}, \lambda), \mathbf{v}, \mathbf{z}).$$

Since  $\phi((\theta, \lambda), \mathbf{v}, \cdot)$  is  $2\lambda$ -strongly concave in  $\mathbf{v}$ , the maximizer

$$\mathbf{z}^*((\theta, \lambda), \mathbf{v}) = \arg \max_{\mathbf{z} \in \mathcal{Z}} \phi((\theta, \lambda), \mathbf{v}, \mathbf{z})$$

is unique. By Danskin's theorem [6, Section 6.11], the function  $g$  is differentiable, and its gradient with respect to  $(\theta, \lambda)$  is

$$\begin{aligned} \nabla_{(\theta, \lambda)} g(\mathbf{v}, (\theta, \lambda)) &= \nabla_{(\theta, \lambda)} \phi((\theta, \lambda), \mathbf{v}, \mathbf{z}^*) \\ &= \begin{bmatrix} \nabla_{\theta} \ell(\mathbf{z}^*, \theta) \\ -\|\mathbf{z}^* - \mathbf{v}\|_2^2 + \rho^2 + 2\tau\lambda \end{bmatrix}, \end{aligned}$$

where  $\mathbf{z}^* = \mathbf{z}^*((\theta, \lambda), \mathbf{v})$ . The Hessian is

$$\begin{aligned} \nabla_{(\theta, \lambda)}^2 g(\mathbf{v}, (\theta, \lambda)) &= \nabla_{(\theta, \lambda)}^2 \phi((\theta, \lambda), \mathbf{v}, \mathbf{z}^*) \\ &= \begin{bmatrix} \nabla_{\theta}^2 \ell(\mathbf{z}^*, \theta) & 0 \\ 0 & 2\tau \end{bmatrix} \succeq \alpha \mathbb{I}. \end{aligned} \quad (18)$$

where  $\alpha = \min(\gamma, 2\tau)$ . This proves strong convexity in  $(\theta, \lambda)$ .

Before we move to the Lipschitz smoothness, we first prove several inequalities. Fix  $\mathbf{z}_1, \mathbf{z}_2 \in \mathcal{Z}$ , we have

$$\begin{aligned} &\| \|\mathbf{z}^*((\theta, \lambda), \mathbf{v}_2) - \mathbf{v}_2\|_2^2 - \|\mathbf{z}^*((\theta, \lambda), \mathbf{v}_1) - \mathbf{v}_1\|_2^2 \|_2 \\ &\stackrel{(a)}{\leq} \| (\|\mathbf{z}^*((\theta, \lambda), \mathbf{v}_2) - \mathbf{v}_2\|_2 + \|\mathbf{z}^*((\theta, \lambda), \mathbf{v}_1) - \mathbf{v}_1\|_2) \\ &\quad \times (\|\mathbf{z}^*((\theta, \lambda), \mathbf{v}_2) - \mathbf{v}_2\|_2 - \|\mathbf{z}^*((\theta, \lambda), \mathbf{v}_1) - \mathbf{v}_1\|_2) \| \\ &\stackrel{(b)}{\leq} 2D \|\mathbf{z}^*((\theta, \lambda), \mathbf{v}_2) - \mathbf{z}^*((\theta, \lambda), \mathbf{v}_1)\|_2 + 2D \|\mathbf{v}_2 - \mathbf{v}_1\|_2 \end{aligned} \quad (19)$$

where (a) uses the difference of squares, and (b) uses the triangular inequality and the boundness of  $\mathcal{Z}$ . Next we prove  $\mathbf{z}^*((\theta, \lambda), \mathbf{v})$  is 1-Lipschitz continuous in  $\mathbf{v}$  as follows.

$$\begin{aligned} &2\lambda \|\mathbf{z}^*((\theta, \lambda), \mathbf{v}_1) - \mathbf{z}^*((\theta, \lambda), \mathbf{v}_2)\|_2^2 \\ &\stackrel{(a)}{\leq} \langle -\nabla_{\mathbf{z}} \phi((\theta, \lambda), \mathbf{v}_2, \mathbf{z}^*((\theta, \lambda), \mathbf{v}_2)) + \nabla_{\mathbf{z}} \phi((\theta, \lambda), \mathbf{v}_2, \mathbf{z}^*((\theta, \lambda), \mathbf{v}_1)), \\ &\quad \mathbf{z}^*((\theta, \lambda), \mathbf{v}_2) - \mathbf{z}^*((\theta, \lambda), \mathbf{v}_1) \rangle \\ &\stackrel{(b)}{\leq} \langle \nabla_{\mathbf{z}} \phi((\theta, \lambda), \mathbf{v}_2, \mathbf{z}^*((\theta, \lambda), \mathbf{v}_1)), \mathbf{z}^*((\theta, \lambda), \mathbf{v}_2) - \mathbf{z}^*((\theta, \lambda), \mathbf{v}_1) \rangle \\ &\stackrel{(c)}{\leq} \langle \nabla_{\mathbf{z}} \phi((\theta, \lambda), \mathbf{v}_2, \mathbf{z}^*((\theta, \lambda), \mathbf{v}_1)) - \nabla_{\mathbf{z}} \phi((\theta, \lambda), \mathbf{v}_1, \mathbf{z}^*((\theta, \lambda), \mathbf{v}_1)), \\ &\quad \mathbf{z}^*((\theta, \lambda), \mathbf{v}_2) - \mathbf{z}^*((\theta, \lambda), \mathbf{v}_1) \rangle \\ &\stackrel{(d)}{\leq} 2\lambda \|\mathbf{v}_1 - \mathbf{v}_2\| \|\mathbf{z}^*((\theta, \lambda), \mathbf{v}_2) - \mathbf{z}^*((\theta, \lambda), \mathbf{v}_1)\| \end{aligned}$$

where (a) uses strong concavity of  $\phi((\theta, \lambda), \mathbf{v}, \cdot)$ , (b) and (c) come from the first order optimality conditions for  $\mathbf{z}^*((\theta, \lambda), \mathbf{v}_1)$  and  $\mathbf{z}^*((\theta, \lambda), \mathbf{v}_2)$ :

$$\langle \nabla_{\mathbf{z}} \phi((\theta, \lambda), \mathbf{v}, \mathbf{z}^*((\theta, \lambda), \mathbf{v})), \mathbf{z} - \mathbf{z}^*((\theta, \lambda), \mathbf{v}) \rangle \leq 0,$$

and (d) uses Cauchy-Schwarz inequality. Hence,

$$\|\mathbf{z}^*((\theta, \lambda), \mathbf{v}_1) - \mathbf{z}^*((\theta, \lambda), \mathbf{v}_2)\|_2 \leq \|\mathbf{v}_1 - \mathbf{v}_2\| \quad (20)$$

Finally we show  $\nabla_{(\theta, \lambda)} g(\mathbf{v}, (\theta, \lambda))$  is Lipschitz in  $\mathbf{v}$ , consider the gradient difference

$$\begin{aligned} &\|\nabla_{(\theta, \lambda)} g(\mathbf{v}_1, (\theta, \lambda)) - \nabla_{(\theta, \lambda)} g(\mathbf{v}_2, (\theta, \lambda))\|_2 \\ &= \left\| \begin{bmatrix} \nabla_{\theta} \ell(\mathbf{z}^*((\theta, \lambda), \mathbf{v}_1), \theta) - \nabla_{\theta} \ell(\mathbf{z}^*((\theta, \lambda), \mathbf{v}_2), \theta) \\ -\|\mathbf{z}^*((\theta, \lambda), \mathbf{v}_2) - \mathbf{v}_2\|_2^2 - \|\mathbf{z}^*((\theta, \lambda), \mathbf{v}_1) - \mathbf{v}_1\|_2^2 \end{bmatrix} \right\|_2 \\ &\stackrel{(a)}{\leq} \|\nabla_{\theta} \ell(\mathbf{z}^*((\theta, \lambda), \mathbf{v}_1), \theta) - \nabla_{\theta} \ell(\mathbf{z}^*((\theta, \lambda), \mathbf{v}_2), \theta)\|_2 \\ &\quad + \left\| \begin{bmatrix} \|\mathbf{z}^*((\theta, \lambda), \mathbf{v}_2) - \mathbf{v}_2\|_2^2 - \|\mathbf{z}^*((\theta, \lambda), \mathbf{v}_1) - \mathbf{v}_1\|_2^2 \end{bmatrix} \right\|_2 \\ &\stackrel{(b)}{\leq} \beta \|\mathbf{z}^*((\theta, \lambda), \mathbf{v}_1) - \mathbf{z}^*((\theta, \lambda), \mathbf{v}_2)\|_2 + \\ &\quad 2D \|\mathbf{z}^*((\theta, \lambda), \mathbf{v}_2) - \mathbf{z}^*((\theta, \lambda), \mathbf{v}_1)\|_2 + 2D \|\mathbf{v}_2 - \mathbf{v}_1\|_2 \\ &\stackrel{(c)}{\leq} (\beta + 4D) \|\mathbf{v}_2 - \mathbf{v}_1\|_2 \end{aligned}$$

where (a) comes from the triangular inequality, (b) uses the  $\beta$ -jointly smoothness of  $\ell(\mathbf{z}, \theta)$  and 19, and (c) uses 20. Hence the gradient  $\nabla_{(\theta, \lambda)} g(\mathbf{v}, (\theta, \lambda))$  is  $(\beta + 4D)$ -Lipschitz in  $\mathbf{v}$ .  $\square$

### E.1 Proof of Theorem 1

*Proof.* Let  $G(\boldsymbol{\theta}_t)$  denote an optimal solution of (RRMP) at iteration  $t$ , i.e.,

$$\boldsymbol{\theta}_{t+1} = G(\boldsymbol{\theta}_t) \in \arg \min_{\boldsymbol{\theta} \in \Theta} \mathbb{J}_{\boldsymbol{\theta}_t}(\boldsymbol{\theta}).$$

where  $\boldsymbol{\theta}_t$  is the current solution and  $\boldsymbol{\theta}_{t+1} \in \Theta$  denotes the optimal solution for next iteration.

We first prove the convergence result for Model 1. Observe that

$$\mathbb{J}_{\boldsymbol{\theta}_t}(\boldsymbol{\theta}) = \mathbb{E}_{\hat{\mathbb{P}}(\boldsymbol{\theta}_t)}[g(\mathbf{Z}, \boldsymbol{\theta})]$$

where  $g(\mathbf{Z}, \bar{\boldsymbol{\theta}}) = \ell(\mathbf{Z}, \boldsymbol{\theta}) + \rho L \|\boldsymbol{\theta}, 1\|_2^2$ .

Fix  $\boldsymbol{\eta}, \boldsymbol{\eta}' \in \Theta$ . Since  $\mathbb{J}_{\boldsymbol{\eta}}(\cdot)$  is  $(\gamma + 2\rho L)$ -strongly convex, where  $2\rho L$  comes from the strong convexity of the regularization term  $\rho L \|\boldsymbol{\theta}, 1\|_2^2$ , we have

$$\begin{aligned} \mathbb{J}_{\boldsymbol{\eta}}(G(\boldsymbol{\eta})) - \mathbb{J}_{\boldsymbol{\eta}}(G(\boldsymbol{\eta}')) &\geq (G(\boldsymbol{\eta}) - G(\boldsymbol{\eta}'))^\top \nabla \mathbb{J}_{\boldsymbol{\eta}}(G(\boldsymbol{\eta}')) + \frac{\gamma + 2\rho L}{2} \|G(\boldsymbol{\eta}) - G(\boldsymbol{\eta}')\|_2^2, \\ \mathbb{J}_{\boldsymbol{\eta}}(G(\boldsymbol{\eta}')) - \mathbb{J}_{\boldsymbol{\eta}}(G(\boldsymbol{\eta})) &\geq \frac{\gamma + 2\rho L}{2} \|G(\boldsymbol{\eta}) - G(\boldsymbol{\eta}')\|_2^2, \end{aligned}$$

where the second inequality follows from the fact that

$$(G(\boldsymbol{\eta}') - G(\boldsymbol{\eta}))^\top \nabla \mathbb{J}_{\boldsymbol{\eta}}(G(\boldsymbol{\eta})) \geq 0$$

in view of the first-order optimality condition in Lemma C.1 since  $G(\boldsymbol{\eta}) \in \arg \min_{\bar{\boldsymbol{\theta}} \in \bar{\Theta}} \mathbb{J}_{\boldsymbol{\eta}}(\bar{\boldsymbol{\theta}})$ . Combining the two inequalities, we obtain

$$\begin{aligned} (\gamma + 2\rho L) \|G(\boldsymbol{\eta}) - G(\boldsymbol{\eta}')\|_2^2 &\leq -(G(\boldsymbol{\eta}) - G(\boldsymbol{\eta}'))^\top \nabla \mathbb{J}_{\boldsymbol{\eta}}(G(\boldsymbol{\eta}')) \\ &\leq (G(\boldsymbol{\eta}) - G(\boldsymbol{\eta}'))^\top [\nabla \mathbb{J}_{\boldsymbol{\eta}'}(G(\boldsymbol{\eta}')) - \nabla \mathbb{J}_{\boldsymbol{\eta}}(G(\boldsymbol{\eta}'))], \end{aligned} \quad (21)$$

where the second inequality follows from the fact that  $(G(\boldsymbol{\eta}) - G(\boldsymbol{\eta}'))^\top \nabla \mathbb{J}_{\boldsymbol{\eta}'}(G(\boldsymbol{\eta}')) \geq 0$  in view of the first-order optimality condition of  $G(\boldsymbol{\eta}')$ . Next, we will upper bound (21) using Cauchy-Schwarz inequality, as follows:

$$\begin{aligned} &(G(\boldsymbol{\eta}) - G(\boldsymbol{\eta}'))^\top [\nabla \mathbb{J}_{\boldsymbol{\eta}'}(G(\boldsymbol{\eta}')) - \nabla \mathbb{J}_{\boldsymbol{\eta}}(G(\boldsymbol{\eta}'))] \\ &\leq \|G(\boldsymbol{\eta}) - G(\boldsymbol{\eta}')\|_2 \|\nabla \mathbb{J}_{\boldsymbol{\eta}'}(G(\boldsymbol{\eta}')) - \nabla \mathbb{J}_{\boldsymbol{\eta}}(G(\boldsymbol{\eta}'))\|_2 \\ &\stackrel{(a)}{=} \|G(\boldsymbol{\eta}) - G(\boldsymbol{\eta}')\|_2 \left\| \mathbb{E}_{\hat{\mathbb{P}}(\boldsymbol{\eta}')} [\nabla g(\tilde{\mathbf{Z}}; G(\boldsymbol{\eta}'))] - \mathbb{E}_{\hat{\mathbb{P}}(\boldsymbol{\eta})} [\nabla g(\tilde{\mathbf{Z}}; G(\boldsymbol{\eta}'))] \right\|_2 \\ &\stackrel{(b)}{\leq} \|G(\boldsymbol{\eta}) - G(\boldsymbol{\eta}')\|_2 \cdot \epsilon \beta \|\boldsymbol{\eta} - \boldsymbol{\eta}'\|_2. \end{aligned}$$

Here, (a) follows from the representation of the loss function, while (b) uses the Kantorovich-Rubinstein Lemma C.2 since the loss function is  $\beta$ -jointly smooth from Lemma E.3 and the map  $\hat{\mathbb{P}}(\boldsymbol{\theta})$  is  $\epsilon$ -sensitive. Combining this bound with (21), we get

$$\|G(\boldsymbol{\eta}) - G(\boldsymbol{\eta}')\|_2 \leq \frac{\epsilon \beta}{\gamma + 2\rho L} \|\boldsymbol{\eta} - \boldsymbol{\eta}'\|_2.$$

Our claim (a) is then established by simply performing the change of variables  $\boldsymbol{\eta} \leftarrow \boldsymbol{\theta}$  and  $\boldsymbol{\eta}' \leftarrow \boldsymbol{\theta}'$ .

To prove claim (b), we observe that  $\boldsymbol{\theta}_t = G(\boldsymbol{\theta}_{t-1})$  by the definition of (RRMP), and  $\boldsymbol{\theta}_{\text{RPS}} = G(\boldsymbol{\theta}_{\text{RPS}})$  by the definition of stability. Applying the result of the claim (a) yields

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{\text{RPS}}\|_2 \leq \frac{\epsilon \beta}{\gamma + 2\rho L} \|\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}_{\text{RPS}}\|_2 \leq \left( \frac{\epsilon \beta}{\gamma + 2\rho L} \right)^t \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_{\text{RPS}}\|_2.$$

Setting the right-hand side expression to be at most  $\delta$  and solving for  $t$  completes the proof for Model 1.

For Model 2. Observe that

$$\mathbb{J}_{\boldsymbol{\theta}_t}^{\mu}(\boldsymbol{\theta}) = \mathbb{E}_{\hat{\mathbb{P}}(\boldsymbol{\theta}_t)}[g(\mathbf{Z}, \boldsymbol{\theta})],$$

where  $g(\mathbf{Z}, \boldsymbol{\theta}) = \ell_{\mu}(\mathbf{Z}, \boldsymbol{\theta}) + \ell_{\mu}^{\text{reg}}(\boldsymbol{\theta})$ . Here  $\ell_{\mu}(\mathbf{Z}, \boldsymbol{\theta})$  is the smoothed loss function defined in (15), and  $\ell_{\mu}^{\text{reg}}(\boldsymbol{\theta})$  is defined in (17). From Corollary E.1, we know that  $g(\mathbf{Z}, \boldsymbol{\theta})$  is  $\rho\alpha$ -strongly convex

in  $\theta$  with  $\alpha = 2 \min_{j \in [J]} \lambda_{\min}(\bar{\mathbf{A}}_j^\top \bar{\mathbf{A}}_j)$ , and the gradient  $\nabla_{\theta} g(\mathbf{Z}, \theta)$  is  $\beta$ -Lipschitz in  $\mathbf{Z}$  for  $\beta = Jdk_3 \left( \frac{k_1 k_2}{\mu} + 1 \right)$  for some constants  $k_1, k_2, k_3 < \infty$ . Using the same techniques as in the proof of Model 1, we can therefore establish the desired result for Model 2.

Finally, for Model 3, one can observe that

$$\mathbb{J}_{\theta_t}^{\tau}(\bar{\theta}) = \mathbb{E}_{\hat{\mathbb{P}}(\theta_t)}[g(\mathbf{Z}, \bar{\theta})],$$

where

$$g(\mathbf{Z}, \bar{\theta}) = \sup_{\mathbf{z} \in \mathcal{Z}} \ell(\mathbf{z}, \theta) - \lambda \|\mathbf{z} - \mathbf{Z}\|_2^2 + \rho^2 \lambda + \tau \lambda^2.$$

From Lemma E.3, we know that  $g(\mathbf{Z}, \bar{\theta})$  is  $\alpha$ -strongly convex in  $\bar{\theta}$  with  $\alpha = \min(\gamma, 2\tau)$ , and that its gradient  $\nabla_{\bar{\theta}} g(\mathbf{Z}, \bar{\theta})$  is  $(\beta + 4D)$ -Lipschitz continuous in  $\mathbf{Z}$ . Hence, by applying the same arguments as in the preceding analysis, we obtain the desired result.  $\square$

## F Deferred Proofs Related to Sub-optimality Guarantees

### E.1 Proof of Sub-optimality Guarantee for Model 1

**Theorem 3.** Suppose the loss function in Model 1 is  $\gamma$ -strongly convex in  $\theta$  (B1) and is  $\beta$ -smooth (B2). Furthermore, assume that the loss function  $\ell(\mathbf{z}, \theta)$  is  $L_z$ -Lipschitz in  $\mathbf{z}$ . Then, the following suboptimality bound holds:

$$\mathbb{J}_{\theta_{\text{RPS}}}(\theta_{\text{RPS}}) - \mathbb{J}_{\theta_{\text{RPO}}}(\theta_{\text{RPO}}) \leq \frac{2\epsilon^2 L_z^2}{\gamma + 2\rho L}.$$

*Proof.* Since  $\mathbb{J}_{\theta_{\text{RPS}}}(\theta)$  is  $(\gamma + 2\rho L)$ -strongly convex in  $\theta$ , we have

$$\frac{\gamma + 2\rho L}{2} \|\theta_{\text{RPO}} - \theta_{\text{RPS}}\|_2^2 \leq \mathbb{J}_{\theta_{\text{RPS}}}(\theta_{\text{RPO}}) - \mathbb{J}_{\theta_{\text{RPS}}}(\theta_{\text{RPS}})$$

since  $(\theta_{\text{RPO}} - \theta_{\text{RPS}})^\top \nabla \mathbb{J}_{\theta_{\text{RPS}}}(\theta_{\text{RPS}}) \geq 0$  by the optimality of  $\theta_{\text{RPS}}$ . Using the fact that  $\mathbb{J}_{\theta_{\text{RPS}}}(\theta_{\text{RPS}}) \geq \mathbb{J}_{\theta_{\text{RPO}}}(\theta_{\text{RPO}})$ , we can further upper bound the right-hand side

$$\begin{aligned} \mathbb{J}_{\theta_{\text{RPS}}}(\theta_{\text{RPO}}) - \mathbb{J}_{\theta_{\text{RPS}}}(\theta_{\text{RPS}}) &\leq \mathbb{J}_{\theta_{\text{RPS}}}(\theta_{\text{RPO}}) - \mathbb{J}_{\theta_{\text{RPO}}}(\theta_{\text{RPO}}) \\ &\leq \epsilon L_z \|\theta_{\text{RPS}} - \theta_{\text{RPO}}\|_2, \end{aligned} \tag{22}$$

where the second inequality holds due the  $\epsilon$ -sensitivity of the distribution map  $\hat{\mathbb{P}}(\cdot)$  and the  $L_z$ -Lipschitz continuity of the loss function in  $\mathbf{z}$ . In summary, we obtain

$$\|\theta_{\text{RPS}} - \theta_{\text{RPO}}\|_2 \leq \frac{2\epsilon L_z}{\gamma + 2\rho L}. \tag{23}$$

Next, we derive a bound on the suboptimality of the robust performatively stable solution  $\theta_{\text{RPS}}$ . We have

$$\begin{aligned} \mathbb{J}_{\theta_{\text{RPS}}}(\theta_{\text{RPS}}) - \mathbb{J}_{\theta_{\text{RPO}}}(\theta_{\text{RPO}}) &\leq \mathbb{J}_{\theta_{\text{RPS}}}(\theta_{\text{RPO}}) - \mathbb{J}_{\theta_{\text{RPO}}}(\theta_{\text{RPO}}) \\ &\leq \epsilon L_z \|\theta_{\text{RPS}} - \theta_{\text{RPO}}\|_2 \\ &\leq \frac{2\epsilon^2 L_z^2}{\gamma + 2\rho L}, \end{aligned}$$

where the first inequality follows from the suboptimality of  $\theta_{\text{RPO}}$  in  $\mathbb{J}_{\theta_{\text{RPS}}}(\theta)$ , the second inequality is from (22), and the last inequality is from (23). This completes the proof.  $\square$

### E.2 Proof of Sub-optimality Guarantee for Model 2

**Theorem 4.** Consider the loss function  $\ell(\mathbf{Z}, \theta)$  from Model 2. Suppose that  $\bar{\mathbf{A}}_j^\top \bar{\mathbf{A}}_j \succ \mathbf{0}$  for all  $j \in [J]$ , and that  $\mathbf{Y}_s \succeq \mathbf{0}$  for all  $s \in [S]$ . Additionally, assume that  $\ell(\mathbf{Z}, \theta)$  is  $L_z$ -Lipschitz in  $\mathbf{Z}$ .

Then, the suboptimality of the stable point  $\theta_{\text{RPS}}^\mu$  of the smoothed robust objective satisfies:

$$\mathbb{J}_{\theta_{\text{RPS}}^\mu}(\theta_{\text{RPS}}^\mu) - \mathbb{J}_{\theta_{\text{RPO}}}(\theta_{\text{RPO}}) \leq \frac{2(\epsilon L_z + 2\mu' \log J)^2}{\rho\alpha}. \quad (24)$$

where  $\alpha = 2 \min_{j \in [J]} \lambda_{\min}(\bar{\mathbf{A}}_j^\top \bar{\mathbf{A}}_j)$ ,  $\mu' \in [0, 1]$  is a constant that satisfies  $\mu \leq \mu' \|\theta_{\text{RPS}}^\mu - \theta_{\text{RPO}}\|_2$ ,  $\mathbb{J}_{\theta_t}^\mu(\theta)$  is the smoothed robust objective defined in (12), and  $\theta_{\text{RPO}}$  denotes the minimizer of the original robust objective  $\mathbb{J}_{\theta_t}(\theta)$ .

*Proof.* By Lemma C.4, we have for any  $\theta$ ,

$$\mathbb{J}_{\theta_t}(\theta) \leq \mathbb{J}_{\theta_t}^\mu(\theta) \leq \mathbb{J}_{\theta_t}(\theta) + 2\mu \log J. \quad (25)$$

Additionally, from Corollary E.1, the function  $\mathbb{J}_{\theta_t}^\mu(\theta)$  is  $\rho\alpha$ -strongly convex, with  $\alpha = 2 \min_{j \in [J]} \lambda_{\min}(\bar{\mathbf{A}}_j^\top \bar{\mathbf{A}}_j)$ . By the definition of strong convexity, we have:

$$\frac{\rho\alpha}{2} \|\theta_{\text{RPO}} - \theta_{\text{RPS}}^\mu\|_2^2 \leq \mathbb{J}_{\theta_{\text{RPS}}^\mu}(\theta_{\text{RPO}}) - \mathbb{J}_{\theta_{\text{RPS}}^\mu}(\theta_{\text{RPS}}^\mu), \quad (26)$$

where the inequality follows from the first-order optimality condition of  $\theta_{\text{RPS}}^\mu$ , i.e.,

$$(\theta_{\text{RPO}} - \theta_{\text{RPS}}^\mu)^\top \nabla \mathbb{J}_{\theta_{\text{RPS}}^\mu}(\theta_{\text{RPS}}^\mu) \geq 0.$$

We next bound the right-hand side of (26). Using the fact that  $\mathbb{J}_{\theta_{\text{RPO}}}(\theta_{\text{RPO}}) \leq \mathbb{J}_{\theta_{\text{RPS}}^\mu}(\theta_{\text{RPS}}^\mu) \leq \mathbb{J}_{\theta_{\text{RPS}}^\mu}(\theta_{\text{RPS}}^\mu)$ , we have

$$\begin{aligned} \mathbb{J}_{\theta_{\text{RPS}}^\mu}(\theta_{\text{RPO}}) - \mathbb{J}_{\theta_{\text{RPS}}^\mu}(\theta_{\text{RPS}}^\mu) &\leq \mathbb{J}_{\theta_{\text{RPS}}^\mu}(\theta_{\text{RPO}}) - \mathbb{J}_{\theta_{\text{RPO}}}(\theta_{\text{RPO}}) \\ &\leq \mathbb{J}_{\theta_{\text{RPS}}^\mu}(\theta_{\text{RPO}}) - \mathbb{J}_{\theta_{\text{RPO}}}(\theta_{\text{RPO}}) + 2\mu \log J \\ &\leq \epsilon L_z \|\theta_{\text{RPS}}^\mu - \theta_{\text{RPO}}\|_2 + 2\mu \log J. \end{aligned} \quad (27)$$

where the second inequality comes from (25), and the last inequality holds due to the Lipschitz continuity of  $\ell(\mathbf{Z}, \theta)$  in  $\mathbf{Z}$ . Substituting (27) into the strong convexity inequality (26) gives:

$$\frac{\rho\alpha}{2} \|\theta_{\text{RPO}} - \theta_{\text{RPS}}^\mu\|_2^2 \leq \epsilon L_z \|\theta_{\text{RPS}}^\mu - \theta_{\text{RPO}}\|_2 + 2\mu \log J. \quad (28)$$

Assuming  $\mu \leq \mu' \|\theta_{\text{RPS}}^\mu - \theta_{\text{RPO}}\|_2$  where  $\mu' \in [0, 1]$ , we can divide both sides by  $\|\theta_{\text{RPS}}^\mu - \theta_{\text{RPO}}\|_2$  to obtain:

$$\|\theta_{\text{RPS}}^\mu - \theta_{\text{RPO}}\|_2 \leq \frac{2(\epsilon L_z + 2\mu' \log J)}{\rho\alpha}. \quad (29)$$

Finally, we derive a bound on the suboptimality of the robust performatively stable solution  $\theta_{\text{RPS}}^\mu$ :

$$\begin{aligned} \mathbb{J}_{\theta_{\text{RPS}}^\mu}(\theta_{\text{RPS}}^\mu) - \mathbb{J}_{\theta_{\text{RPO}}}(\theta_{\text{RPO}}) &\leq \mathbb{J}_{\theta_{\text{RPS}}^\mu}(\theta_{\text{RPO}}) - \mathbb{J}_{\theta_{\text{RPO}}}(\theta_{\text{RPO}}) \\ &\leq (\epsilon L_z + 2\mu' \log J) \|\theta_{\text{RPS}}^\mu - \theta_{\text{RPO}}\|_2 \\ &\leq \frac{2(\epsilon L_z + 2\mu' \log J)^2}{\rho\alpha}, \end{aligned}$$

where the first inequality uses suboptimality of  $\theta_{\text{RPO}}$  in  $\mathbb{J}_{\theta_{\text{RPS}}^\mu}(\theta)$ , the second follows from (27), and the last inequality uses the bound in (29). This concludes the proof.  $\square$

### E.3 Proof of Sub-optimality Guarantee for Model 3

**Theorem 5.** Suppose that  $\mathcal{Z}$  has a finite diameter  $D = \sup_{\mathbf{z}, \mathbf{z}' \in \mathcal{Z}} \|\mathbf{z} - \mathbf{z}'\|_2 < \infty$  and  $\Theta$  is bounded. Assume that the loss function  $\ell(\mathbf{z}, \theta)$  in Theorem (3) is  $L_\theta$  Lipschitz continuous in  $\theta$  and  $L_z$ -Lipschitz in  $\mathbf{z}$ . Let  $\mathbb{J}_{\theta_t}^\tau(\bar{\theta})$  denote the objective defined in problem (7), and let  $\bar{\theta}_{\text{RPS}}^\tau$  denote a robust performatively stable point under this objective. Then the following suboptimality bound holds:

$$\mathbb{J}_{\theta_{\text{RPS}}^\tau}(\bar{\theta}_{\text{RPS}}^\tau) - \mathbb{J}_{\theta_{\text{RPO}}}(\bar{\theta}_{\text{RPO}}) \leq \tau \bar{\lambda}^2 + \frac{(2\tau \bar{\lambda} + \rho^2 + D^2 + L_\theta + \epsilon L_z) 2\epsilon L_z}{\alpha}. \quad (30)$$

where  $\alpha := \min(\gamma, 2\tau)$  is the strong convexity parameter, and  $\bar{\lambda}$  is the upper bound on the optimal  $\lambda$  as given in Lemma F.1.

*Proof.* From Lemma E.3, the objective function  $\mathbb{J}_{\bar{\theta}_t}^{\tau}(\bar{\theta})$  is  $\alpha$ -strongly convex in  $\bar{\theta} = (\theta, \lambda)$  with  $\alpha = \min(\gamma, 2\tau)$ . Applying the strong convexity inequality, we obtain:

$$\frac{\alpha}{2} \left\| \bar{\theta}_{\text{RPO}}^{\tau} - \bar{\theta}_{\text{RPS}}^{\tau} \right\|_2^2 \leq \mathbb{J}_{\bar{\theta}_{\text{RPS}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPO}}^{\tau}) - \mathbb{J}_{\bar{\theta}_{\text{RPS}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPS}}^{\tau}),$$

where the inequality follows from  $(\bar{\theta}_{\text{RPO}}^{\tau} - \bar{\theta}_{\text{RPS}}^{\tau})^{\top} \nabla \mathbb{J}_{\bar{\theta}_{\text{RPS}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPS}}^{\tau}) \geq 0$  by the optimality of  $\bar{\theta}_{\text{RPS}}^{\tau}$ . Using the fact that  $\mathbb{J}_{\bar{\theta}_{\text{RPO}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPO}}^{\tau}) \leq \mathbb{J}_{\bar{\theta}_{\text{RPS}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPS}}^{\tau})$ , we can further upper bound the right-hand side

$$\begin{aligned} \mathbb{J}_{\bar{\theta}_{\text{RPS}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPO}}^{\tau}) - \mathbb{J}_{\bar{\theta}_{\text{RPS}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPS}}^{\tau}) &\leq \mathbb{J}_{\bar{\theta}_{\text{RPS}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPO}}^{\tau}) - \mathbb{J}_{\bar{\theta}_{\text{RPO}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPO}}^{\tau}) \\ &\leq \epsilon L_z \|\bar{\theta}_{\text{RPS}}^{\tau} - \bar{\theta}_{\text{RPO}}^{\tau}\|_2 \\ &\leq \epsilon L_z \|\bar{\theta}_{\text{RPS}}^{\tau} - \bar{\theta}_{\text{RPO}}^{\tau}\|_2, \end{aligned} \quad (31)$$

where the second inequality holds due the  $\epsilon$ -sensitivity of the distribution map  $\hat{\mathbb{P}}(\cdot)$  and the  $L_z$ -Lipschitz continuity of the loss function in  $\mathbf{z}$ . In summary, we obtain

$$\|\bar{\theta}_{\text{RPS}}^{\tau} - \bar{\theta}_{\text{RPO}}^{\tau}\|_2 \leq \frac{2\epsilon L_z}{\alpha}. \quad (32)$$

Now consider the suboptimality decomposition:

$$\begin{aligned} &\mathbb{J}_{\bar{\theta}_{\text{RPS}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPS}}^{\tau}) - \mathbb{J}_{\bar{\theta}_{\text{RPO}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPO}}^{\tau}) \\ &= \left[ \mathbb{J}_{\bar{\theta}_{\text{RPS}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPS}}^{\tau}) - \mathbb{J}_{\bar{\theta}_{\text{RPO}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPO}}^{\tau}) \right] + \left[ \mathbb{J}_{\bar{\theta}_{\text{RPO}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPO}}^{\tau}) - \mathbb{J}_{\bar{\theta}_{\text{RPO}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPO}}) \right]. \end{aligned} \quad (33)$$

We now bound each term:

1. **First term.** Before we start, we first provide a bound on  $\mathbb{J}_{\bar{\theta}_{\text{RPS}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPS}}^{\tau}) - \mathbb{J}_{\bar{\theta}_{\text{RPS}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPO}}^{\tau})$ .

From Lemma (F.2), the function  $f(\mathbf{Z}, \bar{\theta})$  is Lipschitz in  $\bar{\theta}$ . Hence:

$$\begin{aligned} &\mathbb{J}_{\bar{\theta}_{\text{RPS}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPS}}^{\tau}) - \mathbb{J}_{\bar{\theta}_{\text{RPS}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPO}}^{\tau}) \\ &= \mathbb{E}_{\hat{\mathbb{P}}(\bar{\theta}_{\text{RPS}}^{\tau})}[\tau \lambda_{\text{RPS}}^2 + f(\mathbf{Z}, \bar{\theta}_{\text{RPS}}^{\tau})] - \mathbb{E}_{\hat{\mathbb{P}}(\bar{\theta}_{\text{RPO}}^{\tau})}[\tau \lambda_{\text{RPO}}^2 + f(\mathbf{Z}, \bar{\theta}_{\text{RPO}}^{\tau})] \\ &= \tau |(\lambda_{\text{RPS}} + \lambda_{\text{RPO}})(\lambda_{\text{RPS}} - \lambda_{\text{RPO}})| + \mathbb{E}_{\hat{\mathbb{P}}(\bar{\theta}_{\text{RPS}}^{\tau})}[f(\mathbf{Z}, \bar{\theta}_{\text{RPS}}^{\tau}) - f(\mathbf{Z}, \bar{\theta}_{\text{RPO}}^{\tau})] \\ &\leq (2\tau \bar{\lambda} + \rho^2 + D^2) |\lambda_{\text{RPS}} - \lambda_{\text{RPO}}| + L_{\theta} \|\bar{\theta}_{\text{RPS}}^{\tau} - \bar{\theta}_{\text{RPO}}^{\tau}\|_2. \end{aligned} \quad (34)$$

where  $\bar{\lambda}$  is the upper bound on the optimal  $\lambda$  defined in (38). Now we provide a bound on the first term, using the decomposition:

$$\begin{aligned} &\mathbb{J}_{\bar{\theta}_{\text{RPS}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPS}}^{\tau}) - \mathbb{J}_{\bar{\theta}_{\text{RPO}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPO}}^{\tau}) \\ &= \mathbb{J}_{\bar{\theta}_{\text{RPS}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPS}}^{\tau}) - \mathbb{J}_{\bar{\theta}_{\text{RPS}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPO}}^{\tau}) + \mathbb{J}_{\bar{\theta}_{\text{RPS}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPO}}^{\tau}) - \mathbb{J}_{\bar{\theta}_{\text{RPO}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPO}}^{\tau}) \\ &\leq (2\tau \bar{\lambda} + \rho^2 + D^2) |\lambda_{\text{RPS}} - \lambda_{\text{RPO}}| + L_{\theta} \|\bar{\theta}_{\text{RPS}}^{\tau} - \bar{\theta}_{\text{RPO}}^{\tau}\|_2 + \epsilon L_z \|\bar{\theta}_{\text{RPS}}^{\tau} - \bar{\theta}_{\text{RPO}}^{\tau}\|_2 \\ &= (2\tau \bar{\lambda} + \rho^2 + D^2) |\lambda_{\text{RPS}} - \lambda_{\text{RPO}}| + (L_{\theta} + \epsilon L_z) \|\bar{\theta}_{\text{RPS}}^{\tau} - \bar{\theta}_{\text{RPO}}^{\tau}\|_2 \\ &\leq \frac{(2\tau \bar{\lambda} + \rho^2 + D^2 + L_{\theta} + \epsilon L_z) 2\epsilon L_z}{\alpha} \end{aligned} \quad (35)$$

where the first inequality comes from (31) and (34), the last inequality holds because of (32).

2. **Second term.** Using the fact that  $\mathbb{J}_{\bar{\theta}}^{\tau}(\bar{\theta})$  augments  $\mathbb{J}_{\bar{\theta}}(\bar{\theta})$  by  $\tau \lambda^2$ :

$$\mathbb{J}_{\bar{\theta}_{\text{RPO}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPO}}^{\tau}) - \mathbb{J}_{\bar{\theta}_{\text{RPO}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPO}}) \leq \mathbb{J}_{\bar{\theta}_{\text{RPO}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPO}}) - \mathbb{J}_{\bar{\theta}_{\text{RPO}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPO}}) \leq \tau \bar{\lambda}^2 \quad (36)$$

Substituting these into (33), we obtain

$$\mathbb{J}_{\bar{\theta}_{\text{RPS}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPS}}^{\tau}) - \mathbb{J}_{\bar{\theta}_{\text{RPO}}^{\tau}}^{\tau}(\bar{\theta}_{\text{RPO}}) \leq \tau \bar{\lambda}^2 + \frac{(2\tau \bar{\lambda} + \rho^2 + D^2 + L_{\theta} + \epsilon L_z) 2\epsilon L_z}{\alpha}.$$

This concludes the proof.  $\square$

**Lemma F.1.** Suppose that  $\mathcal{Z}$  has a finite diameter  $D = \sup_{\mathbf{z}, \mathbf{z}' \in \mathcal{Z}} \|\mathbf{z} - \mathbf{z}'\|_2 < \infty$  and  $\Theta$  is bounded. Assume that the loss function  $\ell(\mathbf{z}, \boldsymbol{\theta})$  is  $L_\theta$  Lipschitz continuous in  $\boldsymbol{\theta}$  and  $L_z$ -Lipschitz in  $\mathbf{z}$ . Then there exist  $\underline{\ell}, \bar{\ell} \in \mathbb{R}$  such that for all  $(\mathbf{z}, \boldsymbol{\theta}) \in \mathcal{Z} \times \Theta$ ,

$$\underline{\ell} \leq \ell(\mathbf{z}, \boldsymbol{\theta}) \leq \bar{\ell}.$$

Consider the following univariate minimization problem

$$\inf_{\lambda \in \mathbb{R}_+} \rho^2 \lambda + \tau \lambda^2 + \mathbb{E}_{\hat{\mathbb{P}}(\boldsymbol{\theta}_t)} [\ell_c(\mathbf{z}, \boldsymbol{\theta}, \lambda)] \quad (37)$$

where

$$\ell_c(\hat{\mathbf{z}}_s, \boldsymbol{\theta}, \lambda) = \sup_{\mathbf{z} \in \mathcal{Z}} (\ell(\mathbf{z}, \boldsymbol{\theta}) - \lambda \|\mathbf{z} - \hat{\mathbf{z}}_s\|_2^2).$$

Then, the problem (37) admits a minimizer  $\lambda^* \leq \bar{\lambda}$ , where

$$\bar{\lambda} = \frac{\sqrt{\rho^4 + 4\tau(\bar{\ell} - \underline{\ell})} - \rho^2}{2\tau}. \quad (38)$$

*Proof.* First, we observe that for fixed  $(\boldsymbol{\theta}, \hat{\mathbf{z}}_s)$ , the function  $\ell_c(\mathbf{z}, \boldsymbol{\theta}, \lambda)$  is convex and lower semi-continuous in  $\lambda$ , as it is the supremum of functions affine in  $\lambda$ . Since lower semi-continuity is preserved under expectation over a discrete distribution  $\hat{\mathbb{P}}(\boldsymbol{\theta}_t)$ , the objective function in (37) is convex and lower semi-continuous in  $\lambda$ .

Next, we establish a lower bound. Note that for all  $\lambda \geq 0$ ,

$$\ell_c(\hat{\mathbf{z}}_s, \boldsymbol{\theta}, \lambda) \geq \ell(\hat{\mathbf{z}}_s, \boldsymbol{\theta}) \geq \underline{\ell},$$

since the supremum is attained at  $\mathbf{z} = \hat{\mathbf{z}}_s$ . Hence, we can bound the objective from below:

$$\rho^2 \lambda + \tau \lambda^2 + \mathbb{E}_{\hat{\mathbb{P}}(\boldsymbol{\theta}_t)} [\ell_c(\hat{\mathbf{z}}_s, \boldsymbol{\theta}, \lambda)] \geq \rho^2 \lambda + \tau \lambda^2 + \underline{\ell}.$$

Thus, the infimum of problem (37) is attained for some  $\lambda^* \in [0, +\infty)$ . By optimality of  $\lambda^*$ , we then have:

$$\begin{aligned} \rho^2 \lambda^* + \tau \lambda^{*2} + \mathbb{E}_{\hat{\mathbb{P}}(\boldsymbol{\theta}_t)} [\ell(\hat{\mathbf{z}}_s, \boldsymbol{\theta})] &\leq \rho^2 \lambda^* + \tau \lambda^{*2} + \mathbb{E}_{\hat{\mathbb{P}}(\boldsymbol{\theta}_t)} \ell_c(\hat{\mathbf{z}}_s, \boldsymbol{\theta}, \lambda) \\ &\leq \mathbb{E}_{\hat{\mathbb{P}}(\boldsymbol{\theta}_t)} [\ell_c(\hat{\mathbf{z}}_s, \boldsymbol{\theta}, 0)] \\ &\leq \mathbb{E}_{\hat{\mathbb{P}}(\boldsymbol{\theta}_t)} [\sup_{\mathbf{z} \in \mathcal{Z}} \ell(\mathbf{z}, \boldsymbol{\theta})] \leq \bar{\ell}, \end{aligned}$$

where the second inequality comes from evaluating the objective at  $\lambda = 0$ . Rearranging the inequality, we obtain:

$$\rho^2 \lambda^* + \tau \lambda^{*2} \leq \bar{\ell} - \mathbb{E}_{\hat{\mathbb{P}}(\boldsymbol{\theta}_t)} [\ell(\hat{\mathbf{z}}_s, \boldsymbol{\theta})] \leq \bar{\ell} - \underline{\ell}.$$

Solving the quadratic inequality yields the upper bound

$$\lambda^* \leq \bar{\lambda} = \frac{\sqrt{\rho^4 + 4\tau(\bar{\ell} - \underline{\ell})} - \rho^2}{2\tau}.$$

This concludes the proof.  $\square$

**Lemma F.2.** Suppose  $\mathcal{Z}$  has a finite diameter  $D = \sup_{\mathbf{z}, \mathbf{z}' \in \mathcal{Z}} \|\mathbf{z} - \mathbf{z}'\|_2 < \infty$  and the loss function  $\ell(\mathbf{z}, \boldsymbol{\theta})$  is  $L_\theta$  Lipschitz in  $\boldsymbol{\theta}$ . For  $\hat{\mathbf{z}}_s \in \mathcal{Z}$ , define the function

$$f(\hat{\mathbf{z}}_s, \bar{\boldsymbol{\theta}}) = \sup_{\mathbf{z} \in \mathcal{Z}} h(\mathbf{z}, \hat{\mathbf{z}}_s, \bar{\boldsymbol{\theta}}), \quad (39)$$

where  $\bar{\boldsymbol{\theta}} := (\boldsymbol{\theta}, \lambda) \in \Theta \times \mathbb{R}_+ := \bar{\Theta}$ , and

$$h(\mathbf{z}, \hat{\mathbf{z}}_s, \bar{\boldsymbol{\theta}}) = \ell(\mathbf{z}, \boldsymbol{\theta}) - \lambda \|\mathbf{z} - \hat{\mathbf{z}}_s\|_2^2 + \rho^2 \lambda.$$

Then for any  $\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\theta}}' \in \bar{\Theta}$ , the function  $f$  satisfies the Lipschitz bound:

$$|f(\hat{\mathbf{z}}_s, \bar{\boldsymbol{\theta}}) - f(\hat{\mathbf{z}}_s, \bar{\boldsymbol{\theta}}')| \leq (\rho^2 + D^2) |\lambda - \lambda'| + L_\theta \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2.$$

*Proof.* We start with the absolute difference between the two evaluations of  $f$ :

$$|f(\hat{\mathbf{z}}_s, \bar{\boldsymbol{\theta}}) - f(\hat{\mathbf{z}}_s, \bar{\boldsymbol{\theta}}')| = \max\{f(\hat{\mathbf{z}}_s, \bar{\boldsymbol{\theta}}) - f(\hat{\mathbf{z}}_s, \bar{\boldsymbol{\theta}}'), f(\hat{\mathbf{z}}_s, \bar{\boldsymbol{\theta}}') - f(\hat{\mathbf{z}}_s, \bar{\boldsymbol{\theta}})\}.$$

Next, we bound the first term; the second is symmetric. Let

$$\Delta = f(\hat{\mathbf{z}}_s, \bar{\boldsymbol{\theta}}) - f(\hat{\mathbf{z}}_s, \bar{\boldsymbol{\theta}}') = \sup_{\mathbf{z} \in \mathcal{Z}} h(\mathbf{z}, \hat{\mathbf{z}}_s, \bar{\boldsymbol{\theta}}) - \sup_{\mathbf{z}' \in \mathcal{Z}} h(\mathbf{z}', \hat{\mathbf{z}}_s, \bar{\boldsymbol{\theta}}').$$

Using the inequality

$$\sup_{\mathbf{z}} a(\mathbf{z}) - \sup_{\mathbf{z}} b(\mathbf{z}) \leq \sup_{\mathbf{z}} (a(\mathbf{z}) - b(\mathbf{z})),$$

we obtain

$$\begin{aligned} \Delta &\leq \sup_{\mathbf{z} \in \mathcal{Z}} \left[ h(\mathbf{z}, \hat{\mathbf{z}}_s, \bar{\boldsymbol{\theta}}) - h(\mathbf{z}, \hat{\mathbf{z}}_s, \bar{\boldsymbol{\theta}}') \right] \\ &= \sup_{\mathbf{z} \in \mathcal{Z}} [\ell(\mathbf{z}, \boldsymbol{\theta}) - \ell(\mathbf{z}, \boldsymbol{\theta}') - (\lambda - \lambda') \|\mathbf{z} - \hat{\mathbf{z}}_s\|^2 + \rho^2(\lambda - \lambda')] \\ &\leq \sup_{\mathbf{z} \in \mathcal{Z}} |\ell(\mathbf{z}, \boldsymbol{\theta}) - \ell(\mathbf{z}, \boldsymbol{\theta}')| + \sup_{\mathbf{z} \in \mathcal{Z}} |(\rho^2 - \|\mathbf{z} - \hat{\mathbf{z}}_s\|_2^2)(\lambda - \lambda')| \end{aligned}$$

For the first term, since  $\ell$  is  $L_\theta$ -Lipschitz in  $\boldsymbol{\theta}$ :

$$|\ell(\mathbf{z}, \boldsymbol{\theta}) - \ell(\mathbf{z}, \boldsymbol{\theta}')| \leq L_\theta \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2.$$

For the second term, observe that  $\|\mathbf{z} - \hat{\mathbf{z}}_s\|_2 \leq D$  implies

$$|\rho^2 - \|\mathbf{z} - \hat{\mathbf{z}}_s\|_2^2| \leq \rho^2 + D^2.$$

Thus,

$$|(\rho^2 - \|\mathbf{z} - \hat{\mathbf{z}}_s\|_2^2)(\lambda - \lambda')| \leq (\rho^2 + D^2)|\lambda - \lambda'|.$$

By symmetry, the same bound holds for  $f(\hat{\mathbf{z}}_s, \bar{\boldsymbol{\theta}}') - f(\hat{\mathbf{z}}_s, \bar{\boldsymbol{\theta}})$ . This concludes the proof.  $\square$

#### F.4 Proof of Theorem 2

The proof of Theorem 2 follows directly from the results of Theorem 3, Theorem 4 and Theorem 5.

## G Experiment Details

### G.1 Strategic Classification

Following [40, 32], we assume that individuals have linear utilities  $u(\boldsymbol{\theta}, \tilde{\mathbf{x}}) = -\boldsymbol{\theta}^\top \tilde{\mathbf{x}}$  and quadratic costs  $c(\tilde{\mathbf{x}}', \tilde{\mathbf{x}}) = -\frac{1}{2\epsilon} \|\tilde{\mathbf{x}}' - \tilde{\mathbf{x}}\|_2^2$ , where  $\epsilon$  is a positive constant regulating the cost of altering features and thus the sensitivity of the distribution map. In other words, individuals aim to minimize their assigned probability of default but are unable to change their true outcome  $\tilde{y}$ . We select  $S \subseteq [P - 1]$  strategic features, such as the number of open credit lines. Each time an individual manipulates their strategic features as depicted in [40, Section 5], the best response for an individual results in the update

$$\tilde{\mathbf{x}}'_S = \tilde{\mathbf{x}}_S - \epsilon \boldsymbol{\theta}_S$$

where  $\tilde{\mathbf{x}}'_S, \tilde{\mathbf{x}}_S, \boldsymbol{\theta}_S \in \mathbb{R}^{|S|}$ .

**Robust type-1.** Consider the 1-Wasserstein ball, it follows from Theorem (1), at each time  $t$ , we can solve the following Tikhonov regularization problem

$$\mathbb{J}_{\boldsymbol{\theta}_t}(\boldsymbol{\theta}) = \inf_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_{\hat{\mathbb{P}}(\boldsymbol{\theta}_t)} [\log(1 + \exp(-\mathbf{x}^\top \boldsymbol{\theta} \hat{y}))] + \rho L \|\boldsymbol{\theta}\|_2^2.$$

**Robust type-2.** Consider the 2-Wasserstein ball, it follows from Proposition (1), at each time  $t$ , we can solve the following problem

$$\mathbb{J}_{\theta_t}(\theta) = \inf_{\lambda \in \mathbb{R}_+} \frac{1}{S} \sum_{s \in [S]} \sup_{\alpha \in (-1, 0)} \left\{ \alpha \hat{\mathbf{x}}_s^\top \theta \hat{y}_s + \frac{\alpha^2}{4\lambda} \theta^\top \theta - h(\alpha) \right\} + \rho^2 \lambda,$$

where  $h(\alpha) = (\alpha + 1) \log(1 + \alpha) - \alpha \log(-\alpha)$ .

Next, we introduce auxiliary variables  $t_s \forall s \in [S]$  and combine with outer minimization over  $\theta \in \Theta$ , which yield the equivalent formulation for the right hand side of the above inequality

$$\begin{aligned} \inf & \quad \frac{1}{S} \sum_{s \in [S]} t_s + \rho^2 \lambda \\ \text{s. t.} & \quad \theta \in \Theta, \quad \lambda \in \mathbb{R}_+, \quad t_s \in \mathbb{R} \quad \forall s \in [S] \\ & \quad \sup_{\alpha \in (-1, 0)} \left\{ \alpha \hat{\mathbf{x}}_s^\top \theta \hat{y}_s + \frac{\alpha^2}{4\lambda} \theta^\top \theta - h(\alpha) \right\} \leq t_s \quad \forall s \in [S]. \end{aligned} \tag{40}$$

To handle the last constraints, we discretize over some finite set  $\mathcal{S}_\alpha$ , i.e.,

$$\alpha \hat{\mathbf{x}}_s^\top \theta \hat{y}_s + \frac{\alpha^2}{4\lambda} \theta^\top \theta - h(\alpha) \leq t_s \quad \forall \alpha \in \mathcal{S}_\alpha \quad \forall s \in [S].$$

For the experiment we choose  $\mathcal{S}_\alpha = \{-0.9, -0.8, \dots, -0.1\}$ .

**Proposition 1.** Consider the logistic loss

$$\ell(\mathbf{z}, \theta) = \log(1 + \exp(-\mathbf{x}^\top \theta \hat{y})),$$

with  $\mathbf{z} = (\mathbf{x}, \hat{y})$  and  $\theta \in \mathbb{R}^d$ . Consider the 2-Wasserstein ball, where the ground cost  $c$  is given by the Euclidean norm on  $\mathbb{R}^d$ . Assume the support set  $\mathcal{X}$  is convex and closed, then we have

$$\mathbb{J}_{\theta_t}(\theta) = \inf_{\lambda \in \mathbb{R}_+} \frac{1}{S} \sum_{s \in [S]} \sup_{\alpha \in (-1, 0)} \left\{ \alpha \hat{\mathbf{x}}_s^\top \theta \hat{y}_s + \frac{\alpha^2}{4\lambda} \theta^\top \theta - h(\alpha) \right\} + \rho^2 \lambda,$$

where  $h(\alpha) = (\alpha + 1) \log(1 + \alpha) - \alpha \log(-\alpha)$ .

*Proof.* We follow the argument in the proof of Theorem (3). For fixed  $\theta_t$ , we have

$$\mathbb{J}_{\theta_t}(\theta) = \inf_{\lambda \in \mathbb{R}_+} \frac{1}{S} \sum_{s \in [S]} \sup_{\mathbf{z} \in \mathcal{Z}} [\ell(\mathbf{z}, \theta) - \lambda \|\mathbf{z} - \hat{\mathbf{z}}_s\|_2^2 + \rho^2 \lambda].$$

where  $\hat{\mathbf{z}} = \hat{\mathbf{x}} \hat{y}_s$ . Note that label  $\hat{y}_s$  is fixed, we can simplify the inner supremum

$$\sup_{\mathbf{x} \in \mathcal{X}} [\ell(\mathbf{x} \hat{y}_s, \theta) - \lambda \|\mathbf{x} - \hat{\mathbf{x}}_s\|_2^2 + \rho^2 \lambda]$$

Next, substitute the logistic loss

$$\ell(\mathbf{x} \hat{y}_s, \theta) = \log(1 + \exp(-\mathbf{x}^\top \theta \hat{y}_s)).$$

Using the Fenchel conjugate dual formulation [45] of the logistic loss:

$$\log(1 + e^{-u}) = \sup_{\alpha \in (-1, 0)} \{\alpha u - h(\alpha)\}, \quad \text{where } h(\alpha) = (\alpha + 1) \log(1 + \alpha) - \alpha \log(-\alpha),$$

we rewrite the inner supremum as:

$$\begin{aligned} & \sup_{\mathbf{x} \in \mathcal{X}} \log(1 + \exp(-\mathbf{x}^\top \theta \hat{y}_s)) - \lambda \|\mathbf{x} - \hat{\mathbf{x}}_s\|_2^2 \\ &= \sup_{\mathbf{x} \in \mathcal{X}} \sup_{\alpha \in (-1, 0)} \{\alpha \mathbf{x}^\top \theta \hat{y}_s - h(\alpha) - \lambda \|\mathbf{x} - \hat{\mathbf{x}}_s\|_2^2\} \\ &= \sup_{\alpha \in (-1, 0)} \left\{ \sup_{\mathbf{x} \in \mathcal{X}} (\alpha \mathbf{x}^\top \theta \hat{y}_s - \lambda \|\mathbf{x} - \hat{\mathbf{x}}_s\|_2^2) - h(\alpha) \right\}. \end{aligned}$$

We now compute the inner supremum over  $\mathbf{x}$ . For fixed  $\alpha$ ,  $\hat{y}_s$ , and  $\boldsymbol{\theta}$ , the expression

$$\alpha \mathbf{x}^\top \boldsymbol{\theta} \hat{y}_s - \lambda \|\mathbf{x} - \hat{\mathbf{x}}_s\|_2^2$$

is a concave quadratic in  $\mathbf{x}$ . The optimum is achieved at:

$$\mathbf{x}^* = \hat{\mathbf{x}}_s + \frac{\alpha}{2\lambda} \boldsymbol{\theta} \hat{y}_s.$$

Substituting  $\mathbf{x}^*$  back in yields:

$$\alpha \hat{\mathbf{x}}_s^\top \boldsymbol{\theta} \hat{y}_s + \frac{\alpha^2}{4\lambda} \boldsymbol{\theta}^\top \boldsymbol{\theta}.$$

Therefore, the robust objective becomes:

$$\mathbb{J}_{\boldsymbol{\theta}_t}(\boldsymbol{\theta}) = \inf_{\lambda \in \mathbb{R}_+} \frac{1}{S} \sum_{s \in [S]} \sup_{\alpha \in (-1, 0)} \left\{ \alpha \hat{\mathbf{x}}_s^\top \boldsymbol{\theta} \hat{y}_s + \frac{\alpha^2}{4\lambda} \boldsymbol{\theta}^\top \boldsymbol{\theta} - h(\alpha) \right\} + \rho^2 \lambda.$$

□

**Alternative minimization under KL divergence ambiguity set (AMKL)** Consider the Kullback–Leibler (KL) divergence ambiguity set. According to [58, Proposition 3.1], the distributionally robust optimization problem under KL divergence can be reformulated as:

$$\min_{\boldsymbol{\theta}} \inf_{\mu \geq 0} \left\{ \mu \log \mathbb{E}_{\hat{\mathbb{P}}(\boldsymbol{\theta}_t)} \left[ (1 + \exp(-\mathbf{x}^\top \boldsymbol{\theta} \hat{y}))^{1/\mu} \right] + \mu \rho \right\},$$

where  $\rho > 0$  is the radius of the ambiguity set. To solve this optimization problem, [58] propose an alternating minimization approach, which we refer to as AMKL. The procedure alternates between the following two steps:

1.  $\boldsymbol{\theta}$ -step: Fix the robustness parameter  $\mu$ , and minimize the objective with respect to  $\boldsymbol{\theta}$ .
2.  $\mu$ -step: Fix the model parameter  $\boldsymbol{\theta}$ , and minimize the objective with respect to  $\mu$ .

This iterative process is repeated until convergence. In the  $\boldsymbol{\theta}$ -step, we apply the repeated risk minimization algorithm, following the suggestion in [58] that this subproblem can be addressed using any suitable performative risk minimization algorithm.

## G.2 Impact of the Robust Parameter $\rho$

We investigate the impact of the robust parameter  $\rho$  on out-of-sample performance. Specifically, we consider the revenue management problem described in section (4.2) under different quantities of perishable products,  $q \in \{200, 300, 500\}$ .

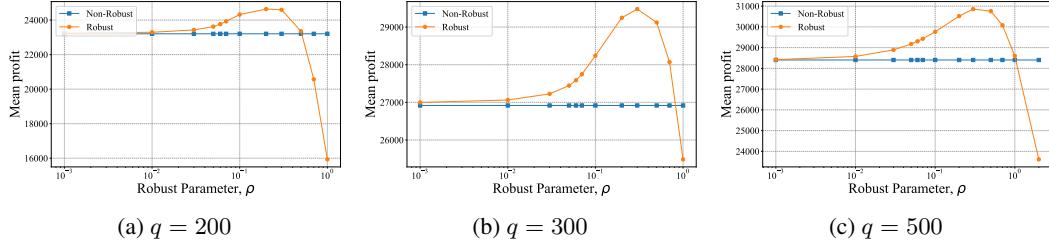


Figure 4: Out-of-sample performance as a function of the robust parameter  $\rho$  and estimated on the basis of 100 simulations.

Figure 4 illustrates the optimal mean profit as a function of  $\rho$ , averaged over 100 independent simulation runs. We observe that the out-of-sample performance improves as  $\rho$  increases up to a critical Wasserstein radius  $\rho^*$ , beyond which it begins to deteriorate. Notably, the robust model outperforms the non-robust counterpart over a wide range of  $\rho$  values. This pattern was consistently observed across all simulation settings and provides an empirical justification for adopting a distributionally robust approach.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: Our claims are supported by both theoretical analysis and empirical results. We present experiments on the standard strategic classification benchmark, as well as two decision-making problems in revenue management and demand response portfolio optimization.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: Limitations are discussed in the conclusion section.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: All theoretical results have been proven rigorously with assumptions clearly stated.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: The main algorithm and the formulations have been clearly described. The details for experiments are included in the main text and appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: The data is from a public dataset [25] or generated synthetically (see details in Section 4). The code is available upon request.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: Experimental details have been discussed in the Experiments section as well as in the Appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [\[Yes\]](#)

Justification: We provide box plots for the out-of-sample performances.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer “Yes” if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: Computational resources are provided in the Experiments section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [\[Yes\]](#)

Justification: We have read and acknowledged the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [\[Yes\]](#)

Justification: We discuss societal impacts in the conclusion. Our framework extends the scope of performative prediction beyond its original focus, enabling its application to a wider range of decision-making problems. In high-stakes settings, adopting a distributionally robust optimization perspective allows our approach to prioritize safe and reliable deployment in the presence of uncertainty and potential adversarial conditions.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: No data or models are released.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The creators of the existing dataset are properly credited in Section 4 and in the reference [25]. We have cited the benchmark algorithms in the Experiments section.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](http://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not introduce new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The paper does not use LLMs for important, original, or non-standard component of the core methods in this research.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.