manuscript

Tractable Robust Supervised Learning Models

Melvyn Sim, Long Zhao and Minglong Zhou
Department of Analytics & Operations (DAO), NUS Business School, National University of Singapore
melvynsim@gmail.com, longzhao@nus.edu.sg, minglong_zhou@u.nus.edu

At the heart of supervised learning is a minimization problem with an objective function that evaluates a set of training data over a loss function that penalizes poor fitting and a regularization function that penalizes over-fitting to the training data. More recently, data-driven robust optimization based learning models provide an intuitive robustness perspective of regularization. However, when the loss function is not Lipschitz continuous, solving the robust learning models exactly can be computationally challenging. We focus on providing tractable approximations for robust regression and classification problems for loss functions derived from Lipschitz continuous functions raised to the power of p. We also show the equivalence of the type-p robust learning models to the pth-root regularization problems when the underlying support sets are unbounded. Inspired by Long et al. (2021), we also propose tractable type-p *robust satisficing* learning models that are specified by target loss parameters. We illustrate that the robust satisficing regression and classification models can be tractably solved for a large class of problems, and we also establish finite sample probabilistic guarantees for limiting losses beyond the specified target. While the family of solutions generated by regularization and robust satisficing can be the same, from empirical studies on popular datasets, the relative targets for reasonably good out-of-sample performance can be found within a narrow range. We also demonstrate in the numerical study that the target-based hyper-parameter is easier to determine via cross-validation and can improve out-of-sample performance compared to standard regularization approaches.

Key words: regression, classification, regularization, robust optimization, robust satisficing
History: December 9, 2021.

1. Introduction

Supervised learning provides a systematic framework for identifying and estimating dependencies from past data (Vapnik 2013). A central question in supervised learning is to design models that would perform well in future or unseen data. In classical models, one aims to find an estimator from a prescribed hypothesis space that minimizes some in-sample loss function. However, such an estimator could overfit the empirical data and may perform terribly when evaluated on another set of data. Regularization, a ubiquitous approach in machine learning, mitigates overfitting by incorporating a penalty to the loss function (*e.g.,* Tikhonov and Arsenin 1977, Bishop 1995, Shalev-Shwartz et al. 2010). For instance, the celebrated LASSO regression penalizes the total magnitude

of the coefficient weights (Tibshirani 1996), which has the effects of mitigating overfitting and obtaining linear estimators that are also sparse.

Robust optimization (Bertsimas and Sim 2004, Ben-Tal et al. 2006, 2013) provides an alternative perspective to regularization. A robust supervised learning model would obtain an estimator that minimizes the worst-case loss function when the training data is subject to some specified amount of perturbations. El Ghaoui and Lebret (1997) have drawn the connection by elucidating the equivalence of the robust least squares regressions to a Tikhonov regularization (*e.g.,* Ridge regression), when the uncertainty set is characterized by a Frobenius norm. Xu et al. (2009, 2010) establish similar equivalences between regularization and supervised learning models, including linear regressions and support vector machines (Cortes and Vapnik 1995). We refer interested readers to Bertsimas and Copenhaver (2018) for a comprehensive characterization of conditions of the equivalence between robustification and regularization.

Similar connections have also been established between regularization and learning models under the distributionally robust optimization framework. Distributionally robust optimization seeks to minimize the worst-case expected loss, where the expectation is taken with respect to the most adversarial distribution within a prescribed ambiguity set (Delage and Ye 2010, Wiesemann et al. 2014, Gao and Kleywegt 2016, Mohajerin Esfahani and Kuhn 2018, Blanchet and Murthy 2019, Chen et al. 2020). It has been studied that distributionally robust optimization with a ϕ-divergence ambiguity set is closely related to variance regularization (Gotoh et al. 2018, 2021). Some works focus on distributionally robust optimization with moment-based ambiguity sets, *e.g.,* Lanckriet et al. (2002), Huang et al. (2004) study minimax probability machines for classification problems using distributionally robust optimization with moment-based ambiguity sets. Farnia and Tse (2016) also investigate the moment-based ambiguity sets with the restriction that the marginal distribution of the data matches the empirical marginal distribution. They also discuss the connections of their distributionally robust learning models to classical regression models.

Inspired by the recent advancement of data-driven robust optimization (Gao and Kleywegt 2016, Mohajerin Esfahani and Kuhn 2018, Blanchet and Murthy 2019), many works in this area focus on data-driven robust learning models with Wasserstein distance-based ambiguity set. The adoption of the statistical distance-based ambiguity set naturally links to the empirical loss minimization problem, and we can associate the Wasserstein distance to established finite sample probabilistic guarantee results (Fournier and Guillin 2015). Shafieezadeh-Abadeh et al. (2015) discuss data-driven robust logistic regression and elucidate its connections to the regularized logistic regression. Shafieezadeh-Abadeh et al. (2019) generalize this work by presenting a unified framework for data-driven robust linear regression and classification. They restrict to Lipschitz continuous loss functions and type-1 Wasserstein distance, and they propose various tractable reformulations

under different loss functions. They also establish that under some technical conditions, the radius of the ambiguity set is equivalent to the penalty parameter in the corresponding regularization model. Specifically, choosing the radius parameter in a data-driven robust learning model is the same as choosing a penalty parameter in its corresponding regularization model. Blanchet et al. (2019) present the data-driven robust linear regression with a square loss function and a type-2 Wasserstein distance ambiguity set, and they establish the equivalence between their model and a square-root LASSO/Ridge regression. Gao et al. (2017) propose an asymptotic equivalence of data-driven robust learning models with a general type-p Wasserstein distance ambiguity set. The asymptotic equivalent model has a clear interpretation in terms of regularization.

When the loss function is not Lipschitz continuous, solving the robust supervised learning problems exactly can be computationally challenging. Hence, in this paper, we focus on providing tractable approximations for robust regression and classification problems for loss functions that are derived from Lipschitz continuous functions raised to the power of p. The type-p robust models recover the known tractable cases of Shafieezadeh-Abadeh et al. (2019) and Blanchet et al. (2019). When the underlying support sets are unbounded, we show their equivalence to the pth-root regularization problems. Inspired by Long et al. (2021), we also propose tractable type-p *robust satisficing* learning models, which are specified by target loss parameters. We illustrate that the robust satisficing regression and classification models can be tractably solved for a large class of problems, and we also establish finite sample probabilistic guarantees for limiting losses beyond the specified target. While the families of solutions generated by regularization and robust satisficing are the same, from empirical studies on popular datasets, the relative targets for reasonably good out-of-sample performance can be found within a narrow range. We also demonstrate in the numerical study that the target-based hyper-parameter is easier to determine via cross-validation and can improve out-of-sample performance compared to standard regularization approaches.

Notation. We use boldface lowercase letters for vectors (*e.g.*, $\boldsymbol{\theta}$), and calligraphic letters for sets (*e.g.*, $\mathcal{X}$). We use $[N]$ to denote the running index $\{1, 2, 3, \ldots, N\}$, where N is a known integer. We adopt the convention that $\inf \emptyset = +\infty$, where $\emptyset$ is the empty set. A random variable $\tilde{v}$ is denoted with a tilde sign such as $\tilde{v} \sim \mathbb{P}, \mathbb{P} \in \mathcal{P}_0$, where $\mathcal{P}_0$ represents the set of all possible distributions. For $\tilde{v}_1, \tilde{v}_2$, we use $\tilde{v}_1 \geq \tilde{v}_2$ to imply $\tilde{v}_1$ state-wise dominates $\tilde{v}_2$. For multivariate random variable, we use $\mathcal{P}_0(\mathcal{Z})$ to represent the set of all distributions for the multivariate random variable that has support $\mathcal{Z} \subseteq \mathbb{R}^N$. Specifically, we use $\tilde{\boldsymbol{z}} \sim \mathbb{P}$, $\mathbb{P} \in \mathcal{P}_0(\mathcal{Z})$ to define $\tilde{\boldsymbol{z}}$ as a multivariate random variable with support $\mathcal{Z}$ and distribution $\mathbb{P}$. We we use the shortcut, $(x)^+$ to represent $\max\{0, x\}$.

2. Robust regression and regularization

We first focus on the linear regression problem, where we predict the scalar response variable $\tilde{y}$ using an affine function of N explanatory variables $\tilde{\boldsymbol{x}}$. Mathematically speaking, we want to solve the following optimization problem,

$$\min_{(\alpha,\boldsymbol{\beta})\in\mathcal{D}} \mathbb{E}_{(\tilde{y},\tilde{\boldsymbol{x}})\sim\mathbb{P}^*}\left[\ell((\alpha+\tilde{\boldsymbol{x}}^\top\boldsymbol{\beta})-\tilde{y})\right],$$

where $\mathcal{D}$ is a convex feasible set for $(\alpha,\boldsymbol{\beta})$, $\mathbb{P}^*\in\mathcal{P}_0(\mathcal{Z})$, $\mathcal{Z}\subseteq\mathbb{R}^{N+1}$ is the unobservable true distribution, and ℓ is the loss function. By defining $\tilde{\boldsymbol{z}}\triangleq(\tilde{\boldsymbol{x}},\tilde{y})$ and $\bar{\boldsymbol{\beta}}\triangleq(\boldsymbol{\beta},-1)$, the optimization above becomes

$$\min_{(\alpha,\bar{\boldsymbol{\beta}})\in\bar{\mathcal{D}}} \mathbb{E}_{\mathbb{P}^*}\left[\ell(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}})\right],$$

where $\bar{\mathcal{D}}=\mathcal{D}\times\{-1\}$. Hereinafter, we use $\mathbb{E}_{\mathbb{P}^*}$ instead of $\mathbb{E}_{\tilde{\boldsymbol{z}}\sim\mathbb{P}^*}$ to simplify the formulas.

Because we only have access to S i.i.d. samples from the unknown true distribution $\mathbb{P}^*$, it is popular to solve the following optimization problem,

$$Z_0=\min_{(\alpha,\bar{\boldsymbol{\beta}})\in\bar{\mathcal{D}}}\left\{\frac{1}{S}\sum_{s\in[S]}\ell(\alpha+\hat{\boldsymbol{z}}_s^\top\bar{\boldsymbol{\beta}})\right\}=\min_{(\alpha,\bar{\boldsymbol{\beta}})\in\bar{\mathcal{D}}}\mathbb{E}_{\hat{\mathbb{P}}}\left[\ell(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}})\right], \tag{1}$$

where $\hat{\boldsymbol{z}}_s\in\mathcal{Z},\ s\in[S]$ are the S samples and we denote $\hat{\mathbb{P}}\in\mathcal{P}_0(\mathcal{Z})$ as the corresponding empirical distribution. It is conceivable that the out-of-sample evaluation of the solution of Problem (1) may perform terribly when $\hat{\mathbb{P}}$ deviates significantly from $\mathbb{P}^*$. The data-driven robust optimization model of Mohajerin Esfahani and Kuhn (2018) can be applied to improve the out-of-sample evaluation using a type-1 Wasserstein distance-based ambiguity set, $\mathcal{W}_1(r)$, where

$$\mathcal{W}_p(r):=\left\{\mathbb{P}\in\mathcal{P}_0(\mathcal{Z})\ \middle|\ \begin{array}{l}\tilde{\boldsymbol{z}}\sim\mathbb{P}\\ \Delta_p(\mathbb{P},\hat{\mathbb{P}})\le r\end{array}\right\},$$

with $\Delta_p(\mathbb{P},\hat{\mathbb{P}})$ being the Wasserstein distance metric (of type-p), defined as follows:

$$\Delta_p(\mathbb{P},\hat{\mathbb{P}}):=\inf_{\mathbb{Q}\in\mathcal{P}_0(\mathcal{Z}^2)}\left\{\left(\mathbb{E}_{\mathbb{Q}}\left[\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|^p\right]\right)^{\frac{1}{p}}\ \middle|\ (\tilde{\boldsymbol{z}},\tilde{\boldsymbol{v}})\sim\mathbb{Q},\tilde{\boldsymbol{z}}\sim\mathbb{P},\tilde{\boldsymbol{v}}\sim\hat{\mathbb{P}}\right\}.$$

The parameter r controls the size of the ambiguity set by limiting the probability distance that can deviate from the empirical distribution. The measure concentration result of Fournier and Guillin (2015) provides the relevant finite sample probabilistic guarantees that justify the Wasserstein distance-based ambiguity set.

THEOREM 1. **(Theorem 2 of Fournier and Guillin 2015).** *Let $\mathbb{P}^S$ denote the distribution that governs the distribution of the independent samples $\hat{\boldsymbol{z}}_1,\ldots,\hat{\boldsymbol{z}}_S$ drawn from $\mathbb{P}^*$ for which the*

empirical distribution $\hat{\mathbb{P}}$ *is constructed. When the true data-generating distribution* $\mathbb{P}^*$, $\tilde{\boldsymbol{z}} \sim \mathbb{P}^*$ *is a light-tailed distribution such that* $\mathbb{E}_{\mathbb{P}^*}[\exp(\|\tilde{\boldsymbol{z}}\|^\alpha)] < \infty$ *for some* $\alpha > p$, $p \geq 1$, *then for all* $r \geq 0$,

$$\mathbb{P}^S \left[\Delta_p(\mathbb{P}^*, \hat{\mathbb{P}}) > r \right] \leq g_p(r)$$

for some function $g_p(r) : \mathbb{R}_+ \mapsto \mathbb{R}_+$ *that decreases to zero at an exponential rate in* r.

However, in practice, the generic probabilistic bound is expected to be loose, and it is typically not used to determine the desired radius, r. Instead, similar to regularization, r could be determined via cross-validation.

To immunize against uncertainty, we can consider solving the following data-driven robust linear regression optimization problem:

$$\begin{aligned} Z_r = \min \; & \sup_{\mathbb{P} \in \mathcal{W}_1(r)} \mathbb{E}_{\mathbb{P}} \left[\ell(\alpha + \tilde{\boldsymbol{z}}^\top \bar{\boldsymbol{\beta}}) \right] \\ \text{s.t.} \; & (\alpha, \bar{\boldsymbol{\beta}}) \in \bar{\mathcal{D}}. \end{aligned} \tag{2}$$

The data-driven robust linear regression optimization problem (2) has well been studied in the literature, and a strong connection with regularization has been established (see, *e.g.*, Shafieezadeh-Abadeh et al. 2019). The robust regression model can be motivated from

$$\mathbb{P}^S \left[\mathbb{E}_{\mathbb{P}^*} \left[\ell(\alpha + \tilde{\boldsymbol{z}}^\top \bar{\boldsymbol{\beta}}) \right] > \sup_{\mathbb{P}: \Delta_1(\mathbb{P}, \hat{\mathbb{P}}) \leq r} \mathbb{E}_{\mathbb{P}} \left[\ell(\alpha + \tilde{\boldsymbol{z}}^\top \bar{\boldsymbol{\beta}}) \right] \right] \leq \mathbb{P}^S \left[\Delta_1(\mathbb{P}^*, \hat{\mathbb{P}}) > r \right], \tag{3}$$

which allows us to relate to the concentration inequality of Theorem 1 for $p = 1$.

To derive a tractable reformulation of Problem (2), we first focus on the class of Lipschitz continuous loss functions defined as follows.

Definition 1 (loss function). A loss function, $\ell : \mathbb{R} \mapsto \mathbb{R}_+$ is a convex and nonnegative function that is also Lipschitz continuous, *i.e.*, there exists a minimum $L > 0$ such that for all $w_1, w_2 \in \mathbb{R}$,

$$|\ell(w_1) - \ell(w_2)| \leq L|w_1 - w_2|.$$

Here, we list several common loss functions used in linear regression:

1. Absolute deviation: $\ell(w) = |w|$.
2. Huber loss with parameter $\delta > 0$: $\ell(w) = \frac{w^2}{2}$ if $|w| \leq \delta$; otherwise, $\ell(w) = \delta(|w| - \frac{\delta}{2})$.
3. Pinball loss with parameter $\delta \in [0, 1]$: $\ell(w) = \max\{-\delta w, (1 - \delta)w\}$.

Note that the Lipschitz continuity would rule out the ubiquitous square loss function used in the ordinary least squares. Now, we could convert the distributionally robust optimization of Problem (2) to an equivalent robust optimization formulation (see, *e.g.*, Mohajerin Esfahani and Kuhn 2018, Gao and Kleywegt 2016):

$$\begin{aligned} Z_r = \min \; & \kappa r + \frac{1}{S} \sum_{s \in [S]} \sup_{\boldsymbol{z}_s \in \mathcal{Z}} \{ \ell(\alpha + \boldsymbol{z}_s^\top \bar{\boldsymbol{\beta}}) - \kappa \|\boldsymbol{z}_s - \hat{\boldsymbol{z}}_s\| \} \\ \text{s.t.} \; & (\alpha, \bar{\boldsymbol{\beta}}) \in \bar{\mathcal{D}}, \; \kappa \geq 0. \end{aligned} \tag{4}$$

Hence, the tractability of the data-driven robust linear regression optimization depends on whether the following set of problems

$$\sup_{\boldsymbol{z}_s \in \mathcal{Z}} \left\{ \ell(\alpha + \boldsymbol{z}_s^\top \bar{\boldsymbol{\beta}}) - \kappa \|\boldsymbol{z}_s - \hat{\boldsymbol{z}}_s\| \right\} \tag{5}$$

can be optimized efficiently for all $s \in [S]$. Unfortunately, as Problem (5) is generally not a convex optimization problem, it can be computationally intractable. Nevertheless, there are useful cases where we can evaluate it tractably as follows.

THEOREM 2. **(Shafieezadeh-Abadeh et al. 2019)**

1. *If* $\ell(w) = \max_{i \in [I]} \{a_i w + b_i\}$, *then Problem* (5) *is equivalent to:*

$$\begin{array}{ll} \min & \max\limits_{i \in [I]} \{a_i \alpha + \hat{\boldsymbol{z}}_s^\top \boldsymbol{\eta}_{i,s} + b_i + \max\limits_{\boldsymbol{z}_s \in \mathcal{Z}} \{(a_i \bar{\boldsymbol{\beta}} - \boldsymbol{\eta}_{i,s})^\top \boldsymbol{z}_s\}\} \\ \text{s.t.} & \|\boldsymbol{\eta}_{i,s}\|_* \leq \kappa & \forall i \in [I] \\ & \boldsymbol{\eta}_{i,s} \in \mathbb{R}^{N+1} & \forall i \in [I]. \end{array}$$

2. *If* $\mathcal{Z} = \mathbb{R}^{N+1}$, *then Problem* (5) *is equivalent to:*

$$\begin{cases} \ell(\alpha + \hat{\boldsymbol{z}}_s^\top \bar{\boldsymbol{\beta}}) & \text{if } L\|\bar{\boldsymbol{\beta}}\|_* \leq \kappa \\ \infty & \text{otherwise.} \end{cases}$$

In particular, if ℓ is a convex piecewise linear loss function with a modest number of linear pieces, then we can formulate Problem (5) as a tractable robust optimization model even if the support set $\mathcal{Z}$ is bounded. Otherwise, if $\mathcal{Z} = \mathbb{R}^{N+1}$, then Problem (2) is equivalent to a regularized regression problem as follows:

$$\min_{(\alpha, \bar{\boldsymbol{\beta}}) \in \bar{\mathcal{D}}} \frac{1}{S} \sum_{s \in [S]} \ell(\alpha + \hat{\boldsymbol{z}}_s^\top \bar{\boldsymbol{\beta}}) + rL\|\bar{\boldsymbol{\beta}}\|_*. \tag{6}$$

We remark that there is a subtle difference between Problem (6) and the regularization in regression (*e.g.,* LASSO and Ridge). Regularization in linear regression often only penalizes the dual norm of $\boldsymbol{\beta}$, while Problem (6) penalizes the dual norm of $\bar{\boldsymbol{\beta}} = (\boldsymbol{\beta}, -1)$. Hence, the finite sample guarantee of (3) can also be extended to the regularization as follows

$$\mathbb{P}^S \left[\mathbb{E}_{\mathbb{P}^*} \left[\ell(\alpha + \tilde{\boldsymbol{z}}^\top \bar{\boldsymbol{\beta}}) \right] > \mathbb{E}_{\hat{\mathbb{P}}} \left[\ell(\alpha + \tilde{\boldsymbol{z}}^\top \bar{\boldsymbol{\beta}}) \right] + \lambda \|\bar{\boldsymbol{\beta}}\|_* \right] \leq \mathbb{P}^S \left[\Delta_1(\mathbb{P}^*, \hat{\mathbb{P}}) > \lambda / L \right], \tag{7}$$

for given $\lambda \geq 0$.

Note that for loss function that is not Lipschitz continuous, Problem (5) may not be bounded if the support set is unbounded. To address this, we would have to consider beyond type-1 Wasserstein metric in characterizing the ambiguity set (see, *e.g.*, Blanchet et al. 2019).

Type-p robust regression model

We now consider the more general regression problem, which is based on minimizing a loss function that is derived from an underlying Lipschitz continuous function raised to the power of p as follows:

$$Z_0 = \min_{(\alpha,\bar{\boldsymbol{\beta}})\in\bar{\mathcal{D}}} \left\{ \left(\mathbb{E}_{\hat{\mathbb{P}}}\left[\ell^p(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}})\right]\right)^{\frac{1}{p}} \right\}, \tag{8}$$

for which case the robust regression problem would be

$$Z_r = \min_{(\alpha,\bar{\boldsymbol{\beta}})\in\bar{\mathcal{D}}} \sup_{\mathbb{P}\in\mathcal{W}_p(r)} \left\{ \left(\mathbb{E}_{\mathbb{P}}\left[\ell^p(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}})\right]\right)^{\frac{1}{p}} \right\}. \tag{9}$$

Similarly, this robust optimization model can be motivated from the following finite sample guarantee associated with the Wasserstein metric,

$$\mathbb{P}^S\left[\left(\mathbb{E}_{\mathbb{P}^*}\left[\ell^p(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}})\right]\right)^{\frac{1}{p}} > \sup_{\mathbb{P}:\Delta_p(\mathbb{P},\hat{\mathbb{P}})\le r}\left(\mathbb{E}_{\mathbb{P}}\left[\ell^p(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}})\right]\right)^{\frac{1}{p}}\right] \le \mathbb{P}^S\left[\Delta_p(\mathbb{P}^*,\hat{\mathbb{P}})>r\right].$$

Unfortunately, tractability results for such models are rather limited. For the case of the ubiquitous ordinary least squares (OLS), we have $p=2$ and the underlying loss function is the absolute deviation $\ell(w)=|w|$. In this special case, Blanchet et al. (2019) show that the data-driven robust linear regression model with type-2 Wasserstein distance corresponds to a square-root regularization problem. Specifically, when the ambiguity set is $\mathcal{W}_2(r)$ and $\mathcal{Z}=\mathbb{R}^{N+1}$, we will obtain

$$\sup_{\mathbb{P}\in\mathcal{W}_2(r)}\left\{\mathbb{E}_{\mathbb{P}}\left[(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}})^2\right]^{1/2}\right\} = \left(\mathbb{E}_{\hat{\mathbb{P}}}\left[\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)^2\right]\right)^{1/2} + r\|\bar{\boldsymbol{\beta}}\|_*.$$

However, Problem (9) might not have a tractable reformulation for other types of loss function or when the support is bounded, such as having $\mathcal{Z}$ as a polytope. We are also not aware of other tractable results for other degree types. Hence, we propose the following type-p robust regression model

$$\begin{aligned} \bar{Z}_r = \min\ & \kappa r + \sup_{\mathbb{Q}\in\mathcal{F}(\hat{\mathbb{P}})}\left\{\left(\mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}})-\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]\right)^{\frac{1}{p}}\right\} \\ \text{s.t. }\ & (\alpha,\bar{\boldsymbol{\beta}})\in\bar{\mathcal{D}}, \kappa\ge 0 \end{aligned} \tag{10}$$

where the ambiguity set of the joint distribution of $(\tilde{\boldsymbol{z}},\tilde{\boldsymbol{v}})$ is given by

$$\mathcal{F}(\hat{\mathbb{P}}) := \left\{ \mathbb{Q}\in\mathcal{P}_0(\mathcal{Z}^2) \;\middle|\; (\tilde{\boldsymbol{z}},\tilde{\boldsymbol{v}})\sim\mathbb{Q}, \tilde{\boldsymbol{v}}\sim\hat{\mathbb{P}} \right\}. \tag{11}$$

The essential advantage of the type-p robust regression model is its computational tractability, which we will elucidate below.

THEOREM 3. *Problem* (10) *is equivalent to the following robust optimization problem with pth-order conic constraints,*

$$\begin{aligned}
\bar{Z}_r = \min\ & \kappa r + \tau \\
\text{s.t.}\ & \sup_{\boldsymbol{z}_s \in \mathcal{Z}} \left\{ \ell\left(\alpha + \boldsymbol{z}_s^\top \bar{\boldsymbol{\beta}}\right) - \kappa \|\boldsymbol{z}_s - \hat{\boldsymbol{z}}_s\| \right\} \leq \tau_s \ \forall s \in [S] \\
& \left(\frac{1}{S} \sum_{s \in [S]} \tau_s^p \right)^{\frac{1}{p}} \leq \tau \\
& \tau_s \in \mathbb{R} \qquad\qquad s \in [S] \\
& (\alpha, \bar{\boldsymbol{\beta}}) \in \bar{\mathcal{D}}, \kappa \geq 0, \tau \in \mathbb{R}.
\end{aligned} \tag{12}$$

In addition, the following explicit formulation holds.

1. *If* $\ell(w) = \max_{i \in [I]} \{a_i w + b_i\}$*, then Problem* (10) *is equivalent to:*

$$\begin{aligned}
\bar{Z}_r = \min\ & \kappa r + \tau \\
\text{s.t.}\ & a_i \alpha + \hat{\boldsymbol{z}}^\top \boldsymbol{\eta}_{i,s} + b_i + \max_{\boldsymbol{z}_s \in \bar{\mathcal{Z}}} \{ (a_i \bar{\boldsymbol{\beta}} - \boldsymbol{\eta}_{i,s})^\top \boldsymbol{z}_s \} \leq \tau_s \quad \forall s \in [S], i \in [I] \\
& \left(\frac{1}{S} \sum_{s \in [S]} \tau_s^p \right)^{\frac{1}{p}} \leq \tau \\
& \|\boldsymbol{\eta}_{i,s}\|_* \leq \kappa \qquad\qquad \forall s \in [S], i \in [I] \\
& \tau_s \in \mathbb{R}, \boldsymbol{\eta}_{i,s} \in \mathbb{R}^{N+1} \qquad\qquad \forall s \in [S], i \in [I] \\
& (\alpha, \bar{\boldsymbol{\beta}}) \in \bar{\mathcal{D}}, \kappa \geq 0, \tau \in \mathbb{R}.
\end{aligned}$$

2. *If* $\mathcal{Z} = \mathbb{R}^{N+1}$*, then Problem* (10) *is equivalent to:*

$$\begin{aligned}
\bar{Z}_r = \min\ & \left(\mathbb{E}_{\hat{\mathbb{P}}} \left[\ell^p \left(\alpha + \tilde{\boldsymbol{z}}^\top \bar{\boldsymbol{\beta}} \right) \right] \right)^{\frac{1}{p}} + L r \|\bar{\boldsymbol{\beta}}\|_* \\
\text{s.t.}\ & (\alpha, \bar{\boldsymbol{\beta}}) \in \bar{\mathcal{D}}.
\end{aligned}$$

When p is a rational number, it has well been known that pth-order conic constraints can well be reformulated with a modest number of second-order conic constraints (see Ben-Tal and Nemirovski 2020, for a great reference), and the type-p robust regression model can then be solved by state-of-the-art solvers such as CPLEX, Gurobi, and Mosek. Besides computational tractability, the type-p robust regression model also closely connects to regularization and finite sample probabilistic guarantees. Moreover, when there is support information available, the type-p robust regression model can provide a tighter bound of the corresponding robust regression problem (9) compared to the pth-root regularization problem.

THEOREM 4. *The following bounds hold for any* $(\alpha, \bar{\boldsymbol{\beta}}) \in \bar{\mathcal{D}}$

1.

$$\begin{aligned}
& \sup_{\mathbb{P} \in \mathcal{W}_p(r)} \left\{ \left(\mathbb{E}_{\mathbb{P}} \left[\ell^p (\alpha + \tilde{\boldsymbol{z}}^\top \bar{\boldsymbol{\beta}}) \right] \right)^{\frac{1}{p}} \right\} \\
\leq & \min_{\kappa \geq 0} \left\{ \kappa r + \sup_{\mathbb{Q} \in \mathcal{F}(\hat{\mathbb{P}})} \left\{ \left(\mathbb{E}_{\mathbb{Q}} \left[\left(\left(\ell(\alpha + \tilde{\boldsymbol{z}}^\top \bar{\boldsymbol{\beta}}) - \kappa \|\tilde{\boldsymbol{z}} - \tilde{\boldsymbol{v}}\| \right)^+ \right)^p \right] \right)^{\frac{1}{p}} \right\} \right\}.
\end{aligned}$$

In addition, the bound is tight in the case of $p=1$, and in the case of $p=2$ with $\mathcal{Z}=\mathbb{R}^{N+1}$ and $\ell(w)=L|w|$.

2.
$$\begin{aligned}&\min_{\kappa\geq 0}\left\{\kappa r+\sup_{\mathbb{Q}\in\mathcal{F}(\hat{\mathbb{P}})}\left\{\left(\mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}})-\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]\right)^{\frac{1}{p}}\right\}\right\}\\ &\leq\left(\mathbb{E}_{\hat{\mathbb{P}}}\left[\ell^p\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)\right]\right)^{\frac{1}{p}}+Lr\|\bar{\boldsymbol{\beta}}\|_*.\end{aligned}$$

In addition, the bound is tight when the support $\mathcal{Z}=\mathbb{R}^{N+1}$.

COROLLARY 1. *Let $\mathbb{P}^S$ denote the distribution that governs the distribution of the independent samples $\hat{\boldsymbol{z}}_1,\ldots,\hat{\boldsymbol{z}}_S$ drawn from $\mathbb{P}^*$ for which the empirical distribution $\hat{\mathbb{P}}$ is constructed. For any solution $(\alpha,\bar{\boldsymbol{\beta}})\in\mathcal{D}$ and $r\geq 0$,*

$$\begin{aligned}&\mathbb{P}^S\left[\left(\mathbb{E}_{\mathbb{P}^*}\left[\ell^p\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)\right]\right)^{\frac{1}{p}}>\left(\mathbb{E}_{\hat{\mathbb{P}}}\left[\ell^p\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)\right]\right)^{\frac{1}{p}}+Lr\|\bar{\boldsymbol{\beta}}\|_*\right]\\ \leq\ &\mathbb{P}^S\left[\left(\mathbb{E}_{\mathbb{P}^*}\left[\ell^p\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)\right]\right)^{\frac{1}{p}}>\min_{\kappa\geq 0}\left\{\kappa r+\sup_{\mathbb{Q}\in\mathcal{F}(\hat{\mathbb{P}})}\left\{\left(\mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}})-\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]\right)^{\frac{1}{p}}\right\}\right\}\right]\\ \leq\ &\mathbb{P}^S\left[\left(\mathbb{E}_{\mathbb{P}^*}\left[\ell^p(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}})\right]\right)^{\frac{1}{p}}>\sup_{\mathbb{P}:\Delta_p(\mathbb{P},\hat{\mathbb{P}})\leq r}\left(\mathbb{E}_{\mathbb{P}}\left[\ell^p(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}})\right]\right)^{\frac{1}{p}}\right]\\ \leq\ &\mathbb{P}^S\left[\Delta_p(\mathbb{P}^*,\hat{\mathbb{P}})>r\right].\end{aligned}$$

Relation to regularization

Due to the very limited tractability in reformulating the exact robust optimization model (9), one cannot easily establish a close relation between this robust regression model and regularization even the support set $\mathcal{Z}$ is unbounded. Nevertheless, the type-p robust regression model (10) is equivalent to a pth-root regularization problem whenever $\mathcal{Z}$ is unbounded.

From Theorem 4, when the support is unbounded, the type-p robust optimization takes the form

$$\bar{Z}_r=\min_{(\alpha,\bar{\boldsymbol{\beta}})\in\bar{\mathcal{D}}}\left(\mathbb{E}_{\hat{\mathbb{P}}}\left[\ell^p\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)\right]\right)^{\frac{1}{p}}+Lr\|\bar{\boldsymbol{\beta}}\|_*,\tag{13}$$

which is a pth-root regularization problem and its objective function can upper bound the actual objective value evaluated under the true distribution with at least $1-\mathbb{P}^S\left[\Delta_p(\mathbb{P}^*,\hat{\mathbb{P}})>r\right]$ probability.

The optimal solution to Problem (13) can be associated to an optimal solution to a typical regularization problem taking the following form,

$$R_\lambda=\min_{(\alpha,\bar{\boldsymbol{\beta}})\in\bar{\mathcal{D}}}\mathbb{E}_{\hat{\mathbb{P}}}\left[\ell^p\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)\right]+\lambda\|\bar{\boldsymbol{\beta}}\|_*,\tag{14}$$

for some $\lambda \geq 0$. To see this, let $(\alpha^*, \bar{\boldsymbol{\beta}}^*)$ be an optimum solution to Problem (13), and $T = \mathbb{E}_{\hat{\mathbb{P}}}\left[\ell^p\left(\alpha^* + \tilde{\boldsymbol{z}}^\top \bar{\boldsymbol{\beta}}^*\right)\right]$. Observe that $(\alpha^*, \bar{\boldsymbol{\beta}}^*)$ is also an optimal solution to the following optimization problem

$$\begin{aligned} C_T = \min\ & \|\bar{\boldsymbol{\beta}}\|_* \\ \text{s.t. } & \mathbb{E}_{\hat{\mathbb{P}}}\left[\ell^p\left(\alpha + \tilde{\boldsymbol{z}}^\top \bar{\boldsymbol{\beta}}\right)\right] \leq T \\ & (\alpha, \bar{\boldsymbol{\beta}}) \in \bar{\mathcal{D}}. \end{aligned} \tag{15}$$

Otherwise, it would contradict the optimality of $(\alpha^*, \bar{\boldsymbol{\beta}}^*)$ in Problem (13).

Proposition 1. *Consider the type-p robust optimization problem with unbounded support and that $\|\cdot\|_* = \|\cdot\|_d$ is an L^d-norm with $d \in [1, \infty]$. Suppose the Slater's conditions apply to Problem* (15)*, then if that $1 < C_T < C_\infty$, there exists $\lambda > 0$ such that the optimal solution of Problem* (15) *is also optimal to Problem* (14)*.*

Hence, the type-p robust optimization model with unbounded support is equivalent to the regularization model with the same loss function and $\|\cdot\|_*$ norm, for some choice of regularization parameter. For instance, let $\|\cdot\|_* = \|\cdot\|_d$, the type-2 robust optimization model is equivalent to LASSO when $d = 1$ and Ridge regression when $d = 2$. If the support is bounded, the type-p robust optimization model will provide a tighter bound than the regularization objective function while achieving the same finite sample guarantee.

3. Robust satisficing regression

We now consider the linear regression model under the robust satisficing framework recently introduced in Long et al. (2021) as follows

$$\begin{aligned} K_T = \min\ & \kappa \\ \text{s.t. } & \mathbb{E}_{\mathbb{P}}\left[\ell(\alpha + \tilde{\boldsymbol{z}}^\top \bar{\boldsymbol{\beta}})\right] \leq T + \kappa \Delta_1(\mathbb{P}, \hat{\mathbb{P}})\ \forall \mathbb{P} \in \mathcal{P}_0(\mathcal{Z}) \\ & (\alpha, \bar{\boldsymbol{\beta}}) \in \bar{\mathcal{D}}, \kappa \geq 0, \end{aligned} \tag{16}$$

In the spirit of satisficing, the model is specified by a parameter $T \geq Z_0$, which represents an acceptable target loss. The interpretation of Problem (16) is to minimize the expected violation of this target loss normalized by the type-1 Wasserstein distance. As before, the robust satisficing regression model (16) can also be converted to the following robust optimization problem:

$$\begin{aligned} K_T = \min\ & \kappa \\ \text{s.t. } & \frac{1}{S}\sum_{s \in [S]} \sup_{\boldsymbol{z}_s \in \mathcal{Z}} \left\{\ell\left(\alpha + \tilde{\boldsymbol{z}}^\top \bar{\boldsymbol{\beta}}\right) - \kappa \|\boldsymbol{z}_s - \hat{\boldsymbol{z}}_s\|\right\} \leq T \\ & (\alpha, \bar{\boldsymbol{\beta}}) \in \bar{\mathcal{D}}, \kappa \geq 0. \end{aligned}$$

Therefore, based on Theorem 2, Problem (16) has a similar tractable reformation as the robust optimization model.

PROPOSITION 2. *The robust satisficing regression model* (16) *has the following tractable reformulation.*

1. *If* $\ell(w) = \max_{i\in[I]}\{a_i w + b_i\}$*, then Problem* (16) *is equivalent to:*

$$\begin{aligned} K_T = \min\ & \kappa \\ \text{s.t. } & \frac{1}{S}\sum_{s\in[S]} \tau_s \le T \\ & a_i\alpha + \hat{\boldsymbol{z}}^\top \boldsymbol{\eta}_{i,s} + b_i + \max_{\boldsymbol{z}_s\in\mathcal{Z}}\{(a_i\bar{\boldsymbol{\beta}} - \boldsymbol{\eta}_{i,s})^\top \boldsymbol{z}_s\} \le \tau_s \quad \forall s\in[S], i\in[I] \\ & \|\boldsymbol{\eta}_{i,s}\|_* \le \kappa \quad \forall s\in[S], i\in[I] \\ & \tau_s \in \mathbb{R}, \boldsymbol{\eta}_{i,s} \in \mathbb{R}^{N+1} \quad \forall s\in[S], i\in[I] \\ & (\alpha,\bar{\boldsymbol{\beta}}) \in \bar{\mathcal{D}}, \kappa \ge 0. \end{aligned}$$

2. *If* $\mathcal{Z} = \mathbb{R}^{N+1}$*, then Problem* (16) *is equivalent to:*

$$\begin{aligned} K_T = \min\ & L\|\bar{\boldsymbol{\beta}}\|_* \\ \text{s.t. } & \frac{1}{S}\sum_{s\in[S]} \ell(\alpha + \hat{\boldsymbol{z}}_s^\top \bar{\boldsymbol{\beta}}) \le T \\ & (\alpha,\bar{\boldsymbol{\beta}}) \in \bar{\mathcal{D}}. \end{aligned}$$

Moreover, the problems are feasible for any $T \ge Z_0$.

Type-p robust satisficing regression model

For a loss function that is not Lipschitz continuous, we would not necessary obtain tractable reformulations under the existing robust satisficing paradigm that we have discussed. To address the more general degree-p loss function of Problem (8), we propose the type-p robust satisficing regression model as follows:

$$\begin{aligned} K_T = \min\ & \kappa \\ \text{s.t. } & \mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell\left(\alpha + \tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right) - \kappa\|\tilde{\boldsymbol{z}} - \tilde{\boldsymbol{v}}\|\right)^+\right)^p\right] \le T^p \quad \forall \mathbb{Q}\in\mathcal{F}(\hat{\mathbb{P}}) \\ & (\alpha,\bar{\boldsymbol{\beta}}) \in \bar{\mathcal{D}}, \kappa \ge 0, \end{aligned} \tag{17}$$

where the joint ambiguity set is given by (11).

Intuitively, the robust satisficing model allows the objective function $(\mathbb{E}_{\mathbb{P}}\left[\ell^p\left(\alpha + \tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)\right])^{1/p}$ evaluated on the distribution, $\mathbb{P}$, $\tilde{\boldsymbol{z}} \sim \mathbb{P}$, to exceed the target loss T if $\mathbb{P}$ should deviate from $\hat{\mathbb{P}}$. As κ decreases, the objective function evaluated under any distribution would exceed the target by a lesser magnitude. We will formalize this with a statistical guarantee in Theorem 6 shortly. The difference from the previous robust satisficing regression model (16) is that the allowable violation of the target loss is no longer directly controlled by the Wasserstein distance. Instead, we directly "compensate" the loss function before raising the power so that they are on the same scale. As a concrete example, when $\ell(w) = |w|$ and $p = 2$, Problem (17) would first adjust the absolute value of the residual before calculating the expected squared (adjusted) error.

THEOREM 5. *Problem* (17) *is equivalent to the following robust optimization problem with pth-order conic constraints,*

$$\begin{aligned} K_T = \min\ & \kappa \\ \text{s.t. } & \frac{1}{S}\sum_{s\in[S]} \tau_s^p \le T^p \\ & \sup_{\boldsymbol{z}_s\in\mathcal{Z}} \left\{\ell\left(\alpha + \boldsymbol{z}_s^\top \bar{\boldsymbol{\beta}}\right) - \kappa\|\boldsymbol{z}_s - \hat{\boldsymbol{z}}_s\|\right\} \le \tau_s \quad \forall s\in[S] \\ & (\alpha,\bar{\boldsymbol{\beta}})\in\bar{\mathcal{D}}, \boldsymbol{\tau}\in\mathbb{R}^S, \kappa\ge 0. \end{aligned} \tag{18}$$

In addition, the following explicit formulation holds.

1. *If* $\ell(w) = \max_{i\in[I]}\{a_i w + b_i\}$, *then Problem* (17) *is equivalent to:*

$$\begin{aligned} K_T = \min\ & \kappa \\ \text{s.t. } & \frac{1}{S}\sum_{s\in[S]} \tau_s^p \le T^p \\ & a_i\alpha + \hat{\boldsymbol{z}}^\top\boldsymbol{\eta}_{i,s} + b_i + \max_{\boldsymbol{z}_s\in\mathcal{Z}}\{(a_i\bar{\boldsymbol{\beta}} - \boldsymbol{\eta}_{i,s})^\top \boldsymbol{z}_s\} \le \tau_s \quad \forall s\in[S], i\in[I] \\ & \|\boldsymbol{\eta}_{i,s}\|_* \le \kappa \quad \forall s\in[S], i\in[I] \\ & \tau_s\in\mathbb{R}, \boldsymbol{\eta}_{i,s}\in\mathbb{R}^{N+1} \quad \forall s\in[S], i\in[I] \\ & (\alpha,\bar{\boldsymbol{\beta}})\in\bar{\mathcal{D}}, \kappa\ge 0. \end{aligned}$$

2. *If* $\mathcal{Z} = \mathbb{R}^{N+1}$, *then Problem* (17) *is equivalent to:*

$$\begin{aligned} K_T = \min\ & L\|\bar{\boldsymbol{\beta}}\|_* \\ \text{s.t. } & \frac{1}{S}\sum_{s\in[S]} \ell^p(\alpha + \hat{\boldsymbol{z}}_s^\top\bar{\boldsymbol{\beta}}) \le T^p \\ & (\alpha,\bar{\boldsymbol{\beta}})\in\bar{\mathcal{D}}. \end{aligned} \tag{19}$$

Moreover, the problems are feasible for any $T \ge Z_0$.

Observe from Theorem 5, the type-1 robust satisficing regression problem recovers Problem (16). Moreover, from Proposition 1, when the support is unbounded, the optimum solution of the type-p robust satisficing problem can also be associated with the optimal solution of the regularization problem (14), for some choice of parameter $\lambda \ge 0$.

Finite sample probabilistic guarantees

Apart from the computational tractable model for all $p \ge 1$, we can relate the robust satisficing model directly to the finite sample probabilistic guarantees of Fournier and Guillin (2015) as follows:

THEOREM 6. *Let* $\mathbb{P}^S$ *denote the distribution that governs the distribution of the independent samples* $\hat{\boldsymbol{z}}_1,\ldots,\hat{\boldsymbol{z}}_S$ *drawn from* $\mathbb{P}^*$ *for which the empirical distribution* $\hat{\mathbb{P}}$ *is constructed. Suppose the solution* $\alpha,\bar{\boldsymbol{\beta}},\kappa$ *is feasible in Problem* (17), *then*

$$\mathbb{P}^S\left[\mathbb{E}_{\mathbb{P}^*}\left[\ell^p\left(\alpha + \tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)\right]^{\frac{1}{p}} > T + \kappa r\right] \le \mathbb{P}^S\left[\Delta_p(\mathbb{P}^*,\hat{\mathbb{P}}) > r\right],$$

for all $r \geq 0$.

Therefore, the solution of the robust satisficing regression problem ensures high confidence that the actual objective function of the regression problem evaluated on the true distribution is within $\kappa r + T$, such that the probability of exceeding the confidence range decreases to zero at an exponential rate in r, accordingly to the measure concentration result of Fournier and Guillin (2015). More importantly, regardless of the tightness of such concentration result, the highest level of robustness is consistent with having the lowest possible κ, which is what the type-p robust satisficing regression model aims to minimize. As in the case of regularization, a good choice of the target parameter $T \geq Z_0$ could vastly improve the out-of-sample performance. In practice, one can also calibrate this hyper-parameter via cross-validation. In our numerical study based on popular datasets, we illustrate how to prescribe this target parameter.

4. Binary classification

We now propose robust models for binary classification. The response variable associated with a binary classification problem has two values representing two different groups; each group, $i \in [2]$ is associated with a set of S_i input samples training represented by $\hat{\boldsymbol{z}}_{i,s},\ s \in [S_i]$. In a binary classification model, we often assume that the prior probability of each group is q_i, $i \in [2]$ is reflected by their frequency of occurrence in the data, *i.e.*, $q_i = S_i/(S_1+S_2)$. However, it may well be possible in applications of classification to specify a different set of prior probabilities. Hence, we provide this flexibility by introducing an ambiguous set of prior probabilities,

$$\mathcal{Q} = \left\{ (q_1, 1-q_1) \in [0,1]^2 \ \mid \ q_1 \in [\underline{q}_1, \bar{q}_1] \right\},$$

which can be a singleton. Indeed, without further information to update the prior, it would be reasonable to have $\mathcal{Q} = \left\{ \left(\frac{S_1}{S_1+S_2}, \frac{S_2}{S_1+S_2} \right) \right\}$.

Unlike regression problems, the response variable is not subjected to uncertainty within its identified group. Hence, we do not need to include the response variable as part of the data that is subject to uncertainty. The goal of a binary classification model is to determine the best separating hyperplane such that

$$\alpha + \boldsymbol{z}_{1,s}^\top \boldsymbol{\beta} \leq 0 \qquad \forall s \in [S_1]$$

and

$$\alpha + \boldsymbol{z}_{2,s}^\top \boldsymbol{\beta} > 0 \qquad \forall s \in [S_2],$$

for some parameters $(\alpha, \boldsymbol{\beta}) \in \mathcal{D}$. This can be achieved through minimizing the loss of misclassification, defined through a one sided loss function ℓ such as

1. Hinge-loss: $\ell(w) = (1-w)^+$.

2. Logexp loss: $\ell(w) = \log(1+\exp(-w))$.

In particular, the hinge-loss function would lead to the linear SVM model, while the logexp-loss would be associated with the logistic regression model. As before, we consider Lipschitz continuous loss function, which we will subsequently extend to the pth degree.

The binary classification model that minimizes the empirical loss and is robust to the uncertain priors $\boldsymbol{q} \in \mathcal{Q}$ can be written as:

$$\begin{aligned} Z_0 = \min \max_{\boldsymbol{q}\in\mathcal{Q}} & \left\{ \sum_{i\in[2]} q_i \mathbb{E}_{\hat{\mathbb{P}}_i} \left[\ell(b_i(\alpha + \tilde{\boldsymbol{z}}_i^\top \boldsymbol{\beta}) \right] \right\} \\ \text{s.t. } & (\alpha, \boldsymbol{\beta}) \in \mathcal{D}, \end{aligned} \tag{20}$$

where $b_1 = -1$ and $b_2 = 1$, and we use $\hat{\mathbb{P}}_i \in \mathcal{P}_0(\mathcal{Z}_i), \tilde{\boldsymbol{z}}_i \sim \hat{\mathbb{P}}$ to denote the empirical distribution of $\tilde{\boldsymbol{z}}_i$, such that $\hat{\mathbb{P}}_i[\tilde{\boldsymbol{z}}_i = \hat{\boldsymbol{z}}_{i,s}] = 1/S_i$, for all $s \in [S_i]$. This new formulation takes into account of the uncertainty in the prior probabilities and it could also be applied to the multinominial logistic regression. Indeed, we can formulate the negative log-likelihood based classification problem for K groups as follows,

$$\begin{aligned} Z_0 = \min \max_{\boldsymbol{q}\in\mathcal{Q}} & \left\{ \sum_{i\in[K]} q_i \mathbb{E}_{\hat{\mathbb{P}}_i} \left[\log \left(\sum_{j\in[K]} \exp\left(\alpha_j - \alpha_i + \tilde{\boldsymbol{z}}^\top (\boldsymbol{\beta}_j - \boldsymbol{\beta}_i)\right) \right) \right] \right\} \\ \text{s.t. } & (\alpha_j, \boldsymbol{\beta}_j)_{j\in[K-1]} \in \mathcal{D}, \\ & \alpha_K = 0,\ \boldsymbol{\beta}_K = \mathbf{0}. \end{aligned}$$

Here, we set the last choice as the benchmark.

As a generalization of the loss function to the pth degree, which includes the squared hinge loss, we consider the following classification problem with uncertain priors as follows

$$Z_0 = \min_{(\alpha,\boldsymbol{\beta})\in\mathcal{D}} \max_{\boldsymbol{q}\in\mathcal{Q}} \left\{ \left(\sum_{i\in[2]} q_i \mathbb{E}_{\hat{\mathbb{P}}_i} \left[\ell^p \left(b_i \left(\alpha + \tilde{\boldsymbol{z}}_i^\top \boldsymbol{\beta} \right) \right) \right] \right)^{\frac{1}{p}} \right\}. \tag{21}$$

Type-p robust classification model

We now consider the robust classification model as follows:

$$\begin{aligned} Z_r = \min \max_{\boldsymbol{q}\in\mathcal{Q}} & \left\{ \left(\sum_{i\in[2]} q_i \sup_{\mathbb{P}_i \in \mathcal{U}_{i,p}(r)} \mathbb{E}_{\mathbb{P}_i} \left[\ell^p (b_i(\alpha + \tilde{\boldsymbol{z}}_i^\top \boldsymbol{\beta}) \right] \right)^{\frac{1}{p}} \right\} \\ \text{s.t. } & (\alpha, \boldsymbol{\beta}) \in \mathcal{D}, \end{aligned} \tag{22}$$

where

$$\mathcal{U}_{i,p}(r) := \left\{ \mathbb{P}_i \in \mathcal{P}_0(\mathcal{Z}_i) \;\middle|\; \begin{array}{l} \tilde{\boldsymbol{z}}_i \sim \mathbb{P}_i \\ \Delta_p(\mathbb{P}_i, \hat{\mathbb{P}}_i) \le r \end{array} \right\}, \qquad i \in [2].$$

Note that we provide a different approach of formulating the robust classification models from those discussed in the literature, such as Shafieezadeh-Abadeh et al. (2019), Gao et al. (2017), among others. Apart from introducing the ambiguity set for group probabilities, we also have separate support sets associated with each classification group. We believe that this an important information that can be elicited from data. Specifically, for the ith group, $i \in [2]$, the set $\mathcal{Z}_i \subseteq \mathbb{R}^N$ represents the support of the corresponding input variables, which is a convex set and can be bounded or unbounded.

THEOREM 7. *When $p=1$, Problem* (22) *have the following equivalent representation:*

$$
\begin{aligned}
Z_r = \min\ & u_1 - u_2 \underline{q}_1 + u_3 \bar{q}_1 \\
\text{s.t.}\ & u_1 - u_2 + u_3 \geq \kappa_1 r + v_1 \\
& u_1 \geq \kappa_2 r + v_2 \\
& v_i \geq \frac{1}{S_i} \sum_{s\in[S_i]} \tau_{i,s}. && \forall i \in [2] \\
& \tau_{i,s} \geq \sup_{\boldsymbol{z}_{i,s}\in\mathcal{Z}_i} \{\ell(b_i(\alpha + \boldsymbol{z}_i^\top \boldsymbol{\beta}) - \kappa_i \|\boldsymbol{z}_{i,s} - \hat{\boldsymbol{z}}_{i,s}\|\} && \forall i \in [2], s \in [S_i] \\
& \tau_{i,s} \in \mathbb{R} && \forall i \in [2], s \in [S_i] \\
& (\alpha, \boldsymbol{\beta}) \in \mathcal{D}, \kappa_1 \geq 0, \kappa_2 \geq 0, v_1 \in \mathbb{R}, v_2 \in \mathbb{R}, u_1 \in \mathbb{R}, u_2 \geq 0, u_3 \geq 0.
\end{aligned}
$$

When $\ell(w) = \max_{i\in[I]}\{a_i w + b_i\}$ or $\mathcal{Z} = \mathbb{R}^{N+1}$, the explicit formulations would follow directly from Theorem 2 and Theorem 7. For brevity, we do not present the explicit formulations in the rest of the paper.

However, similarly as the regression setting, the generalization to above pth degree robust classification model is not computationally tractable in general when $p > 1$. For computational tractability, we propose the following type-p robust classification model,

$$
\begin{aligned}
\bar{Z}_r = \min \max_{\boldsymbol{q}\in\mathcal{Q}} \Bigg\{ & (\kappa_1 q_1 + \kappa_2 q_2) r + \left(\sum_{i\in[2]} q_i \sup_{\mathbb{Q}_i \in \mathcal{F}_i(\hat{\mathbb{P}}_i)} \mathbb{E}_{\mathbb{Q}_i} \left[\left(\left(\ell(b_i(\alpha + \tilde{\boldsymbol{z}}^\top \boldsymbol{\beta})) - \kappa_i \|\tilde{\boldsymbol{z}} - \tilde{\boldsymbol{v}}\| \right)^+ \right)^p \right] \right)^{\frac{1}{p}} \Bigg\} \\
\text{s.t.}\ & (\alpha, \boldsymbol{\beta}) \in \mathcal{D}, \kappa_1 \geq 0, \kappa_2 \geq 0,
\end{aligned}
\tag{23}
$$

where the joint ambiguity set is given by

$$
\mathcal{F}_i(\hat{\mathbb{P}}_i) := \left\{ \mathbb{Q}_i \in \mathcal{P}_0(\mathcal{Z}_i^2) \ \middle|\ (\tilde{\boldsymbol{z}}_i, \tilde{\boldsymbol{v}}_i) \sim \mathbb{Q}_i, \tilde{\boldsymbol{v}}_i \sim \hat{\mathbb{P}}_i \right\}, \quad i \in [2].
$$

THEOREM 8. *The following bounds hold:*

1. *$Z_r \leq \bar{Z}_r$. In addition, the bound is tight in the case of $p=1$.*

2.
$$\bar{Z}_r \leq \min_{(\alpha,\boldsymbol{\beta})\in\mathcal{D}} \left\{ \max_{\boldsymbol{q}\in\mathcal{Q}} \left\{ \left(\sum_{i\in[2]} q_i \mathbb{E}_{\hat{\mathbb{P}}_i} \left[\ell^p \left(b_i(\alpha + \tilde{\boldsymbol{z}}^\top \boldsymbol{\beta}) \right) \right] \right)^{\frac{1}{p}} \right\} + Lr\|\boldsymbol{\beta}\|_* \right\}.$$

In addition, the bound is tight when the support $\mathcal{Z} = \mathbb{R}^{N+1}$.

We can draw the same conclusion as in the case of regression. The type-p robust classification model for unbounded support is equivalent to the regularization model with the same loss function and $\|\cdot\|_*$ norm.

THEOREM 9. *For* $p > 1$, *Problem* (23) *is equivalent to the following robust optimization problem with p-order conic constraints,*

$$\begin{aligned}
\bar{Z}_r = \min\ & p^{\frac{p}{1-p}}(p-1)u_1 + u_2 - u_3\underline{q}_1 + u_4\bar{q}_1 \\
\text{s.t.}\ & v_i \geq \left(\frac{1}{S_i} \sum_{s\in[S_i]} \tau_{i,s}^p \right)^{\frac{1}{p}} && \forall i \in [2] \\
& \tau_{i,s} \geq \sup_{\boldsymbol{z}_{i,s}\in\mathcal{Z}_i} \{ \ell(b_i(\alpha + \boldsymbol{z}_i^\top \boldsymbol{\beta}) - \kappa_i \|\boldsymbol{z}_{i,s} - \hat{\boldsymbol{z}}_{i,s}\| \} && \forall i \in [2], s \in [S_i] \\
& u_1(v_1/u_1)^p \leq u_2 - u_3 + u_4 - \kappa_1 r \\
& u_1(v_2/u_1)^p \leq u_2 - \kappa_2 r \\
& \tau_{i,s} \in \mathbb{R} && \forall i \in [2], s \in [S_i] \\
& u_1 \geq 0, u_2 \in \mathbb{R}, u_3 \geq 0, u_4 \geq 0, \\
& (\alpha, \boldsymbol{\beta}) \in \mathcal{D}, \kappa_1 \geq 0, \kappa_2 \geq 0, v_1 \geq 0, v_2 \geq 0.
\end{aligned}$$

Type-p robust satisficing classification model

Finally, we also propose the type-p robust satisficing classification model as follows:

$$\begin{aligned}
K_T = \min\ & \kappa \\
\text{s.t.}\ & \sum_{i\in[2]} q_i \mathbb{E}_{\mathbb{Q}_i} \left[\left(\left(\ell \left(b_i \left(\alpha + \tilde{\boldsymbol{z}}_i^\top \boldsymbol{\beta} \right) \right) - \kappa \|\tilde{\boldsymbol{z}}_i - \tilde{\boldsymbol{v}}_i\| \right)^+ \right)^p \right] \leq T^p \\
& \qquad \forall \boldsymbol{q} \in \mathcal{Q}, \mathbb{Q}_i \in \mathcal{F}_i(\hat{\mathbb{P}}_i), i \in [2] \\
& (\alpha, \boldsymbol{\beta}) \in \mathcal{D}, \kappa \geq 0,
\end{aligned} \tag{24}$$

where the joint ambiguity set is given by

$$\mathcal{F}_i(\hat{\mathbb{P}}_i) := \left\{ \mathbb{Q}_i \in \mathcal{P}_0(\mathcal{Z}_i^2) \;\middle|\; (\tilde{\boldsymbol{z}}_i, \tilde{\boldsymbol{v}}_i) \sim \mathbb{Q}_i, \tilde{\boldsymbol{v}}_i \sim \hat{\mathbb{P}}_i \right\}, \quad i \in [2].$$

THEOREM 10. *Problem* (24) *has the following equivalent robust optimization representation with pth-order conic constraints,*

$$\begin{aligned} K_T = \min\ & \kappa \\ \text{s.t. } & \frac{\underline{q}_1}{S_1}\sum_{s\in[S_1]}\tau_{1,s}^p + \frac{1-\underline{q}_1}{S_2}\sum_{s\in[S_2]}\tau_{2,s}^p \le T^p \\ & \frac{\bar{q}_1}{S_1}\sum_{s\in[S_1]}\tau_{1,s}^p + \frac{1-\bar{q}_1}{S_2}\sum_{s\in[S_2]}\tau_{2,s}^p \le T^p \\ & \tau_{i,s} \ge \sup_{\boldsymbol{z}_{i,s}\in\mathcal{Z}_i}\{\ell(b_i(\alpha+\boldsymbol{z}_i^\top\boldsymbol{\beta})-\kappa\|\boldsymbol{z}_{i,s}-\hat{\boldsymbol{z}}_{i,s}\|\} \ \forall i\in[2], s\in[S_i] \\ & \tau_{i,s}\in\mathbb{R} \qquad\qquad \forall i\in[2], s\in[S_i] \\ & (\alpha,\boldsymbol{\beta})\in\mathcal{D}, \kappa\ge 0. \end{aligned}$$

In addition, when $\mathcal{Z}_i = \mathbb{R}^N$*, then Problem* (24) *is equivalent to:*

$$\begin{aligned} K_\rho = \min\ & L\|\boldsymbol{\beta}\|_* \\ \text{s.t. } & \frac{\underline{q}_1}{S_1}\sum_{s\in[S_1]}\tau_{1,s}^p + \frac{1-\underline{q}_1}{S_2}\sum_{s\in[S_2]}\tau_{2,s}^p \le T^p \\ & \frac{\bar{q}_1}{S_1}\sum_{s\in[S_1]}\tau_{1,s}^p + \frac{1-\bar{q}_1}{S_2}\sum_{s\in[S_2]}\tau_{2,s}^p \le T^p \\ & \tau_{i,s} \ge \ell(b_i(\alpha+\hat{\boldsymbol{z}}_i^\top\boldsymbol{\beta}) \qquad \forall i\in[2], s\in[S_i] \\ & \tau_{i,s}\in\mathbb{R} \qquad\qquad \forall i\in[2], s\in[S_i] \\ & (\alpha,\boldsymbol{\beta})\in\mathcal{D}. \end{aligned}$$

THEOREM 11. *Let* $\mathbb{P}^S$ *denote the distribution that governs the distribution of the independent samples* $\hat{\boldsymbol{z}}_1,\ldots,\hat{\boldsymbol{z}}_S$ *drawn from* $\mathbb{P}^*$ *for which the empirical distributions* $\hat{\mathbb{P}}_1$ *and* $\hat{\mathbb{P}}_2$ *are constructed. Suppose the true proportion of group 1,* q_1^**, satisfies* $q_1^*\in[\underline{q}_1,\bar{q}_1]$ *and the solution* $\alpha,\boldsymbol{\beta},\kappa$ *is feasible in Problem* (24)*, then*

$$\mathbb{P}^S\left[\sum_{i\in[2]} q_i^*\left(\mathbb{E}_{\mathbb{P}_i^*}\left[\ell^p\left(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta})\right)\right]^{\frac{1}{p}}\right) > T+\kappa r\right] \le \sum_{i\in[2]}\mathbb{P}^S\left[\Delta_p(\mathbb{P}_i^*,\hat{\mathbb{P}}_i) > r\right],$$

for all $r\ge 0$.

Hence, analogous to the type-p robust satisficing regression model, the solution of the type-p robust satisficing classification model ensures high confidence that the actual objective function of classification problem evaluated on the true distribution is within $T+\kappa r$, such that the probability of exceeding the confidence range decreases to zero in r.

5. Computational studies

In this section, we conduct experiments to compare our robust satisficing models with LASSO (ℓ_1-norm regularization) and Ridge (ℓ_2-norm regularization). We choose square loss for regression

and squared hinge loss for binary classification. For robust satisficing methods, we set $p=2$, choose the dual norm as ℓ_1- or ℓ_2-norm, and let the support set be unbounded.

Our experiments are based on about twenty common public datasets used for regression and classification from UCI Machine Learning repository (Dua and Graff 2017) including forest fire, diabetes, and wine quality datasets. For some datasets that only contain a small number of features (*e.g.,* airfoil noise and fish toxicity datasets), we also apply a second-order polynomial transformation of the original features and generate more features.

Cross-validation

For all models, we use cross-validation to choose the parameters. We use ten folds for regression and five folds for classification. For the robust satisficing regression models, we propose the following way to select the candidates for cross-validation. Recall that the type-p robust satisficing regression model with unbounded support can be written as:

$$\begin{aligned} K_T = \min\ & L\|\bar{\boldsymbol{\beta}}\|_* \\ \text{s.t.}\ & \frac{1}{S}\sum_{s\in[S]} \ell^p(\alpha + \hat{\boldsymbol{z}}_s^\top \bar{\boldsymbol{\beta}}) \le T^p \\ & (\alpha, \bar{\boldsymbol{\beta}}) \in \bar{\mathcal{D}}, \end{aligned}$$

for some target parameter T. A simple way to define this parameter is to let $T = \rho Z_0$ and introduce the normalized target parameter $\rho \ge 1$. It could be interpreted as a proportionally inflated empirical loss.

As suggested by Proposition 1 and the discussion following Theorem 5, the robust satisficing model with unbounded support has the same family of solutions as the corresponding regularization model in linear regression. However, they often perform differently even after an identical cross-validation procedure. The main reason is that the penalty candidates are fixed across different validations in regularization models while the targets candidates are adaptive via Z_0 in robust satisficing models. Moreover, because the normalized target parameter is interpretable for robust satisficing models, it is easy to incorporate prior knowledge to exclude problematic candidates. This would also contribute to the model performance of the robust satisficing models. In all experiments, we use $\rho \in [1,2]$.

Evaluation procedure

We randomly split the data into training and testing data for every dataset. The training data size is determined by the number of features, N. Specifically, we let the size of training data be $S = sN$, for $s \in \{2,3,5\}$. For brevity, we only present the results for the instances with $s=3$, and other results convey a similar message. This corresponds to a small to medium data setting where training data is relatively limited. We use the cross-validation procedure for all models on

the training data to choose the parameters, and then we solve the models again using the entire training data with the selected parameters. Finally, we test the out-of-sample performance of the solution on the testing data. We repeat the above procedures 50 times and evaluate the average performance over different initial train-test data splits. For classification tasks, the splitting of training and testing data is done in a stratified fashion so that the proportion of samples of each label in both training and testing datasets resembles the proportion in the original dataset.

Regression

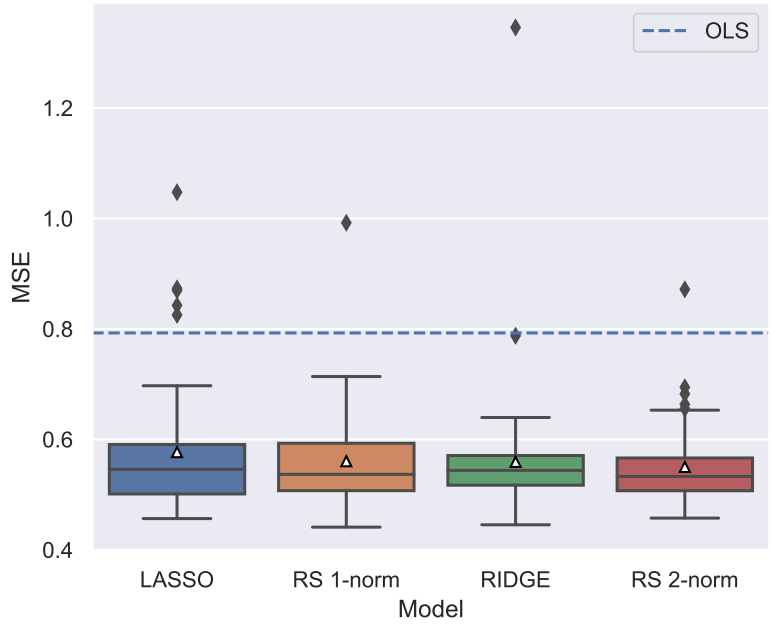

Figure 1 **Out-of-sample performance comparison: Red wine quality ($s = 3$).**

We first look at the computational results of the popular red wine quality dataset. The performance comparison is summarized in Figure 1 which contains the box plots of the out-of-sample mean squared error (MSE) of the models. The triangle marks the average MSE. We include the average MSE of the ordinary least-squares regression model as a reference, marked as the dashed horizontal line. In this experiment, robust satisficing (RS) with ℓ_1-norm achieves a lower median and mean of the MSE than the LASSO regression; similarly, RS with ℓ_2-norm outperforms the Ridge regression. On average, RS 1-norm and 2-norm models lead to 2.9% and 1.7% improvements in MSE compared to LASSO and Ridge regression, respectively. The box plots also demonstrate that the RS models could achieve a better overall distribution of the out-of-sample MSE (over

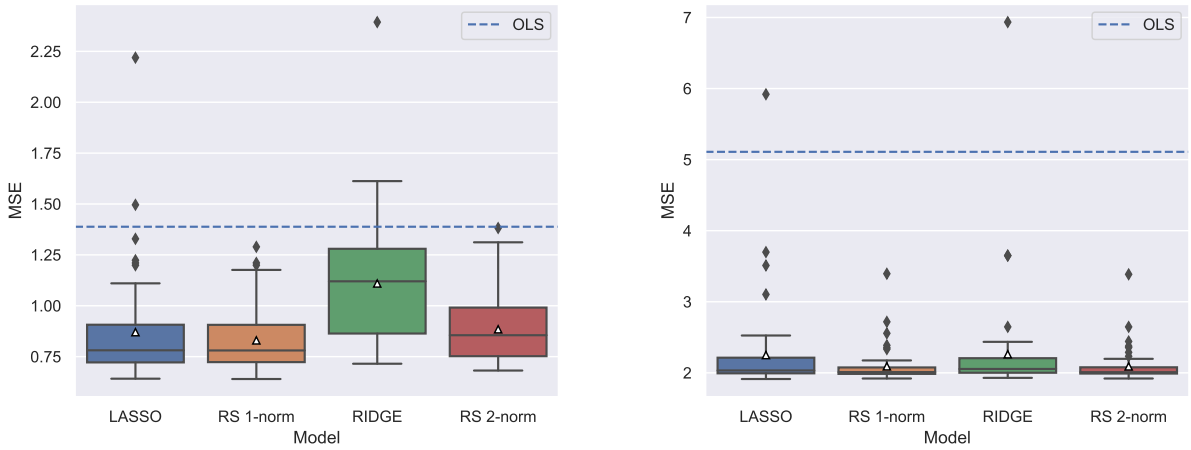

Figure 2 **Out-of-sample performance comparison: QSAR bioconcentration (left) and forest fire (right).**

different random splits of training and testing data) in terms of spread and other quantiles. The results are similar for the white wine dataset, which is relegated to Appendix B (Figure 7).

In Figure 2, we provide an example where the RS models could outperform the benchmark regularization models by a more considerable margin than the previous experiment. The observations from the box plots are consistent with the previous instance, *i.e.,* the RS models produce better distributions of the out-of-sample MSE over random data splits than the benchmark models. In the left subfigure, LASSO and Ridge regression have 4.9% and 25.4% higher mean MSE compared to the RS models. In the right subfigure, the percentage gaps with respect to the RS models are 7.6% and 8.2%. Improvements are observed in most of our testing datasets. We include three more examples on the concrete dataset, Auto MPG dataset, fish toxity dataset, and diabetes dataset in Appendix B (Figure 8–Figure 9), illustrating the potentiality of our approach.

From the above result, we elucidate that the RS models can outperform the classical LASSO and Ridge regression models in many instances. Nevertheless, we do not claim that the RS models always outperform benchmarks. There is also a minority of the tested instances where our model does not outperform the benchmark models. Here, we provide an example on a power consumption dataset (Figure 3). In Figure 3, LASSO regression has a 4.1% lower mean MSE than the RS model, while Ridge regression has a 1.4% higher mean MSE than the RS model. In our experiments, we do not observe any instances where benchmark models significantly dominate the RS model. Hence, we believe that our RS model provides a competitive alternative to regularization models.

In the above discussion, we illustrate the potentiality of the RS models when implemented with cross-validation. Because of the interpretability of the normalized target parameter ρ, it is possible to specify it using prior knowledge and circumvent the need of cross-validation. Next, we fix ρ

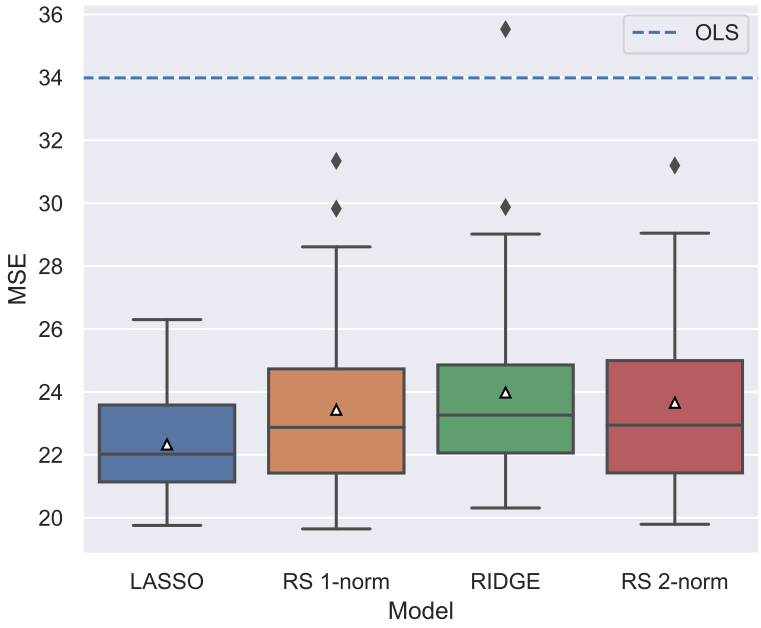

Figure 3 **Out-of-sample performance comparison: Power consumption ($s = 3$).**

for the training data and evaluate the solutions on the test. Figure 4 contains the out-of-sample average MSE with respect to ρ for RS models on twelve real-world datasets. The performance of the RS models is quite stable over the range of $[1, 2]$ and the RS models can often outperform OLS which is set to be 1. For the twelve selected datasets in Figure 4, one would consistently get a nice solution when $\rho \in [1, 1.5]$. From the management perspective, this observation indicates that the manager could prescribe the target parameter based on an acceptable level of loss without paying much cost in performance.

Results for binary classification

Let us first look at the experiments with banknote authentication dataset and haberman's survival dataset. We compare the test accuracy across models in Figure 5 using box plots. A higher value indicates a better result. In this experiment, we note that the RS models lead to higher mean and median accuracy than the benchmark models. In the left subfigure, the RS models outperform LASSO by 4.3% and Ridge by 1.0%. In the right subfigure, the test accuracy of the RS models is higher than that of benchmark models by 0.9% and 2.4%, respectively.

Similar to the experiments with regression tasks, we also observe cases where the RS models do not outperform the benchmark models. As an example, we summarize the performance on the Pima Indians diabetes dataset in Figure 6. The RS models lead to slightly lower test accuracy than

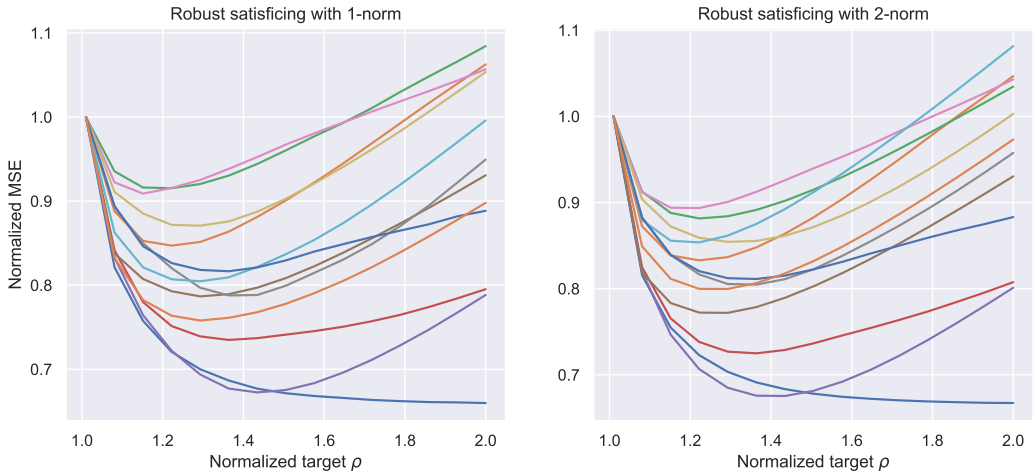

Figure 4 **Normalized average out-of-sample performance w.r.t. model parameters ($s = 3$).**

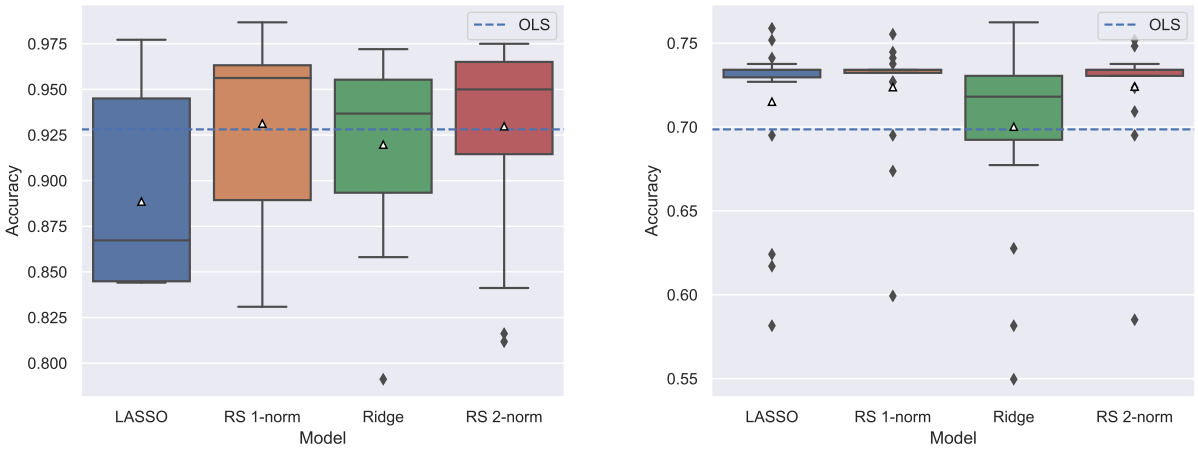

Figure 5 **Out-of-sample performance comparison: Banknote authentication (left) and Haberman's survival (right).**

benchmark models in this instance. Overall, the RS model is not dominated by benchmark models in our experiments.

6. Conclusion

In this paper, we provide tractable approximations for robust regression and classification problems when loss functions are derived from Lipschitz continuous functions raised to the power of p. We present a tractable type-p robust optimization approach and establishes its connections to regularization in regression and classification. We also propose a new perspective to regularization in supervised learning via the lens of robust satisficing. The robust satisficing framework leverages

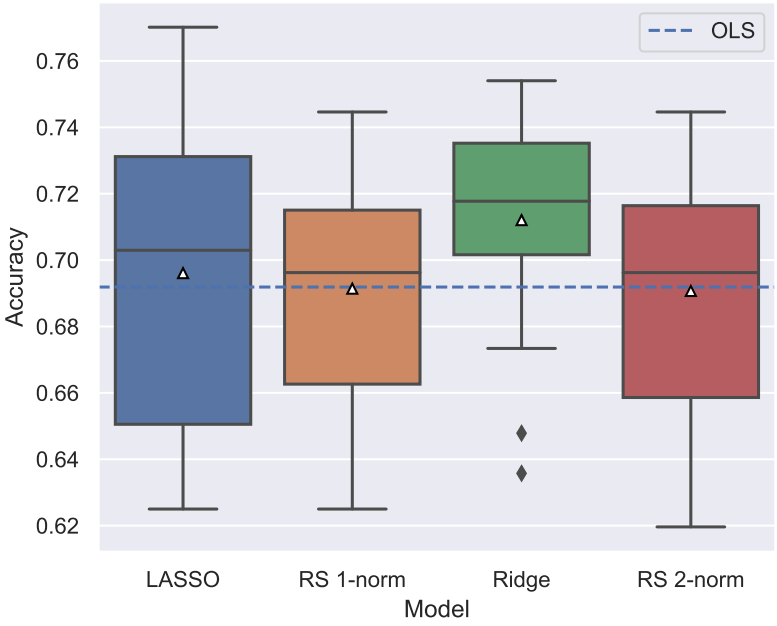

Figure 6 **Out-of-sample performance comparison: Pima Indians diabetes ($s = 3$).**

a novel model parameter that specifies the target objective relative to the optimal objective of the nominal model. It provides a new understanding of regularization in supervised learning via the trade-off between the acceptable level of in-sample loss and robustness in out-of-sample performance. This differentiates our work from previous studies on regularization via robust optimization. We introduce a type-p robust satisficing model that is closely related to a pth-root regularization model. We establish the tractability results and a finite sample guarantee. When the underlying support is unbounded, the family of solutions of the type-p robust satisficing model coincides with the pth-root regularization model. Nevertheless, from the empirical experiments, we elucidate that the target-based hyper-parameter of our robust satisficing model can be easier to determine via cross-validation in many instances compared to standard regularization approaches.

References

Ben-Tal, Aharon, Stephen Boyd, Arkadi Nemirovski. 2006. Extending scope of robust optimization: Comprehensive robust counterparts of uncertain problems. *Mathematical Programming* **107**(1) 63–89.

Ben-Tal, Aharon, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, Gijs Rennen. 2013. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* **59**(2) 341–357.

Ben-Tal, Aharon, A Nemirovski. 2020. *Lectures on modern convex optimization*. MOS-SIAM Series on Optimization.

Bertsimas, Dimitris, Martin S Copenhaver. 2018. Characterization of the equivalence of robustification and regularization in linear and matrix regression. *European Journal of Operational Research* **270**(3) 931–942.

Bertsimas, Dimitris, Melvyn Sim. 2004. The price of robustness. *Operations research* **52**(1) 35–53.

Bishop, Chris M. 1995. Training with noise is equivalent to tikhonov regularization. *Neural computation* **7**(1) 108–116.

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

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

Chen, Zhi, Melvyn Sim, Peng Xiong. 2020. Robust stochastic optimization made easy with rsome. *Management Science* **66**(8) 3329–3339.

Cortes, Corinna, Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* **20**(3) 273–297.

Delage, Erick, Yinyu Ye. 2010. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research* **58**(3) 595–612.

Dua, Dheeru, Casey Graff. 2017. UCI machine learning repository. URL `http://archive.ics.uci.edu/ml`.

El Ghaoui, Laurent, Hervé Lebret. 1997. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on matrix analysis and applications* **18**(4) 1035–1064.

Farnia, Farzan, David Tse. 2016. A minimax approach to supervised learning. *Advances in Neural Information Processing Systems* **29** 4240–4248.

Fournier, Nicolas, Arnaud Guillin. 2015. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields* **162**(3) 707–738.

Gao, Rui, Xi Chen, Anton J Kleywegt. 2017. Distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050* .

Gao, Rui, Anton J Kleywegt. 2016. Distributionally robust stochastic optimization with wasserstein distance. *arXiv preprint arXiv:1604.02199* .

Gotoh, Jun-ya, Michael Jong Kim, Andrew EB Lim. 2018. Robust empirical optimization is almost the same as mean–variance optimization. *Operations research letters* **46**(4) 448–452.

Gotoh, Jun-ya, Michael Jong Kim, Andrew EB Lim. 2021. Calibration of distributionally robust empirical optimization models. *Operations Research* .

Huang, Kaizhu, Haiqin Yang, Irwin King, Michael R Lyu, Laiwan Chan. 2004. The minimum error minimax probability machine. *Journal of Machine Learning Research* **5**(Oct) 1253–1286.

Lanckriet, Gert RG, Laurent El Ghaoui, Chiranjib Bhattacharyya, Michael I Jordan. 2002. A robust minimax approach to classification. *Journal of Machine Learning Research* **3**(Dec) 555–582.

Long, Daniel Zhuoyu, Melvyn Sim, Minglong Zhou. 2021. Robust satisficing. *Operations Research (forthcoming)* .

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

Shafieezadeh-Abadeh, Soroosh, Peyman Mohajerin Esfahani, Daniel Kuhn. 2015. Distributionally robust logistic regression. *Proceedings of the 28th International Conference on Neural Information Processing Systems*, vol. 1. 1576–1584.

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

Shalev-Shwartz, Shai, Ohad Shamir, Nathan Srebro, Karthik Sridharan. 2010. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research* **11** 2635–2670.

Sion, Maurice. 1958. On general minimax theorems. *Pacific Journal of Mathematics* **8**(1) 171–176.

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

Tikhonov, Andrey N, Vasiliy Y Arsenin. 1977. Solutions of ill-posed problems. *New York* **1**(30) 487.

Vapnik, Vladimir. 2013. *The nature of statistical learning theory*. Springer science & business media.

Wiesemann, W., D. Kuhn, M. Sim. 2014. Distributionally robust convex optimization. *Operations Research* **62**(6) 1358–1376.

Xu, Huan, Constantine Caramanis, Shie Mannor. 2009. Robustness and regularization of support vector machines. *Journal of machine learning research* **10**(7).

Xu, Huan, Constantine Caramanis, Shie Mannor. 2010. Robust regression and lasso. *IEEE Transactions on Information Theory* **56**(7) 3561–3574.

A. Proof of Results

Proof of Theorem 3. Note that $(w)^{\frac{1}{p}}$ is a non-decreasing function for $w \geq 0$; hence,

$$\begin{aligned}&\sup_{\mathbb{Q}\in\mathcal{F}(\hat{\mathbb{P}})}\left\{\left(\mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}})-\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]\right)^{\frac{1}{p}}\right\}\\ =&\left(\sup_{\mathbb{Q}\in\mathcal{F}(\hat{\mathbb{P}})}\left\{\mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}})-\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]\right\}\right)^{\frac{1}{p}}\end{aligned}$$

Now, we look at the maximization problem on the right-hand-side of above equation. We have

$$\sup_{\mathbb{Q}\in\mathcal{F}(\hat{\mathbb{P}})}\mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)-\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]=\frac{1}{S}\sum_{s\in[S]}\sup_{\boldsymbol{z}_s\in\mathcal{Z}}\left\{\left(\left(\ell\left(\alpha+\boldsymbol{z}_s^\top\bar{\boldsymbol{\beta}}\right)-\kappa\|\boldsymbol{z}_s-\hat{\boldsymbol{z}}_s\|\right)^+\right)^p\right\}.$$

Note that $\left(\left(\ell\left(\cdot\right)\right)^+\right)^p$ is a nondecreasing function; hence, we can write the above as:

$$\sup_{\mathbb{Q}\in\mathcal{F}(\hat{\mathbb{P}})}\mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)-\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]=\frac{1}{S}\sum_{s\in[S]}\left(\left(\sup_{\boldsymbol{z}_s\in\mathcal{Z}}\left\{\ell\left(\alpha+\boldsymbol{z}_s^\top\bar{\boldsymbol{\beta}}\right)-\kappa\|\boldsymbol{z}_s-\hat{\boldsymbol{z}}_s\|\right\}\right)^+\right)^p.$$

Because the loss function ℓ is nonnegative, we have

$$\sup_{\boldsymbol{z}_s\in\mathcal{Z}}\left\{\ell\left(\alpha+\boldsymbol{z}_s^\top\bar{\boldsymbol{\beta}}\right)-\kappa\|\boldsymbol{z}_s-\hat{\boldsymbol{z}}_s\|\right\}\geq\ell\left(\alpha+\hat{\boldsymbol{z}}_s^\top\bar{\boldsymbol{\beta}}\right)\geq 0.$$

Hence, we can equivalently write Problem (10) as Problem (12). The two explicit formulations then follow from Problem (12) and Theorem 2. □

Proof of Theorem 4. Observe that for any $\kappa \geq 0$,

$$\begin{aligned}&\sup_{\mathbb{P}\in\mathcal{W}_p(r)}\left\{\left(\mathbb{E}_{\mathbb{P}}\left[\ell^p(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}})\right]\right)^{\frac{1}{p}}\right\}\\ =&\sup_{\mathbb{P}\in\mathcal{P}_0(\mathcal{Z})}\left\{\min_{k\geq 0}\left\{\left(\mathbb{E}_{\mathbb{P}}\left[\ell^p(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}})\right]\right)^{\frac{1}{p}}-\kappa(\Delta_p(\mathbb{P},\hat{\mathbb{P}})-r)\right\}\right\}\\ \leq&\,\kappa r+\sup_{\mathbb{P}\in\mathcal{P}_0(\mathcal{Z})}\left\{\left(\mathbb{E}_{\mathbb{P}}\left[\ell^p(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}})\right]\right)^{\frac{1}{p}}-\kappa\Delta_p(\mathbb{P},\hat{\mathbb{P}})\right\}\\ =&\,\kappa r+\sup_{\mathbb{Q}\in\mathcal{F}(\hat{\mathbb{P}})}\left\{\left(\mathbb{E}_{\mathbb{Q}}\left[\ell^p(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}})\right]\right)^{\frac{1}{p}}-\kappa\left(\mathbb{E}_{\mathbb{Q}}\left[\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|^p\right]\right)^{\frac{1}{p}}\right\}\\ =&\,\kappa r+\sup_{\mathbb{Q}\in\mathcal{F}(\hat{\mathbb{P}})}\left\{\left(\mathbb{E}_{\mathbb{Q}}\left[((\ell(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}))^+)^p\right]\right)^{\frac{1}{p}}-\left(\mathbb{E}_{\mathbb{Q}}\left[((\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|)^+)^p\right]\right)^{\frac{1}{p}}\right\}\\ \leq&\,\kappa r+\sup_{\mathbb{Q}\in\mathcal{F}(\hat{\mathbb{P}})}\left\{\left(\mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}})-\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]\right)^{\frac{1}{p}}\right\}.\end{aligned}$$

The last inequality follows from the property of $\left(\mathbb{E}_{\mathbb{Q}}\left[((\cdot)^+)^p\right]\right)^{1/p}$ for $p \geq 1$. Specifically, by the convexity and positive homogeneity properties of $\left(\mathbb{E}_{\mathbb{Q}}\left[((\cdot)^+)^p\right]\right)^{1/p}$ for $p \geq 1$, we have subadditivity, and hence,

$$\begin{aligned}&\mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)\right)^+\right)^p\right]^{\frac{1}{p}}\\ =&\,\mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)-\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|+\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]^{\frac{1}{p}}\\ \leq&\,\mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)-\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]^{\frac{1}{p}}+\mathbb{E}_{\mathbb{Q}}\left[\left((\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|)^+\right)^p\right]^{\frac{1}{p}}.\end{aligned}$$

When $p=1$, the strong duality holds (Shafieezadeh-Abadeh et al. 2019); hence,

$$\begin{aligned}
&\sup_{\mathbb{P}\in\mathcal{W}_1(r)} \left\{ \left(\mathbb{E}_{\mathbb{P}}\left[\ell(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}})\right]\right)\right\}\\
&= \sup_{\mathbb{P}\in\mathcal{P}_0(\mathcal{Z})} \left\{ \min_{k\geq 0}\left\{ \mathbb{E}_{\mathbb{P}}\left[\ell(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}})\right] - \kappa(\Delta_1(\mathbb{P},\hat{\mathbb{P}})-r)\right\}\right\}\\
&= \min_{\kappa\geq 0}\left\{ \kappa r + \sup_{\mathbb{P}\in\mathcal{P}_0(\mathcal{Z})}\left\{ \mathbb{E}_{\mathbb{P}}\left[\ell(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}})\right] - \kappa\Delta_1(\mathbb{P},\hat{\mathbb{P}})\right\}\right\}\\
&= \min_{\kappa\geq 0}\left\{ \kappa r + \sup_{\mathbb{Q}\in\mathcal{F}(\hat{\mathbb{P}})}\left\{ \mathbb{E}_{\mathbb{Q}}\left[\ell(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}})\right] - \mathbb{E}_{\mathbb{Q}}\left[\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right]\right\}\right\}\\
&= \min_{\kappa\geq 0}\left\{ \kappa r + \frac{1}{S}\sum_{s\in[S]}\sup_{\boldsymbol{z}_s\in\mathcal{Z}}\left\{ \ell(\alpha+\boldsymbol{z}_s^\top\bar{\boldsymbol{\beta}}) - \kappa\|\boldsymbol{z}_s-\hat{\boldsymbol{z}}_s\|\right\}\right\}\\
&= \min_{\kappa\geq 0}\left\{ \kappa r + \frac{1}{S}\sum_{s\in[S]}\left(\sup_{\boldsymbol{z}_s\in\mathcal{Z}}\left\{ \ell(\alpha+\boldsymbol{z}_s^\top\bar{\boldsymbol{\beta}}) - \kappa\|\boldsymbol{z}_s-\hat{\boldsymbol{z}}_s\|\right\}\right)^+\right\}\\
&= \min_{\kappa\geq 0}\left\{ \kappa r + \frac{1}{S}\sum_{s\in[S]}\sup_{\boldsymbol{z}_s\in\mathcal{Z}}\left\{ \left(\ell(\alpha+\boldsymbol{z}_s^\top\bar{\boldsymbol{\beta}}) - \kappa\|\boldsymbol{z}_s-\hat{\boldsymbol{z}}_s\|\right)^+\right\}\right\}\\
&= \min_{\kappa\geq 0}\left\{ \kappa r + \sup_{\mathbb{Q}\in\mathcal{F}(\hat{\mathbb{P}})}\left\{ \mathbb{E}_{\mathbb{Q}}\left[\left(\ell(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}) - \kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right]\right\}\right\},
\end{aligned}$$

where the third last equality follows because $\sup_{\boldsymbol{z}_s\in\mathcal{Z}}\left\{\ell(\alpha+\boldsymbol{z}_s^\top\bar{\boldsymbol{\beta}}) - \kappa\|\boldsymbol{z}_s-\hat{\boldsymbol{z}}_s\|\right\}$ is non-negative, and the penultimate equality follows because $(\cdot)^+$ is a non-decreasing function. The last equality holds because the worst-case distribution is a singleton distribution (see *e.g.,* Mohajerin Esfahani and Kuhn 2018).

We first prove the second part of the theorem before coming back to the case of $p=2$ when $\mathcal{Z}=\mathbb{R}^{N+1}$ and $\ell(w)=L|w|$. For the second part, note that $|\ell(\alpha+\boldsymbol{z}^\top\bar{\boldsymbol{\beta}})-\ell(\alpha+\boldsymbol{v}^\top\bar{\boldsymbol{\beta}})|\leq L|(\boldsymbol{z}-\boldsymbol{v})^\top\bar{\boldsymbol{\beta}}|\leq L\|\boldsymbol{z}-\boldsymbol{v}\|\|\bar{\boldsymbol{\beta}}\|_*$. In addition, by the subadditivity of $\left(\mathbb{E}_{\mathbb{Q}}\left[((\cdot)^+)^p\right]\right)^{1/p}$ for $p\geq 1$, for any $\mathbb{Q}\in\mathcal{F}(\hat{\mathbb{P}})$, we have

$$\begin{aligned}
&\left(\mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}) - \kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]\right)^{\frac{1}{p}}\\
&\leq \left(\mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}) - \ell(\alpha+\tilde{\boldsymbol{v}}^\top\bar{\boldsymbol{\beta}}) - \kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]\right)^{\frac{1}{p}} + \left(\mathbb{E}_{\hat{\mathbb{P}}}\left[\ell^p(\alpha+\tilde{\boldsymbol{v}}^\top\bar{\boldsymbol{\beta}})\right]\right)^{\frac{1}{p}}\\
&\leq \left(\mathbb{E}_{\mathbb{Q}}\left[\left(\left(L\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\|\bar{\boldsymbol{\beta}}\|_* - \kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]\right)^{\frac{1}{p}} + \left(\mathbb{E}_{\hat{\mathbb{P}}}\left[\ell^p(\alpha+\tilde{\boldsymbol{v}}^\top\bar{\boldsymbol{\beta}})\right]\right)^{\frac{1}{p}}.
\end{aligned}$$

Note that $\kappa=L\|\bar{\boldsymbol{\beta}}\|_*$ is a feasible dual solution; hence,

$$\begin{aligned}
&\min_{\kappa\geq 0}\left\{\kappa r + \sup_{\mathbb{Q}\in\mathcal{F}(\hat{\mathbb{P}})}\left\{\left(\mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}) - \kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]\right)^{\frac{1}{p}}\right\}\right\}\\
&\leq \left(\mathbb{E}_{\hat{\mathbb{P}}}\left[\ell^p\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)\right]\right)^{\frac{1}{p}} + Lr\|\bar{\boldsymbol{\beta}}\|_*.
\end{aligned}$$

Now we prove the equivalence when $\mathcal{Z} = \mathbb{R}^{N+1}$. From Theorem 3, we have

$$\sup_{\mathbb{Q}\in\mathcal{F}(\hat{\mathbb{P}})} \left\{ \left(\mathbb{E}_{\mathbb{Q}} \left[\left(\left(\ell(\alpha + \tilde{\boldsymbol{z}}^\top \bar{\boldsymbol{\beta}}) - \kappa \|\tilde{\boldsymbol{z}} - \tilde{\boldsymbol{v}}\| \right)^+ \right)^p \right] \right)^{\frac{1}{p}} \right\}$$
$$= \left(\frac{1}{S} \sum_{s\in[S]} \left(\sup_{\boldsymbol{z}_s \in \mathcal{Z}} \left\{ \ell\left(\alpha + \boldsymbol{z}_s^\top \bar{\boldsymbol{\beta}}\right) - \kappa \|\boldsymbol{z}_s - \hat{\boldsymbol{z}}_s\| \right\} \right)^p \right)^{\frac{1}{p}}.$$

When $\mathcal{Z} = \mathbb{R}^{N+1}$, for any $\kappa < L\|\bar{\boldsymbol{\beta}}\|_*$, we have $\sup_{\boldsymbol{z}_s\in\mathcal{Z}} \left\{ \ell\left(\alpha + \boldsymbol{z}_s^\top \bar{\boldsymbol{\beta}}\right) - \kappa\|\boldsymbol{z}_s - \hat{\boldsymbol{z}}_s\| \right\} = +\infty$; for any $\kappa \geq L\|\bar{\boldsymbol{\beta}}\|_*$, we have $\sup_{\boldsymbol{z}_s\in\mathcal{Z}} \left\{ \ell\left(\alpha + \boldsymbol{z}_s^\top \bar{\boldsymbol{\beta}}\right) - \kappa\|\boldsymbol{z}_s - \hat{\boldsymbol{z}}_s\| \right\} = \ell\left(\alpha + \hat{\boldsymbol{z}}_s^\top \bar{\boldsymbol{\beta}}\right)$ (Gao et al. 2017, Shafieezadeh-Abadeh et al. 2019). Hence, the bound is tight at $\kappa = L\|\bar{\boldsymbol{\beta}}\|_*$. Finally, in case of $p = 2$ when $\mathcal{Z} = \mathbb{R}^{N+1}$ and $\ell(w) = L|w|$, the first bound is tight follows because of the second part of this theorem and the equivalence result of Blanchet et al. (2019). □

Proof of Proposition 1. Observe that $\|\bar{\boldsymbol{\beta}}\|_d = \|(\boldsymbol{\beta}, -1)\|_d \geq 1$. Since $C_T > 1$, the optimal solution of Problem (15) is also optimal to the following problem

$$\begin{aligned} \min\ & \|\boldsymbol{\beta}\|_d \\ \text{s.t. } & \mathbb{E}_{\hat{\mathbb{P}}}\left[\ell^p(\alpha + \tilde{\boldsymbol{z}}^\top \bar{\boldsymbol{\beta}})\right] \leq T \\ & (\alpha, \bar{\boldsymbol{\beta}}) \in \bar{\mathcal{D}}. \end{aligned}$$

Given the Slater's conditions apply, it is equivalent to the following problem

$$\max_{\mu\geq 0} \min_{(\alpha,\bar{\boldsymbol{\beta}})\in\bar{\mathcal{D}}} \left\{ \|\boldsymbol{\beta}\|_d + \mu\left(\mathbb{E}_{\hat{\mathbb{P}}}\left[\ell^p(\alpha + \tilde{\boldsymbol{z}}^\top \bar{\boldsymbol{\beta}})\right] - T\right) \right\}.$$

We now argue that at optimality, we must have $\mu > 0$. Otherwise, if $\mu = 0$, we would have $C_T = C_\infty$. Hence, we choose the regularization parameter as $\lambda = 1/\mu$ to obtain the same optimal solution for the regularization problem. □

Proof of Theorem 5. First, we have

$$\sup_{\mathbb{Q}\in\mathcal{F}(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}} \left[\left(\left(\ell\left(\alpha + \tilde{\boldsymbol{z}}^\top \bar{\boldsymbol{\beta}}\right) - \kappa\|\tilde{\boldsymbol{z}} - \tilde{\boldsymbol{v}}\| \right)^+ \right)^p \right] = \frac{1}{S} \sum_{s\in[S]} \sup_{\boldsymbol{z}_s\in\mathcal{Z}} \left\{ \left(\left(\ell\left(\alpha + \boldsymbol{z}_s^\top \bar{\boldsymbol{\beta}}\right) - \kappa\|\boldsymbol{z}_s - \hat{\boldsymbol{z}}_s\| \right)^+ \right)^p \right\}.$$

Note that $\left(\left(\ell(\cdot) \right)^+ \right)^p$ is a nondecreasing function; hence, we can write the above as:

$$\sup_{\mathbb{Q}\in\mathcal{F}(\hat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}} \left[\left(\left(\ell\left(\alpha + \tilde{\boldsymbol{z}}^\top \bar{\boldsymbol{\beta}}\right) - \kappa\|\tilde{\boldsymbol{z}} - \tilde{\boldsymbol{v}}\| \right)^+ \right)^p \right] = \frac{1}{S} \sum_{s\in[S]} \left(\left(\sup_{\boldsymbol{z}_s\in\mathcal{Z}} \left\{ \ell\left(\alpha + \boldsymbol{z}_s^\top \bar{\boldsymbol{\beta}}\right) - \kappa\|\boldsymbol{z}_s - \hat{\boldsymbol{z}}_s\| \right\} \right)^+ \right)^p.$$

Because the loss function ℓ is nonnegative, we have

$$\sup_{\boldsymbol{z}_s\in\mathcal{Z}} \left\{ \ell\left(\alpha + \boldsymbol{z}_s^\top \bar{\boldsymbol{\beta}}\right) - \kappa\|\boldsymbol{z}_s - \hat{\boldsymbol{z}}_s\| \right\} \geq \ell\left(\alpha + \hat{\boldsymbol{z}}_s^\top \bar{\boldsymbol{\beta}}\right) \geq 0.$$

Hence, we can equivalently write Problem (17) as Problem (18). The two explicit formulations then follow from Problem (18) and Proposition 2. Finally, feasibility follows because the optimal solution to the empirical loss minimization problem (8) is always feasible in Problem (19) for $T \geq Z_0$, even if the support is unbounded. Hence, for the general case when the support can be bounded, Problem (19) would also be feasible for $T \geq Z_0$. □

Proof of Theorem 6. Observe that

$$\mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)-\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right] \leq T^p \qquad \forall \mathbb{Q}\in\mathcal{F}(\hat{\mathbb{P}}),$$

implies

$$\mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)-\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right] \leq T^p \qquad \forall \mathbb{Q}\in\mathcal{F}(\mathbb{P}^*,\hat{\mathbb{P}}),$$

where

$$\mathcal{F}(\mathbb{P}^*,\hat{\mathbb{P}}) \triangleq \left\{\mathbb{Q}\in\mathcal{P}_0(\mathcal{Z}^2) \;\middle|\; (\tilde{\boldsymbol{z}},\tilde{\boldsymbol{v}})\sim\mathbb{Q}, \tilde{\boldsymbol{z}}\sim\mathbb{P}^*, \tilde{\boldsymbol{v}}\sim\hat{\mathbb{P}}\right\}.$$

Equivalently, we have

$$\mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)-\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]^{\frac{1}{p}} \leq T \qquad \forall \mathbb{Q}\in\mathcal{F}(\mathbb{P}^*,\hat{\mathbb{P}}).$$

By the convexity and positive homogeneity properties of $\left(\mathbb{E}_{\mathbb{Q}}\left[((\cdot)^+)^p\right]\right)^{1/p}$ for $p\geq 1$, we have sub-additivity, and hence,

$$\begin{aligned}
&\mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)\right)^+\right)^p\right]^{\frac{1}{p}}\\
=\;&\mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)-\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|+\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]^{\frac{1}{p}}\\
\leq\;&\mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)-\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]^{\frac{1}{p}}+\mathbb{E}_{\mathbb{Q}}\left[\left((\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|)^+\right)^p\right]^{\frac{1}{p}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)\right)^+\right)^p\right]^{\frac{1}{p}}-\mathbb{E}_{\mathbb{Q}}\left[\left((\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|)^+\right)^p\right]^{\frac{1}{p}}\leq T \qquad \forall \mathbb{Q}\in\mathcal{F}(\mathbb{P}^*,\hat{\mathbb{P}})\\
\iff\;&\mathbb{E}_{\mathbb{P}^*}\left[\ell^p\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)\right]^{\frac{1}{p}}-\kappa\inf_{\mathbb{Q}\in\mathcal{F}(\mathbb{P}^*,\hat{\mathbb{P}})}\mathbb{E}_{\mathbb{Q}}\left[\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|^p\right]^{\frac{1}{p}}\leq T\\
\iff\;&\mathbb{E}_{\mathbb{P}^*}\left[\ell^p\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)\right]^{\frac{1}{p}}-T\leq\kappa\Delta_p(\mathbb{P}^*,\hat{\mathbb{P}}).
\end{aligned}$$

Hence, for all $r\geq 0$,

$$\mathbb{P}^S\left[\mathbb{E}_{\mathbb{P}^*}\left[\ell^p\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)\right]^{\frac{1}{p}}-T>\kappa r\right]\leq\mathbb{P}^S[\Delta_p(\mathbb{P}^*,\hat{\mathbb{P}})>r].$$

□

Proof of Theorem 7. We first focus on the inner maximization problems in Model (22):

$$\sup_{\mathbb{P}_i\in\mathcal{U}_i(r),i\in[2]}\max_{\boldsymbol{q}\in\mathcal{Q}}\left\{\sum_{i\in[2]}q_i\mathbb{E}_{\mathbb{P}_i}\left[\ell(b_i(\alpha+\tilde{\boldsymbol{z}}_i^\top\boldsymbol{\beta})\right]\right\}=\max_{\boldsymbol{q}\in\mathcal{Q}}\left\{\sum_{i\in[2]}q_i\sup_{\mathbb{P}_i\in\mathcal{U}_i(r)}\mathbb{E}_{\mathbb{P}_i}\left[\ell(b_i(\alpha+\tilde{\boldsymbol{z}}_i^\top\boldsymbol{\beta})\right]\right\}.$$

By classical results (*e.g.,* Shafieezadeh-Abadeh et al. 2015, Mohajerin Esfahani and Kuhn 2018), the inner maximization can be written as:

$$\sup_{\mathbb{P}_i\in\mathcal{U}_i(r)}\mathbb{E}_{\mathbb{P}_i}\left[\ell(b_i(\alpha+\tilde{\boldsymbol{z}}_i^\top\boldsymbol{\beta})\right]=\inf_{\kappa\geq0}\left\{\kappa r+\frac{1}{S_i}\sum_{s\in[S_i]}\sup_{\boldsymbol{z}_{i,s}\in\mathcal{Z}_i}\{\ell(b_i(\alpha+\boldsymbol{z}_i^\top\boldsymbol{\beta})-\kappa\|\boldsymbol{z}_{i,s}-\hat{\boldsymbol{z}}_{i,s}\|\}\right\}.$$

By above results, we have

$$\begin{aligned}
&\min_{(\alpha,\boldsymbol{\beta})\in\mathcal{D}}\sup_{\mathbb{P}_i\in\mathcal{U}_i(r),i\in[2]}\max_{\boldsymbol{q}\in\mathcal{Q}}\left\{\sum_{i\in[2]}q_i\mathbb{E}_{\mathbb{P}_i}\left[\ell(b_i(\alpha+\tilde{\boldsymbol{z}}_i^\top\boldsymbol{\beta})\right]\right\}\\
=&\min_{(\alpha,\boldsymbol{\beta})\in\mathcal{D}}\max_{\boldsymbol{q}\in\mathcal{Q}}\left\{\sum_{i\in[2]}q_i\sup_{\mathbb{P}_i\in\mathcal{U}_i(r)}\mathbb{E}_{\mathbb{P}_i}\left[\ell(b_i(\alpha+\tilde{\boldsymbol{z}}_i^\top\boldsymbol{\beta})\right]\right\}\\
=&\min_{(\alpha,\boldsymbol{\beta})\in\mathcal{D}}\max_{\boldsymbol{q}\in\mathcal{Q}}\left\{\sum_{i\in[2]}q_i\inf_{\kappa\geq0}\left\{\kappa r+\frac{1}{S_i}\sum_{s\in[S_i]}\sup_{\boldsymbol{z}_{i,s}\in\mathcal{Z}_i}\{\ell(b_i(\alpha+\boldsymbol{z}_i^\top\boldsymbol{\beta})-\kappa\|\boldsymbol{z}_{i,s}-\hat{\boldsymbol{z}}_{i,s}\|\}\right\}\right\}\\
=&\min_{(\alpha,\boldsymbol{\beta})\in\mathcal{D}}\max_{\boldsymbol{q}\in\mathcal{Q}}\inf_{\kappa_1,\kappa_2\geq0}\left\{\sum_{i\in[2]}\left(q_i\kappa_i r+\frac{1}{S_i}q_i\sum_{s\in[S_i]}\sup_{\boldsymbol{z}_{i,s}\in\mathcal{Z}_i}\{\ell(b_i(\alpha+\boldsymbol{z}_i^\top\boldsymbol{\beta})-\kappa_i\|\boldsymbol{z}_{i,s}-\hat{\boldsymbol{z}}_{i,s}\|\}\right)\right\}\\
=&\min_{(\alpha,\boldsymbol{\beta})\in\mathcal{D}}\inf_{\kappa_1,\kappa_2\geq0}\max_{\boldsymbol{q}\in\mathcal{Q}}\left\{\sum_{i\in[2]}\left(q_i\kappa_i r+\frac{1}{S_i}q_i\sum_{s\in[S_i]}\sup_{\boldsymbol{z}_{i,s}\in\mathcal{Z}_i}\{\ell(b_i(\alpha+\boldsymbol{z}_i^\top\boldsymbol{\beta})-\kappa_i\|\boldsymbol{z}_{i,s}-\hat{\boldsymbol{z}}_{i,s}\|\}\right)\right\},
\end{aligned}$$

where the last equality follows because the objective function is affine in $\boldsymbol{q}$ and convex in $\boldsymbol{\kappa}$, and set $\mathcal{Q}$ is convex and compact (Sion 1958). The optimization problem in the last equation can be written as

$$\begin{aligned}
Z_r=\min\ &\max_{\boldsymbol{q}\in\mathcal{Q}}\left\{(\kappa_1q_1+\kappa_2q_2)r+q_1v_1+q_2v_2\right\}\\
\text{s.t. }\ &v_i\geq\frac{1}{S_i}\sum_{s\in[S_i]}\tau_{i,s}. && \forall i\in[2]\\
&\tau_{i,s}\geq\sup_{\boldsymbol{z}_{i,s}\in\mathcal{Z}_i}\{\ell(b_i(\alpha+\boldsymbol{z}_i^\top\boldsymbol{\beta})-\kappa_i\|\boldsymbol{z}_{i,s}-\hat{\boldsymbol{z}}_{i,s}\|\} && \forall i\in[2],s\in[S_i]\\
&\tau_{i,s}\in\mathbb{R} && \forall i\in[2],s\in[S_i]\\
&(\alpha,\boldsymbol{\beta})\in\mathcal{D},\kappa_1\geq0,\kappa_2\geq0,v_1\in\mathbb{R},v_2\in\mathbb{R},
\end{aligned}$$

where we introduce epigraph variables $\boldsymbol{v}$ and $\boldsymbol{\tau}$.

Now, we focus on the maximization problem in its objective function:

$$\begin{aligned}
\max\ & (\kappa_1 q_1 + \kappa_2 q_2) r + q_1 v_1 + q_2 v_2 \\
\text{s.t.}\ & q_1 + q_2 = 1 \\
& q_1 \geq \underline{q}_1 \\
& q_1 \leq \bar{q}_1 \\
& q_1, q_2 \geq 0.
\end{aligned}$$

The above problem is a linear optimization problem and is equivalent to its dual problem:

$$\begin{aligned}
\min\ & u_1 - u_2 \underline{q}_1 + u_3 \bar{q}_1 \\
\text{s.t.}\ & u_1 - u_2 + u_3 \geq \kappa_1 r + v_1 \\
& u_1 \geq \kappa_2 r + v_2 \\
& u_1 \in \mathbb{R}, u_2 \geq 0, u_3 \geq 0.
\end{aligned}$$

Replacing the maximization problem with its dual minimization problem, we have

$$\begin{aligned}
Z_r = \min\ & u_1 - u_2 \underline{q}_1 + u_3 \bar{q}_1 \\
\text{s.t.}\ & u_1 - u_2 + u_3 \geq \kappa_1 r + v_1 \\
& u_1 \geq \kappa_2 r + v_2 \\
& v_i \geq \frac{1}{S_i} \sum_{s \in [S_i]} \tau_{i,s}. && \forall i \in [2] \\
& \tau_{i,s} \geq \sup_{\boldsymbol{z}_{i,s} \in \mathcal{Z}_i} \{\ell(b_i(\alpha + \boldsymbol{z}_i^\top \boldsymbol{\beta}) - \kappa_i \|\boldsymbol{z}_{i,s} - \hat{\boldsymbol{z}}_{i,s}\|\} && \forall i \in [2], s \in [S_i] \\
& \tau_{i,s} \in \mathbb{R} && \forall i \in [2], s \in [S_i] \\
& u_1 \in \mathbb{R}, u_2 \geq 0, u_3 \geq 0 \\
& (\alpha, \boldsymbol{\beta}) \in \mathcal{D}, \kappa_1 \geq 0, \kappa_2 \geq 0, v_1 \in \mathbb{R}, v_2 \in \mathbb{R}.
\end{aligned}$$

□

Proof of Theorem 8. Observe that for any $\boldsymbol{q} \in \mathcal{Q}$,

$$
\begin{aligned}
&\sup_{\mathbb{P}_i \in \mathcal{U}_{i,p}(r), i\in[2]} \left\{ \left(\sum_{i\in[2]} q_i \mathbb{E}_{\mathbb{P}_i} \left[\ell^p(b_i(\alpha + \tilde{\boldsymbol{z}}_i^\top \boldsymbol{\beta}) \right] \right)^{\frac{1}{p}} \right\} \\
&= \sup_{\mathbb{P}_i \in \mathcal{P}_0(\mathcal{Z}_i), i\in[2]} \left\{ \min_{\kappa_1,\kappa_2 \geq 0} \left\{ \left(\sum_{i\in[2]} q_i \mathbb{E}_{\mathbb{P}} \left[\ell^p(b_i(\alpha + \tilde{\boldsymbol{z}}^\top \boldsymbol{\beta})) \right] \right)^{\frac{1}{p}} - \sum_{i\in[2]} \kappa_i (\Delta_p(\mathbb{P}_i, \hat{\mathbb{P}}_i) - r) \right\} \right\} \\
&\leq \min_{\kappa_1,\kappa_2 \geq 0} \left\{ (\kappa_1+\kappa_2) r + \sup_{\mathbb{P}_i \in \mathcal{P}_0(\mathcal{Z}_i), i\in[2]} \left\{ \left(\sum_{i\in[2]} q_i \mathbb{E}_{\mathbb{P}} \left[\ell^p(b_i(\alpha + \tilde{\boldsymbol{z}}^\top \boldsymbol{\beta})) \right] \right)^{\frac{1}{p}} - \sum_{i\in[2]} \kappa_i \Delta_p(\mathbb{P}_i, \hat{\mathbb{P}}_i) \right\} \right\} \\
&= \min_{\kappa_1,\kappa_2 \geq 0} \left\{ (\kappa_1+\kappa_2) r + \sup_{\mathbb{Q}_i \in \mathcal{F}_i(\hat{\mathbb{P}}_i), i\in[2]} \left\{ \left(\sum_{i\in[2]} q_i \mathbb{E}_{\mathbb{Q}_i} \left[\ell^p(b_i(\alpha + \tilde{\boldsymbol{z}}^\top \boldsymbol{\beta})) \right] \right)^{\frac{1}{p}} \right. \right. \\
&\qquad \left. \left. - \sum_{i\in[2]} \kappa_i \left(\mathbb{E}_{\mathbb{Q}_i} \left[\|\tilde{\boldsymbol{z}} - \tilde{\boldsymbol{v}}\|^p \right] \right)^{\frac{1}{p}} \right\} \right\} \\
&= \min_{\kappa_1,\kappa_2 \geq 0} \left\{ (\kappa_1+\kappa_2) r + \sup_{\mathbb{Q}_i \in \mathcal{F}_i(\hat{\mathbb{P}}_i), i\in[2]} \left\{ \left(\sum_{i\in[2]} q_i \mathbb{E}_{\mathbb{Q}_i} \left[((\ell(b_i(\alpha + \tilde{\boldsymbol{z}}^\top \boldsymbol{\beta})))^+)^p \right] \right)^{\frac{1}{p}} \right. \right. \\
&\qquad \left. \left. - \sum_{i\in[2]} \left(\mathbb{E}_{\mathbb{Q}_i} \left[((\kappa_i \|\tilde{\boldsymbol{z}} - \tilde{\boldsymbol{v}}\|)^+)^p \right] \right)^{\frac{1}{p}} \right\} \right\} \\
&= \min_{\kappa_1,\kappa_2 \geq 0} \left\{ (q_1\kappa_1+q_2\kappa_2) r + \sup_{\mathbb{Q}_i \in \mathcal{F}_i(\hat{\mathbb{P}}_i), i\in[2]} \left\{ \left(\sum_{i\in[2]} q_i \mathbb{E}_{\mathbb{Q}_i} \left[((\ell(b_i(\alpha + \tilde{\boldsymbol{z}}^\top \boldsymbol{\beta})))^+)^p \right] \right)^{\frac{1}{p}} \right. \right. \\
&\qquad \left. \left. - \sum_{i\in[2]} \left(\mathbb{E}_{\mathbb{Q}_i} \left[((q_i\kappa_i \|\tilde{\boldsymbol{z}} - \tilde{\boldsymbol{v}}\|)^+)^p \right] \right)^{\frac{1}{p}} \right\} \right\} \\
&\leq \min_{\kappa_1,\kappa_2 \geq 0} \left\{ \sum_{i\in[2]} q_i\kappa_i r + \sup_{\substack{\mathbb{Q}_i \in \mathcal{F}_i(\hat{\mathbb{P}}_i) \\ i\in[2]}} \left\{ \left(\sum_{i\in[2]} q_i \mathbb{E}_{\mathbb{Q}_i} \left[\left(\left(\ell(b_i(\alpha + \tilde{\boldsymbol{z}}^\top \boldsymbol{\beta})) - \kappa_i \|\tilde{\boldsymbol{z}} - \tilde{\boldsymbol{v}}\| \right)^+ \right)^p \right] \right)^{\frac{1}{p}} \right\} \right\}.
\end{aligned}
$$

We now show that the last equality and the last inequality in above derivation indeed hold.

We first focus on the last equality. Suppose $q_1, q_2 \neq 0$, then the last equality follows by a change of variable. Suppose $q_1 = 1$ and $q_2 = 0$, then the right-hand-side of the second last equality above can be reduced as

$$
\begin{aligned}
&\min_{\kappa_1,\kappa_2 \geq 0} \left\{ \sum_{i\in[2]} \kappa_i r + \sup_{\substack{\mathbb{Q}_i \in \mathcal{F}_i(\hat{\mathbb{P}}_i) \\ i\in[2]}} \left\{ \left(\mathbb{E}_{\mathbb{Q}_1} ((\ell(b_i(\alpha + \tilde{\boldsymbol{z}}^\top \boldsymbol{\beta})))^+)^p \right)^{\frac{1}{p}} - \sum_{i\in[2]} \left(\mathbb{E}_{\mathbb{Q}_i} \left[((\kappa_i \|\tilde{\boldsymbol{z}} - \tilde{\boldsymbol{v}}\|)^+)^p \right] \right)^{\frac{1}{p}} \right\} \right\} \\
&= \min_{\kappa_1,\kappa_2 \geq 0} \left\{ \sum_{i\in[2]} \kappa_i r + \sup_{\mathbb{Q}_1 \in \mathcal{F}_1(\hat{\mathbb{P}}_1)} \left\{ \left(\mathbb{E}_{\mathbb{Q}_1} ((\ell(b_i(\alpha + \tilde{\boldsymbol{z}}^\top \boldsymbol{\beta})))^+)^p \right)^{\frac{1}{p}} - \left(\mathbb{E}_{\mathbb{Q}_1} ((\kappa_1 \|\tilde{\boldsymbol{z}} - \tilde{\boldsymbol{v}}\|)^+)^p \right)^{\frac{1}{p}} \right\} \right\} \\
&= \min_{\kappa_1 \geq 0} \left\{ \kappa_1 r + \sup_{\mathbb{Q}_1 \in \mathcal{F}_1(\hat{\mathbb{P}}_1)} \left\{ \left(\mathbb{E}_{\mathbb{Q}_1} ((\ell(b_i(\alpha + \tilde{\boldsymbol{z}}^\top \boldsymbol{\beta})))^+)^p \right)^{\frac{1}{p}} - \left(\mathbb{E}_{\mathbb{Q}_1} ((\kappa_1 \|\tilde{\boldsymbol{z}} - \tilde{\boldsymbol{v}}\|)^+)^p \right)^{\frac{1}{p}} \right\} \right\}.
\end{aligned}
$$

We could similarly derive the result when $p_1 = 0, p_2 = 1$. Hence, the final equality holds.

Now, we focus on the last inequality. By the convexity and positive homogeneity properties of $\left(\mathbb{E}_{\mathbb{Q}}\left[((\cdot)^+)^p\right]\right)^{1/p}$ for $p \geq 1$, we have subadditivity, and hence,

$$\begin{aligned}
&\mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)\right)^+\right)^p\right]^{\frac{1}{p}} \\
&= \mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)-\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|+\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]^{\frac{1}{p}} \\
&\leq \mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell\left(\alpha+\tilde{\boldsymbol{z}}^\top\bar{\boldsymbol{\beta}}\right)-\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]^{\frac{1}{p}} + \mathbb{E}_{\mathbb{Q}}\left[\left((\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|)^+\right)^p\right]^{\frac{1}{p}}.
\end{aligned}$$

Now, we show that the bound is tight when $p=1$. When $p=1$, the strong duality holds (Shafieezadeh-Abadeh et al. 2019); hence,

$$\begin{aligned}
&\sup_{\mathbb{P}_i\in\mathcal{U}_{i,p}(r),i\in[2]}\left\{\sum_{i\in[2]} q_i\mathbb{E}_{\mathbb{P}_i}\left[\ell(b_i(\alpha+\tilde{\boldsymbol{z}}_i^\top\boldsymbol{\beta})\right]\right\} \\
&= \sup_{\mathbb{P}_i\in\mathcal{P}_0(\mathcal{Z}_i),i\in[2]}\left\{\min_{\kappa_1,\kappa_2\geq 0}\left\{\sum_{i\in[2]} q_i\mathbb{E}_{\mathbb{P}}\left[\ell(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta}))\right]-\sum_{i\in[2]}\kappa_i(\Delta_1(\mathbb{P}_i,\hat{\mathbb{P}}_i)-r)\right\}\right\} \\
&= \min_{\kappa_1,\kappa_2\geq 0}\left\{(\kappa_1+\kappa_2)r+\sup_{\mathbb{P}_i\in\mathcal{P}_0(\mathcal{Z}_i),i\in[2]}\left\{\sum_{i\in[2]} q_i\mathbb{E}_{\mathbb{P}}\left[\ell(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta}))\right]-\sum_{i\in[2]}\kappa_i\Delta_1(\mathbb{P}_i,\hat{\mathbb{P}}_i)\right\}\right\} \\
&= \min_{\kappa_1,\kappa_2\geq 0}\left\{(\kappa_1+\kappa_2)r+\sup_{\mathbb{Q}_i\in\mathcal{F}_i(\hat{\mathbb{P}}_i),i\in[2]}\left\{\sum_{i\in[2]} q_i\mathbb{E}_{\mathbb{Q}_i}\left[\ell(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta}))\right]-\sum_{i\in[2]}\kappa_i\mathbb{E}_{\mathbb{Q}_i}\left[\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right]\right\}\right\} \\
&= \min_{\kappa_1,\kappa_2\geq 0}\left\{(q_1\kappa_1+q_2\kappa_2)r+\sup_{\mathbb{Q}_i\in\mathcal{F}_i(\hat{\mathbb{P}}_i),i\in[2]}\left\{\sum_{i\in[2]} q_i\mathbb{E}_{\mathbb{Q}_i}\left[\ell(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta}))-\kappa_i\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right]\right\}\right\}
\end{aligned}$$

The last equality follows similarly by our argument above. In addition,

$$\begin{aligned}
&\max_{\boldsymbol{q}\in\mathcal{Q}}\min_{\kappa_1,\kappa_2\geq 0}\left\{(q_1\kappa_1+q_2\kappa_2)r+\sup_{\mathbb{Q}_i\in\mathcal{F}_i(\hat{\mathbb{P}}_i),i\in[2]}\left\{\sum_{i\in[2]} q_i\mathbb{E}_{\mathbb{Q}_i}\left[\ell(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta}))-\kappa_i\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right]\right\}\right\} \\
&= \min_{\kappa_1,\kappa_2\geq 0}\max_{\boldsymbol{q}\in\mathcal{Q}}\left\{(q_1\kappa_1+q_2\kappa_2)r+\sup_{\mathbb{Q}_i\in\mathcal{F}_i(\hat{\mathbb{P}}_i),i\in[2]}\left\{\sum_{i\in[2]} q_i\mathbb{E}_{\mathbb{Q}_i}\left[\ell(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta}))-\kappa_i\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right]\right\}\right\} \\
&= \min_{\kappa_1,\kappa_2\geq 0}\max_{\boldsymbol{q}\in\mathcal{Q}}\left\{\sum_{i\in[2]} q_i\kappa_i r+\sup_{\mathbb{Q}_i\in\mathcal{F}_i(\hat{\mathbb{P}}_i),i\in[2]}\left\{\sum_{i\in[2]} q_i\mathbb{E}_{\mathbb{Q}_i}\left[\left(\ell(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta}))-\kappa_i\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right]\right\}\right\}.
\end{aligned}$$

The first equality follows from Sion's minimax theorem (Sion 1958) because the objective is affine in $\boldsymbol{q}$ and convex in $\boldsymbol{\kappa}$ and set $\mathcal{Q}$ is convex and compact. The last equality follows similarly as the corresponding proof of Theorem 4.

For the second part, note that $|\ell(b_i(\alpha+\boldsymbol{z}^\top\boldsymbol{\beta}))-\ell(b_i(\alpha+\boldsymbol{v}^\top\boldsymbol{\beta}))|\le L|(\boldsymbol{z}-\boldsymbol{v})^\top\boldsymbol{\beta}|\le L\|\boldsymbol{z}-\boldsymbol{v}\|\|\boldsymbol{\beta}\|_*$. In addition, by the subadditivity of $\left(\mathbb{E}_{\mathbb{Q}}\left[((\cdot)^+)^p\right]\right)^{1/p}$ for $p\ge 1$, for any $\mathbb{Q}_i\in\mathcal{F}_i(\hat{\mathbb{P}}_i)$, $i\in[2]$, we have

$$\begin{aligned}
&\left(\sum_{i\in[2]} q_i\mathbb{E}_{\mathbb{Q}_i}\left[\left(\left(\ell(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta}))-\kappa_i\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]\right)^{\frac{1}{p}}\\
\le&\left(\sum_{i\in[2]} q_i\mathbb{E}_{\mathbb{Q}_i}\left[\left(\left(\ell(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta}))-\ell(b_i(\alpha+\tilde{\boldsymbol{v}}^\top\boldsymbol{\beta}))-\kappa_i\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]\right)^{\frac{1}{p}}\\
&+\sum_{i\in[2]} q_i\left(\mathbb{E}_{\hat{\mathbb{P}}_i}\left[\ell^p(\alpha+\tilde{\boldsymbol{v}}^\top\bar{\boldsymbol{\beta}})\right]\right)^{\frac{1}{p}}\\
\le&\left(\sum_{i\in[2]} q_i\mathbb{E}_{\mathbb{Q}_i}\left[\left(\left(L\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\|\bar{\boldsymbol{\beta}}\|_*-\kappa_i\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]\right)^{\frac{1}{p}}+\sum_{i\in[2]} q_i\left(\mathbb{E}_{\hat{\mathbb{P}}_i}\left[\ell^p(\alpha+\tilde{\boldsymbol{v}}^\top\boldsymbol{\beta})\right]\right)^{\frac{1}{p}}.
\end{aligned}$$

Note that $\kappa_1=\kappa_2=L\|\bar{\boldsymbol{\beta}}\|_*$ is a feasible dual solution; hence,

$$\begin{aligned}
&\min_{\kappa_1,\kappa_2\ge 0}\left\{\sum_{i\in[2]} q_i\kappa_i r+\max_{\boldsymbol{q}\in\mathcal{Q}}\sup_{\mathbb{Q}_i\in\mathcal{F}_i(\hat{\mathbb{P}}_i),i\in[2]}\left\{\left(\sum_{i\in[2]}\mathbb{E}_{\mathbb{Q}_i}\left[\left(\left(\ell(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta}))-\kappa_i\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]\right)^{\frac{1}{p}}\right\}\right\}\\
\le&\max_{\boldsymbol{q}\in\mathcal{Q}}\left\{\left(\sum_{i\in[2]} q_i\mathbb{E}_{\hat{\mathbb{P}}_i}\left[\ell^p\left(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta})\right)\right]\right)^{\frac{1}{p}}\right\}+Lr\|\boldsymbol{\beta}\|_*.
\end{aligned}$$

When $\mathcal{Z}=\mathbb{R}^{N+1}$, for any $\kappa<L\|\boldsymbol{\beta}\|_*$, $\sup_{\mathbb{Q}_i\in\mathcal{F}_i(\hat{\mathbb{P}}_i)}\mathbb{E}_{\mathbb{Q}_i}\left[\left((L\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\|\boldsymbol{\beta}\|_*-\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|)^+\right)^p\right]=+\infty$ (Gao et al. 2017). Hence, the bound is tight at $\kappa_1=\kappa_2=L\|\bar{\boldsymbol{\beta}}\|_*$. □

Proof of Theorem 9. By previous proofs (*e.g.,* proof of Theorem 3), we have

$$\begin{aligned}
&\sup_{\mathbb{Q}_i\in\mathcal{F}_i(\hat{\mathbb{P}}_i)}\mathbb{E}_{\mathbb{Q}_i}\left[\left(\left(\ell\left(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta})\right)-\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]\\
=&\frac{1}{S_i}\sum_{s\in[S_i]}\left(\sup_{\boldsymbol{z}_s\in\mathcal{Z}_i}\left\{\ell\left(b_i(\alpha+\boldsymbol{z}_s^\top\boldsymbol{\beta})\right)-\kappa\|\boldsymbol{z}_s-\hat{\boldsymbol{z}}_{i,s}\|\right\}\right)^p.
\end{aligned}$$

Hence, we can equivalently write Problem (23) as Problem (25) by incorporating epigraph variables:

$$\begin{aligned}
\bar{Z}_r=\min\ &\max_{\boldsymbol{q}\in\mathcal{Q}}\left\{(\kappa_1q_1+\kappa_2q_2)r+\left(\sum_{i\in[2]} q_iv_i^p\right)^{\frac{1}{p}}\right\}\\
\text{s.t. }\ &v_i\ge\left(\frac{1}{S_i}\sum_{s\in[S_i]}\tau_{i,s}^p\right)^{\frac{1}{p}} && \forall i\in[2]\\
&\tau_{i,s}\ge\sup_{\boldsymbol{z}_{i,s}\in\mathcal{Z}_i}\{\ell(b_i(\alpha+\boldsymbol{z}_i^\top\boldsymbol{\beta})-\kappa_i\|\boldsymbol{z}_{i,s}-\hat{\boldsymbol{z}}_{i,s}\|\} && \forall i\in[2],s\in[S_i]\\
&\tau_{i,s}\in\mathbb{R} && \forall i\in[2],s\in[S_i]\\
&(\alpha,\boldsymbol{\beta})\in\mathcal{D},\kappa_1\ge 0,\kappa_2\ge 0,v_1\in\mathbb{R},v_2\in\mathbb{R}.
\end{aligned}\tag{25}$$

Note that $p \geq 1$ and $\tau_i \geq 0$ for $i \in [2]$. Now, we focus on the maximization problem in the objective function:

$$\begin{aligned} \max\ & (\kappa_1 q_1 + \kappa_2 q_2) r + \gamma \\ \text{s.t.}\ & \gamma^p \leq \sum_{i\in[2]} q_i v_i^p \\ & q_1 + q_2 = 1 \\ & q_1 \geq \underline{q}_1 \\ & q_1 \leq \bar{q}_1 \\ & q_1, q_2 \geq 0, \gamma \in \mathbb{R}. \end{aligned}$$

The above problem is a convex optimization problem. When $\tau_1 + \tau_2 \neq 0$, the Slater's condition holds. When $\tau_1 = \tau_2 = 0$, the optimal value to γ is zero and the above becomes a linear optimization model. The strong duality holds in either case. The dual of the problem is given by:

$$\begin{aligned} \min\ & \max_{\boldsymbol{q}\geq \mathbf{0},\gamma} \left\{ (\kappa_1 q_1 + \kappa_2 q_2) r + \gamma + u_1 \left(\sum_{i\in[2]} q_i v_i^p - \gamma^p \right) + u_2(1 - q_1 - q_2) - u_3(\underline{q}_1 - q_1) + u_4(\bar{q}_1 - q_1) \right\} \\ \text{s.t.}\ & u_1 \geq 0, u_2 \in \mathbb{R}, u_3 \geq 0, u_4 \geq 0. \end{aligned}$$

Now, the inner maximization problem can be written as:

$$\max_{\boldsymbol{q}\geq \mathbf{0},\gamma} \left\{ (\kappa_1 r + u_1 v_1^p - u_2 + u_3 - u_4) q_1 + (\kappa_2 r + u_1 v_2^p - u_2) q_2 + \gamma - u_1 \gamma^p + u_2 - u_3 \underline{q}_1 + u_4 \bar{q}_1 \right\}$$

The above equals to

$$\begin{cases} \max\limits_{\gamma} \{\gamma - u_1 \gamma^p\} + u_2 - u_3 \underline{q}_1 + u_4 \bar{q}_1 & \text{if } \kappa_1 r + u_1 v_1^p - u_2 + u_3 - u_4 \leq 0 \text{ and } \kappa_2 r + u_1 v_2^p - u_2 \leq 0; \\ +\infty & \text{otherwise.} \end{cases}$$

When $p = 1$, we have $\max_\gamma \{\gamma - u_1 \gamma^p\} = 0$ when we restrict $u_1 = 1$ and it equals to $+\infty$ otherwise. Hence, when $p = 1$, the final reformulation of Problem (25) reduces to the formulation in Theorem 7.

When $p > 1$, by first order condition, we have

$$\max_{\gamma} \{\gamma - u_1 \gamma^p\} = p^{\frac{p}{1-p}} (p-1) u_1^{1/(1-p)} = p^{\frac{p}{1-p}} (p-1) u_1,$$

where we use a change of variable and treat $u_1^{1/(1-p)}$ as the new variable u_1. Finally, the inner maximization problem becomes:

$$\begin{cases} p^{\frac{p}{1-p}} (p-1) u_1 + u_2 - u_3 \underline{q}_1 + u_4 \bar{q}_1 & \text{if } \kappa_1 r + u_1 (\frac{v_1}{u_1})^p - u_2 + u_3 - u_4 \leq 0,\ \kappa_2 r + u_1 (\frac{v_2}{u_1})^p - u_2 \leq 0; \\ +\infty & \text{otherwise.} \end{cases}$$

The entire dual problem is then given by:

$$\begin{aligned} \min\ & p^{\frac{p}{1-p}} (p-1) u_1 + u_2 - u_3 \underline{q}_1 + u_4 \bar{q}_1 \\ \text{s.t.}\ & \kappa_1 r + u_1 (v_1/u_1)^p - u_2 + u_3 - u_4 \leq 0 \\ & \kappa_2 r + u_1 (v_2/u_1)^p - u_2 \leq 0 \\ & u_1 \geq 0, u_2 \in \mathbb{R}, u_3 \geq 0, u_4 \geq 0. \end{aligned}$$

Hence, when $p>1$, Problem (25) can be reformulated as

$$\begin{aligned}
\bar{Z}_r = \min\ & p^{\frac{p}{1-p}}(p-1)u_1 + u_2 - u_3\underline{q}_1 + u_4\bar{q}_1 \\
\text{s.t. } & v_i \geq \left(\frac{1}{S_i}\sum_{s\in[S_i]}\tau_{i,s}^p\right)^{\frac{1}{p}} && \forall i\in[2] \\
& \tau_{i,s} \geq \sup_{\boldsymbol{z}_{i,s}\in\mathcal{Z}_i}\{\ell(b_i(\alpha+\boldsymbol{z}_i^\top\boldsymbol{\beta}) - \kappa_i\|\boldsymbol{z}_{i,s}-\hat{\boldsymbol{z}}_{i,s}\|\} && \forall i\in[2], s\in[S_i] \\
& \kappa_1 r + u_1(v_1/u_1)^p - u_2 + u_3 - u_4 \leq 0 \\
& \kappa_2 r + u_1(v_2/u_1)^p - u_2 \leq 0 \\
& \tau_{i,s}\in\mathbb{R} && \forall i\in[2], s\in[S_i] \\
& v_1\geq 0, u_2\in\mathbb{R}, u_3\geq 0, u_4\geq 0, \\
& (\alpha,\boldsymbol{\beta})\in\mathcal{D}, \kappa_1\geq 0, \kappa_2\geq 0, v_1\geq 0, v_2\geq 0.
\end{aligned}$$

□

Proof of Theorem 10. For convenience, we define $\underline{\boldsymbol{q}} := (\underline{q}_1, 1-\underline{q}_1)$ and $\bar{\boldsymbol{q}} := (\bar{q}_1, 1-\bar{q}_1)$. First, we have

$$\begin{aligned}
&\sup_{\mathbb{Q}_i\in\mathcal{F}_i(\hat{\mathbb{P}}_i),i\in[2]} \max_{\boldsymbol{q}\in\mathcal{Q}} \left\{\sum_{i\in[2]} q_i \mathbb{E}_{\mathbb{Q}_i}\left[\left(\left(\ell(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta})) - \kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]\right\} \\
&= \sup_{\mathbb{Q}_i\in\mathcal{F}_i(\hat{\mathbb{P}}_i),i\in[2]} \max_{\boldsymbol{q}\in\{\underline{\boldsymbol{q}},\bar{\boldsymbol{q}}\}} \left\{\sum_{i\in[2]} q_i \mathbb{E}_{\mathbb{Q}_i}\left[\left(\left(\ell(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta})) - \kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]\right\} \\
&= \max_{\boldsymbol{q}\in\{\underline{\boldsymbol{q}},\bar{\boldsymbol{q}}\}} \sup_{\mathbb{Q}_i\in\mathcal{F}_i(\hat{\mathbb{P}}_i),i\in[2]} \left\{\sum_{i\in[2]} q_i \mathbb{E}_{\mathbb{Q}_i}\left[\left(\left(\ell(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta})) - \kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]\right\} \\
&= \max_{\boldsymbol{q}\in\{\underline{\boldsymbol{q}},\bar{\boldsymbol{q}}\}} \left\{\sum_{i\in[2]} q_i \sup_{\mathbb{Q}_i\in\mathcal{F}_i(\hat{\mathbb{P}}_i),i\in[2]} \mathbb{E}_{\mathbb{Q}_i}\left[\left(\left(\ell(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta})) - \kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]\right\}.
\end{aligned}$$

The first equality follows because the objective function of the inner maximization problem is a linear function of q_1. The second equality is due to the interchangeability of the two maximization problems.

By previous proofs (*e.g.,* Theorem 5), we have

$$\begin{aligned}
&\sup_{\mathbb{Q}_i\in\mathcal{F}_i(\hat{\mathbb{P}}_i)} \mathbb{E}_{\mathbb{Q}_i}\left[\left(\left(\ell\left(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta})\right) - \kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right] \\
&= \frac{1}{S_i}\sum_{s\in[S_i]}\left(\sup_{\boldsymbol{z}_s\in\mathcal{Z}_i}\left\{\ell\left(b_i(\alpha+\boldsymbol{z}_s^\top\boldsymbol{\beta})\right) - \kappa\|\boldsymbol{z}_s-\hat{\boldsymbol{z}}_{i,s}\|\right\}\right)^p.
\end{aligned}$$

Hence, we can equivalently write the above as

$$\begin{aligned}
&\max_{\boldsymbol{q}\in\{\underline{\boldsymbol{q}},\bar{\boldsymbol{q}}\}} \left\{\sum_{i\in[2]} q_i \sup_{\mathbb{Q}_i\in\mathcal{F}_i(\hat{\mathbb{P}}_i),i\in[2]} \mathbb{E}_{\mathbb{Q}_i}\left[\left(\left(\ell(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta})) - \kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]\right\} \\
&= \max_{\boldsymbol{q}\in\{\underline{\boldsymbol{q}},\bar{\boldsymbol{q}}\}} \left\{\sum_{i\in[2]} q_i \frac{1}{S_i}\sum_{s\in[S_i]}\left(\sup_{\boldsymbol{z}_s\in\mathcal{Z}_i}\left\{\ell\left(b_i(\alpha+\boldsymbol{z}_s^\top\boldsymbol{\beta})\right) - \kappa\|\boldsymbol{z}_s-\hat{\boldsymbol{z}}_{i,s}\|\right\}\right)^p\right\}.
\end{aligned}$$

Hence, Problem (24) admits the equivalent formulation as stated in the Theorem.

Now, suppose $\mathcal{Z}_i = \mathbb{R}^N$. For $i \in \{1,2\}$, note that $\|b_i\boldsymbol{\beta}\| = \|\boldsymbol{\beta}\|$ because $|b_i| = 1$. By a similar argument as in Theorem 2, we have

$$\sup_{\boldsymbol{z}_s \in \mathcal{Z}_i} \left\{ \ell\left(b_i(\alpha + \boldsymbol{z}_s^\top \boldsymbol{\beta})\right) - \kappa \|\boldsymbol{z}_s - \hat{\boldsymbol{z}}_{i,s}\| \right\} = \begin{cases} \ell(b_i(\alpha + \hat{\boldsymbol{z}}_{i,s}^\top \boldsymbol{\beta})) & \text{if } L\|\boldsymbol{\beta}\|_* \leq \kappa \\ \infty & \text{otherwise.} \end{cases}$$

Hence,

$$\begin{aligned} &\max_{\boldsymbol{q} \in \{\underline{\boldsymbol{q}}, \bar{\boldsymbol{q}}\}} \left\{ \sum_{i \in [2]} q_i \frac{1}{S_i} \sum_{s \in [S_i]} \left(\sup_{\boldsymbol{z}_s \in \mathcal{Z}_i} \left\{ \ell\left(b_i(\alpha + \boldsymbol{z}_s^\top \boldsymbol{\beta})\right) - \kappa \|\boldsymbol{z}_s - \hat{\boldsymbol{z}}_{i,s}\| \right\} \right)^p \right\} \\ &= \begin{cases} \max\limits_{\boldsymbol{q} \in \{\underline{\boldsymbol{q}}, \bar{\boldsymbol{q}}\}} \left\{ \sum\limits_{i \in [2]} q_i \frac{1}{S_i} \sum\limits_{s \in [S_i]} \ell^p(b_i(\alpha + \hat{\boldsymbol{z}}_{i,s}^\top \boldsymbol{\beta})) \right\} & \text{if } L\|\boldsymbol{\beta}\|_* \leq \kappa \\ \infty & \text{otherwise.} \end{cases} \end{aligned}$$

By above results, the final reformulation follows. □

Proof of Theorem 11. Observe that

$$\sum_{i \in [2]} q_i \mathbb{E}_{\mathbb{Q}_i} \left[\left(\left(\ell\left(b_i\left(\alpha + \tilde{\boldsymbol{z}}_i^\top \boldsymbol{\beta}\right)\right) - \kappa \|\tilde{\boldsymbol{z}}_i - \tilde{\boldsymbol{v}}_i\| \right)^+ \right)^p \right] \leq T^p \qquad \forall \boldsymbol{q} \in \mathcal{Q}, \mathbb{Q}_i \in \mathcal{F}_i(\hat{\mathbb{P}}_i), i \in [2],$$

implies

$$\sum_{i \in [2]} q_i^* \mathbb{E}_{\mathbb{Q}_i} \left[\left(\left(\ell\left(b_i(\alpha + \tilde{\boldsymbol{z}}^\top \boldsymbol{\beta})\right) - \kappa \|\tilde{\boldsymbol{z}} - \tilde{\boldsymbol{v}}\| \right)^+ \right)^p \right] \leq T^p \qquad \forall \mathbb{Q}_i \in \mathcal{F}_i(\mathbb{P}_i^*, \hat{\mathbb{P}}_i),$$

where

$$\mathcal{F}_i(\mathbb{P}^*, \hat{\mathbb{P}}) \triangleq \left\{ \mathbb{Q} \in \mathcal{P}_0(\mathcal{Z}_i^2) \;\middle|\; (\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{v}}) \sim \mathbb{Q}, \tilde{\boldsymbol{z}} \sim \mathbb{P}^*, \tilde{\boldsymbol{v}} \sim \hat{\mathbb{P}} \right\}.$$

Equivalently, we have

$$\left(\sum_{i \in [2]} q_i^* \mathbb{E}_{\mathbb{Q}_i} \left[\left(\left(\ell\left(b_i(\alpha + \tilde{\boldsymbol{z}}^\top \boldsymbol{\beta})\right) - \kappa \|\tilde{\boldsymbol{z}} - \tilde{\boldsymbol{v}}\| \right)^+ \right)^p \right] \right)^{\frac{1}{p}} \leq T \qquad \forall \mathbb{Q}_i \in \mathcal{F}_i(\mathbb{P}_i^*, \hat{\mathbb{P}}_i), i \in [2].$$

By the concavity of $(\cdot)^{\frac{1}{p}}$ for $p \geq 1$ and non-negative input, we have

$$\begin{aligned} &\left(\sum_{i \in [2]} q_i^* \mathbb{E}_{\mathbb{Q}_i} \left[\left(\left(\ell\left(b_i(\alpha + \tilde{\boldsymbol{z}}^\top \boldsymbol{\beta})\right) - \kappa \|\tilde{\boldsymbol{z}} - \tilde{\boldsymbol{v}}\| \right)^+ \right)^p \right] \right)^{\frac{1}{p}} \\ &\geq \sum_{i \in [2]} q_i^* \left(\mathbb{E}_{\mathbb{Q}_i} \left[\left(\left(\ell\left(b_i(\alpha + \tilde{\boldsymbol{z}}^\top \boldsymbol{\beta})\right) - \kappa \|\tilde{\boldsymbol{z}} - \tilde{\boldsymbol{v}}\| \right)^+ \right)^p \right] \right)^{\frac{1}{p}}. \end{aligned}$$

By the convexity and positive homogeneity properties of $\left(\mathbb{E}_{\mathbb{Q}}\left[((\cdot)^+)^p\right]\right)^{1/p}$ for $p \geq 1$, we have subadditivity, and hence,

$$\begin{aligned}
&\mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell\left(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta})\right)\right)^+\right)^p\right]^{\frac{1}{p}} \\
&= \mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell\left(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta})\right) - \kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\| + \kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]^{\frac{1}{p}} \\
&\leq \mathbb{E}_{\mathbb{Q}}\left[\left(\left(\ell\left(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta})\right) - \kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]^{\frac{1}{p}} + \mathbb{E}_{\mathbb{Q}}\left[\left(\left(\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]^{\frac{1}{p}}.
\end{aligned}$$

Then, for any $\mathbb{Q}_1, \mathbb{Q}_2$, we have

$$\begin{aligned}
&\left(\sum_{i\in[2]} q_i^* \mathbb{E}_{\mathbb{Q}_i}\left[\left(\left(\ell\left(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta})\right) - \kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]\right)^{\frac{1}{p}} \\
&\geq \sum_{i\in[2]} q_i^* \left(\mathbb{E}_{\mathbb{Q}_i}\left[\left(\left(\ell\left(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta})\right) - \kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]\right)^{\frac{1}{p}} \\
&\geq \sum_{i\in[2]} q_i^* \left(\mathbb{E}_{\mathbb{Q}_i}\left[\left(\left(\ell\left(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta})\right)\right)^+\right)^p\right]^{\frac{1}{p}} - \mathbb{E}_{\mathbb{Q}_i}\left[\left(\left(\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]^{\frac{1}{p}}\right).
\end{aligned}$$

For any $i \in [2]$:

$$\begin{aligned}
&\sup_{\mathbb{Q}_i\in\mathcal{F}_i(\mathbb{P}_i^*,\hat{\mathbb{P}}_i)} \left\{\mathbb{E}_{\mathbb{Q}_i}\left[\left(\left(\ell\left(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta})\right)\right)^+\right)^p\right]^{\frac{1}{p}} - \mathbb{E}_{\mathbb{Q}_i}\left[\left(\left(\kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]^{\frac{1}{p}}\right\} \\
&= \mathbb{E}_{\mathbb{P}_i^*}\left[\ell^p\left(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta})\right)\right]^{\frac{1}{p}} - \kappa \inf_{\mathbb{Q}_i\in\mathcal{F}_i(\mathbb{P}_i^*,\hat{\mathbb{P}}_i)} \mathbb{E}_{\mathbb{Q}_i}\left[\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|^p\right]^{\frac{1}{p}} \\
&= \mathbb{E}_{\mathbb{P}_i^*}\left[\ell^p\left(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta})\right)\right]^{\frac{1}{p}} - \kappa\Delta_p(\mathbb{P}_i^*,\hat{\mathbb{P}}_i).
\end{aligned}$$

Combining above results, we have

$$\begin{aligned}
&\left(\sum_{i\in[2]} q_i^* \mathbb{E}_{\mathbb{Q}_i}\left[\left(\left(\ell\left(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta})\right) - \kappa\|\tilde{\boldsymbol{z}}-\tilde{\boldsymbol{v}}\|\right)^+\right)^p\right]\right)^{\frac{1}{p}} \leq T \qquad \forall \mathbb{Q}_i \in \mathcal{F}_i(\mathbb{P}_i^*,\hat{\mathbb{P}}_i), i\in[2] \\
\implies &\sum_{i\in[2]} q_i^* \left(\mathbb{E}_{\mathbb{P}_i^*}\left[\ell^p\left(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta})\right)\right]^{\frac{1}{p}} - \kappa\Delta_p(\mathbb{P}_i^*,\hat{\mathbb{P}}_i)\right) \leq T \\
\iff &\sum_{i\in[2]} q_i^* \left(\mathbb{E}_{\mathbb{P}_i^*}\left[\ell^p\left(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta})\right)\right]^{\frac{1}{p}}\right) - T \leq \kappa \sum_{i\in[2]} q_i^*\Delta_p(\mathbb{P}_i^*,\hat{\mathbb{P}}_i)
\end{aligned}$$

Hence, for all $r \geq 0$,

$$\begin{aligned}
\mathbb{P}^S\left[\sum_{i\in[2]} q_i^* \left(\mathbb{E}_{\mathbb{P}_i^*}\left[\ell^p\left(b_i(\alpha+\tilde{\boldsymbol{z}}^\top\boldsymbol{\beta})\right)\right]^{\frac{1}{p}}\right) - T > \kappa r\right] &\leq \mathbb{P}^S\left[\sum_{i\in[2]} q_i^*\Delta_p(\mathbb{P}_i^*,\hat{\mathbb{P}}_i) > r\right] \\
&\leq \mathbb{P}^S\left[\max_{i\in[2]}\{\Delta_p(\mathbb{P}_i^*,\hat{\mathbb{P}}_i)\} > r\right]. \\
&\leq \sum_{i\in[2]} \mathbb{P}^S\left[\Delta_p(\mathbb{P}_i^*,\hat{\mathbb{P}}_i) > r\right]
\end{aligned}$$

□

B. Additional Simulation Results for Linear Regression

We include some additional simulation results on real-world datasets below.

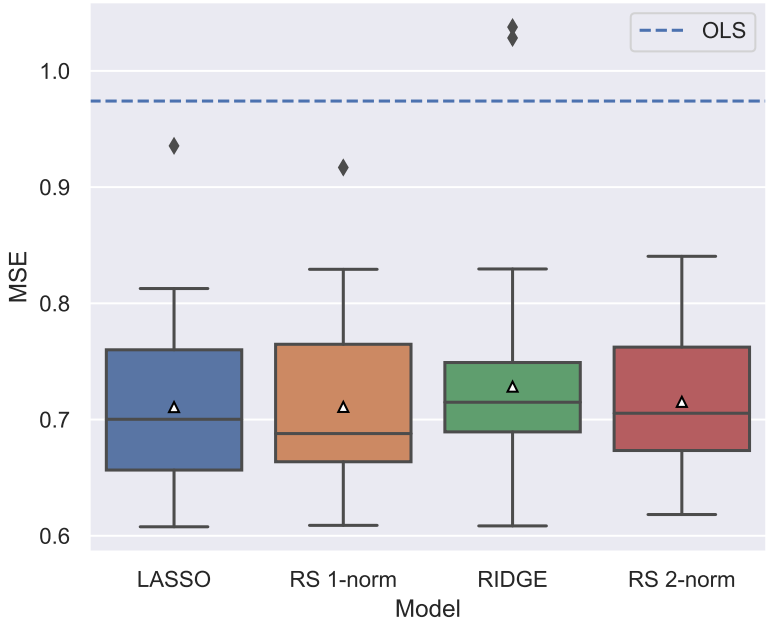

Figure 7 **Out-of-sample performance comparison: White wine quality ($s = 3$).**

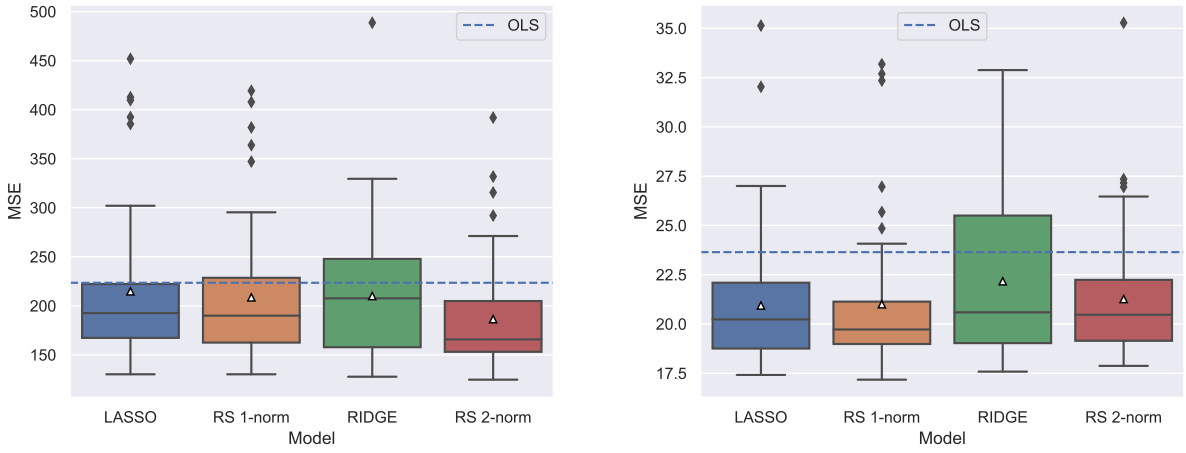

Figure 8 **Out-of-sample performance comparison: Concrete (left) and Auto MPG (right).**

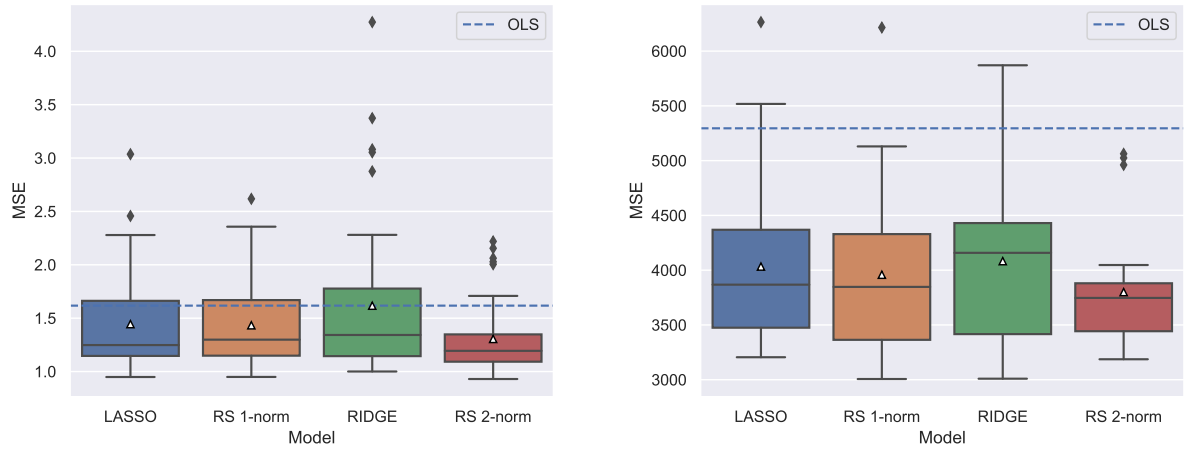

Figure 9 **Out-of-sample performance comparison: Fish toxicity (left) and diabetes (right).**