Model-Uncertainty-Aware Residuals-Based Sample Average Approximation

Man Yiu Tsang, Karmel S. Shehadeh, Güzin Bayraksan

Abstract

We consider a contextual stochastic optimization (CSO) problem, where one has observations of the uncertain parameters together with concurrent observations of covariates, and the goal is to choose decisions that minimize expected cost conditioned on new covariate observations. The empirical residuals-based sample average approximation (ER-SAA) of the CSO problem constructs scenarios of uncertainty by combining regression predictions with residuals. This approach implicitly assumes that the regression model captures the relationship among covariates, uncertain factors, and decisions with reasonable accuracy. However, since this relationship is fundamentally unknown, identifying a suitable regression model to approximate it remains a central challenge, commonly referred to as *model uncertainty*. Overlooking such uncertainty can yield poor approximations and suboptimal decisions. To address this, we introduce a model-uncertainty–aware residuals-based SAA framework (MUR-SAA), formulated as a distributionally robust optimization problem on the regression model. By treating the regression function as a random element and optimizing over the worst-case expected cost across all plausible distributions of the unknown regression function, MUR-SAA produces solutions that are robust to model misspecification. We establish, for the first time, efficient mechanisms for quantifying the stability of general residuals-based stochastic programs and our MUR-SAA model. Moreover, we identify conditions under which our MUR-SAA model enjoys desirable asymptotic convergence and finite-sample properties. We derive equivalent tractable reformulations of MUR-SAA and develop a decomposition algorithm to solve them. We illustrate our theoretical results and demonstrate the advantages of our MUR-SAA model using a contextual portfolio optimization problem.

Keywords: Contextual stochastic optimization; Model uncertainty; Distributionally robust optimization; Stability analysis

1. Introduction

Consider the following contextual stochastic optimization (CSO) problem

$$v^\star(\boldsymbol{z}) := \inf_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^\star}[c(\boldsymbol{x},\boldsymbol{\xi}) \mid \boldsymbol{Z} = \boldsymbol{z}], \tag{1}$$

Man Yiu Tsang, Department of Industrial, Manufacturing, and Systems Engineering, Texas Tech University
Karmel S. Shehadeh (corresponding author), Daniel J. Epstein Department of Industrial and Systems Engineering, University of Southern California
Güzin Bayraksan, Integrated Systems Engineering Department, Ohio State University

where $\boldsymbol{x} \in \mathbb{R}^{d_x}$ is a vector of first-stage decisions, $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ is a set of deterministic constraints on $\boldsymbol{x}$, $\boldsymbol{\xi}$ is a random vector that contains the uncertain quantities of interest with a closed convex support $\Xi \subseteq \mathbb{R}^{d_\xi}$, $\boldsymbol{Z}$ is a random vector of covariates (also known as side or contextual information) that have some predictive power on $\boldsymbol{\xi}$ with a compact support $\mathcal{Z} \subseteq \mathbb{R}^{d_z}$, $\mathbb{P}^\star$ is the true distribution of $\boldsymbol{\xi}$, and c is the cost of the system for a given decision $\boldsymbol{x} \in \mathcal{X}$ and a realization of $\boldsymbol{\xi} \in \Xi$. Problem (1) finds a decision $\boldsymbol{x}$ that minimizes the expected cost given the observed covariates $\boldsymbol{z}$. We define the set of optimal solutions to problem (1) corresponding to the covariate observation $\boldsymbol{Z} = \boldsymbol{z}$ as $\mathcal{X}^\star(\boldsymbol{z}) := \arg\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}^\star}[c(\boldsymbol{x},\boldsymbol{\xi}) \mid \boldsymbol{Z} = \boldsymbol{z}]$.

We consider settings in which the decision-maker has access to a set of N historical observations $\{\widehat{\boldsymbol{\xi}}^1, \ldots, \widehat{\boldsymbol{\xi}}^N\}$ of $\boldsymbol{\xi}$ and together with concurrent observations of covariates $\{\boldsymbol{z}^1, \ldots, \boldsymbol{z}^N\}$ that may carry predictive information about $\boldsymbol{\xi}$. For example, in vehicle routing, one may collect historical travel-time observations, along with contextual features such as weather conditions and/or time of day, to improve predictions. Throughout the paper, we use $\mathcal{D}_N := \{(\boldsymbol{z}^i, \widehat{\boldsymbol{\xi}}^i)\}_{i=1}^N$ to represent the set of available data.

Recently, there has been growing interest in CSO approaches, which integrate machine learning and optimization techniques to approximate solutions to the CSO problem (1) using the dataset $\mathcal{D}_N$. The recent surveys by Qi and Shen (2022) and Sadana et al. (2025) provide comprehensive overviews of the CSO literature. One stream of this literature employs the decision rule approach, in which a parametrized class of mappings—referred to as decision rules (e.g., linear models, neural networks)—is used, and the parameters are learned by minimizing the in-sample cost over $\mathcal{D}_N$ (Ban and Rudin, 2019; Bertsimas and Koduri, 2022; Oroojlooyjadid et al., 2020). Another stream focuses on integrating learning and optimization (ILO). Specifically, ILO is an end-to-end framework that includes a prediction model (which maps the covariate to a predicted distribution), an optimization model that takes as input a prediction and returns a decision, and a task-based loss function that captures the downstream optimization problem; see, e.g., Costa and Iyengar (2023); Donti et al. (2017); Elmachtoub and Grigas (2022); Muñoz et al. (2022); Qi et al. (2025). In particular, instead of minimizing the (in-sample) estimation error, parameters of the prediction model of the conditional distribution are trained to maximize the prescriptive performance in the downstream optimization task.

Another notable approach is sequential learning and optimization (SLO), which has received significant attention in recent years. In contrast to the ILO paradigm, SLO first approximates the conditional expectation in the CSO problem (1) using a predictive model trained on the available data $\mathcal{D}_N$, and then solves the resulting optimization problem to determine the decisions (Bertsimas and Kallus, 2020; Bertsimas and Van Parys, 2022; Chen and Paschalidis, 2019; Deng and Sen, 2018; Kannan et al., 2025). In this paper, we address one of the fundamental challenges associated with a popular class of sequential methods—namely, residual-based approaches (Ban et al., 2019; Deng and Sen, 2018, 2022; Kannan et al., 2024, 2025; Zhu et al., 2024)—which rely on estimating conditional

distributions from regression residuals. Specifically, these methods first estimate the structural relationship between $\boldsymbol{\xi}$ and $\boldsymbol{z}$ by performing regression over a pre-specified model class $\mathcal{F}$. The resulting estimated regression function $\widehat{f}_N \in \mathcal{F}$ and its associated residuals $\{\widehat{\boldsymbol{\varepsilon}}_N^i := \widehat{\boldsymbol{\xi}}^i - \widehat{f}_N(\boldsymbol{z}^i)\}_{i=1}^N$ are then used to approximate the conditional expectation in the CSO problem (1). In particular, the empirical residuals-based sample average approximation (ER-SAA) approach replaces the true conditional distribution in the CSO problem (1) with an empirical distribution constructed from scenarios $\{\text{proj}_{\Xi}(\widehat{f}_N(\boldsymbol{z}) + \widehat{\boldsymbol{\varepsilon}}_N^i)\}_{i=1}^N$, where $\text{proj}_{\Xi}(\cdot)$ is the orthogonal projection onto Ξ.

Despite its nice theoretical properties, the ER-SAA approach suffers from the following shortcomings. First, the quality and performance of optimal solutions to the ER-SAA problem depend on the accuracy of the estimated regression function $\widehat{f}_N$, where inaccuracies may arise due to limited data, noise, or overfitting. Observe that obtaining accurate estimators is challenging, and this could happen even if, for example, the model class $\mathcal{F}$ is correctly specified. For instance, the standard errors of estimated regression coefficients tend to be larger with smaller sample sizes. As another example, the dataset $\mathcal{D}_N$ may be small relative to the complexity of the regression method or inadequate to provide an accurate estimate of the regression function. In particular, in high-dimensional settings where the number of predictors exceeds the number of samples, overfitting becomes a significant concern (see, e.g., Babyak, 2004; McNeish, 2015). Some studies have also shown that overfitting can persist even in low-dimensional settings (Subramanian and Simon, 2013). In such cases, the estimated regression function $\widehat{f}_N$ may be optimistically biased and generalize poorly, i.e., yield inaccurate predictions under unseen data. Furthermore, overfitting results in smaller residuals, leading to scenarios that do not fully capture the uncertainty in the problem; see Yurdakul and Bayraksan (2024) for a discussion of drawbacks associated with overfitting in ER-SAA for a real-world power systems application.

Second, the ER-SAA approach implicitly assumes that the chosen regression model correctly captures the true (unknown) structural relationship between the uncertain vector $\boldsymbol{\xi}$ and covariates $\boldsymbol{z}$; that is, it assumes, for example, the model class $\mathcal{F}$ is correctly specified. However, it is well known that this relationship contains aspects that are not known with certainty (Chatfield, 1995; Draper, 1995; Fan and Lv, 2010). Prior studies have demonstrated significant variations in prediction performance when using different training methods (e.g., different function class $\mathcal{F}$ and loss function ℓ), in particular when the sample size is small and/or under sparse regression settings (Bertsimas et al., 2020; Fan et al., 2024; Hastie et al., 2020; Riley et al., 2021; Sarwar et al., 2020; Wang et al., 2020; Zhang et al., 2021). Such uncertainty will typically remain even after cross-validation. Moreover, with small samples of data—precisely when structural uncertainty is most significant—cross-validation may not be feasible because there are too few data points with which to carry out the modeling and validation steps in a stable manner. Thus, selecting a suitable regression model to approximate the relationship between $\boldsymbol{\xi}$ and $\boldsymbol{z}$ remains a key challenge, known as "*model uncertainty*". Note also that incorporating more data does not necessarily eliminate model uncertainty, especially

when the functional form of the regression function is misspecified.

While several studies have addressed the first shortcoming by proposing approaches to mitigate estimation errors in approximating the true conditional distribution, the fundamental issue of model uncertainty has, to our knowledge, not been explicitly addressed within the ER-SAA framework. For example, Kannan et al. (2025) introduced a jackknife variant of the ER-SAA approach that leverages leave-one-out residuals to reduce overfitting in small samples. Kannan et al. (2024) developed a distributionally robust optimization (DRO) extension of the ER-SAA approach, denoted as ER-DRO, that accounts for approximation errors in both the regression function $\widehat{f}_N$ and its residuals $\{\widehat{\boldsymbol{\varepsilon}}_N^i\}_{i=1}^N$. While the ER-DRO framework can partially hedge against misspecification of the regression function by robustifying the empirical conditional distribution, it does not fully hedge against model uncertainty. Dou and Anitescu (2019) studied a class of the CSO problem in which the uncertainty $\boldsymbol{\xi}$ follows a vector autoregression (VAR) model, and proposed a DRO model that addresses parameter estimation errors and distributional ambiguity in the noise term of the VAR model. However, their approach does not account for the potential structural misspecification of the VAR model itself. Zhu et al. (2022) proposed a joint estimation and robustness optimization model to mitigate estimation uncertainty. Their model incorporates an uncertainty set, whose size depends on the statistical accuracy of the employed estimation procedure. However, this approach lacks desirable asymptotic and finite-sample guarantees and is primarily designed for deterministic optimization models. More recently, Sim et al. (2025) examined parameter estimation uncertainty within the contextual robust satisficing framework. Specifically, they propose a predict–optimize–satisfice–fortify framework that integrates robust satisficing principles to derive decisions under both distributional ambiguity and parameter uncertainty, and established probabilistic confidence bounds on the gap between the true expected cost of the implemented solution and a target benchmark.

A key limitation shared by these approaches is that they treat the regression model as fixed after estimation and account only for uncertainty in parameters, residuals, or induced distributions conditional on that model. As a result, they do not capture uncertainty arising from misspecification of the model class or variability across competing regression models. Consequently, even robustified residuals-based approaches may yield decisions that are sensitive to the choice of the regression model. Addressing this limitation requires a framework that explicitly incorporates uncertainty in the regression function itself, rather than conditioning on a single estimated model. To this end, we introduce and analyze a new model-uncertainty-aware residuals-based SAA (MUR-SAA) framework.

1.1. Summary of main contributions

The following summarizes the main contributions of this paper:

- We introduce a new MUR-SAA approach that protects against model uncertainty within the ER-SAA framework. We formulate our MUR-SAA model as a DRO problem on the regression model. Specifically, we treat the unknown regression function f as a random element governed by

an unknown distribution and construct ambiguity sets over this distribution. This distributional perspective enables our framework to capture structural uncertainty across diverse model classes. As a result, MUR-SAA simultaneously accounts for parameter estimation error and model misspecification, yielding robust decisions even when the underlying structural relationship between $\boldsymbol{\xi}$ and $\boldsymbol{z}$ is unknown.

- We conduct quantitative stability analysis for the general residuals-based stochastic programming (R-SP) and the ER-SAA models, characterizing how the choice of regression function and residual distribution affects optimal values and solution sets for these models. Specifically, we establish, for the first time, upper bounds that provide mechanisms for quantifying these effects. Building on this analysis, we further examine how perturbations in the ambiguity set governing the distribution of the regression function influence the optimal value and solutions of the MUR-SAA model.
- We identify conditions under which the MUR-SAA exhibits desirable asymptotic convergence and finite-sample properties. Specifically, we establish that, under mild assumptions, the optimal value and the set of optimal solutions converge to their true counterparts in the CSO problem (1) as the sample size $N \to \infty$. Furthermore, we derive a nontrivial lower bound on the probability that solutions obtained from the MUR-SAA model are near-optimal, thereby providing finite-sample performance guarantees.
- We analyze the computational tractability of the MUR-SAA models and show that, under standard modeling assumptions, they admit equivalent reformulations with a robust optimization structure that is amenable to decomposition methods.
- We illustrate our theoretical results through a contextual portfolio optimization model. Our findings demonstrate the effectiveness of our MUR-SAA model in producing robust solutions under model uncertainty in both limited- and rich-data settings. We also provide insights into the relative performance of the MUR-SAA, ER-SAA, empirical residuals-based model averaging (ER-MA), and ER-DRO models. Notably, the results indicate that MUR-SAA is a strong alternative to the ER-DRO framework in limited-data settings, often outperforming it. It also produces solutions with superior out-of-sample performance compared to both ER-DRO and ER-SAA in relatively data-rich settings.

1.2. Structure of the Paper

The remainder of the paper is organized as follows. In Section 2, we introduce our MUR-SAA model. In Section 3, we present our quantitative stability analyses. In Section 4, we study the asymptotic convergence and finite-sample guarantees of the MUR-SAA model. In Section 5, we analyze the computational tractability of our MUR-SAA models. Then, in Section 6, we present our numerical experiments demonstrating the potential benefits of our MUR-SAA model. We conclude and discuss future directions in Section 7.

1.3. Notation

We introduce notation used throughout the paper; additional notation is defined where needed. For any $N \in \mathbb{N}$, we define the set $[N] := \{1, 2, \ldots, N\}$. The symbol $\|\cdot\|$ denotes a generic norm defined on a vector space unless specified otherwise. In the Euclidean space $\mathbb{R}^n$, boldface letters (e.g., $\boldsymbol{s}$) represent column vectors. The vectors $\mathbf{1}$ and $\mathbf{0}$ denote vectors of ones and zeros, respectively. We define $d(\boldsymbol{s}, \mathcal{S}) = \inf_{\boldsymbol{s}' \in \mathcal{S}} \|\boldsymbol{s} - \boldsymbol{s}'\|$ as the distance between a point $\boldsymbol{s} \in \mathbb{R}^n$ and a set $\mathcal{S} \subseteq \mathbb{R}^n$ and $D(\mathcal{S}_1, \mathcal{S}_2) = \sup_{\boldsymbol{s}_1 \in \mathcal{S}_1} \inf_{\boldsymbol{s}_2 \in \mathcal{S}_2} \|\boldsymbol{s}_1 - \boldsymbol{s}_2\|$ as the distance between two sets $\mathcal{S}_1 \subseteq \mathbb{R}^n$ and $\mathcal{S}_2 \subseteq \mathbb{R}^n$. For a vector-valued function $f : \mathcal{S}_1 \to \mathcal{S}_2$ mapping from vector space $\mathcal{S}_1$ to $\mathcal{S}_2$, its supremum norm is $\|f\|_\infty = \sup_{\boldsymbol{s} \in \mathcal{S}_1} \|f(\boldsymbol{s})\|$. We write "$\xrightarrow{p}$" for convergence in probability (with respect to the true data-generating distribution) as the sample size N tends to infinity. The abbreviations "a.s." and "a.e." are shorthand for "almost surely" and "almost everywhere," with the underlying probability measure clear from context. We denote by Δ_M the probability simplex in $\mathbb{R}^M$. For any $a > 0$, we adopt the convention $a/0 = \infty$.

Let (Ω, d_Ω) be a Polish (metric) space, where Ω is a general vector space and d_Ω is a continuous metric on Ω. Denote $\mathcal{P}(\Omega)$ as the set of probability measures $\mathbb{P}_\Omega$ defined on the measurable space $(\Omega, \mathcal{B}_\Omega)$, where $\mathcal{B}_\Omega$ is the Borel σ-field of Ω; see Garling (2018) and Parthasarathy (2005) for details on probability measures on Polish spaces. For any $\omega \in \Omega$, we denote by δ_ω the Dirac measure on $\{\omega\}$. Let $\mathcal{G}$ be a set of measurable functions defined on $(\Omega, \mathcal{B}_\Omega)$. For any two probability measures $\mathbb{P} \in \mathcal{P}(\Omega)$ and $\mathbb{Q} \in \mathcal{P}(\Omega)$, define the *integral probability metric* (a.k.a. metric with a ζ-structure) as

$$\mathsf{d}_{\mathcal{G}}(\mathbb{P}, \mathbb{Q}) := \sup_{g \in \mathcal{G}} \left| \mathbb{E}_{\mathbb{P}}[g(\omega)] - \mathbb{E}_{\mathbb{Q}}[g(\omega)] \right|. \tag{2}$$

If we use the following set of measurable functions in (2):

$$\mathcal{G}_{\mathrm{K}} := \left\{ g : \Omega \to \mathbb{R} \;\middle|\; g \text{ is Lipschitz continuous with modulus } L \leq 1 \right\}, \tag{3}$$

then the resulting metric, denoted as d_{K}, is called the Kantorovich metric. For $p \in [1, \infty)$, the p-Wasserstein distance between $\mathbb{P}$ and $\mathbb{Q}$ is defined as

$$\mathsf{d}_{W,p}(\mathbb{P}, \mathbb{Q}) := \left(\inf_{\pi \in \Pi} \int_{\Omega \times \Omega} d_\Omega(\omega, \omega')^p \, \pi(\mathrm{d}\omega, \mathrm{d}\omega') \right)^{\frac{1}{p}},$$

where Π is the set of all joint probability measures of (ω, ω') over $\Omega \times \Omega$ with marginals $\mathbb{P}$ and $\mathbb{Q}$. By the Kantorovich-Rubinstein duality, d_{K} equals the Wasserstein metric with $p = 1$, i.e., $\mathsf{d}_{\mathrm{K}} = \mathsf{d}_{W,1}$. For a given metric $\mathsf{d}_{\mathcal{G}}$ and two sets of probability measures $\mathcal{P}_1 \subseteq \mathcal{P}(\Omega)$ and $\mathcal{P}_2 \subseteq \mathcal{P}(\Omega)$, we define the *Hausdorff distance* as $\mathbb{H}(\mathcal{P}_1, \mathcal{P}_2; \mathsf{d}_{\mathcal{G}}) := \max\left\{ \sup_{\mathbb{P}_1 \in \mathcal{P}_1} \inf_{\mathbb{P}_2 \in \mathcal{P}_2} \mathsf{d}_{\mathcal{G}}(\mathbb{P}_1, \mathbb{P}_2), \sup_{\mathbb{P}_2 \in \mathcal{P}_2} \inf_{\mathbb{P}_1 \in \mathcal{P}_1} \mathsf{d}_{\mathcal{G}}(\mathbb{P}_1, \mathbb{P}_2) \right\}$.

2. The MUR-SAA Model

In this section, we formally introduce our proposed MUR-SAA model designed to mitigate model uncertainty. To facilitate the subsequent technical discussion, we begin by introducing the generic residuals-based stochastic programming (R-SP) model and the ER-SAA model in Section 2.1. Then, we present our proposed MUR-SAA model in Section 2.2.

2.1. The R-SP and ER-SAA Formulations

In the residuals-based approach, one assumes that the true structural relationship between $\boldsymbol{\xi}$ and $\boldsymbol{z}$ takes the form $\boldsymbol{\xi} = f^\star(\boldsymbol{z}) + \boldsymbol{\varepsilon}$, where $f^\star(\boldsymbol{z}) = \mathbb{E}_{\mathbb{P}^\star}[\boldsymbol{\xi} \mid \boldsymbol{Z} = \boldsymbol{z}]$ is the (true) regression function and $\boldsymbol{\varepsilon}$ is a zero-mean random error with the (true) distribution $\mathbb{P}^\star_{\boldsymbol{\varepsilon}}$ that is independent of $\boldsymbol{Z}$[1]. Under this setting, the CSO problem (1) for a given $\boldsymbol{z} \in \mathcal{Z}$ is equivalent to the following R-SP problem

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \Big\{ h(f, \mathbb{P}_{\boldsymbol{\varepsilon}}; \boldsymbol{x}) := \mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}}\big[c\big(\boldsymbol{x}, \mathrm{proj}_\Xi(f(\boldsymbol{z}) + \boldsymbol{\varepsilon})\big)\big] \Big\} \tag{4}$$

with $f = f^\star$ and $\mathbb{P}_{\boldsymbol{\varepsilon}} = \mathbb{P}^\star_{\boldsymbol{\varepsilon}}$. The expectation in (4) is computed with respect to the distribution $\mathbb{P}_{\boldsymbol{\varepsilon}}$ of $\boldsymbol{\varepsilon}$. We suppress the dependence of h on $\boldsymbol{z}$ for notational simplicity. In the R-SP model (4), the true relationship between the random vector $\boldsymbol{\xi}$ and the covariates $\boldsymbol{z}$ is given by $\boldsymbol{\xi} = \mathrm{proj}_\Xi\big(f(\boldsymbol{z}) + \boldsymbol{\varepsilon}\big)$. Consequently, the distribution of $\boldsymbol{\xi}$ is characterized by the regression function $f \in \mathcal{F}$ and the residual distribution $\mathbb{P}_{\boldsymbol{\varepsilon}}$.

As mentioned earlier, true regression function $f^\star$ is generally unknown. Hence, studies often obtain an estimator $\widehat{f}_N$ of $f^\star$ using a regression method applied to the data $\mathcal{D}_N = \{(\boldsymbol{z}^i, \widehat{\boldsymbol{\xi}}^i)\}_{i=1}^N$. In particular, given data $\mathcal{D}_N$, one may estimate $f^\star$, for instance, using an M-estimator of the form

$$\widehat{f}_N \in \arg\min_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^N \ell(\widehat{\boldsymbol{\xi}}^i, f(\boldsymbol{z}^i)), \tag{5}$$

where ℓ is a loss function and $\mathcal{F}$ is a class of regression functions. A popular example of the set $\mathcal{F}$ is the set of linear regression functions, i.e., $\mathcal{F} = \{\boldsymbol{\theta} \in \mathbb{R}^{d_\xi}, \boldsymbol{\Theta} \in \mathbb{R}^{d_\xi \times d_z} \mid f(\boldsymbol{z}) = \boldsymbol{\theta} + \boldsymbol{\Theta}\boldsymbol{z}\}$, although other function classes and nonparametric models can be used. The estimator $\widehat{f}_N$ and the empirical residuals $\{\widehat{\boldsymbol{\varepsilon}}^i_N := \widehat{\boldsymbol{\xi}}^i - \widehat{f}_N(\boldsymbol{z}^i)\}_{i=1}^N$ are then used to approximate a solution to (1) by solving the following ER-SAA model

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{N}\sum_{i=1}^N c\big(\boldsymbol{x}, \mathrm{proj}_\Xi\big(\widehat{f}_N(\boldsymbol{z}) + \widehat{\boldsymbol{\varepsilon}}^i_N\big)\big) = \inf_{\boldsymbol{x}\in\mathcal{X}} \frac{1}{N}\sum_{i=1}^N c\big(\boldsymbol{x}, \widetilde{\boldsymbol{\xi}}^i(\boldsymbol{z})\big), \tag{6}$$

where $\mathrm{proj}_\Xi(\cdot)$ is the orthogonal projection onto Ξ and $\widetilde{\boldsymbol{\xi}}^i(\boldsymbol{z}) := \mathrm{proj}_\Xi\big(\widehat{f}_N(\boldsymbol{z}) + \widehat{\boldsymbol{\varepsilon}}^i_N\big)$ for all $i \in [N]$. Note that $\{\widehat{f}_N(\boldsymbol{z}) + \widehat{\boldsymbol{\varepsilon}}^i_N\}_{i=1}^N$ may lie outside the support Ξ. The projection operator ensures that the

[1]It is possible to generalize to heteroscedastic cases where the additive error depends on the covariates; see, e.g., Kannan et al. (2025).

constructed scenarios $\{\widetilde{\boldsymbol{\xi}}^i(\boldsymbol{z})\}_{i=1}^N$ belong to Ξ. Observe that the ER-SAA model is a special case of the R-SP model (4) obtained by setting $f = \widehat{f}_N$ and choosing $\mathbb{P}_{\boldsymbol{\varepsilon}}$ as the empirical distribution of the estimated residuals $\{\widehat{\boldsymbol{\varepsilon}}_N^i\}_{i=1}^N$ obtained based on $\widehat{f}_N$.

The R-SP model and similarly the ER-SAA formulation implicitly assume that the chosen regression model $f \in \mathcal{F}$ correctly specifies the structural relationship between $\boldsymbol{\xi}$ and $\boldsymbol{z}$. As discussed in Section 1, this assumption overlooks model uncertainty arising from incomplete knowledge of the true (unknown) structural relationship between $\boldsymbol{\xi}$ and $\boldsymbol{z}$. In the next section, we introduce our MUR-SAA framework, which explicitly addresses model uncertainty.

2.2. The MUR-SAA Model

We are now ready to present our MUR-SAA model, formulated as a general DRO problem over the regression function. Specifically, we model the unknown regression function f as a random element of the set $\mathcal{F}$ governed by an unknown distribution $\mathbb{P}_f$. We define $\mathcal{P}^f$ as an ambiguity set for $\mathbb{P}_f$. Using this notation, our MUR-SAA model can now be stated as follows:

$$\widehat{v}_N^{\mathrm{MUR}}(\boldsymbol{z}) := \inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}_f\in\mathcal{P}^f} \mathbb{E}_{\mathbb{P}_f}\left[\widehat{h}_N(f;\boldsymbol{x}) := \frac{1}{N}\sum_{i=1}^N c\Big(\boldsymbol{x}, \mathrm{proj}_\Xi\left(f(\boldsymbol{z}) + [\widehat{\boldsymbol{\xi}}^i - f(\boldsymbol{z}^i)]\right)\Big)\right]. \tag{7}$$

Here, $\widehat{h}_N$ depends on $\boldsymbol{z}$ and f; for notational simplicity we suppress the explicit dependence on $\boldsymbol{z}$. Throughout the paper, whenever we construct the set $\mathcal{P}^f$ from the data $\mathcal{D}_N$, we denote it by $\mathcal{P}_N^f$ to make this dependence explicit. Unlike the ER-SAA approach, which relies on a single estimator of the regression function, our MUR-SAA model accounts for regression model uncertainty by optimizing over an ambiguity set of distributions $\mathcal{P}^f$ for the regression function, thereby yielding solutions that are robust to model uncertainty.

The set $\mathcal{P}^f$ is a key ingredient of our MUR-SAA formulation. Before we introduce the types of ambiguity set $\mathcal{P}^f$ that we consider in our investigations, we next discuss a data-driven construction based on multiple candidate estimators of the regression function. This approach is motivated by the statistical literature on model selection, which shows that selecting a single model from a large pool of candidates may discard useful information and ignore variability across competing models (Zhang and Liu, 2023). A widely used alternative is *model averaging*, which incorporates information from the set of candidate models by constructing a weighted average of their predictions (Hansen, 2007; Liao et al., 2022; Song et al., 2026). Specifically, suppose there are M estimated regression functions $\{\widehat{f}_{m,N}\}_{m=1}^M$ and let $\widehat{\boldsymbol{w}}_N = (\widehat{w}_{1,N}, \ldots, \widehat{w}_{M,N})^\top \in \Delta_M$ denote reference weights on these estimators. For example, $\widehat{f}_{1,N}$ may be a linear regression model, $\widehat{f}_{2,N}$ a quadratic regression model, and so forth. The weight $\widehat{\boldsymbol{w}}_N$ may be computed using any statistical approach (see, e.g., Chen et al., 2022; Hansen, 2007; Liao et al., 2022). Using $\{\widehat{f}_{m,N}\}_{m=1}^M$ and $\widehat{\boldsymbol{w}}_N$, we construct the empirical distribution of the random regression function f as $\widehat{\mathbb{P}}_{f,N} = \sum_{m=1}^M \widehat{w}_{m,N}\delta_{\widehat{f}_{m,N}}$.

In the remainder of this paper, we consider two MUR ambiguity sets introduced in Defini-

tions 2.1–2.2 below. We construct data-driven variants of these sets using the model-averaging framework described above.

Definition 2.1 (Sample robust MUR ambiguity set)**.** Given $\{\widetilde{f}_m\}_{m=1}^M \subseteq \mathcal{F}$ and $\{r_m\}_{m=1}^M \subseteq \mathbb{R}_+$, we define the sample robust MUR ambiguity set as

$$\mathcal{P}^{\mathrm{SR}} := \left\{ \mathbb{P}_f = \sum_{m=1}^M w_m \delta_{f_m} \;\middle|\; f_m \in \mathcal{F}_m,\ \forall m \in [M] \right\}, \tag{8}$$

where $\mathcal{F}_m = \{f \in \mathcal{F} \mid d_{\mathcal{F}}(f, \widetilde{f}_m) \leq r_m\}$. We refer to this set as a *sample robust* MUR ambiguity set because it resembles a sample robust ambiguity set in traditional DRO literature (cf. Bertsimas et al., 2022). In particular, the set $\mathcal{P}^{\mathrm{SR}}$ permits the perturbation of each of the regression functions f_m within a certain distance of $\widetilde{f}_m$, while keeping the associated weights fixed.

Definition 2.2 (p-Wasserstein MUR ambiguity set)**.** Given $p \in [1, \infty)$ and radius $r \geq 0$, define

$$\mathcal{P}^{W,p} := \left\{ \mathbb{P}_f \in \mathcal{P}(\mathcal{F}) \mid \mathsf{d}_{W,p}(\mathbb{P}_f, \mathbb{Q}_f) \leq r \right\} \tag{9}$$

where $\mathsf{d}_{W,p}$ is the p-Wasserstein distance between two probability measures defined on the metric space $(\mathcal{F}, d_{\mathcal{F}})$ for $p \in [1, \infty)$ (cf. Mohajerin Esfahani and Kuhn, 2018).

The set $\mathcal{P}^{W,p}$ in (9) can be constructed using the model averaging approach described above by setting $\mathbb{Q}_f = \widehat{\mathbb{P}}_{f,N} = \sum_{m=1}^M \widehat{w}_{m,N} \delta_{\widehat{f}_{m,N}}$. Under this specification, the p-Wasserstein MUR ambiguity set is centered at the empirical distribution of the regression estimators. Similarly, a data-driven version of $\mathcal{P}^{\mathrm{SR}}$ in (8), denoted $\mathcal{P}_N^{\mathrm{SR}}$, can be obtained by setting $\widetilde{f}_m = \widehat{f}_{m,N}$, $w_m = \widehat{w}_N$, and $\mathcal{F}_m = \widehat{\mathcal{F}}_{m,N} = \{f \in \mathcal{F} \mid d_{\mathcal{F}}(f, \widehat{f}_{m,N}) \leq r_m\}$ for all $m \in [M]$. Substituting $\mathcal{P}^f = \mathcal{P}_N^{\mathrm{SR}}$ into the MUR-SAA formulation (7) yields

$$\inf_{\boldsymbol{x} \in \mathcal{X}} \sum_{m=1}^M \widehat{w}_{m,N} \sup_{f_m \in \widehat{\mathcal{F}}_{m,N}} \left\{ \frac{1}{N} \sum_{i=1}^N c\Big(\boldsymbol{x}, \mathrm{proj}_{\Xi}\left(f_m(\boldsymbol{z}) + [\widehat{\boldsymbol{\xi}}^i - f_m(\boldsymbol{z}^i)]\right)\Big) \right\}, \tag{10}$$

Formulation (10) resembles the form of multimodal DRO models (see, e.g., Hanasusanto et al., 2015; Yu and Basciftci, 2026), where $\boldsymbol{\xi}$ follows different distributions with certain probabilities. Classical multimodal DRO models, however, do not incorporate contextual information. The MUR-SAA formulation (10) generalizes classical multimodal DRO models by incorporating contextual information via the residuals-based approach.

Observe that when $\mathcal{P}^f = \{\widehat{\mathbb{P}}_{f,N}\}$ is a singleton, the MUR-SAA model reduces to the following *empirical residuals-based model averaging* (ER-MA) model:

$$\widehat{v}_N^{\text{ER-MA}}(\boldsymbol{z}) := \inf_{\boldsymbol{x} \in \mathcal{X}} \sum_{m=1}^M \widehat{w}_{m,N} \left\{ \frac{1}{N} \sum_{i=1}^N c\Big(\boldsymbol{x}, \mathrm{proj}_{\Xi}\left(\widehat{f}_{m,N}(\boldsymbol{z}) + [\widehat{\boldsymbol{\xi}}^i - \widehat{f}_{m,N}(\boldsymbol{z}^i)]\right)\Big) \right\}. \tag{11}$$

When there is only one candidate model (i.e., $M = 1$), the ER-MA model (11) reduces further to the ER-SAA model (6). Hence, both the ER-MA and ER-SAA are special cases of our MUR-SAA model.

Finally, we note that when the model class is parameterized, the MUR-SAA formulation can be written as an optimization problem over distributions on the parameter. Specifically, if $\mathcal{F} = \{f(\boldsymbol{z};\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \mathfrak{T}\}$, problem (7) reduces to

$$\inf_{\boldsymbol{x}\in\mathcal{X}} \sup_{\mathbb{P}_{\boldsymbol{\theta}}\in\mathcal{P}^{\theta}} \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}}\left[\frac{1}{N}\sum_{i=1}^{N} c\Big(\boldsymbol{x}, \mathrm{proj}_{\Xi}\left(f(\boldsymbol{z};\boldsymbol{\theta}) + [\widehat{\boldsymbol{\xi}}^{i} - f(\boldsymbol{z}^{i};\boldsymbol{\theta})]\right)\Big)\right], \tag{12}$$

where $\mathcal{P}^{\theta}$ is an ambiguity set consisting of distributions $\mathbb{P}_{\boldsymbol{\theta}}$ of the random parameter $\boldsymbol{\theta}$. Model (12) resembles the *Bayesian risk optimization model* (Wu et al., 2018). However, the classical Bayesian risk optimization model does not incorporate contextual information. Thus, the MUR-SAA formulation (12) generalizes the classical Bayesian risk optimization model by incorporating contextual information via the residuals-based CSO approach.

3. Quantitative Stability Analysis

As mentioned earlier, the choice of model parameters, particularly the regression function, in residuals-based models significantly affects the models' optimal values and optimal solutions. However, the existing literature lacks effective mechanisms to quantify this impact. To address this fundamental knowledge gap, we conduct a quantitative stability analysis in this section for the R-SP model and our MUR-SAA model. Our results extend the existing quantitative stability results (e.g., Rachev and Römisch, 2002; Römisch and Schultz, 1991) in the stochastic optimization literature for the R-SP models. In Sections 3.1 and 3.2, we conduct quantitative stability analysis on a general R-SP model and our MUR-SAA model, respectively, for a fixed covariate $\boldsymbol{z} \in \mathcal{Z}$. Later, in Section 4, we leverage these stability results to analyze asymptotic and finite-sample properties of our MUR-SAA model (7).

3.1. Quantitative Stability Analysis of the R-SP Model

Given the setup in Section 2, there are two main components of the randomness that appear in the R-SP model: the true regression function $f^{\star}$ (evaluated at the new covariate $\boldsymbol{z}$) and the distribution of the additive error $\boldsymbol{\varepsilon}$. However, $f^{\star}$ is not known and is typically replaced by an estimator. Our aim in this section is to quantify the impact of the choice of the regression function and the error distribution on the optimal value and set of optimal solutions to the R-SP model. We adopt the following assumptions to ensure that the R-SP model is well-defined.

Assumption 1. The cost function c satisfies the following conditions for any $\mathbb{P}_{\boldsymbol{\xi}} \in \mathcal{P}(\Xi)$:

(a) For each fixed $\boldsymbol{\xi} \in \Xi$, the function $c(\cdot, \boldsymbol{\xi})$ is Lipschitz continuous with Lipschitz constant $L_x(\boldsymbol{\xi})$ satisfying $\mathbb{E}_{\mathbb{P}_{\boldsymbol{\xi}}}[L_x(\boldsymbol{\xi})] < \infty$;

(b) The function $c(\boldsymbol{x},\cdot)$ is uniformly Lipschitz over $\boldsymbol{x}\in\mathcal{X}$; i.e., for all $\boldsymbol{x}\in\mathcal{X}$, there exists $L_\xi>0$ such that $|c(\boldsymbol{x},\boldsymbol{\xi}_1)-c(\boldsymbol{x},\boldsymbol{\xi}_2)|\le L_\xi\|\boldsymbol{\xi}_1-\boldsymbol{\xi}_2\|$ for any $\{\boldsymbol{\xi}_1,\boldsymbol{\xi}_2\}\subseteq\Xi$;

(c) There exists $\boldsymbol{x}_0\in\mathcal{X}$ such that $\mathbb{E}_{\mathbb{P}_{\boldsymbol{\xi}}}\big|c(\boldsymbol{x}_0,\boldsymbol{\xi})\big|<\infty$.

Assumption 2. The feasible set $\mathcal{X}$ is compact.

Assumption 1 is standard in the stochastic optimization literature (see, e.g., Sun and Xu, 2016). Specifically, Assumptions 1(a) and 1(b) ensure the continuity of the objective function. Assumption 1(c) ensures the finiteness of the objective function at some feasible solution (Gao, 2023; Pichler and Xu, 2022). Assumption 1(b) aligns with conditions commonly imposed in prior work on quantitative stability analysis (see, e.g., Pichler and Xu, 2022; Römisch and Schultz, 1991; Xu and Zhang, 2023). For example, let $c(\boldsymbol{x},\boldsymbol{\xi})=\min_{\boldsymbol{y}\in\mathcal{Y}(\boldsymbol{x},\boldsymbol{\xi})}\{\boldsymbol{q}^\top\boldsymbol{y}\}$ with $\mathcal{Y}(\boldsymbol{x},\boldsymbol{\xi}):=\{\boldsymbol{y}\in\mathbb{R}^m_+\mid \boldsymbol{T}\boldsymbol{x}+\boldsymbol{W}\boldsymbol{y}\ge\boldsymbol{h}-\boldsymbol{C}\boldsymbol{\xi}\}$ be the recourse function of a two-stage stochastic programming (SP) model. It is easy to verify that $c(\boldsymbol{x},\cdot)$ is uniformly Lipschitz if $c(\boldsymbol{x},\boldsymbol{\xi})$ is finite for any $\boldsymbol{x}\in\mathcal{X}$ and $\boldsymbol{\xi}\in\Xi$. Similarly, Assumption 2 is a common assumption adopted in the stochastic optimization literature (Duchi et al., 2021; Shapiro et al., 2014) to ensure the existence of an optimal solution. Together, Assumptions 1 and 2 ensure that the optimal value of the R-SP model is finite and attained for any regression function $f\in\mathcal{F}$ and distribution of the residuals $\mathbb{P}_{\boldsymbol{\varepsilon}}\in\mathcal{P}(\mathbb{R}^{d_\xi})$.

Remark 1. In Assumption 1, conditions (a) and (c) are required to hold for any residual distribution $\mathbb{P}_{\boldsymbol{\xi}}\in\mathcal{P}(\Xi)$. This ensures that the R-SP model (4) is well-defined for any regression function $f\in\mathcal{F}$ and distribution of the residuals $\mathbb{P}_{\boldsymbol{\varepsilon}}\in\mathcal{P}(\mathbb{R}^{d_\xi})$. If the R-SP model (4) is defined on a smaller class of regression functions and distribution of the residuals, this requirement can be relaxed by assuming that (a) and (c) hold for any $\mathbb{P}_{\boldsymbol{\xi}}\in\mathcal{P}'\subset\mathcal{P}(\Xi)$. For example, suppose $\mathbb{P}_{\boldsymbol{\varepsilon}}\in\mathcal{P}_1(\mathbb{R}^{d_\xi})$, where $\mathcal{P}_1(\mathbb{R}^{d_\xi})$ is the set of probability distributions on $\mathbb{R}^{d_\xi}$ with finite first moment. In this case, one could show that $\mathbb{P}_{\boldsymbol{\xi}}\in\mathcal{P}_1(\Xi)$. To see this, fix any $\boldsymbol{\xi}_0\in\Xi$ with $\|\boldsymbol{\xi}_0\|<\infty$. For any $\boldsymbol{\xi}\in\Xi$, we have $\mathbb{E}_{\mathbb{P}_{\boldsymbol{\xi}}}\|\boldsymbol{\xi}\|=\mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}}\|\mathrm{proj}_\Xi(f(\boldsymbol{z})+\boldsymbol{\varepsilon})\|=\mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}}\|\mathrm{proj}_\Xi(f(\boldsymbol{z})+\boldsymbol{\varepsilon})-\boldsymbol{\xi}_0+\boldsymbol{\xi}_0\|\le\mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}}\|\mathrm{proj}_\Xi(f(\boldsymbol{z})+\boldsymbol{\varepsilon})-\mathrm{proj}_\Xi(\boldsymbol{\xi}_0)\|+\mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}}\|\boldsymbol{\xi}_0\|\le\mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}}\|f(\boldsymbol{z})+\boldsymbol{\varepsilon}-\boldsymbol{\xi}_0\|+\boldsymbol{\xi}_0\le\|f(\boldsymbol{z})\|+2\|\boldsymbol{\xi}_0\|+\mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}}\|\boldsymbol{\varepsilon}\|<\infty$. Thus, we could set $\mathcal{P}'=\mathcal{P}_1(\Xi)$.

In Theorem 1, we quantify the impact of the choice of $(f,\mathbb{P}_{\boldsymbol{\varepsilon}})$ on the optimal value and the set of optimal solutions to the R-SP model (4). Recall that d_{K} denotes the Kantorovich (or 1-Wasserstein) metric; see (3).

Theorem 1. *Let $\upsilon^\star(f,\mathbb{P}_{\boldsymbol{\varepsilon}})$ (resp. $\upsilon^\star(\widetilde{f},\widetilde{\mathbb{P}}_{\boldsymbol{\varepsilon}})$) and $\mathcal{X}^\star(f,\mathbb{P}_{\boldsymbol{\varepsilon}})$ (resp. $\mathcal{X}^\star(\widetilde{f},\widetilde{\mathbb{P}}_{\boldsymbol{\varepsilon}})$) be the optimal value and set of optimal solutions to the R-SP model (4) with regression function f (resp. $\widetilde{f}$) and residual distribution $\mathbb{P}_{\boldsymbol{\varepsilon}}$ (resp. $\widetilde{\mathbb{P}}_{\boldsymbol{\varepsilon}}$). Under Assumptions 1–2, the following assertions hold:*

(i) $\big|\upsilon^\star(f,\mathbb{P}_{\boldsymbol{\varepsilon}})-\upsilon^\star(\widetilde{f},\widetilde{\mathbb{P}}_{\boldsymbol{\varepsilon}})\big|\le L_\xi\big[\|f(\boldsymbol{z})-\widetilde{f}(\boldsymbol{z})\|+\mathsf{d}_K(\mathbb{P}_{\boldsymbol{\varepsilon}},\widetilde{\mathbb{P}}_{\boldsymbol{\varepsilon}})\big]$.

(ii) If, in addition, $h(f,\mathbb{P}_{\boldsymbol{\varepsilon}};\boldsymbol{x})$ satisfies the second order growth condition at $\mathcal{X}^\star(f,\mathbb{P}_{\boldsymbol{\varepsilon}})$, i.e., there exists $\tau>0$ such that $h(f,\mathbb{P}_{\boldsymbol{\varepsilon}};\boldsymbol{x}) \geq v^\star(f,\mathbb{P}_{\boldsymbol{\varepsilon}}) + \tau\big[d\big(\boldsymbol{x},\mathcal{X}^\star(f,\mathbb{P}_{\boldsymbol{\varepsilon}})\big)\big]^2$ for all $\boldsymbol{x}\in\mathcal{X}$, then

$$D\big(\mathcal{X}^\star(\widetilde{f},\widetilde{\mathbb{P}}_{\boldsymbol{\varepsilon}}),\mathcal{X}^\star(f,\mathbb{P}_{\boldsymbol{\varepsilon}})\big) \leq \sqrt{\frac{3}{\tau}\Big\{L_\xi\big[\|f(\boldsymbol{z})-\widetilde{f}(\boldsymbol{z})\| + \mathsf{d}_K(\mathbb{P}_{\boldsymbol{\varepsilon}},\widetilde{\mathbb{P}}_{\boldsymbol{\varepsilon}})\big]\Big\}}.$$

Theorem 1 provides a mechanism to quantify the sensitivity of the optimal value and solution set to the R-SP model (4) with respect to perturbations in the regression function and the residual distribution. In particular, it establishes the Lipchitz continuity of the optimal value $v^\star(f,\mathbb{P}_{\boldsymbol{\varepsilon}})$ and the Hölder continuity of the set of optimal solutions $\mathcal{X}^\star(f,\mathbb{P}_{\boldsymbol{\varepsilon}})$ with respect to distance D. Suppose we characterize the distribution of $\boldsymbol{\xi}$ using two different regression functions, f and $\widetilde{f}$, and residual distributions, $\mathbb{P}_{\boldsymbol{\varepsilon}}$ and $\widetilde{\mathbb{P}}_{\boldsymbol{\varepsilon}}$. Theorem 1 implies that the optimal value and solution to (4) may differ significantly if either $\|f(\boldsymbol{z})-\widetilde{f}(\boldsymbol{z})\|$ or $\mathsf{d}_{\mathrm{K}}(\mathbb{P}_{\boldsymbol{\varepsilon}},\widetilde{\mathbb{P}}_{\boldsymbol{\varepsilon}})$ or both are large. The second order growth condition is standard in the stochastic optimization literature and is satisfied , for example, when $h(f,\mathbb{P}_{\boldsymbol{\varepsilon}};\boldsymbol{x})$ is strongly convex in $\boldsymbol{x}$; see, Liu and Xu (2013) and Pichler and Xu (2022) for further discussions. Theorem 1 generalizes existing quantitative stability results for SP models by incorporating contextual information through a residuals-based approach. In particular, when the distribution of $\boldsymbol{\xi}$ is independent of the covariates $\boldsymbol{z}$, i.e., $\boldsymbol{\xi}=\boldsymbol{C}_f+\boldsymbol{\varepsilon}$ for some constant vector $\boldsymbol{C}_f\in\mathbb{R}^{d_\xi}$, the term $\|f(\boldsymbol{z})-\widetilde{f}(\boldsymbol{z})\|$ vanishes. In this special case, the results in Theorem 1 reduce to classical quantitative stability results for SP models (cf. Rachev and Römisch, 2002; Römisch and Schultz, 1991).

Next, we leverage the results in Theorem 1 to analyze the stability of the ER-SAA model, which is a special case of the R-SP model with $f=\widehat{f}_N$ and $\mathbb{P}_{\boldsymbol{\varepsilon}}=N^{-1}\sum_{i=1}^N \delta_{\widehat{\boldsymbol{\varepsilon}}_N^i}$. Observe that in this case, for a fixed dataset $\mathcal{D}_N=\{(\widehat{\boldsymbol{\xi}}^i,\boldsymbol{z}^i)\}_{i=1}^N$, the distribution of the residuals is automatically determined from the estimated regression function: $\mathbb{P}_{\boldsymbol{\varepsilon}}=N^{-1}\sum_{i=1}^N \delta_{\mathrm{proj}_\Xi(\widehat{\boldsymbol{\xi}}^i-\widehat{f}_N(\boldsymbol{z}^i))}$. Therefore, the results in the following theorem only focus on the change in the regression function. The following theorem establishes upper bounds on the difference between the optimal values and optimal solutions to the ER-SAA models under different estimated regression functions.

Theorem 2. *Let $v^{\text{ER-SAA}}(\widehat{f}_{N,j})$ and $\mathcal{X}^{\text{ER-SAA}}(\widehat{f}_{N,j})$ be the optimal value and the set of optimal solutions to the ER-SAA model (6) with estimated regression function $\widehat{f}_{N,j}$ for $j\in\{1,2\}$. Under Assumptions 1–2, the following assertions hold:*

(i) $\big|v^{\text{ER-SAA}}(\widehat{f}_{N,1}) - v^{\text{ER-SAA}}(\widehat{f}_{N,2})\big| \leq 2L_\xi\|\widehat{f}_{N,1}-\widehat{f}_{N,2}\|_\infty$.

(ii) Let $\widehat{h}_{N,1}(\boldsymbol{x}) = N^{-1}\sum_{i=1}^N c\big(\boldsymbol{x},\mathrm{proj}_\Xi(\widehat{f}_{N,1}(\boldsymbol{z})+\boldsymbol{\varepsilon}_{i,1}^N)\big)$ be the objective function of the ER-SAA model with estimated regression function $\widehat{f}_{N,1}$, where $\boldsymbol{\varepsilon}_{i,1}^N = \widehat{\boldsymbol{\xi}}^i - \widehat{f}_{N,1}(\boldsymbol{z})$ for all $i\in[N]$. If $\widehat{h}_{N,1}(\boldsymbol{x})$ satisfies the second order growth condition at $\mathcal{X}^{\text{ER-SAA}}(\widehat{f}_{N,1})$, i.e., there exists $\tau>0$

such that $\widehat{h}_{N,1}(\boldsymbol{x}) \geq \upsilon^{\text{ER-SAA}}(\widehat{f}_{N,1}) + \tau\big[d\big(\boldsymbol{x}, \mathcal{X}^{\text{ER-SAA}}(\widehat{f}_{N,1})\big)\big]^2$ *for all* $\boldsymbol{x} \in \mathcal{X}$*, then*

$$D\big(\mathcal{X}^{\text{ER-SAA}}(\widehat{f}_{N,2}), \mathcal{X}^{\text{ER-SAA}}(\widehat{f}_{N,1})\big) \leq \sqrt{\frac{6L_\xi}{\tau}\|\widehat{f}_{N,1} - \widehat{f}_{N,2}\|_\infty}.$$

The bounds in Theorem 2 provide, for the first time, quantitative characterizations of how model misspecification affects the performance of the ER-SAA model. If the true regression function $f^\star$ is far from the model class $\mathcal{F}$ (i.e., if $\inf_{f\in\mathcal{F}}\|f - f^\star\|_\infty$ is large), then part (i) of Theorem 2 indicate that the estimation error $\|f^\star - \widehat{f}_N\|_\infty$ will also be large. Consequently, the optimal values of the corresponding ER-SAA problems may differ substantially. An analogous conclusion holds for the distance between their optimal solution sets. Thus, the bound in Theorem 2 provides a quantitative characterization of how model misspecification affects the performance of the ER-SAA framework.

3.2. Quantitative Stability Analysis of the MUR-SAA Model

We now build on the stability results in the previous section to analyze the stability of our MUR-SAA model (7). Observe that the optimal value and the set of optimal solutions to (7) are functions of the ambiguity set $\mathcal{P}^f$. Next, we analyze the impact of perturbations in the parameters of the set $\mathcal{P}^f$ on the optimal value and the set of optimal solutions to (7). We use the following additional notation in our analyses. Let $\mathcal{F}$ be the set of continuous functions mapping from the compact metric space $\mathcal{Z}$ to $\mathbb{R}^{d_\xi}$ equipped with the uniform metric $d_{\mathcal{F}}(f_1, f_2) := \|f_1 - f_2\|_\infty$. Denote $\mathcal{P}(\mathcal{F})$ as the set of probability measures $\mathbb{P}_f$ defined on the measurable space $(\mathcal{F}, \mathcal{B}_{\mathcal{F}})$, where $\mathcal{B}_{\mathcal{F}}$ is the Borel σ-field of $\mathcal{F}$. In addition to Assumption 2, we make the following assumption to ensure the MUR-SAA model (7) is well-defined.

Assumption 3. For a given realization of the dataset $\mathcal{D}_N$ and $\boldsymbol{z} \in \mathcal{Z}$, the associated function $\widehat{h}_N$ satisfies the following conditions:

(a) For each fixed $f \in \mathcal{F}$, the cost function $\widehat{h}_N(f;\cdot)$ is Lipschitz continuous with Lipschitz constant $L_x(f)$ satisfying $\sup_{\mathbb{P}_f\in\mathcal{P}^f} \mathbb{E}_{\mathbb{P}_f}[L_x(f)] < \infty$.

(b) There exists $\boldsymbol{x}_0 \in \mathcal{X}$ such that $\sup_{\mathbb{P}_f\in\mathcal{P}^f} \big|\mathbb{E}_{\mathbb{P}_f}[\widehat{h}_N(f;\boldsymbol{x}_0)]\big| < \infty$.

Assumptions 2–3 ensure that the optimal value of the MUR-SAA model (7) is finite and that an optimal solution exists. In the following proposition, we provide a sufficient condition under which Assumption 3 holds.

Proposition 1. *Suppose that cost function c satisfies the following conditions:*

(a) For each fixed $\boldsymbol{\xi} \in \Xi$, the cost function $c(\cdot, \boldsymbol{\xi})$ is Lipschitz continuous with Lipschitz constant $L_x(\boldsymbol{\xi})$ satisfying $\sup_{\boldsymbol{\xi}\in\Xi} L_x(\boldsymbol{\xi}) < \infty$.

(b) There exists $\boldsymbol{x}_0 \in \mathcal{X}$ such that $\sup_{\boldsymbol{\xi}\in\Xi} \big|c(\boldsymbol{x}_0, \boldsymbol{\xi})\big| < \infty$.

Then, Assumption 3 holds. Consequently, if Assumption 1 holds and Ξ is compact, then Assumption 3 holds.

We are now ready to analyze the stability of the MUR-SAA model. First, in Theorem 3, we quantify the impact of the nominal regression functions and the radii $\{(\widetilde{f}_m, r_m)\}_{m=1}^M$ on the optimal value and the set of optimal solutions to the MUR-SAA model (7) formed with the sample robust MUR ambiguity set (8). Specifically, we derive upper bounds on the difference between the model's optimal values and the set of optimal solutions under two different $\{(\widetilde{f}_m, r_m)\}_{m=1}^M$ in $\mathcal{P}^{\mathrm{SR}}$.

Theorem 3. *Let v_j^{SR} and $\mathcal{X}_j^{\mathrm{SR}}$ be the optimal value and the set of optimal solutions to the MUR-SAA model (7) formed with the sample robust MUR ambiguity set $\mathcal{P}_j^{\mathrm{SR}}$ characterized by $\{(\widetilde{f}_{j,m}, r_{j,m})\}_{m=1}^M$ for $j \in \{1,2\}$. Under Assumptions 1(b), 2, and 3, the following assertions hold:*

(i) We have

$$\left|v_1^{\mathrm{SR}} - v_2^{\mathrm{SR}}\right| \leq 2L_\xi \max_{m\in[M]} \left\{ d_{\mathcal{F}}(\widetilde{f}_{1,m}, \widetilde{f}_{2,m}) + |r_{1,m} - r_{2,m}| \right\}.$$

(ii) If, in addition, $\sup_{\mathbb{P}_f \in \mathcal{P}_1^{\mathrm{SR}}} \widehat{h}_N(f;\boldsymbol{x})$ satisfies the second order growth condition at $\mathcal{X}_1^{\mathrm{SR}}$, i.e., there exists $\tau > 0$ such that $\sup_{\mathbb{P}_f \in \mathcal{P}_1^{\mathrm{SR}}} \widehat{h}_N(f;\boldsymbol{x}) \geq v_1^{\mathrm{SR}} + \tau\left[d\big(\boldsymbol{x}, \mathcal{X}_1^{\mathrm{SR}}\big)\right]^2$ for all $\boldsymbol{x} \in \mathcal{X}$, then

$$D\big(\mathcal{X}_2^{\mathrm{SR}}, \mathcal{X}_1^{\mathrm{SR}}\big) \leq \sqrt{\frac{6L_\xi}{\tau} \max_{m\in[M]} \left\{ d_{\mathcal{F}}(\widetilde{f}_{1,m}, \widetilde{f}_{2,m}) + |r_{1,m} - r_{2,m}| \right\}}.$$

Next, in Theorem 4, we quantify the impact of the choice of $(\mathbb{Q}_f, r)$ on the MUR-SAA model (7) formed with the p-Wasserstein MUR ambiguity set (9). Specifically, we derive upper bounds on the difference between the model's optimal values and the sets of solutions under two different $(\mathbb{Q}_f, r)$ in $\mathcal{P}^{W,p}$.

Theorem 4. *Let $v_j^{W,p}$ and $\mathcal{X}_j^{W,p}$ be the optimal value and the set of optimal solutions to the MUR-SAA model (7) formed with the p-Wasserstein MUR ambiguity set $\mathcal{P}_j^{W,p}$ characterized by $(\mathbb{Q}_{j,f}, r_j)$ for $j \in \{1,2\}$. Under Assumptions 1(b), 2, and 3, the following assertions hold:*

(i) We have

$$\left|v_1^{W,p} - v_2^{W,p}\right| \leq 2L_\xi \left[\max\left\{ \left[r_1 + \mathsf{d}_{W,p}(\mathbb{Q}_{f,1}, \mathbb{Q}_{f,2})\right]^p - r_2^p, \left[r_2 + \mathsf{d}_{W,p}(\mathbb{Q}_{f,1}, \mathbb{Q}_{f,2})\right]^p - r_1^p \right\} \right]^{\frac{1}{p}}.$$

(ii) If, in addition, $\sup_{\mathbb{P}_f \in \mathcal{P}_1^{W,p}} \widehat{h}_N(f;\boldsymbol{x})$ satisfies the second order growth condition at $\mathcal{X}_1^{W,p}$, i.e., there exists $\tau > 0$ such that $\sup_{\mathbb{P}_f \in \mathcal{P}_1^{W,p}} \widehat{h}_N(f;\boldsymbol{x}) \geq v_1^{W,p} + \tau\left[d\big(\boldsymbol{x}, \mathcal{X}_1^{W,p}\big)\right]^2$ for all $\boldsymbol{x} \in \mathcal{X}$, then

$$D\big(\mathcal{X}_2^{W,p}, \mathcal{X}_1^{W,p}\big) \leq \sqrt{\frac{6L_\xi}{\tau} \left[\max\left\{ \left[r_1 + \mathsf{d}_{W,p}(\mathbb{Q}_{f,1}, \mathbb{Q}_{f,2})\right]^p - r_2^p, \left[r_2 + \mathsf{d}_{W,p}(\mathbb{Q}_{f,1}, \mathbb{Q}_{f,2})\right]^p - r_1^p \right\} \right]^{\frac{1}{p}}}.$$

Remark 2. As mentioned in Section 1, the set of functions $\mathcal{F}$ may be characterized by some parameter $\boldsymbol{\theta} \in \mathfrak{T}$. This may include those considered in parametric regressions (e.g., regression coefficients

in linear regression functions) and nonparametric regressions (e.g., bandwidth in kernel regression). In such cases, one can easily modify Theorems 3 and 4 to obtain the corresponding quantitative stability results. Specifically, assume that $\mathfrak{T}$ is equipped with a metric $d_{\mathfrak{T}}$ such that $(\mathfrak{T}, d_{\mathfrak{T}})$ forms a Polish space. Suppose that $\widehat{h}_N(\boldsymbol{\theta};\boldsymbol{x}) = \widehat{h}_N(f(\cdot,\boldsymbol{\theta});\boldsymbol{x})$ is uniformly Lipschitz continuous in $\boldsymbol{\theta}$ over $\boldsymbol{x} \in \mathcal{X}$ with Lipschitz constant L_θ. Then, the assertions in Theorem 3 hold by replacing the Lipschitz constant $2L_\xi$ with L_θ and $d_{\mathcal{F}}(\widetilde{f}_{1,m}, \widetilde{f}_{2,m})$ with $d_{\mathfrak{T}}(\widetilde{\theta}_{1,m}, \widetilde{\theta}_{2,m})$, where $\{\widetilde{\theta}_{j,m}\}_{m=1}^M$ are the (estimated) parameters of the candidate models for $j \in \{1,2\}$. Similarly, the assertions in Theorem 4 remain valid by replacing the Lipschitz constant $2L_\xi$ with L_θ and $d_{W,p}(\mathbb{Q}_{f,1}, \mathbb{Q}_{f,2})$ (defined on the set of probability distributions on $\mathcal{F}$) with $d_{W,p}(\mathbb{Q}_{\boldsymbol{\theta},1}, \mathbb{Q}_{\boldsymbol{\theta},2})$ (defined on the set of probability distributions on $\mathfrak{T}$), where $\mathbb{Q}_{\boldsymbol{\theta},j}$ is the center of the Wasserstein ball for $j \in \{1,2\}$.

4. Asymptotic and Finite-Sample Properties

In this section, we analyze the asymptotic and finite-sample properties of the MUR-SAA model. We leverage the following chain of inequalities in our analysis:

$$
\begin{aligned}
&\left|\widehat{v}_N^{\mathrm{MUR}}(\boldsymbol{z}) - v^\star(\boldsymbol{z})\right| \\
&\leq \sup_{\boldsymbol{x}\in\mathcal{X}} \left| \sup_{\mathbb{P}_f \in \mathcal{P}_N^f} \mathbb{E}_{\mathbb{P}_f}[\widehat{h}_N(f;\boldsymbol{x})] - \mathbb{E}_{\mathbb{P}^\star(\boldsymbol{z})}[c(\boldsymbol{x},\boldsymbol{\xi})] \right| \\
&\leq \underbrace{\sup_{\boldsymbol{x}\in\mathcal{X}} \left| \sup_{\mathbb{P}_f \in \mathcal{P}_N^f} \mathbb{E}_{\mathbb{P}_f}[\widehat{h}_N(f;\boldsymbol{x})] - \mathbb{E}_{\widehat{\mathbb{P}}_{f,N}}[\widehat{h}_N(f;\boldsymbol{x})] \right|}_{=:A_N} + \underbrace{\sup_{\boldsymbol{x}\in\mathcal{X}} \left| \mathbb{E}_{\widehat{\mathbb{P}}_{f,N}}[\widehat{h}_N(f;\boldsymbol{x})] - \mathbb{E}_{\mathbb{P}^\star(\boldsymbol{z})}[c(\boldsymbol{x},\boldsymbol{\xi})] \right|}_{=:B_N}. \qquad (13)
\end{aligned}
$$

Above, recall that $v^\star(\boldsymbol{z})$ and $\widehat{v}_N^{\mathrm{MUR}}(\boldsymbol{z})$ were defined in (1) and (7), respectively, and $\mathbb{P}^\star(\boldsymbol{z})$ denotes the true conditional distribution of $\boldsymbol{\xi}$ given $\boldsymbol{z}$. Furthermore, $\widehat{\mathbb{P}}_{f,N}$ denotes the empirical distribution of the random regression function f; we will shortly discuss this in more detail below. Here the ambiguity set $\mathcal{P}_N^f$ depends on the dataset $\mathcal{D}_N$, which is reflected in the subscript N. To facilitate subsequent analysis, let us denote the first and second terms in (13) as A_N and B_N, respectively. To establish the desired finite-sample and asymptotic properties of the estimator, it suffices to study the corresponding properties of A_N and B_N. We leverage the quantitative stability results for the R-SP model (Section 3.1) and the MUR-SAA model (Section 3.2) to analyze the terms B_N and A_N, respectively.

As discussed in Section 2.2, model averaging is a popular approach to account for model selection bias. Hence, we focus on the following general form of the empirical distribution: $\widehat{\mathbb{P}}_{f,N} = \sum_{m=1}^M \widehat{w}_{m,N}\delta_{\widehat{f}_{m,N}}$, where $\{\widehat{f}_{m,N}\}_{m=1}^M$ are M different candidate models, and $\{\widehat{w}_{m,N}\}_{m=1}^M$ are the corresponding weights associated with each candidate model. In Section 4.1, we first analyze the asymptotic convergence of the optimal value and the set of optimal solutions to the MUR-SAA model. Then, in Section 4.2, we analyze the finite-sample properties of the MUR-SAA model.

4.1. Asymptotic Convergence

In this section, we analyze the asymptotic properties of the MUR-SAA model. We first consider the ER-MA model (10) introduced in Section 2.2, which is a special case of our MUR-SAA model obtained by setting $\mathcal{P}_N^f = \{\widehat{\mathbb{P}}_{f,N}\}$. Under this specification, the term A_N in (13) vanishes, so it suffices to study the convergence of B_N to establish the desired convergence results for the ER-MA model. Our results generalize the convergence results of Kannan et al. (2025), who studied a special case of the ER-MA model with $M = 1$. In addition, our analysis techniques differ in that they involve investigating the convergence of the estimated weights $\widehat{\boldsymbol{w}}_N$ and of the M candidate regression functions. This requires a different proof technique to handle the joint randomness of the estimated weights and regression functions. We begin our analysis with the following assumptions.

Assumption 4. Let $\mathcal{M}_0 \subseteq [M]$ be the set of indices corresponding to correctly specified models.

(a) The estimated regression function $\widehat{f}_{m,N}$ converges uniformly to some limit $f_m^\star \in \mathcal{F}$ in probability, i.e., $\|\widehat{f}_{m,N} - f_m^\star\|_\infty \xrightarrow{p} 0$, with $\|f_m^\star - f^\star\|_\infty < \infty$ for all $m \in [M]$. Moreover, for any $m \in \mathcal{M}_0$, we have $\widehat{f}_{m,N}$ converges uniformly to the true regression function $f^\star$ in probability, i.e., $\|\widehat{f}_{m,N} - f^\star\|_\infty \xrightarrow{p} 0$.

(b) There exists a candidate model that is correctly specified, i.e., $\mathcal{M}_0 \neq \emptyset$, and the weights are asymptotically consistent in the sense that $\sum_{m\in\mathcal{M}_0} \widehat{w}_{m,N} \xrightarrow{p} 1$.

Assumption 5. For any $f \in \mathcal{F}$, the uniform law of large numbers holds for $\widehat{h}_N(f;\boldsymbol{x})$, i.e., $\sup_{\boldsymbol{x}\in\mathcal{X}} \left|\widehat{h}_N(f;\boldsymbol{x}) - h^\star(f;\boldsymbol{x})\right| \xrightarrow{p} 0$, where $h^\star(f;\boldsymbol{x}) = \mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}}[c(\boldsymbol{x}, \mathrm{proj}_\Xi(f(\boldsymbol{z}) + \boldsymbol{\varepsilon}))]$ and $\mathbb{P}_{\boldsymbol{\varepsilon}}$ is the distribution of $\boldsymbol{\xi} - f(\boldsymbol{Z})$.

The first part of Assumption 4(a) requires that the estimated regression function $\widehat{f}_{m,N}$ converges uniformly to some (finite) limit $f_m^\star$, regardless of whether the model is correctly specified. This is true for regression functions estimated using standard statistical procedures, such as the maximum likelihood estimation or the least-squares approach (White, 1981, 1982). For example, in parametric regressions, if we apply the least-squares approach to estimate the unknown parameter $\boldsymbol{\theta}$, Corollary 2.2 of White (1981) ensures that the resulting estimator $\widehat{\boldsymbol{\theta}}_N$ satisfies $\widehat{\boldsymbol{\theta}}_N \xrightarrow{p} \boldsymbol{\theta}^\star$, where $\boldsymbol{\theta}^\star$ is the (unique) parameter that minimizes the prediction mean squared error. The desired uniform convergence follows immediately under some stochastic equicontinuity assumption on f (see, e.g., Newey, 1991). The second part of Assumption 4(a) is similar to Assumption 2(a) in Kannan et al. (2025), which ensures that if the model is correctly specified, the estimated regression function converges to the true one. The first part of Assumption 4(b) is a standard assumption in the literature and ensures that there exists a correctly specified model (Chen et al., 2022; Hurvich and Tsai, 1989; Liao et al., 2022). The second part of Assumption 4(b) requires that, as the sample size increases, the weights eventually concentrate on the correctly specified model(s). Recent studies have developed statistical procedures to obtain weights that are asymptotic consistent; see, e.g.,

Chen et al., 2022; Song et al., 2026; Zhang, 2015; Zhang and Liu, 2023. Finally, Assumption 5 is standard in the stochastic optimization literature, which guarantees the uniform convergence of $\widehat{h}_N(f;\boldsymbol{x})$ over $\boldsymbol{x} \in \mathcal{X}$ (Kannan et al., 2025; Shapiro et al., 2014). This assumption holds, for example, if $\{(\boldsymbol{z}^i, \widehat{\boldsymbol{\xi}}^i)\}_{i=1}^N$ is i.i.d. and there exists a measurable function E with $\mathbb{E}_{\mathbb{P}_{\boldsymbol{\xi}}}[E(\boldsymbol{\xi})] < \infty$ such that $\sup_{\boldsymbol{x}\in\mathcal{X}} |c(\boldsymbol{x},\boldsymbol{\xi})| \leq E(\boldsymbol{\xi})$ for all $\boldsymbol{\xi} \in \Xi$ (see, e.g., Theorem 7.53 of Shapiro et al., 2014).

Theorem 5 establishes that, under the stated assumptions, the optimal value and the set of optimal solutions to the ER-MA model converge in probability to the true optimal value $\upsilon^\star$ and the set of optimal solutions $\mathcal{X}^\star$ to (1), respectively.

Theorem 5. *For a given covariate $\boldsymbol{z} \in \mathcal{Z}$, let $\widehat{\upsilon}_N^{\text{ER-MA}}(\boldsymbol{z})$ and $\widehat{\mathcal{X}}_N^{\text{ER-MA}}(\boldsymbol{z})$ denote the optimal value and the set of optimal solutions to the ER-MA model. Under Assumptions 1, 2, 4, and 5, we have (i) $\widehat{\upsilon}_N^{\text{ER-MA}}(\boldsymbol{z}) \xrightarrow{p} \upsilon^\star(\boldsymbol{z})$, (ii) $D(\widehat{\mathcal{X}}_N^{\text{ER-MA}}(\boldsymbol{z}), \mathcal{X}^\star(\boldsymbol{z})) \xrightarrow{p} 0$, and (iii) $\sup_{\boldsymbol{x}\in\widehat{\mathcal{X}}_N^{\text{ER-MA}}(\boldsymbol{z})} \mathbb{E}_{\mathbb{P}^\star(\boldsymbol{z})}[c(\boldsymbol{x},\boldsymbol{\xi})] \xrightarrow{p} \upsilon^\star(\boldsymbol{z})$ for a.e. $\boldsymbol{z} \in \mathcal{Z}$.*

We now make two remarks regarding the above asymptotic result. First, Theorem 5 generalizes the convergence results established by Kannan et al. (2025). In particular, by restricting to a single (correctly specified) regression model with $M = 1$ (i.e., ER-SAA model), our results recover Theorem 1 of Kannan et al. (2025). Second, as shown in the proof of Theorem 5, the difference $|\widehat{\upsilon}_N^{\text{ER-MA}}(\boldsymbol{z}) - \upsilon^\star(\boldsymbol{z})|$ is upper bounded by the sum of three different sources of errors: the estimation error (i.e., the difference between $\widehat{f}_{m,N}$ and $f_m^\star$), the statistical error (arising from having only a finite sample), and the model misspecification error (i.e., the difference between $f_m^\star$ and $f^\star$); see the bounds on $T_{1,N}$, $T_{2,N}$, and $T_{3,N}$ in (A.19)–(A.21). In Example 4.1, we leverage this upper bound to demonstrate the value of incorporating multiple candidate models (i.e., $M > 1$) in the residuals-based approach with carefully chosen weights $\widehat{\boldsymbol{w}}_N$, particularly in the finite-sample regime.

Example 4.1. Consider two candidate models with estimators $\widehat{f}_{1,N}$ and $\widehat{f}_{2,N}$. Suppose that model $m = 1$ is misspecified, but it converges quickly to its limit $f_1^\star$ with a small $\|f_1^\star - f^\star\|_\infty > 0$. Suppose that model $m = 2$ is correctly specified, but it converges to its limit $f_2^\star = f^\star$ slowly. For example, model $m = 1$ is a (parametric) linear regression model that omits some covariates necessary for accurate prediction, while model $m = 2$ is a (nonparametric) kernel regression model. Let $\|\widehat{f}_{m,N} - f_m^\star\|_\infty = O_p(N^{-\alpha_m})$ for $m \in \{1,2\}$ with $\alpha_1 \gg \alpha_2$. Also, let $\sup_{\boldsymbol{x}\in\mathcal{X}} \left|\widehat{h}_N(f_m^\star;\boldsymbol{x}) - h(f_m^\star, \mathbb{P}_{\boldsymbol{\varepsilon}}^m;\boldsymbol{x})\right| = O_p(N^{-1/2})$ for $m \in \{1,2\}$, where $\mathbb{P}_{\boldsymbol{\varepsilon}}^m$ is the distribution of $\boldsymbol{\xi} - f_m^\star(\boldsymbol{Z})$. Writing $(\widehat{w}_{1,N}, \widehat{w}_{2,N})$ as $(w, 1-w)$ and using the bounds on $T_{1,N}$, $T_{2,N}$, and $T_{3,N}$ from (A.19)–(A.21), we have

$$
\begin{aligned}
|\widehat{\upsilon}_N^{\text{ER-MA}} - \upsilon^\star| &\leq 2L_\xi\big[wO_p(N^{-\alpha_1}) + (1-w)O_p(N^{-\alpha_2})\big] + O_p(N^{-1/2}) + 2L_\xi w\|f_1^\star - f^\star\|_\infty \\
&= \underbrace{2L_\xi w\big[O_p(N^{-\alpha_1}) - O_p(N^{-\alpha_2}) + \|f_1^\star - f^\star\|_\infty\big]}_{=:D_{1,N}} + \underbrace{\big[O_p(N^{-1/2}) + O_p(N^{-\alpha_2})\big]}_{=:D_{2,N}}.
\end{aligned} \tag{14}
$$

The upper bound in (14) has the following two implications. First, (14) quantifies the effect of the use of the misspecified model $m = 1$. Specifically, observe that the upper bound (14) is linear in w: the term $D_{2,N}$ provides an upper bound on $|\widehat{v}_N^{\text{ER-MA}} - v^\star|$ when only the correctly specified model is employed (i.e., $w = 0$), and the term $D_{1,N}$ is the adjustment of the upper bound when incorporating the misspecified model. Since $\alpha_1 \gg \alpha_2$ and $\|f_1^\star - f^\star\|_\infty > 0$ is small, the term $D_{1,N}$ is potentially negative for (small) finite $N \in \mathbb{N}$. Hence, combining the misspecified model with the correctly specified model could potentially lead to a tighter upper bound in the finite-sample regime. This indicates that, in certain settings, incorporating multiple candidate models can yield tighter finite-sample bounds within the MUR framework.

We now turn our attention to the asymptotic behavior of the MUR-SAA model. In what follows, we say Assumption 3 holds if $\widehat{h}_N$ satisfies Assumptions 3(a)–(b) a.s. for all $N \in \mathbb{N}$. The convergence of the term A_N in (13) depends in part on the choice of the ambiguity set $\mathcal{P}_N^f$. First, we establish convergence results for the MUR-SAA model formed with the sample robust MUR ambiguity set (8). In particular, we consider the case where the weights $w_m = \widehat{w}_{m,N}$ and the uncertainty sets are defined by $\mathcal{F}_{m,N} = \{f \in \mathcal{F} \mid \|f_m - \widehat{f}_{m,N}\|_\infty \le r_{m,N}\}$ with $r_{m,N} \ge 0$ for all $m \in [M]$. Under this construction,

$$\mathcal{P}_N^{\text{SR}} = \left\{ \mathbb{P}_f = \sum_{m=1}^M \widehat{w}_{m,N} \delta_{f_m} \;\middle|\; \|f_m - \widehat{f}_{m,N}\|_\infty \le r_{m,N},\ \forall m \in [M] \right\}. \tag{15}$$

In Theorem 6, establish the asymptotic convergence of the optimal value and set of optimal solutions to the MUR-SAA model (7) formed with $\mathcal{P}_N^{\text{SR}}$.

Theorem 6. *For a given covariate $\boldsymbol{z} \in \mathcal{Z}$, let $\widehat{v}_N^{\text{SR}}(\boldsymbol{z})$ and $\widehat{\mathcal{X}}_N^{\text{SR}}(\boldsymbol{z})$ be the optimal value and the set of optimal solutions to the MUR-SAA model (7) formed with the sample robust MUR ambiguity set $\mathcal{P}_N^{SR}$ defined in (15). In addition to Assumptions 1–5, suppose that $C_N := \sup_{\boldsymbol{x}\in\mathcal{X}} \sup_{f \in \bigcup_{m\in[M]} \mathcal{F}_{m,N}} |\widehat{h}_N(f;\boldsymbol{x})| < \infty$ a.s. for all $N \in \mathbb{N}$, where $\mathcal{F}_{m,N} := \{f \in \mathcal{F} \mid \|f - \widehat{f}_{m,N}\|_\infty \le r_{m,N}\}$ for $m \in [M]$. Moreover, assume that $r_{m,N} = o_p(1)$ for all $m \in [M]$. Then, we have (i) $\widehat{v}_N^{\text{SR}}(\boldsymbol{z}) \xrightarrow{p} v^\star(\boldsymbol{z})$, (ii) $D(\widehat{\mathcal{X}}_N^{\text{SR}}(\boldsymbol{z}), \mathcal{X}^\star(\boldsymbol{z})) \xrightarrow{p} 0$, and (iii) $\sup_{\boldsymbol{x}\in\widehat{\mathcal{X}}_N^{\text{SR}}(\boldsymbol{z})} \mathbb{E}_{\mathbb{P}^\star(\boldsymbol{z})}[c(\boldsymbol{x},\boldsymbol{\xi})] \xrightarrow{p} v^\star(\boldsymbol{z})$ for a.e. $\boldsymbol{z} \in \mathcal{Z}$.*

Next, we consider the p-Wasserstein MUR ambiguity set (9) with $\mathbb{Q}_f = \widehat{\mathbb{P}}_{f,N}$ and $r = r_N$ for some $r_N \ge 0$, where $\widehat{\mathbb{P}}_{f,N} = \sum_{m=1}^M \widehat{w}_{m,N} \delta_{\widehat{f}_{m,N}}$. That is, we consider

$$\mathcal{P}_N^{W,p} = \left\{ \mathbb{P}_f \in \mathcal{P}(\mathcal{F}) \;\middle|\; \mathsf{d}_{W,p}(\mathbb{P}_f, \widehat{\mathbb{P}}_{f,N}) \le r_N \right\}, \tag{16}$$

In Theorem 7, we establish the asymptotic convergence of the MUR-SAA model (7) formed with $\mathcal{P}_N^{W,p}$.

Theorem 7. *For a given covariate $\boldsymbol{z} \in \mathcal{Z}$, let $\widehat{v}_N^{W,p}(\boldsymbol{z})$ and $\widehat{\mathcal{X}}_N^{W,p}(\boldsymbol{z})$ be the optimal value and the set of optimal solutions to the MUR-SAA model (7) formed with the p-Wasserstein MUR ambiguity*

set $\mathcal{P}_N^{W,p}$ *in* (16). *In addition to Assumptions 1–5, assume that* $r_N = o_p(1)$. *Then, we have (i)* $\widehat{\upsilon}_N^{W,p}(\boldsymbol{z}) \xrightarrow{p} \upsilon^\star(\boldsymbol{z})$, *(ii)* $D(\widehat{\mathcal{X}}_N^{W,p}(\boldsymbol{z}), \mathcal{X}^\star(\boldsymbol{z})) \xrightarrow{p} 0$, *and (iii)* $\sup_{\boldsymbol{x}\in\widehat{\mathcal{X}}_N^{W,p}(\boldsymbol{z})} \mathbb{E}_{\mathbb{P}^\star(\boldsymbol{z})}[c(\boldsymbol{x},\boldsymbol{\xi})] \xrightarrow{p} \upsilon^\star(\boldsymbol{z})$ *for a.e.* $\boldsymbol{z}\in\mathcal{Z}$.

4.2. Finite-Sample Guarantees

In this section, we analyze the finite-sample properties of the MUR-SAA model. We use "Prob" to denote the probability generated from the data $\mathcal{D}_N$ and suppress its dependence on N. We once again begin by establishing finite-sample guarantees for the ER-MA model, which provides a basis for finite-sample results for the MUR models. We require the following standard assumptions in our analysis; see Kannan et al. (2025) for conditions under which these assumptions hold for various parametric and nonparametric regression methods.

Assumption 6. The estimated regression functions $\{\widehat{f}_{m,N}\}_{m=1}^M$ possess the following finite-sample properties: for any constant $\kappa>0$ and $N\in\mathbb{N}$, there exist constants $K_m(\kappa)>0$ and $\beta_m(N,\kappa)>0$ with $\lim_{N\to\infty}\beta_m(N,\kappa)=\infty$ such that $\text{Prob}\left(\|\widehat{f}_{m,N}-f_m^\star\|_\infty>\kappa\right)\le K_m(\kappa)\exp\left\{-\beta_m(N,\kappa)\right\}$ for all $m\in[M]$.

Assumption 7. For any $f\in\mathcal{F}$, the function $\widehat{h}_N(f;\boldsymbol{x})$ possesses the following finite-sample properties: for any constant $\kappa>0$ and $N\in\mathbb{N}$, there exist constants $K_f(\kappa,\boldsymbol{z})>0$ and $\beta_f(N,\kappa,\boldsymbol{z})>0$ with $\lim_{N\to\infty}\beta_f(N,\kappa,\boldsymbol{z})=\infty$ such that $\text{Prob}\left(\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\widehat{h}_N(f;\boldsymbol{x})-h^\star(f;\boldsymbol{x})\right|>\kappa\right)\le K_f(\kappa,\boldsymbol{z})\exp\left\{-\beta_f(N,\kappa,\boldsymbol{z})\right\}$, where $h^\star(f;\boldsymbol{x})=\mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}}[c(\boldsymbol{x},\text{proj}_\Xi(f(\boldsymbol{z})+\boldsymbol{\varepsilon}))]$ and $\mathbb{P}_{\boldsymbol{\varepsilon}}$ is the distribution of $\boldsymbol{\xi}-f(\boldsymbol{Z})$.

In the following theorem, we establish an upper bound on $\text{Prob}(|\widehat{\upsilon}_N^{\text{ER-MA}}(\boldsymbol{z})-\upsilon^\star(\boldsymbol{z})|>\kappa)$ under Assumptions 1, 2, 6, and 7.

Theorem 8. *For a given covariate* $\boldsymbol{z}\in\mathcal{Z}$, *let* $\widehat{\upsilon}_N^{\text{ER-MA}}(\boldsymbol{z})$ *be the optimal value of the ER-MA model. Under Assumptions 1, 2, 6, and 7, for any* $\kappa>0$ *and a.e.* $\boldsymbol{z}\in\mathcal{Z}$, *there exist constants* $\overline{K}(\kappa,\boldsymbol{z})>0$ *and* $\overline{\beta}(N,\kappa,\boldsymbol{z})>0$ *with* $\lim_{N\to\infty}\overline{\beta}(N,\kappa,\boldsymbol{z})=\infty$ *such that*

$$\text{Prob}\left(|\widehat{\upsilon}_N^{\text{ER-MA}}(\boldsymbol{z})-\upsilon^\star(\boldsymbol{z})|>\kappa\right)\le\overline{K}(\kappa,\boldsymbol{z})\exp\left\{-\overline{\beta}(N,\kappa,\boldsymbol{z})\right\}+\text{Prob}\left(\sum_{m\notin\mathcal{M}_0}\widehat{w}_{m,N}>\frac{\kappa}{3V(\boldsymbol{z})}\right),\quad(17)$$

where $V(\boldsymbol{z})=L_\xi\max_{m\notin\mathcal{M}_0}\left\{\|f_m^\star(\boldsymbol{z})-f^\star(\boldsymbol{z})\|+\mathsf{d}_K(\mathbb{P}_{\boldsymbol{\varepsilon}}^m,\widetilde{\mathbb{P}}_{\boldsymbol{\varepsilon}}^\star)\right\}$.

The upper bound on $\text{Prob}\left(|\widehat{\upsilon}_N^{\text{ER-MA}}(\boldsymbol{z})-\upsilon^\star(\boldsymbol{z})|>\kappa\right)$ in (17) consists of two terms. The first term decays exponentially with a rate $\overline{\beta}(N,\kappa,\boldsymbol{z})$ as the sample size N grows. This rate matches the convergence behavior of ER-SAA models with a *single correctly specified* model. The second term represents the probability that the total weight assigned to misspecified models exceeds a constant threshold of $\kappa/3M$. This term is specific to our ER-MA model. This additional term can be interpreted as the cost incurred in the finite-sample regime due to model uncertainty, reflecting the possibility of assigning non-negligible weight to misspecified models.

In Theorems 9 and 10, we build on the results for the ER-MA model to establish finite-sample guarantees for the MUR-SAA model formed with the sample robust MUR ambiguity set $\mathcal{P}_N^{\mathrm{SR}}$ in (15) and the p-Wasserstein MUR ambiguity set $\mathcal{P}_N^{W,p}$ in (16), respectively. In particular, Theorems 9 and 10 demonstrate that, under appropriate choices of the parameters defining the ambiguity sets $\mathcal{P}_N^{\mathrm{SR}}$ and $\mathcal{P}_N^{W,p}$, the optimal value of the MUR-SAA model attains the same convergence rate as that of the ER-MA model (compare to Theorem 8).

Theorem 9. *For a given covariate $\boldsymbol{z} \in \mathcal{Z}$, let $\widehat{v}_N^{\mathrm{SR}}(\boldsymbol{z})$ be the optimal value of the MUR-SAA model formed with the sample robust MUR ambiguity set $\mathcal{P}_N^{SR}$ in (15). In addition to Assumptions 1, 2, 3, 6, and 7, suppose that $C_N := \sup_{\boldsymbol{x}\in\mathcal{X}} \sup_{f\in\bigcup_{m\in[M]}\mathcal{F}_{m,N}} \left|\widehat{h}_N(f;\boldsymbol{x})\right| < \infty$ a.s., where $\mathcal{F}_{m,N} = \left\{f \in \mathcal{F} \mid \|f - \widehat{f}_{m,N}\|_\infty \le r_{m,N}\right\}$ for $m \in [M]$. For any $\kappa > 0$, if $r_{m,N} \le \kappa/2L_\xi - \epsilon$ a.s. for all $m \in [M]$ and for some small $\epsilon > 0$, then for a.e. $\boldsymbol{z} \in \mathcal{Z}$, there exist constants $\overline{K}(\kappa,\boldsymbol{z}) > 0$ and $\overline{\beta}(N,\kappa,\boldsymbol{z}) > 0$ with $\lim_{N\to\infty}\overline{\beta}(N,\kappa,\boldsymbol{z}) = \infty$ such that*

$$\mathrm{Prob}\left(|\widehat{v}_N^{\mathrm{SR}}(\boldsymbol{z}) - v^\star(\boldsymbol{z})| > \kappa\right) \le \overline{K}(R_N^c,\boldsymbol{z})\exp\left\{-\overline{\beta}(N,R_N^c,\boldsymbol{z})\right\} + \mathrm{Prob}\left(\sum_{m\notin\mathcal{M}_0}\widehat{w}_{m,N} > \frac{R_N^c}{3V(\boldsymbol{z})}\right), \tag{18}$$

where $R_N^c = \kappa - 2L_\xi \operatorname{ess\,sup}\max_{m\in[M]}\{r_{m,N}\} > 0$ and $V(\boldsymbol{z}) = L_\xi \max_{m\notin\mathcal{M}_0}\left\{\|f_m^\star(\boldsymbol{z}) - f^\star(\boldsymbol{z})\| + \mathsf{d}_K(\mathbb{P}_{\boldsymbol{\epsilon}}^m, \widetilde{\mathbb{P}}_{\boldsymbol{\epsilon}}^\star)\right\}$.

Theorem 10. *For a given covariate $\boldsymbol{z} \in \mathcal{Z}$, let $\widehat{v}_N^{W,p} = \widehat{v}_N^{W,p}(\boldsymbol{z})$ and $\widehat{\mathcal{X}}_N^{W,p} = \widehat{\mathcal{X}}_N^{W,p}(\boldsymbol{z})$ be the optimal value and the set of optimal solutions to the MUR-SAA model formed with the p-Wasserstein MUR ambiguity set $\mathcal{P}_N^{W,p}$ in (16). Moreover, suppose that Assumptions 1, 2, 3, 6, and 7 hold. For any $\kappa > 0$, if $r_N \le \kappa/2L_\xi - \epsilon$ a.s. for some small $\epsilon > 0$, then for a.e. $\boldsymbol{z} \in \mathcal{Z}$, there exist constants $\overline{K}(\kappa,\boldsymbol{z}) > 0$ and $\overline{\beta}(N,\kappa,\boldsymbol{z}) > 0$ with $\lim_{N\to\infty}\overline{\beta}(N,\kappa,\boldsymbol{z}) = \infty$ such that*

$$\mathrm{Prob}(|\widehat{v}_N^{W,p} - v^\star| > \kappa) \le \overline{K}(R_N^c,\boldsymbol{z})\exp\left\{-\overline{\beta}(N,R_N^c,\boldsymbol{z})\right\} + \mathrm{Prob}\left(\sum_{m\notin\mathcal{M}_0}\widehat{w}_{m,N} > \frac{R_N^c}{3V(\boldsymbol{z})}\right), \tag{19}$$

where $R_N^c = \kappa - 2L_\xi \operatorname{ess\,sup} r_N > 0$ and $V(\boldsymbol{z}) = L_\xi \max_{m\notin\mathcal{M}_0}\left\{\|f_m^\star(\boldsymbol{z}) - f^\star(\boldsymbol{z})\| + \mathsf{d}_K(\mathbb{P}_{\boldsymbol{\epsilon}}^m, \widetilde{\mathbb{P}}_{\boldsymbol{\epsilon}}^\star)\right\}$.

5. Computational Tractability

We now turn our attention to the tractability of the MUR-SAA problem (7). The computational complexity of this problem depends on several modeling components, including the regression function class $\mathcal{F}$, ambiguity set $\mathcal{P}_N^f$, support set Ξ, and the cost function c. In general, problem (7) is an infinite-dimensional optimization problem due to the ambiguity set over distributions of regression functions. As such, it is not directly solvable in the presented form. In this section, we show that, under standard modeling assumptions, the MUR-SAA problem admits equivalent reformulations that exhibit a robust optimization structure amenable to decomposition methods.

We consider a generic function class of the form $\mathcal{F} := \{f(\boldsymbol{z}) = f(\boldsymbol{z};\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \mathfrak{T}\}$, which consists of functions characterized by a parameter $\boldsymbol{\theta}$ within a defined parameter space $\mathfrak{T}$. This representation covers a board-range of parametric and nonparametric regression functions. For instance, f may represent a linear regression model with $\boldsymbol{\theta}$ denoting regression coefficients, or a kernel regression model with $\boldsymbol{\theta}$ being the kernel function's bandwidth. Under this parameterization, ambiguity over regression functions can be represented as ambiguity over the parameter $\boldsymbol{\theta}$.

We begin by considering the MUR-SAA model (7) formed with the sample robust MUR ambiguity set (8), i.e., $\mathcal{P}^{\mathrm{SR}} = \left\{\mathbb{P}_f = \sum_{m=1}^M w_m f(\boldsymbol{z};\boldsymbol{\theta}_m) \;\middle|\; \boldsymbol{\theta}_m \in \mathfrak{T}_m \;\forall m \in [M]\right\}$, where $\mathfrak{T}_m \subseteq \mathfrak{T}$ for all $m \in [M]$. Substituting $\mathcal{P}^f = \mathcal{P}^{\mathrm{SR}}$ into (7), yields

$$\begin{aligned}\widehat{v}_N^{\mathrm{MUR}}(\boldsymbol{z}) &= \inf_{\boldsymbol{x}\in\mathcal{X}} \left\{ \sum_{m=1}^M w_m \sup_{\boldsymbol{\theta}_m\in\mathfrak{T}_m} \frac{1}{N}\sum_{i=1}^N c\Big(\boldsymbol{x}, \mathrm{proj}_{\Xi}\left(f(\boldsymbol{z};\boldsymbol{\theta}_m) + (\widehat{\boldsymbol{\xi}}^i - f(\boldsymbol{z}^i;\boldsymbol{\theta}_m))\right)\Big)\right\} \\ &= \inf_{\boldsymbol{x}\in\mathcal{X}} \left\{ \sum_{m=1}^M w_m \sup_{\boldsymbol{\theta}_m\in\mathfrak{T}_m} \frac{1}{N}\sum_{i=1}^N c\Big(\boldsymbol{x}, \mathrm{proj}_{\Xi}\left([f(\boldsymbol{z};\boldsymbol{\theta}_m) - f(\boldsymbol{z}^i;\boldsymbol{\theta}_m)] + \widehat{\boldsymbol{\xi}}^i\right)\Big)\right\}. \end{aligned} \tag{20}$$

Next, we consider the MUR-SAA model (7) formed with the p-Wasserstein MUR ambiguity set (9), i.e., $\mathcal{P}^{W,p} = \left\{\mathbb{P}_{\boldsymbol{\theta}} \in \mathcal{P}(\mathfrak{T}) \,\middle|\, \mathsf{d}_{W,p}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{Q}_{\boldsymbol{\theta}}) \le r\right\}$, where $\mathbb{Q}_{\boldsymbol{\theta}} = \sum_{m=1}^M w_m \delta_{\widehat{\boldsymbol{\theta}}_m}$. Setting $\mathcal{P}^f = \mathcal{P}^{W,p}$, our MUR-SAA model (7) becomes

$$\widehat{v}_N^{\mathrm{MUR}}(\boldsymbol{z}) = \inf_{\boldsymbol{x}\in\mathcal{X}} \left\{ \sup_{\mathbb{P}_{\boldsymbol{\theta}}\in\mathcal{P}^{W,p}} \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}}\left[\frac{1}{N}\sum_{i=1}^N c\Big(\boldsymbol{x}, \mathrm{proj}_{\Xi}\left([f(\boldsymbol{z};\boldsymbol{\theta}) - f(\boldsymbol{z}^i;\boldsymbol{\theta})] + \widehat{\boldsymbol{\xi}}^i\right)\Big)\right]\right\}. \tag{21}$$

Applying the duality theory for Wasserstein DRO models (see, e.g., Gao and Kleywegt, 2023), we derive the following equivalent reformulation of (21):

$$\inf_{\boldsymbol{x}\in\mathcal{X},\,\lambda\ge 0} \left\{ r^p\lambda + \sum_{m=1}^M w_m \sup_{\boldsymbol{\theta}_m\in\mathfrak{T}} \left\{ \frac{1}{N}\sum_{i=1}^N c\Big(\boldsymbol{x}, \mathrm{proj}_{\Xi}\left([f(\boldsymbol{z};\boldsymbol{\theta}_m) - f(\boldsymbol{z}^i;\boldsymbol{\theta}_m)] + \widehat{\boldsymbol{\xi}}^i\right)\Big) - \lambda\|\boldsymbol{\theta}_m - \widehat{\boldsymbol{\theta}}_m\|^p \right\}\right\}. \tag{22}$$

Notably, (20) and (22) exhibit the structure of a robust optimization problem. The computational tractability of these formulations depends critically on the structure of the cost function c, the support set Ξ, and $\mathfrak{T}$. In particular, the presence of the projection operator and the dependence of $c(\boldsymbol{x},\boldsymbol{\xi})$ on $\boldsymbol{\xi}$ generally prevent direct reformulation as a single-level optimization problem. While such reformulations may be derived in special cases, problems with general cost structures, such as two-stage stochastic programs, typically do not admit tractable single-level reformulations. In such general settings, one can employ decomposition techniques, such as column-and-constraint generation (C&CG) or cutting-plane methods, to solve (20) and (22). For completeness, in Appendix B, we provide a detailed implementation of a generic C&CG framework, including tractable reformulations of the resulting subproblems.

6. Numerical Experiments

In this section, we use a mean-risk portfolio optimization problem to illustrate the theoretical properties of our proposed MUR-SAA model and demonstrate its potential benefits over existing residuals-based models. In Section 6.1, discuss our experimental setup. In Section 6.2, we present results demonstrating the asymptotic convergence of our MUR-SAA models. Finally, in Sections 6.3 and 6.4, we respectively compare the quality and out-of-sample costs of the solutions produced by our MUR-SAA models with those of existing approaches.

6.1. Experimental Setup

We consider the mean-risk portfolio optimization problem from the literature (Kannan et al., 2024; Mohajerin Esfahani and Kuhn, 2018). Specifically, we suppose that there are d_ξ trading assets with random return vector $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_{d_\xi})^\top$ and consider an investor who wants to decide the proportion $\boldsymbol{x} = (x_1, x_2, \ldots, x_{d_\xi})^\top$ of capital to invest in these assets. The goal is to minimize the sum of the expected portfolio loss $\mathbb{E}_{\mathbb{P}}(-\boldsymbol{x}^\top\boldsymbol{\xi})$ and portfolio risk measured using conditional value-at-risk $\mathbb{P}\text{-}\,\mathrm{CVaR}_\beta(-\boldsymbol{x}^\top\boldsymbol{\xi})$. This problem can be formulated as follows.

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \mathbb{E}_{\mathbb{P}}(-\boldsymbol{x}^\top\boldsymbol{\xi}) + \rho\,\mathbb{P}\text{-}\,\mathrm{CVaR}_\beta(-\boldsymbol{x}^\top\boldsymbol{\xi}) \tag{23a}$$

$$\text{subject to} \quad \mathbf{1}^\top\boldsymbol{x} = 1, \quad \boldsymbol{x} \geq \mathbf{0}, \tag{23b}$$

where $\rho > 0$ is the weight on the portfolio risk and

$$\mathbb{P}\text{-}\,\mathrm{CVaR}_\beta(-\boldsymbol{x}^\top\boldsymbol{\xi}) := \min_{\tau\geq 0}\left\{\tau + \frac{1}{1-\beta}\mathbb{E}_{\mathbb{P}}\big[(-\boldsymbol{x}^\top\boldsymbol{\xi} - \tau)^+\big]\right\}$$

for some $\beta \in (0,1)$ and $(a)^+ := \max\{a, 0\}$ for any $a \in \mathbb{R}$. We consider settings where the decision maker has access to contextual information that has predictive power on the return. Given the covariates $\boldsymbol{Z} = \boldsymbol{z}$, we formulate the following *contextual portfolio optimization model*:

$$\underset{\boldsymbol{x},\,\tau}{\text{minimize}} \quad \rho\tau + \mathbb{E}_{\mathbb{P}}\left[-\boldsymbol{x}^\top\boldsymbol{\xi} + \frac{\rho}{1-\beta}(-\boldsymbol{x}^\top\boldsymbol{\xi} - \tau)^+ \;\middle|\; \boldsymbol{Z} = \boldsymbol{z}\right] \tag{24a}$$

$$\text{subject to} \quad \mathbf{1}^\top\boldsymbol{x} = 1, \quad \boldsymbol{x} \geq 0, \tag{24b}$$

As in Kannan et al. (2024) and Mohajerin Esfahani and Kuhn (2018), we set $\beta = 0.8$, $\rho = 10$, and $d_\xi = 10$. Following the setup in Kannan et al. (2024), we model the relationship between returns $\boldsymbol{\xi}$ and covariates $\boldsymbol{Z}$ as follows:

$$\xi_j = \varphi_j^\star + \sum_{\ell\in\mathcal{L}^\star} \mu_{q,jl}^\star (Z_\ell)^q + \varepsilon_j + \omega \qquad \forall j \in [d_\xi], \tag{25}$$

where $q \in \{0.5, 1, 2\}$, $\varepsilon_j \sim N(0, 0.025j)$, $\omega \sim N(0, 0.02)$, and $\mathcal{L}^\star \subseteq \mathcal{L}$ with $|\mathcal{L}^\star| = 3$. We further assume that covariates follow a multivariate folded-normal distribution: $\boldsymbol{Z} = |\widetilde{\boldsymbol{Z}}|$, where $\widetilde{\boldsymbol{Z}} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$

and $\mathbf{\Sigma}$ is a correlation matrix generated using the vine method (Lewandowski et al., 2009). For illustrative purposes, we consider $d_z = 10$. We set $\varphi_j^\star = 0.005j$, $\mu_{q,j1}^\star = (0.025j)s_q - \mu_{q,j2}^\star - \mu_{q,j3}^\star$, $\mu_{q,j2}^\star = (0.0075j)s_q \cdot U(0.8, 1.2)$, $\mu_{q,j3}^\star = (0.005j)s_q \cdot U(0.8, 1.2)$ for all $j \in [d_\xi]$ with the scaling factor $s_q = (1.25, 1.22, 1)$ when $q = (1, 0.5, 2)$, respectively. These coefficients result in an expected return of asset j of $\mathbb{E}(\xi_j) = 0.03j$ for all $j \in [d_\xi]$.

We consider two different regression training methods in our MUR-SAA model: (a) LASSO regression with predictors $\{1, Z_1, Z_2, \ldots, Z_{d_z}\}$ and (b) a weighted average of model (a) and the LASSO regression with squared predictors $\{1, Z_1^2, Z_2^2, \ldots, Z_{d_z}^2\}$. For model (b), we employ the K-fold cross-validation approach in Zhang and Liu (2023) to obtain the weights. Specifically, we first partition the data $\mathcal{D}_N = \{(\boldsymbol{z}^i, \widehat{\boldsymbol{\xi}}^i)\}_{i=1}^N$ into K subsets of (roughly) equal size, denoted as $\{\mathcal{D}_N^k\}_{k=1}^K$. We then construct the training and testing sets as follows: $\{(\mathcal{D}_{N,\text{test}}^k = \mathcal{D}_N^k, \mathcal{D}_{N,\text{train}}^k = \bigcup_{k' \neq k} \mathcal{D}_N^{k'})\}_{k=1}^K$. For each $k \in [K]$, we train the $M = 2$ regression models $\{\widehat{f}_m^k\}_{m=1}^M$ using the $\mathcal{D}_{N,\text{train}}^k$ and obtain the residuals $\{\boldsymbol{e}_s^{k,m}\}_{s=1}^{S_k}$ using the testing data $\mathcal{D}_{N,\text{test}}^k$, where $S_k = |\mathcal{D}_{N,\text{test}}^k|$. Finally, we compute the weight $\widehat{\boldsymbol{w}}_N$ by solving the following quadratic program:

$$\underset{\boldsymbol{w} \in \Delta_M}{\text{minimize}} \quad \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{s=1}^{S_k} \sum_{j \in \mathcal{J}} (w_m e_{s,j}^{k,m})^2, \tag{26}$$

where $\Delta_M = \{\boldsymbol{w} \in \mathbb{R}_+^M \mid \mathbf{1}^\top \boldsymbol{w} = 1\}$. The objective function (26) represents the sum over all the squared element-wise residuals. For the ambiguity set $\mathcal{P}_f$, we use the sample robust MUR ambiguity set and 1-Wasserstein MUR ambiguity set (see Section 5) equipped with the vectorized ℓ_1 norm, i.e., $\|\mathbf{\Theta}\| = \|\text{vec}(\mathbf{\Theta})\|_1$. We will specify the corresponding radii later.

We consider the following models in our experiments:

- ER-SAA: We denote by `ER-SAA-a` the ER-SAA model trained using regression method (a). Since the ER-SAA model is designed for a single regression model, only regression method (a) is applicable.
- ER-DRO: We denote by `ER-DRO-S-a` the ER-DRO model constructed with the sample robust ambiguity set using regression method (a). Similar to the ER-SAA model, only regression method (a) can be used.
- ER-MA: We denote by `ER-MA-b` the ER-MA model trained using regression method (b).
- MUR-SAA: We denote the MUR-SAA models by `MUR-SAA-S/W-a/b`, where `S`/`W` indicates the MUR ambiguity set (sample-robust or 1-Wasserstein, respectively), and `a`/`b` specifies the regression model.

Taken together, we consider seven models: `ER-SAA-a`, `MUR-SAA-S-a`, `MUR-SAA-W-a`, `ER-DRO-S-a`, `ER-MA-b`, `MUR-SAA-S-b`, and `MUR-SAA-W-b`. In Appendix C, we present a tailored reformulation of our MUR-SAA models for this contextual mean-risk portfolio optimization problem. We do not include the ER-DRO model formed with the 1-Wasserstein ambiguity set, as it is equivalent to our

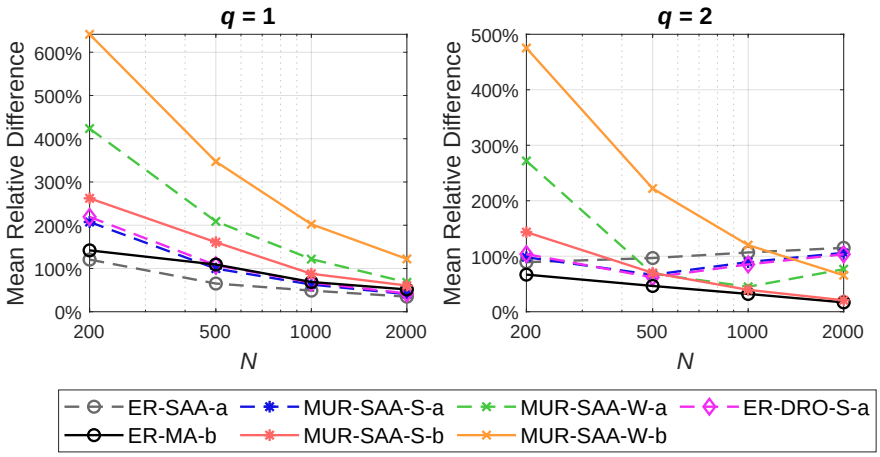

Figure 1: Average relative difference for different sample sizes N under the true model (25) with $q = 1$ and $q = 2$

MUR-SAA model formed with the 1-Wasserstein MUR ambiguity set.

We trained the regression models in the R programming language using `glmnet` with default settings and five-fold cross-validation. All residual-based models were implemented and analyzed in MATLAB and solved using Gurobi (version 11.0.2). We conducted all experiments on a computer equipped with a 3.80 GHz AMD Ryzen 7 5800X processor and 32 GB of RAM.

6.2. Convergence

We start by demonstrating our asymptotic convergence results for the MUR-SAA models presented in Section 4.1. We consider four different sample sizes: $N \in \{200, 500, 1000, 2000\}$. For each N, we generate $R = 30$ different sets of $\{(\boldsymbol{z}^i, \widehat{\boldsymbol{\xi}}^i)\}_{i=1}^N$. Then, for each replication $r \in [R]$, we solve all the seven models described in Section 6.1 using the same randomly generated covariate realization $\boldsymbol{z}$. Our observations are consistent across different covariate realizations. For illustrative purposes, we set the radii in the ambiguity sets as $200/N = o(1)$ in line with the asymptotic convergence results presented in Section 4.1. For each model (mod) and replication (r), we record the optimal value $\{\upsilon_{r,N}^{\text{mod}}\}_{r=1}^R$. Then, we obtain an estimator of the true optimal value $\upsilon^\star$ by solving the SAA model with $N' = 10,000$ scenarios generated from the true model (25). We denote this estimated optimal value as $\hat{\upsilon}^\star$. Finally, we compute the relative difference $\left\{\left|(\upsilon_{r,N}^{\text{mod}} - \hat{\upsilon}^\star)/\hat{\upsilon}^\star\right|\right\}_{r=1}^R$ between each $\upsilon_{r,N}^{\text{mod}}$ and the estimator of the true optimal value $\upsilon^\star$.

Figure 1 illustrates the average relative difference across $R = 30$ replications for increasing sample sizes N under the true model (25) with $q = 1$ and $q = 2$. The case $q = 1$ corresponds to the setting where $\boldsymbol{\xi}$ depends on the linear covariates $\{Z_1, Z_2, Z_3\}$, whereas $q = 2$ corresponds to the setting where $\boldsymbol{\xi}$ depends on the squared covariates $\{Z_1^2, Z_2^2, Z_3^2\}$. We observe the following from Figure 1. First, for $q = 1$ (left panel), the mean relative difference converges to zero as the number of samples N increases for all models. This is expected because, in this case, training methods (a) and (b) both contain the true linear model. Second, for $q = 2$ (right panel), models trained with method (a) do not converge. Notably, their average relative differences remain large even for very

large values of N (e.g., 75% to 115% at $N = 2,000$) since method (a) uses only linear covariates and therefore excludes the true model with squared covariates. In contrast, models trained with method (b) exhibit mean relative differences that converge to zero as N grows because method (b) includes the true model with squared covariates. These observations are consistent with asymptotic convergence results presented in Section 4.1 and demonstrate the benefits of our proposed MUR-SAA model that allows for the use of multiple candidate models to hedge against model-selection uncertainty.

6.3. Upper Confidence Bound (UCB) Analysis

We evaluate the quality of solutions from different models using an estimator of the normalized upper bound of the 99% confidence interval (UCB) for their out-of-sample optimality gaps. The 99% UCB provides decision-makers with a conservative estimate of the optimality gaps of the solutions produced by different models. We compute these UCBs as follows. We first generate $R = 30$ different datasets $\mathcal{D}_r = \{(\boldsymbol{z}^i, \widehat{\boldsymbol{\xi}}^i)\}_{i=1}^N$, one for each $r \in [R]$, using model (25) with $q \in \{1, 2\}$ as described in Section 6.1. For each replication $r \in [R]$, we solve the `ER-SAA-a`, `MUR-SAA-S-a`, `MUR-SAA-W-a`, `ER-DRO-S-a`, `ER-MA-b`, `MUR-SAA-S-b`, and `MUR-SAA-W-b` models with data $\mathcal{D}_r = \{(\boldsymbol{z}^i, \widehat{\boldsymbol{\xi}}^i)\}_{i=1}^N$ and a new randomly generated covariate $\boldsymbol{z}^r_{\text{new}}$. For the `MUR-SAA-S-a`, `ER-DRO-S-a`, `MUR-SAA-S-b`, and `MUR-SAA-W-b` models, we consider a wide range of radii given by $\{10^{-a} + 0.2(10)^{-a}b \mid a \in \{2, 1, 0\}, b \in \{0, 1, 2, \ldots, 44\}\} \cup \{0.2(10)^{-2}b \mid b \in \{0, 1, 2, 3, 4\}\} \cup \{10\}$. For each model (mod) and replication, we record the optimal solution $\boldsymbol{x}^{\text{mod}}_{r,N}$, yielding the collection $\{\boldsymbol{x}^{\text{mod}}_{r,N}\}_{r=1}^R$ for each model. Next, for each solution $\boldsymbol{x}^{\text{mod}}_{r,N}$, we compute the associated normalized UCB. To this end, we generate $S = 30$ different sets of residuals $\{(\boldsymbol{\varepsilon}^{s,i}, \boldsymbol{\omega}^{s,i})\}_{i=1}^{N'}$ as described in Section 6.1, each of size $N' = 10,000$. Then, for each $s \in [S]$, we compute the out-of-sample cost of $\boldsymbol{x}^{\text{mod}}_{r,N}$, denoted by $\widehat{v}^s$, as the average objective value over the scenarios $\{f^\star(\boldsymbol{z}^r_{\text{new}}) + \boldsymbol{\varepsilon}^{s,i} + \boldsymbol{\omega}^{s,i}\}_{i=1}^{N'}$, where $f^\star$ is the true regression function. We also solve the SAA counterpart of (23) for each $s \in [S]$ using the same scenarios $\{f^\star(\boldsymbol{z}^r_{\text{new}}) + \boldsymbol{\varepsilon}^{s,i} + \boldsymbol{\omega}^{s,i}\}_{i=1}^{N'}$ and record their optimal values $\{\overline{v}^s\}_{s=1}^S$. Using $\{(\widehat{v}^s, \overline{v}^s)\}_{s=1}^S$, we compute estimates of the optimality gap $\{\widehat{g}^s := \widehat{v}^s - \overline{v}^s\}_{s=1}^S$. Finally, we compute the normalized 99% UCB as $\texttt{UCB} = \frac{100}{|\widehat{m}|} \cdot \left(\widehat{\mu}_g + 2.462\sqrt{\widehat{\sigma}_g^2/S}\right)$, where $\widehat{m}$ is the average of $\{\widehat{v}^s\}_{s=1}^S$, and $\widehat{\mu}_g$ and $\widehat{\sigma}_g^2$ are the average and variance of $\{\widehat{g}^s\}_{s=1}^S$, respectively.

Figure 2 presents the average 99% UCB (across the $R = 30$ replications, excluding a small number of outliers with $|\widehat{m}| \leq 0.2$ that produce extreme UCB values) for the seven models under four settings: $q \in \{1, 2\}$, $N \in \{50, 500\}$. Here, $N = 50$ (top panel) and $N = 500$ (bottom panel) represent limited (small) data and relatively data-rich settings, respectively. We observe the following from this figure. First, the UCBs obtained from all models coincide for sufficiently large radii because the optimal solutions of all models converge to the equal-weighted portfolio as the radius increases. Second, in the limited-data setting ($N = 50$), the ER-SAA and ER-MA models produce solutions with significantly large UCBs, with the UCBs of solutions to the `ER-SAA-a` model being the largest among all models. This is expected, as these models are susceptible to estimation

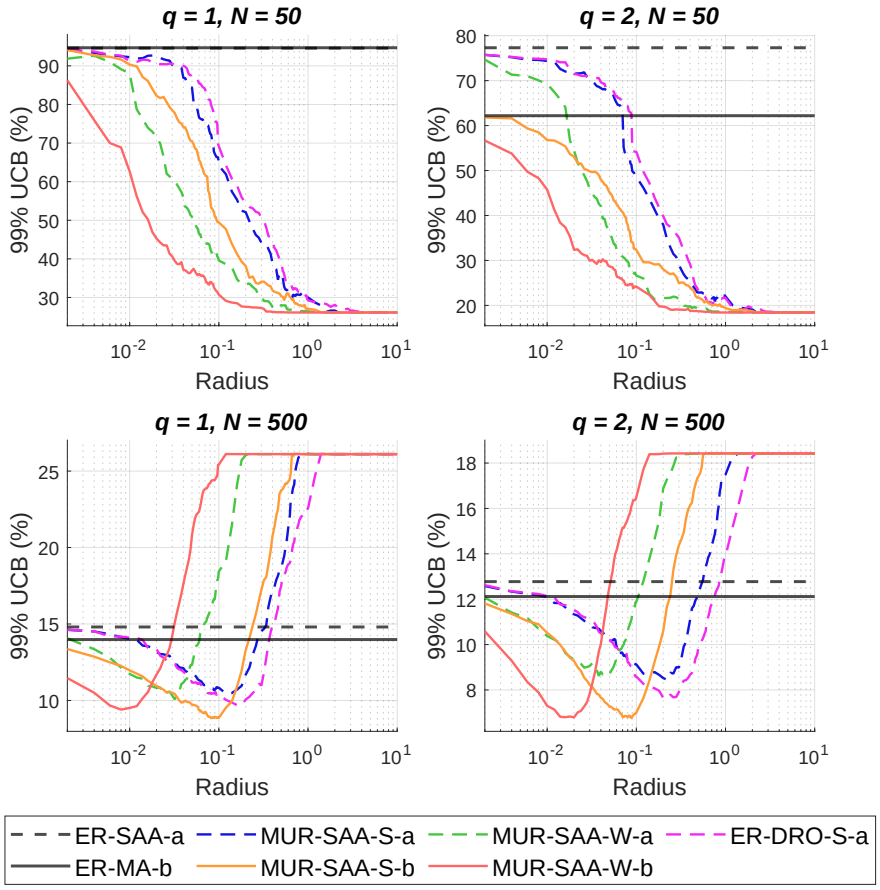

Figure 2: Average UCB for different sample sizes N under the true model (25) with $q = 1$ and $q = 2$

uncertainty in this regime. However, across all settings, the `ER-MA-b` model consistently yields smaller UCBs than the `ER-SAA-a` model. This difference becomes especially pronounced in the data-scarce and model-misspecification settings (i.e., under $N = 50$ and $q = 2$), where `ER-SAA-a` produces the largest UCBs among the models. This shows that adopting model averaging alone during the training stage, a new element introduced to the ER-SAA framework in our paper, can mitigate model uncertainty within this class of ER-SAA formulations.

Third, UCBs of the MUR-SAA and `ER-DRO-S-a` models exhibit different patterns under $N = 50$ and $N = 500$. Let us first examine the limited-data setting in more detail. When $N = 50$, the UCBs of optimal solutions to the `ER-DRO-a` and all the MUR-SAA models decrease as the radius of the ambiguity sets employed in these models increases. Notably, for a fixed small-to-moderate radius (before the optimal solutions converge), the UCBs of solutions to the MUR-SAA models are smaller than those of solutions to the `ER-DRO-S-a` model, with UCBs of the `MUR-SAA-W-b` solutions being the best among all models. Surprisingly, these observations hold for both cases, when the model is correctly specified ($q = 1$) and misspecified ($q = 2$) during the training step. Moreover, MUR-SAA models trained with method (b), which employ model averaging to combine linear and quadratic regression models, generally produce solutions with smaller UCBs compared with MUR-

SAA models trained with method (a) for small values of the radius, especially when $q = 2$. This is consistent with our hypothesis and results in Section 6.2: when $q = 2$, models trained with method (a) do not include the true model, whereas models trained with method (b) capture the true one, resulting in smaller UCBs. These results suggest that our MUR-SAA model can produce solutions that are robust to model uncertainty even in limited-data settings. It also demonstrates that MUR-SAA models trained using method (b) provide an attractive alternative to the ER-DRO model in this setting.

In contrast, UCB values are substantially smaller in the relatively data-rich setting with $N = 500$, ranging from about 7% to 26%, compared with much larger values in the limited-data setting with $N = 50$ (where UCBs range from about 26% to 95%). Moreover, UCBs of the optimal solutions obtained from `ER-DRO-S-a` and MUR-SAA models exhibit a non-monotonic pattern: they first decrease as the radius increases, then increase, and eventually stabilize. The radius at which the UCBs of `ER-DRO-S-a` and MUR-SAA models first start to grow and then level off varies across models. This behavior is expected in the data-rich regime, where the training process can extract more informative structure directly from the data. Consequently, a moderate level of robustification helps mitigate learning errors. However, excessive robustification may lead to overly conservative solutions and degraded performance. Nevertheless, we again observe that MUR-SAA models trained with method (b) (i.e., `MUR-SAA-S-b` and `MUR-SAA-W-b`) and solved with small radii produce solutions with smaller UCBs than the `ER-DRO-a` model and the MUR-SAA models trained with method (a). For example, under $q = 1$, solutions to the `MUR-SAA-S-b` obtained with radius 0.084 have the lowest UCBs among all models. When $q = 2$, solutions to the `MUR-SAA-S-b` obtained with radius 0.088 have the lowest UCBs among all models. Interestingly, even when $q = 1$ and both methods (a) and (b) contain the true linear model, the UCBs of the `MUR-SAA-S-b`/`MUR-SAA-W-b` solutions can be smaller than those of the `MUR-SAA-S-a`/`MUR-SAA-W-a` solutions. This suggests that the additional quadratic candidate model incorporated in method (b), although different from the true linear model, still provides useful information and therefore yields tighter UCBs.

6.4. Out-of-Sample Cost

Finally, we compare the out-of-sample performance of solutions from different models. We follow the same procedure described in Section 6.3 to generate $R = 30$ different datasets using model (25) with $q \in \{1, 2\}$ and use them to obtain $\{\boldsymbol{x}_{r,N}^{\text{mod}}\}_{r=1}^{R}$ to each model. To evaluate the out-of-sample cost (objective value) of these solutions, we generate a set of residuals $\{(\boldsymbol{\varepsilon}^i, \boldsymbol{\omega}^i)\}_{i=1}^{N'}$ of size $N' = 10,000$. For each solution, we then compute its average out-of-sample cost over the scenarios $\{f_{\text{out}}(\boldsymbol{z}_{\text{new}}^r) + \boldsymbol{\varepsilon}^i + \boldsymbol{\omega}^i\}_{i=1}^{N'}$, where f_{out} denotes the out-of-sample regression function (the true model observed in practice). We consider the following models for f_{out}: model (25) with $q' \in \{1, 2, 0.5\}$. When $q' = q$, this corresponds to the perfect distributional information settings. In contrast, $q' \neq q$ represents cases where f_{out} differs from the regression model used to generate the data for optimization, thereby mimicking settings with distributional shift.

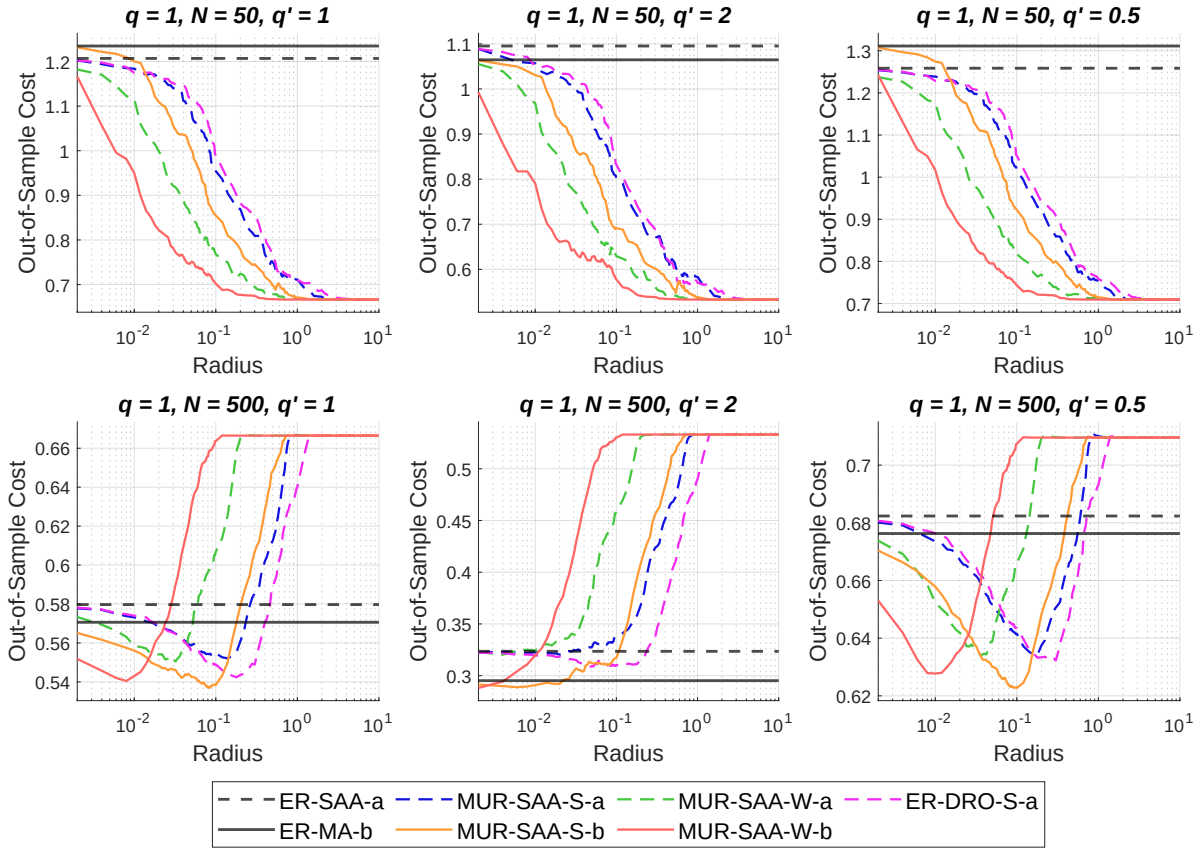

Figure 3: Average out-of-sample cost for different sample sizes N when the data is generated from the true model (25) with $q = 1$

Figure 3 presents the average out-of-sample cost (OC) (across the $R = 30$ replications) for the seven models with $q = 1$ under six settings: $N \in \{50, 500\}$, $q' \in \{0.5, 1, 2\}$. Again, $N = 50$ (top panel) and $N = 500$ (bottom panel) represent limited (small) data and relatively data-rich settings, respectively. We observe the following from Figure 3. First, similar to UCBs, OCs obtained from all models coincide for sufficiently large radii. Second, in the limited-data setting ($N = 50$), the OCs associated with solutions to the `ER-DRO-a` and all the MUR-SAA models decrease as the radius of the ambiguity sets employed in these models increases. Again, for a fixed small-to-moderate radius, OCs of solutions to the MUR-SAA models are smaller than those of solutions to the `ER-DRO-a` model. Interestingly, the patterns remain the same both with distributional shift ($q' \in \{2, 0.5\}$) and without ($q' = 1$). These results are consistent with our results in Section 6.3.

Third, in the data-rich setting ($N = 500$), OCs are substantially smaller. Moreover, similar to UCBs, OCs of the optimal solutions obtained from `ER-DRO-a` and MUR-SAA models exhibit a non-monotonic pattern: they first decrease as the radius increases, then increase, and eventually stabilize. Again, the radius at which the out-of-sample costs of `ER-DRO-a` and MUR-SAA models first start to grow and then level off varies across models. However, we observe that MUR-SAA models trained with method (b) (i.e., `MUR-SAA-S-b` and `MUR-SAA-W-b`) and solved with small radii produce solutions

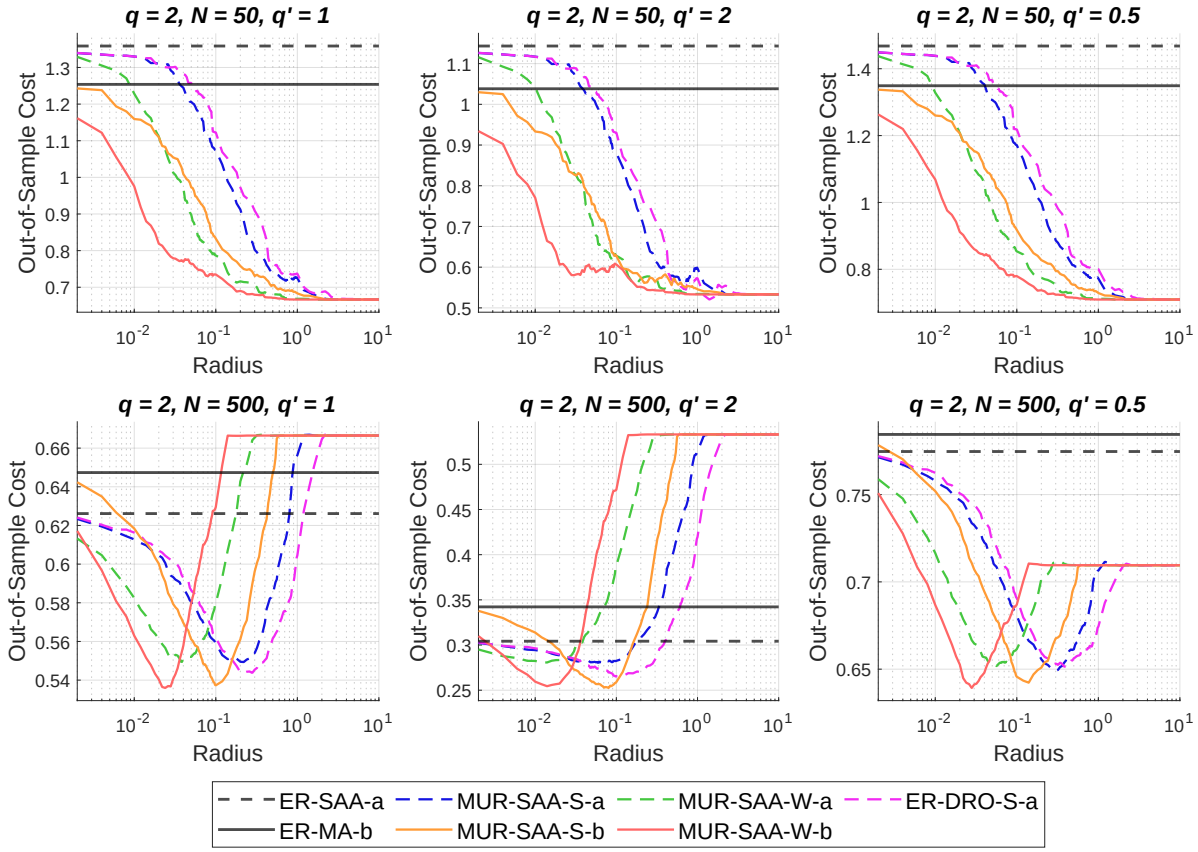

Figure 4: Average out-of-sample cost for different sample sizes N when the data is generated from the true model (25) with $q = 2$

with smaller OCs than the `ER-DRO-a` model and the MUR-SAA models trained with method (a). Notably, the improvements are more pronounced when there is model misspecification in out-of-sample data (i.e., when $q' \in \{2, 0.5\}$). In particular, when $q' = 2$, the `ER-MA-b` solution yields a lower out-of-sample cost than models trained with method (a). We have similar observations for settings when $q = 2$ (see Figure 4).

These results and those in Section 6.2–6.3 demonstrate the strength of the proposed MUR-SAA framework: by incorporating multiple candidate models and hedging against model uncertainty, it consistently delivers high-quality solutions in both data-rich and limited-data settings that remain robust even under significant model uncertainty and distributional shift.

7. Conclusion

We introduce a new data-driven, model-uncertainty-aware residuals-based SAA framework (MUR-SAA). By treating the unknown regression function f as a random element with unknown distribution $\mathbb{P}_f$ and optimizing against the worst-case expected cost over an ambiguity set of $\mathbb{P}_f$, our MUR-SAA produces solutions that are robust to model misspecification. Our MUR-SAA framework represents a substantial generalization of the popular empirical-residuals-based SAA (ER-

SAA), systematically incorporating multiple candidate models to hedge against model uncertainty and thereby enhancing the reliability of decision-making in data-driven optimization. We introduce innovative mechanisms for quantifying the stability of general residual-based SP models and our MUR-SAA model. We identify conditions under which our MUR-SAA models exhibit desirable convergence and finite-sample properties. Moreover, we derive equivalent reformulations of the MUR-SAA model as robust optimization problems, which can be solved using decomposition algorithms. Our numerical results based on a contextual portfolio optimization problem illustrate the benefits of our MUR-SAA models over existing approaches. Specifically, by incorporating multiple candidate regression models, our MUR-SAA models yield robust solutions with smaller optimality gaps and out-of-sample costs, demonstrating their ability to mitigate model uncertainty. Moreover, the results indicate that the MUR-SAA framework is a strong alternative to the ER-DRO framework in the limited-data setting, often outperforming it. It also produces solutions with superior out-of-sample performance compared to both ER-DRO and ER-SAA in relatively data-rich settings.

Our proposed MUR-SAA framework serves as a fundamental building block for several promising extensions. First, we aim to develop a DRO counterpart to MUR-SAA that simultaneously hedges against model uncertainty and distributional ambiguity. Another important direction is to generalize MUR-SAA to risk-averse settings to capture decision-makers' risk preferences. In addition, extending the framework to account for uncertainty or missingness in covariates would substantially enhance its practical relevance.

Acknowledgment

Karmel S. Shehadeh is supported by the Air Force Office of Scientific Research under grant number FA9550-25-1-0253. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Air Force Office of Scientific Research.

Appendix A. Mathematical Proofs

Appendix A.1. Proof of Theorem 1

Proof. We first prove part (i). Note that

$$
\begin{aligned}
&\left|\upsilon^\star(f,\mathbb{P}_{\boldsymbol{\varepsilon}}) - \upsilon^\star(\widetilde{f},\widetilde{\mathbb{P}}_{\boldsymbol{\varepsilon}})\right| \\
&\leq \sup_{\boldsymbol{x}\in\mathcal{X}} \left|\mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}}\left[c\big(\boldsymbol{x},\mathrm{proj}_{\Xi}\left(f(\boldsymbol{z})+\boldsymbol{\varepsilon}\right)\big)\right] - \mathbb{E}_{\widetilde{\mathbb{P}}_{\boldsymbol{\varepsilon}}}\left[c\big(\boldsymbol{x},\mathrm{proj}_{\Xi}\left(\widetilde{f}(\boldsymbol{z})+\boldsymbol{\varepsilon}\right)\big)\right]\right| \\
&\leq \sup_{\boldsymbol{x}\in\mathcal{X}} \left|\mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}}\left[c\big(\boldsymbol{x},\mathrm{proj}_{\Xi}\left(f(\boldsymbol{z})+\boldsymbol{\varepsilon}\right)\big)\right] - \mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}}\left[c\big(\boldsymbol{x},\mathrm{proj}_{\Xi}\left(\widetilde{f}(\boldsymbol{z})+\boldsymbol{\varepsilon}\right)\big)\right]\right| && \text{(A.1a)} \\
&\quad + \sup_{\boldsymbol{x}\in\mathcal{X}} \left|\mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}}\left[c\big(\boldsymbol{x},\mathrm{proj}_{\Xi}\left(\widetilde{f}(\boldsymbol{z})+\boldsymbol{\varepsilon}\right)\big)\right] - \mathbb{E}_{\widetilde{\mathbb{P}}_{\boldsymbol{\varepsilon}}}\left[c\big(\boldsymbol{x},\mathrm{proj}_{\Xi}\left(\widetilde{f}(\boldsymbol{z})+\boldsymbol{\varepsilon}\right)\big)\right]\right| && \text{(A.1b)} \\
&=: T_1 + T_2, && \text{(A.1c)}
\end{aligned}
$$

where T_1 and T_2 denote the terms (A.1a) and (A.1b), respectively. We next derive upper bounds on the terms T_1 and T_2. Note that

$$
\begin{aligned}
T_1 &\leq \sup_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}}\left|c\big(\boldsymbol{x},\mathrm{proj}_{\Xi}\left(f(\boldsymbol{z})+\boldsymbol{\varepsilon}\right)\big) - c\big(\boldsymbol{x},\mathrm{proj}_{\Xi}\left(\widetilde{f}(\boldsymbol{z})+\boldsymbol{\varepsilon}\right)\big)\right| \\
&\leq L_\xi\, \mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}}\left\|\mathrm{proj}_{\Xi}\left(f(\boldsymbol{z})+\boldsymbol{\varepsilon}\right) - \mathrm{proj}_{\Xi}\left(\widetilde{f}(\boldsymbol{z})+\boldsymbol{\varepsilon}\right)\right\| && \text{(A.2a)} \\
&\leq L_\xi\, \mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}}\left\|\left(f(\boldsymbol{z})+\boldsymbol{\varepsilon}\right) - \left(\widetilde{f}(\boldsymbol{z})+\boldsymbol{\varepsilon}\right)\right\| && \text{(A.2b)} \\
&\leq L_\xi \|f(\boldsymbol{z}) - \widetilde{f}(\boldsymbol{z})\|.
\end{aligned}
$$

Here, (A.2a) follows from the uniform Lipschitz continuity of c in Assumption 1(b) and (A.2b) follows from the fact that the orthogonal projection is a non-expansive operator. Next, note that

$$
T_2 \leq \sup_{\boldsymbol{x}\in\mathcal{X}} L_\xi \left\{ \sup_{g\in\mathcal{G}_{\mathrm{K}}} \left|\mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}}[g(\boldsymbol{\xi})] - \mathbb{E}_{\widetilde{\mathbb{P}}_{\boldsymbol{\varepsilon}}}[g(\boldsymbol{\xi})]\right| \right\} = L_\xi\, \mathsf{d}_{\mathrm{K}}(\mathbb{P}_{\boldsymbol{\varepsilon}}, \widetilde{\mathbb{P}}_{\boldsymbol{\varepsilon}}), \tag{A.3}
$$

where the inequality follows from Assumption 1(b) that $c(\boldsymbol{x},\cdot)$ is uniformly Lipschitz with Lipschitz constant L_ξ and the definition of the Kantorovich metric $\mathsf{d}_{\mathrm{K}}(\mathbb{P}_{\boldsymbol{\varepsilon}}, \widetilde{\mathbb{P}}_{\boldsymbol{\varepsilon}})$. The desired inequality then follows from (A.1c), (A.2), and (A.3).

Now, we prove part (ii). Let $\mathbb{P}_{\boldsymbol{\xi}}$ be the distribution of $\mathrm{proj}_{\Xi}(f(\boldsymbol{z})+\boldsymbol{\varepsilon})$ with $\boldsymbol{\varepsilon}\sim\mathbb{P}_{\boldsymbol{\varepsilon}}$, $\widetilde{\mathbb{P}}_{\boldsymbol{\xi}}$ be the distribution of $\mathrm{proj}_{\Xi}(\widetilde{f}(\boldsymbol{z})+\boldsymbol{\varepsilon})$ with $\boldsymbol{\varepsilon}\sim\widetilde{\mathbb{P}}_{\boldsymbol{\varepsilon}}$, and $\mathcal{G} := \{c(\cdot,\boldsymbol{\xi}) \mid \boldsymbol{x}\in\mathcal{X}\}$. Since $h(f,\mathbb{P}_{\boldsymbol{\varepsilon}};\boldsymbol{x})$ satisfies the second order growth condition, it follows from Proposition EC.4 of Tsang and Shehadeh (2025) that

$$
\begin{aligned}
D\big(\mathcal{X}^\star(\widetilde{f},\widetilde{\mathbb{P}}_{\boldsymbol{\varepsilon}}), \mathcal{X}^\star(f,\mathbb{P}_{\boldsymbol{\varepsilon}})\big) &\leq \sqrt{\frac{3}{\tau}\mathbb{H}\big(\{\mathbb{P}_{\boldsymbol{\xi}}\},\{\widetilde{\mathbb{P}}_{\boldsymbol{\xi}}\};d_{\mathcal{G}}\big)} \\
&= \sqrt{\frac{3}{\tau}\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}}\left[c\big(\boldsymbol{x},\mathrm{proj}_{\Xi}\left(f(\boldsymbol{z})+\boldsymbol{\varepsilon}\right)\big)\right] - \mathbb{E}_{\widetilde{\mathbb{P}}_{\boldsymbol{\varepsilon}}}\left[c\big(\boldsymbol{x},\mathrm{proj}_{\Xi}\left(\widetilde{f}(\boldsymbol{z})+\boldsymbol{\varepsilon}\right)\big)\right]\right|}.
\end{aligned}
$$

The desired inequality then follows from (A.1), (A.2), and (A.3) in the proof of part (i). □

Appendix A.2. Proof of Theorem 2

Proof. In view of Theorem 1, since $\|\widehat{f}_{N,1}(\boldsymbol{z}) - \widehat{f}_{N,2}(\boldsymbol{z})\| \le \|\widehat{f}_{N,1} - \widehat{f}_{N,2}\|_\infty$ for all $\boldsymbol{z} \in \mathcal{Z}$, it suffices to show that $\mathsf{d}_{\mathrm{K}}\big(\widehat{\mathbb{P}}_{\boldsymbol{\varepsilon},1}, \widehat{\mathbb{P}}_{\boldsymbol{\varepsilon},2}\big) \le \|\widehat{f}_{N,1} - \widehat{f}_{N,2}\|_\infty$, where $\widehat{\mathbb{P}}_{\boldsymbol{\varepsilon},j} = N^{-1}\sum_{i=1}^N \delta_{\mathrm{proj}_\Xi(\widehat{\boldsymbol{\xi}}^i - \widehat{f}_{N,j}(\boldsymbol{z}^i))}$ for $j \in \{1,2\}$. Indeed, by definition of d_{K}, we have

$$\begin{aligned}
\mathsf{d}_{\mathrm{K}}\big(\widehat{\mathbb{P}}_{\boldsymbol{\varepsilon},1}, \widehat{\mathbb{P}}_{\boldsymbol{\varepsilon},2}\big) &= \sup_{g\in\mathcal{G}_{\mathrm{K}}} \Big|\mathbb{E}_{\widehat{\mathbb{P}}_{\boldsymbol{\varepsilon},1}}\big[g(\boldsymbol{\xi})\big] - \mathbb{E}_{\widehat{\mathbb{P}}_{\boldsymbol{\varepsilon},2}}\big[g(\boldsymbol{\xi})\big]\Big| \\
&= \sup_{g\in\mathcal{G}_{\mathrm{K}}} \left|\frac{1}{N}\sum_{i=1}^N g\big(\widehat{\boldsymbol{\xi}}^i - \widehat{f}_{N,1}(\boldsymbol{z}^i)\big) - \frac{1}{N}\sum_{i=1}^N g\big(\widehat{\boldsymbol{\xi}}^i - \widehat{f}_{N,2}(\boldsymbol{z}^i)\big)\right| \\
&\le \frac{1}{N}\sum_{i=1}^N \Big\|\widehat{f}_{N,1}(\boldsymbol{z}^i) - \widehat{f}_{N,2}(\boldsymbol{z}^i)\Big\| \\
&\le \|\widehat{f}_{N,1} - \widehat{f}_{N,2}\|_\infty,
\end{aligned}$$

where the first inequality follows from the definition of $\mathcal{G}_{\mathrm{K}}$. This completes the proof. □

Appendix A.3. Proof of Proposition 1

Proof. Consider a fixed dataset $\mathcal{D}_N$. First, we show that Assumption 3(a) holds. Let $\widetilde{\boldsymbol{\xi}}^i = \mathrm{proj}_\Xi\big(f(\boldsymbol{z}) + [\widehat{\boldsymbol{\xi}}^i - f(\boldsymbol{z}^i)]\big)$ for $i \in [N]$, where we suppress its dependence on the regression function f for notational simplicity. Note that

$$\big|\widehat{h}_N(f;\boldsymbol{x}) - \widehat{h}_N(f;\widetilde{\boldsymbol{x}})\big| \le \frac{1}{N}\sum_{i=1}^N \big|c(\boldsymbol{x},\widetilde{\boldsymbol{\xi}}^i) - c(\widetilde{\boldsymbol{x}},\widetilde{\boldsymbol{\xi}}^i)\big| \le \frac{1}{N}\sum_{i=1}^N L_x(\widetilde{\boldsymbol{\xi}}^i)\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|,$$

where the last inequality follows from assumption (a). Thus, $\widehat{h}_N(f;\cdot)$ is Lipschitz continuous with Lipschitz constant $L_x(f) = N^{-1}\sum_{i=1}^N L_x(\widetilde{\boldsymbol{\xi}}^i)$. In particular, we have

$$\sup_{\mathbb{P}_f\in\mathcal{P}^f} \mathbb{E}_{\mathbb{P}_f}[L_x(f)] = \sup_{\mathbb{P}_f\in\mathcal{P}^f} \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mathbb{P}_f}[L_x(\widetilde{\boldsymbol{\xi}}^i)] \le \sup_{\boldsymbol{\xi}\in\Xi} L_x(\boldsymbol{\xi}) < \infty,$$

where the last inequality follows from assumption (a). Therefore, Assumption 3(a) holds.

Next, we show that Assumption 3(b) holds. Note that

$$\sup_{\mathbb{P}_f\in\mathcal{P}^f} \Big|\mathbb{E}_{\mathbb{P}_f}[\widehat{h}_N(f;\boldsymbol{x}_0)]\Big| \le \sup_{\mathbb{P}_f\in\mathcal{P}^f} \frac{1}{N}\sum_{i=1}^N \Big|\mathbb{E}_{\mathbb{P}_f}[c(\boldsymbol{x}_0,\widetilde{\boldsymbol{\xi}}^i)]\Big| \le \sup_{\boldsymbol{\xi}\in\Xi} \big|c(\boldsymbol{x}_0,\boldsymbol{\xi})\big| < \infty,$$

where the last inequality follows from assumption (b). Therefore, Assumption 3(b) holds. □

Appendix A.4. Proof of Theorem 3

Proof. We first prove part (i). Define the set of functions $\mathcal{G} = \{\widehat{h}_N(\cdot;\boldsymbol{x}) \mid \boldsymbol{x} \in \mathcal{X}\}$. Following the proof of Proposition EC.4 of Tsang and Shehadeh (2025), we have

$$\left|v_1^{\mathrm{SR}} - v_2^{\mathrm{SR}}\right| \leq \mathbb{H}\big(\mathcal{P}_1^{\mathrm{SR}}, \mathcal{P}_2^{\mathrm{SR}}; \mathsf{d}_{\mathcal{G}}\big). \tag{A.4}$$

To bound $\mathbb{H}\big(\mathcal{P}_1^{\mathrm{SR}}, \mathcal{P}_2^{\mathrm{SR}}; \mathsf{d}_{\mathcal{G}}\big)$ in (A.4), note that for any $\mathbb{P}_1 \in \mathcal{P}_1^{\mathrm{SR}}$ and $\mathbb{P}_2 \in \mathcal{P}_2^{\mathrm{SR}}$, we have

$$\begin{aligned}
\mathsf{d}_{\mathcal{G}}(\mathbb{P}_1, \mathbb{P}_2) &= \sup_{\boldsymbol{x}\in\mathcal{X}} \left|\mathbb{E}_{\mathbb{P}_1}[\widehat{h}_N(f;\boldsymbol{x})] - \mathbb{E}_{\mathbb{P}_2}[\widehat{h}_N(f;\boldsymbol{x})]\right| \\
&= \sup_{\boldsymbol{x}\in\mathcal{X}} \left|\sum_{m=1}^{M} w_m \widehat{h}_N(f_{1,m};\boldsymbol{x}) - \sum_{m=1}^{M} w_m \widehat{h}_N(f_{2,m};\boldsymbol{x})\right| \\
&\leq \sum_{m=1}^{M} w_m \sup_{\boldsymbol{x}\in\mathcal{X}} \left|\widehat{h}_N(f_{1,m};\boldsymbol{x}) - \widehat{h}_N(f_{2,m};\boldsymbol{x})\right| \\
&\leq \sum_{m=1}^{M} w_m \big[L_\xi \|f_{1,m}(\boldsymbol{z}) - f_{2,m}(\boldsymbol{z})\| + \mathsf{d}_{\mathrm{K}}(\mathbb{P}_{\boldsymbol{\varepsilon},1,m}, \mathbb{P}_{\boldsymbol{\varepsilon},2,m})\big], \qquad \text{(A.5a)} \\
&\leq 2L_\xi \sum_{m=1}^{M} w_m d_{\mathcal{F}}(f_{1,m}, f_{2,m}), \qquad \text{(A.5b)}
\end{aligned}$$

where $f_{j,m} \in \mathcal{F}_{j,m}$ for all $m \in [M]$ and $j \in \{1,2\}$. Here, (A.5a) follows from a similar argument in the proof of Theorem 1 and (A.5b) follows from a similar argument in the proof of Theorem 2. Therefore, we have

$$\begin{aligned}
&\sup_{\mathbb{P}_2\in\mathcal{P}_2^{\mathrm{SR}}} \inf_{\mathbb{P}_1\in\mathcal{P}_1^{\mathrm{SR}}} \mathsf{d}_{\mathcal{G}}(\mathbb{P}_1, \mathbb{P}_2) \\
&= \sup_{f_{2,m}\in\mathcal{F}_{2,m}, \forall m\in[M]} \left\{2L_\xi \inf_{f_{1,m}\in\mathcal{F}_{1,m}, \forall m\in[M]} \sum_{m=1}^{M} w_m d_{\mathcal{F}}(f_{1,m}, f_{2,m})\right\} \\
&\leq 2L_\xi \sum_{m=1}^{M} w_m \left\{\sup_{f_{2,m}\in\mathcal{F}_{2,m}} \inf_{f_{1,m}\in\mathcal{F}_{1,m}} d_{\mathcal{F}}(f_{1,m}, f_{2,m})\right\}. \qquad \text{(A.6)}
\end{aligned}$$

We claim that

$$D_{1,m} := \sup_{f_{2,m}\in\mathcal{F}_{2,m}} \inf_{f_{1,m}\in\mathcal{F}_{1,m}} d_{\mathcal{F}}(f_{1,m}, f_{2,m}) \leq d_{\mathcal{F}}(\widetilde{f}_{1,m}, \widetilde{f}_{2,m}) + |r_{1,m} - r_{2,m}| \tag{A.7}$$

for all $m \in [M]$. For the trivial case when $\mathcal{F}_{2,m} \subseteq \mathcal{F}_{1,m}$, we have $D_{1,m} = 0$. Thus, assume that $\mathcal{F}_{2,m} \not\subseteq \mathcal{F}_{1,m}$. Let $\overline{f} \in \mathcal{F}_{2,m} \setminus \mathcal{F}_{1,m}$ and $\widehat{\lambda} = r_{1,m}/d_{\mathcal{F}}(\overline{f}, \widetilde{f}_{1,m}) \in [0,1)$. Define $g = \widehat{\lambda}\overline{f} + (1-\widehat{\lambda})\widetilde{f}_{1,m}$. Note that $g \in \mathcal{F}_{1,m}$. Indeed, $d_{\mathcal{F}}(g, \widetilde{f}_{1,m}) = d_{\mathcal{F}}\big(\widehat{\lambda}\overline{f} + (1-\widehat{\lambda})\widetilde{f}_{1,m}, \widetilde{f}_{1,m}\big) \leq \widehat{\lambda} d_{\mathcal{F}}(\overline{f}, \widetilde{f}_{1,m}) = r_{1,m}$.

Therefore, we have

$$\begin{aligned}\inf_{f\in\mathcal{F}_{1,m}} d_{\mathcal{F}}(f,\overline{f}) \le d_{\mathcal{F}}(g,\overline{f}) &= d_{\mathcal{F}}\big(\widehat{\lambda}\overline{f} + (1-\widehat{\lambda})\widetilde{f}_{1,m}, \overline{f}\big) \\ &= (1-\widehat{\lambda}) d_{\mathcal{F}}(\widetilde{f}_{1,m}, \overline{f}) \\ &= d_{\mathcal{F}}(\overline{f}, \widetilde{f}_{1,m}) - r_{1,m} \qquad \text{(A.8a)} \\ &\le d_{\mathcal{F}}(\overline{f}, \widetilde{f}_{2,m}) + d_{\mathcal{F}}(\widetilde{f}_{2,m}, \widetilde{f}_{1,m}) - r_{1,m} \qquad \text{(A.8b)} \\ &\le (r_{2,m} - r_{1,m}) + d_{\mathcal{F}}(\widetilde{f}_{1,m}, \widetilde{f}_{2,m}), \qquad \text{(A.8c)}\end{aligned}$$

where (A.8a) follows from the definition of $\widehat{\lambda}$, (A.8b) follows from the triangular inequality of the norm $d_{\mathcal{F}}$, and (A.8c) follows from $\overline{f} \in \mathcal{F}_{2,m}$. It follows that $D_{1,m} \le d_{\mathcal{F}}(\widetilde{f}_{1,m}, \widetilde{f}_{2,m}) + |r_{1,m} - r_{2,m}|$, which gives (A.7). Thus, combining (A.6)–(A.7), we obtain

$$\begin{aligned}&\sup_{\mathbb{P}_2\in\mathcal{P}_2^{\mathrm{SR}}} \inf_{\mathbb{P}_1\in\mathcal{P}_1^{\mathrm{SR}}} \mathsf{d}_{\mathcal{G}}(\mathbb{P}_1,\mathbb{P}_2) \\ &\le 2L_\xi \sum_{m=1}^{M} w_m \Big\{ d_{\mathcal{F}}(\widetilde{f}_{1,m}, \widetilde{f}_{2,m}) + |r_{1,m} - r_{2,m}| \Big\} \\ &\le 2L_\xi \max_{m\in[M]} \Big\{ d_{\mathcal{F}}(\widetilde{f}_{1,m}, \widetilde{f}_{2,m}) + |r_{1,m} - r_{2,m}| \Big\}. \qquad \text{(A.9)}\end{aligned}$$

Next, for any $\mathbb{P}_1 \in \mathcal{P}_1^{\mathrm{SR}}$ and $\mathbb{P}_2 \in \mathcal{P}_2^{\mathrm{SR}}$, following the same argument in (A.5), we have

$$\mathsf{d}_{\mathcal{G}}(\mathbb{P}_1,\mathbb{P}_2) \le 2L_\xi \sum_{m=1}^{M} w_m d_{\mathcal{F}}(f_{1,m}, f_{2,m}) \qquad \text{(A.10)}$$

Therefore, we have

$$\begin{aligned}&\sup_{\mathbb{P}_1\in\mathcal{P}_1^{\mathrm{SR}}} \inf_{\mathbb{P}_2\in\mathcal{P}_2^{\mathrm{SR}}} \mathsf{d}_{\mathcal{G}}(\mathbb{P}_1,\mathbb{P}_2) \\ &= \sup_{f_{1,m}\in\mathcal{F}_{1,m}, \forall m\in[M]} \Big\{ 2L_\xi \inf_{f_{2,m}\in\mathcal{F}_{2,m}, \forall m\in[M]} \sum_{m=1}^{M} w_m d_{\mathcal{F}}(f_{1,m}, f_{2,m}) \Big\} \\ &\le 2L_\xi \sum_{m=1}^{M} w_m \Big\{ \sup_{f_{1,m}\in\mathcal{F}_{1,m}} \inf_{f_{2,m}\in\mathcal{F}_{2,m}} d_{\mathcal{F}}(f_{1,m}, f_{2,m}) \Big\} \\ &\le 2L_\xi \sum_{m=1}^{M} w_m \Big[d_{\mathcal{F}}(\widetilde{f}_{1,m}, f_{2,m}) + |r_{1,m} - r_{2,m}| \Big] \qquad \text{(A.11a)} \\ &\le 2L_\xi \max_{m\in[M]} \Big\{ d_{\mathcal{F}}(\widetilde{f}_{1,m}, f_{2,m}) + |r_{1,m} - r_{2,m}| \Big\}. \qquad \text{(A.11b)}\end{aligned}$$

Here, we can follow the same argument used for deriving (A.7) to derive the inequality (A.11a).

Combining (A.9) and (A.11b), we have

$$\mathbb{H}\big(\mathcal{P}_1^{\mathrm{SR}}, \mathcal{P}_2^{\mathrm{SR}}; \mathsf{d}_{\mathcal{G}}\big) \leq 2L_\xi \max_{m\in[M]} \Big\{ d_{\mathcal{F}}(\widetilde{f}_{1,m}, f_{2,m}) + |r_{1,m} - r_{2,m}| \Big\}. \tag{A.12}$$

and the desired inequality immediately follows from (A.4).

Next, we prove part (ii). Given the second-order growth condition, following the proof of Proposition EC.4 of Tsang and Shehadeh (2025), we have

$$D\big(\mathcal{X}_2^{\mathrm{SR}}, \mathcal{X}_1^{\mathrm{SR}}\big) \leq \sqrt{\frac{3}{\tau}\mathbb{H}\big(\mathcal{P}_1^{\mathrm{SR}}, \mathcal{P}_2^{\mathrm{SR}}; \mathsf{d}_{\mathcal{G}}\big)}.$$

The desired inequality follows from (A.12). □

Appendix A.5. Proof of Theorem 4

Proof. We first prove part (i). Again, define the set of functions $\mathcal{G} = \{\widehat{h}_N(\cdot;\boldsymbol{x}) \mid \boldsymbol{x} \in \mathcal{X}\}$. Following the proof of Proposition EC.4 of Tsang and Shehadeh (2025), we have

$$\big|\upsilon_1^{W,p} - \upsilon_2^{W,p}\big| \leq \mathbb{H}\big(\mathcal{P}_1^{W,p}, \mathcal{P}_2^{W,p}; \mathsf{d}_{\mathcal{G}}\big). \tag{A.13}$$

Using the uniformly Lipschitz continuity of $c(\boldsymbol{x},\cdot)$, it follows from Theorem 2 that $\widehat{h}_N(\cdot;\boldsymbol{x})$ is uniformly Lipschitz over $\boldsymbol{x} \in \mathcal{X}$ with Lipschitz constant $2L_\xi$. Therefore, for any $\mathbb{P}_1 \in \mathcal{P}_1^{W,p}$ and $\mathbb{P}_2 \in \mathcal{P}_2^{W,p}$, we have

$$\mathsf{d}_{\mathcal{G}}(\mathbb{P}_1,\mathbb{P}_2) \leq 2L_\xi \mathsf{d}_K(\mathbb{P}_1,\mathbb{P}_2) = 2L_\xi \mathsf{d}_{W,1}(\mathbb{P}_1,\mathbb{P}_2) \leq 2L_\xi \mathsf{d}_{W,p}(\mathbb{P}_1,\mathbb{P}_2), \tag{A.14}$$

where the first inequality follows from $\mathcal{G} \subseteq 2L_\xi \mathcal{G}_{\mathrm{K}}$, the equality follows from the Kantorovich-Rubinstein duality (see, e.g., Corollary 21.2.3 of Garling, 2018), and the last inequality follows from $\mathsf{d}_{W,1} \leq \mathsf{d}_{W,p}$ for any $p \geq 1$ (see, e.g., Proposition 21.1.3 of Garling, 2018). Combining (A.13)–(A.14), we obtain

$$\big|\upsilon_1^{W,p} - \upsilon_2^{W,p}\big| \leq 2L_\xi \mathbb{H}\big(\mathcal{P}_1^{W,p}, \mathcal{P}_2^{W,p}; \mathsf{d}_{W,p}\big). \tag{A.15}$$

Next, we follow the idea of the proof of Theorem 2 of Pichler and Xu (2022) to derive an upper bound on $\mathbb{H}\big(\mathcal{P}_1^{W,p}, \mathcal{P}_2^{W,p}; \mathsf{d}_{W,p}\big)$ in (A.15). Note that different from Pichler and Xu (2022) where they consider probability distributions on real number spaces (as in the classical DRO setting), we consider probability distributions on a general Polish space, which requires us to leverage different theoretical results. We claim that

$$\mathbb{H}\big(\mathcal{P}_1^{W,p}, \mathcal{P}_2^{W,p}; \mathsf{d}_{W,p}\big) \leq \Big[\max\Big\{\big[r_1 + \mathsf{d}_{W,p}(\mathbb{Q}_{f,1},\mathbb{Q}_{f,2})\big]^p - r_2^p, \big[r_2 + \mathsf{d}_{W,p}(\mathbb{Q}_{f,1},\mathbb{Q}_{f,2})\big]^p - r_1^p\Big\}\Big]^{\frac{1}{p}} \tag{A.16}$$

To show (A.16), consider $\mathbb{D}(\mathcal{P}_2^{W,p}, \mathcal{P}_1^{W,p}) = \sup_{\mathbb{P}_2\in\mathcal{P}_2^{W,p}} \inf_{\mathbb{P}_1\in\mathcal{P}_1^{W,p}} \mathsf{d}_{\mathcal{G}}(\mathbb{P}_1,\mathbb{P}_2)$. If $\mathcal{P}_2^{W,p} \subseteq \mathcal{P}_1^{W,p}$,

then $\mathbb{D}(\mathcal{P}_2^{W,p}, \mathcal{P}_1^{W,p}) = 0$. Now consider $\mathcal{P}_2^{W,p} \not\subseteq \mathcal{P}_1^{W,p}$. Let $\mathbb{M} \in \mathcal{P}_2^{W,p} \setminus \mathcal{P}_1^{W,p}$ and define $\widehat{\lambda} = r_1/\mathsf{d}_{W,p}(\mathbb{M}, \mathbb{P}_{f,1}) \in [0,1)$. Consider $\mathbb{M}' = \widehat{\lambda}^p \mathbb{M} + (1-\widehat{\lambda}^p)\mathbb{P}_{f,1}$. Note that $\mathbb{M}' \in \mathcal{P}_1^{W,p}$. Indeed, $\mathsf{d}_{W,p}(\mathbb{M}', \mathbb{P}_{f,1})^p = \mathsf{d}_{W,p}\big(\widehat{\lambda}^p \mathbb{M} + (1-\widehat{\lambda}^p)\mathbb{P}_{f,1}, \mathbb{P}_{f,1}\big)^p \leq \widehat{\lambda}^p \mathsf{d}_{W,p}(\mathbb{M}, \mathbb{P}_{f,1})^p = r_1^p$. Here, the inequality follows from the p-convexity of the p-Wasserstein distance, which could be verified following the proof of Lemma 2.10 of Pflug and Pichler (2014) with the existence of the optimal transport plan in general Polish space guaranteed by Corollary 20.2.2 of Garling (2018). Therefore, we have

$$\begin{aligned}
\inf_{\mathbb{P}_1 \in \mathcal{P}_1^{W,p}} \mathsf{d}_{W,p}(\mathbb{P}_1, \mathbb{M})^p \leq \mathsf{d}_{W,p}(\mathbb{M}', \mathbb{M})^p &= \mathsf{d}_{W,p}\big(\widehat{\lambda}^p \mathbb{M} + (1-\widehat{\lambda}^p)\mathbb{P}_{f,1}, \mathbb{M}\big)^p \\
&\leq (1-\widehat{\lambda}^p)\mathsf{d}_{W,p}(\mathbb{P}_{f,1}, \mathbb{M})^p \\
&= \mathsf{d}_{W,p}(\mathbb{P}_{f,1}, \mathbb{M})^p - r_1^p \qquad \text{(A.17a)} \\
&\leq \big[\mathsf{d}_{W,p}(\mathbb{M}, \mathbb{P}_{f,2}) + \mathsf{d}_{W,p}(\mathbb{P}_{f,2}, \mathbb{P}_{f,1})\big]^p - r_1^p \\
&\leq \big[r_2 + \mathsf{d}_{W,p}(\mathbb{P}_{f,2}, \mathbb{P}_{f,1})\big]^p - r_1^p, \qquad \text{(A.17b)}
\end{aligned}$$

where (A.17a) follows from the definition of $\widehat{\lambda}$ and (A.17b) follows from $\mathbb{M} \in \mathcal{P}_2^{W,p}$. This shows that $\mathbb{D}(\mathcal{P}_2^{W,p}, \mathcal{P}_1^{W,p}) \leq \big\{\big[r_2 + \mathsf{d}_{W,p}(\mathbb{P}_{f,2}, \mathbb{P}_{f,1})\big]^p - r_1^p\big\}^{1/p}$. Interchanging the roles played by $\mathbb{P}_{f,2}$ and $\mathbb{P}_{f,1}$, we also have $\mathbb{D}(\mathcal{P}_1^{W,p}, \mathcal{P}_2^{W,p}) \leq \big\{\big[r_1 + \mathsf{d}_{W,p}(\mathbb{P}_{f,2}, \mathbb{P}_{f,1})\big]^p - r_2^p\big\}^{1/p}$. Combining the last two inequalities on $\mathbb{D}(\mathcal{P}_2^{W,p}, \mathcal{P}_1^{W,p})$ and $\mathbb{D}(\mathcal{P}_1^{W,p}, \mathcal{P}_2^{W,p})$ gives (A.16). Finally, combining inequalities (A.15) and (A.16) gives the desired inequality.

Next, we prove part (ii). Given the second-order growth condition, following the proof of Proposition EC.4 of Tsang and Shehadeh (2025), we have

$$D(\mathcal{X}_2^{W,p}, \mathcal{X}_1^{W,p}) \leq \sqrt{\frac{3}{\tau}\mathbb{H}(\mathcal{P}_1^{W,p}, \mathcal{P}_2^{W,p}; \mathsf{d}_{\mathcal{G}})} \leq \sqrt{\frac{6L_\xi}{\tau}\mathbb{H}(\mathcal{P}_1^{W,p}, \mathcal{P}_2^{W,p}; \mathsf{d}_{W,p})},$$

where the last inequality follows from (A.14). The desired inequality then follows from (A.16). □

Appendix A.6. Proof of Theorem 5

Proof. We first prove part (i). Let $\widehat{\boldsymbol{\varepsilon}}_{m,i} = \widehat{\boldsymbol{\xi}}^i - \widehat{f}_{m,N}(\boldsymbol{z}^i)$ and $\boldsymbol{\varepsilon}^\star_{m,i} = \widehat{\boldsymbol{\xi}}^i - f^\star_m(\boldsymbol{z}^i)$ for all $i \in [N]$ and $m \in [M]$. For notational simplicity, we suppress the dependence of $\widehat{\boldsymbol{\varepsilon}}_{m,i}$ and $\boldsymbol{\varepsilon}^\star_{m,i}$ on N. Define $\mathbb{P}^\star_{\boldsymbol{\varepsilon}}$ as the distribution of $\boldsymbol{\xi} - f^\star(\boldsymbol{Z})$, i.e., the true distribution of the residuals, and $\mathbb{P}^m_{\boldsymbol{\varepsilon}}$ as the distribution of $\boldsymbol{\xi} - f^\star_m(\boldsymbol{Z})$ for $m \in [M]$. By the triangle inequality, we have

$$\begin{aligned}
B_N &= \sup_{\boldsymbol{x} \in \mathcal{X}} \Big| \mathbb{E}_{\widehat{\mathbb{P}}_{f,N}}[\widehat{h}_N(f; \boldsymbol{x})] - \mathbb{E}_{\mathbb{P}^\star(\boldsymbol{z})}[c(\boldsymbol{x}, \boldsymbol{\xi})] \Big| \\
&\leq \sup_{\boldsymbol{x} \in \mathcal{X}} \left| \sum_{m=1}^M \widehat{w}_{m,N} \left[\frac{1}{N} \sum_{i=1}^N c\big(\boldsymbol{x}, \mathrm{proj}_\Xi(\widehat{f}_{m,N}(\boldsymbol{z}) + \widehat{\boldsymbol{\varepsilon}}_{m,i})\big) \right] - \sum_{m=1}^M \widehat{w}_{m,N} \left[\frac{1}{N} \sum_{i=1}^N c\big(\boldsymbol{x}, \mathrm{proj}_\Xi(f^\star_m(\boldsymbol{z}) + \boldsymbol{\varepsilon}^\star_{m,i})\big) \right] \right| \qquad \text{(A.18a)}
\end{aligned}$$

$$+\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\sum_{m=1}^{M}\widehat{w}_{m,N}\left[\frac{1}{N}\sum_{i=1}^{N}c\big(\boldsymbol{x},\mathrm{proj}_{\Xi}(f_m^\star(\boldsymbol{z})+\boldsymbol{\varepsilon}_{m,i}^\star)\big)\right]-\sum_{m=1}^{M}\widehat{w}_{m,N}\mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}^m}\Big[c\big(\boldsymbol{x},\mathrm{proj}_{\Xi}(f_m^\star(\boldsymbol{z})+\boldsymbol{\varepsilon})\big)\Big]\right| \tag{A.18b}$$

$$+\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\sum_{m=1}^{M}\widehat{w}_{m,N}\mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}^m}\Big[c\big(\boldsymbol{x},\mathrm{proj}_{\Xi}(f_m^\star(\boldsymbol{z})+\boldsymbol{\varepsilon})\big)\Big]-\mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}^\star}\big[c(\boldsymbol{x},f^\star(\boldsymbol{z})+\boldsymbol{\varepsilon})\big]\right|. \tag{A.18c}$$

We denote the terms in (A.18a), (A.18b), and (A.18c) as $T_{1,N}$, $T_{2,N}$, and $T_{3,N}$, respectively. Next, we analyze the convergence of these terms. First, from Theorem 2, we have

$$T_{1,N}\le 2L_\xi\sum_{m=1}^{M}\widehat{w}_{m,N}\|\widehat{f}_{m,N}-f_m^\star\|_\infty\le 2L_\xi\max_{m\in[M]}\|\widehat{f}_{m,N}-f_m^\star\|_\infty\xrightarrow{p}0, \tag{A.19}$$

where the second inequality follows from the fact that $0\le\widehat{w}_{m,N}\le 1$ for all $m\in[M]$, and the convergence follows from Assumption 4(a). Second, for $T_{2,N}$, we have

$$T_{2,N}\le\sum_{m=1}^{M}\widehat{w}_{m,N}\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\frac{1}{N}\sum_{i=1}^{N}c\big(\boldsymbol{x},\mathrm{proj}_{\Xi}(f_m^\star(\boldsymbol{z})+\boldsymbol{\varepsilon}_{m,i}^\star)\big)-\mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}^m}\Big[c\big(\boldsymbol{x},\mathrm{proj}_{\Xi}(f_m^\star(\boldsymbol{z})+\boldsymbol{\varepsilon})\big)\Big]\right|\xrightarrow{p}0, \tag{A.20}$$

where the convergence follows from Assumption 5 and a similar argument in (A.19). Finally, for $T_{3,N}$, we have

$$\begin{aligned}
T_{3,N}&=\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\sum_{m=1}^{M}\widehat{w}_{m,N}\mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}^m}\Big[c\big(\boldsymbol{x},\mathrm{proj}_{\Xi}(f_m^\star(\boldsymbol{z})+\boldsymbol{\varepsilon})\big)\Big]-\mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}^\star}\big[c(\boldsymbol{x},f^\star(\boldsymbol{z})+\boldsymbol{\varepsilon})\big]\right|\\
&\le\sum_{m\in\mathcal{M}_0}\widehat{w}_{m,N}\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}^m}\Big[c\big(\boldsymbol{x},\mathrm{proj}_{\Xi}(f_m^\star(\boldsymbol{z})+\boldsymbol{\varepsilon})\big)\Big]-\mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}^\star}\big[c(\boldsymbol{x},f^\star(\boldsymbol{z})+\boldsymbol{\varepsilon})\big]\right|\\
&\quad+\sum_{m\notin\mathcal{M}_0}\widehat{w}_{m,N}\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}^m}\Big[c\big(\boldsymbol{x},\mathrm{proj}_{\Xi}(f_m^\star(\boldsymbol{z})+\boldsymbol{\varepsilon})\big)\Big]-\mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}^\star}\big[c(\boldsymbol{x},f^\star(\boldsymbol{z})+\boldsymbol{\varepsilon})\big]\right|\\
&=\sum_{m\notin\mathcal{M}_0}\widehat{w}_{m,N}\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}^m}\Big[c\big(\boldsymbol{x},\mathrm{proj}_{\Xi}(f_m^\star(\boldsymbol{z})+\boldsymbol{\varepsilon})\big)\Big]-\mathbb{E}_{\mathbb{P}_{\boldsymbol{\varepsilon}}^\star}\big[c(\boldsymbol{x},f^\star(\boldsymbol{z})+\boldsymbol{\varepsilon})\big]\right| &\text{(A.21a)}\\
&\le\sum_{m\notin\mathcal{M}_0}\widehat{w}_{m,N}L_\xi\big[\|f_m^\star(\boldsymbol{z})-f^\star(\boldsymbol{z})\|+\mathsf{d}_{\mathrm{K}}(\mathbb{P}_{\boldsymbol{\varepsilon}}^m,\mathbb{P}_{\boldsymbol{\varepsilon}}^\star)\big] &\text{(A.21b)}\\
&\le L_\xi\max_{m\notin\mathcal{M}_0}\Big\{\|f_m^\star(\boldsymbol{z})-f^\star(\boldsymbol{z})\|+\mathsf{d}_{\mathrm{K}}(\mathbb{P}_{\boldsymbol{\varepsilon}}^m,\mathbb{P}_{\boldsymbol{\varepsilon}}^\star)\Big\}\cdot\sum_{m\notin\mathcal{M}_0}\widehat{w}_{m,N}\\
&\xrightarrow{p}0. &\text{(A.21c)}
\end{aligned}$$

Here, (A.21a) follows from Assumption 4 that $f_m^\star$ equals $f^\star$ for any $m\in\mathcal{M}_0$; (A.21b) follows from the proof of Theorem 1; (A.21c) follows from Assumption 4(a) that $\|f_m^\star-f^\star\|_\infty$ is finite for all $m\in[M]$ and Assumption 4(b) that $\sum_{m\notin\mathcal{M}_0}\widehat{w}_{m,N}\xrightarrow{p}0$. Since $B_N\le T_{1,N}+T_{2,N}+T_{3,N}$, it

follows from (A.18)–(A.21) that $B_N \xrightarrow{p} 0$, and thus $\widehat{v}_N^{\text{ER-MA}}(\boldsymbol{z}) \xrightarrow{p} v^\star(\boldsymbol{z})$. This completes the proof of part (i).

Next, we prove part (iii). In the following, we simply denote $v^\star(\boldsymbol{z})$ as $v^\star$, $\widehat{v}_N^{\text{ER-MA}}(\boldsymbol{z})$ as $\widehat{v}_N^{\text{ER-MA}}$, $\widehat{\mathcal{X}}_N^{\text{ER-MA}}(\boldsymbol{z})$ as $\widehat{\mathcal{X}}_N^{\text{ER-MA}}$, $\mathcal{X}^\star(\boldsymbol{z})$ as $\mathcal{X}^\star$, and $\mathbb{P}^\star(\boldsymbol{z})$ as $\mathbb{P}^\star$ by suppressing their dependence on $\boldsymbol{z}$ for ease of notation. Consider any $\boldsymbol{z} \in \mathcal{Z}$ such that $B_N \xrightarrow{p} 0$. For any $\varepsilon > 0$, we have

$$
\begin{aligned}
&\text{Prob}\left(\sup_{\boldsymbol{x}\in\widehat{\mathcal{X}}_N^{\text{ER-MA}}} \mathbb{E}_{\mathbb{P}^\star}[c(\boldsymbol{x},\boldsymbol{\xi})] - v^\star > \varepsilon\right) \\
&\le \text{Prob}\left(\sup_{\boldsymbol{x}\in\widehat{\mathcal{X}}_N^{\text{ER-MA}}} \mathbb{E}_{\mathbb{P}^\star}[c(\boldsymbol{x},\boldsymbol{\xi})] - \sup_{\boldsymbol{x}\in\mathcal{X}^\star} \mathbb{E}_{\widehat{\mathbb{P}}_{f,N}}[\widehat{h}_N(f;\boldsymbol{x})] > \frac{\varepsilon}{2}\right) + \text{Prob}\left(\sup_{\boldsymbol{x}\in\mathcal{X}^\star} \mathbb{E}_{\widehat{\mathbb{P}}_{f,N}}[\widehat{h}_N(f;\boldsymbol{x})] - v^\star > \frac{\varepsilon}{2}\right) \\
&\le \text{Prob}\left(\sup_{\boldsymbol{x}\in\widehat{\mathcal{X}}_N^{\text{ER-MA}}} \mathbb{E}_{\mathbb{P}^\star}[c(\boldsymbol{x},\boldsymbol{\xi})] - \widehat{v}_N^{\text{ER-MA}} > \frac{\varepsilon}{2}\right) + \text{Prob}\left(\sup_{\boldsymbol{x}\in\mathcal{X}^\star} \mathbb{E}_{\widehat{\mathbb{P}}_{f,N}}[\widehat{h}_N(f;\boldsymbol{x})] - v^\star > \frac{\varepsilon}{2}\right) \\
&\le \text{Prob}\left(\sup_{\boldsymbol{x}\in\widehat{\mathcal{X}}_N^{\text{ER-MA}}} \left|\mathbb{E}_{\mathbb{P}^\star}[c(\boldsymbol{x},\boldsymbol{\xi})] - \mathbb{E}_{\widehat{\mathbb{P}}_{f,N}}[\widehat{h}_N(f;\boldsymbol{x})]\right| > \frac{\varepsilon}{2}\right) + \text{Prob}\left(\sup_{\boldsymbol{x}\in\mathcal{X}^\star} \left|\mathbb{E}_{\widehat{\mathbb{P}}_{f,N}}[\widehat{h}_N(f;\boldsymbol{x})] - \mathbb{E}_{\mathbb{P}^\star}[c(\boldsymbol{x},\boldsymbol{\xi})]\right| > \frac{\varepsilon}{2}\right) \\
&\le 2\,\text{Prob}\left(B_N > \frac{\varepsilon}{2}\right) \\
&\xrightarrow{p} 0,
\end{aligned}
$$

where "Prob" denotes the probability generated from the data $\mathcal{D}_N$ and we suppress its dependence on N for simplicity. Here, the first inequality follows from the union bound; the second inequality follows from $\widehat{v}_N^{\text{ER-MA}} \le \sup_{\boldsymbol{x}\in\mathcal{X}^\star} \mathbb{E}_{\widehat{\mathbb{P}}_{f,N}}[\widehat{h}_N(f;\boldsymbol{x})]$; the third inequality follows from the facts that for any $\boldsymbol{x} \in \widehat{\mathcal{X}}_N^{\text{ER-MA}}$, we have $\mathbb{E}_{\mathbb{P}^\star}[c(\boldsymbol{x},\boldsymbol{\xi})] - \widehat{v}_N^{\text{ER-MA}} = \mathbb{E}_{\mathbb{P}^\star}[c(\boldsymbol{x},\boldsymbol{\xi})] - \mathbb{E}_{\widehat{\mathbb{P}}_{f,N}}[\widehat{h}_N(f;\boldsymbol{x})] \le \left|\mathbb{E}_{\mathbb{P}^\star}[c(\boldsymbol{x},\boldsymbol{\xi})] - \mathbb{E}_{\widehat{\mathbb{P}}_{f,N}}[\widehat{h}_N(f;\boldsymbol{x})]\right|$ and for any $\boldsymbol{x} \in \mathcal{X}^\star$, we have $\mathbb{E}_{\widehat{\mathbb{P}}_{f,N}}[\widehat{h}_N(f;\boldsymbol{x})] - v^\star = \mathbb{E}_{\widehat{\mathbb{P}}_{f,N}}[\widehat{h}_N(f;\boldsymbol{x})] - \mathbb{E}_{\mathbb{P}^\star}[c(\boldsymbol{x},\boldsymbol{\xi})] \le \left|\mathbb{E}_{\widehat{\mathbb{P}}_{f,N}}[\widehat{h}_N(f;\boldsymbol{x})] - \mathbb{E}_{\mathbb{P}^\star}[c(\boldsymbol{x},\boldsymbol{\xi})]\right|$; the last inequality follows from replacing the supremum over $\boldsymbol{x} \in \widehat{\mathcal{X}}_N^{\text{ER-MA}}$ and $\boldsymbol{x} \in \mathcal{X}^\star$ with supremum over $\boldsymbol{x} \in \mathcal{X}$. This completes the proof of part (iii).

Finally, we prove part (ii). Similarly to the above, we again suppress the dependence of the sets $\widehat{\mathcal{X}}_N^{\text{ER-MA}}$ and $\mathcal{X}^\star$, optimal value $v^\star$, and the true conditional distribution $\mathbb{P}^\star$ on $\boldsymbol{z}$ for ease of notation. Again, consider any $\boldsymbol{z} \in \mathcal{Z}$ such that $B_N \xrightarrow{p} 0$. Suppose, on the contrary, that $D(\widehat{\mathcal{X}}_N^{\text{ER-MA}}, \mathcal{X}^\star) \not\xrightarrow{p} 0$. That is, there exist $\delta > 0$ and $\beta > 0$ such that $\text{Prob}\left(D(\widehat{\mathcal{X}}_{N_r}^{\text{ER-MA}}, \mathcal{X}^\star) > \delta\right) \ge \beta$ for some subsequence $\{N_r\}_{r\in\mathbb{N}} \subseteq \mathbb{N}$. Note that by Assumptions 1 and 2, $c(\cdot,\boldsymbol{\xi})$ and thus $\mathbb{E}_{\mathbb{P}^\star}[c(\cdot,\boldsymbol{\xi})]$ are continuous and the set $\mathcal{X}_\delta^\star := \{\boldsymbol{x} \in \mathcal{X} \mid d(\boldsymbol{x}, \mathcal{X}^\star) \ge \delta\}$ is closed and non-empty. Therefore, if $D(\widehat{\mathcal{X}}_{N_r}^{\text{ER-MA}}, \mathcal{X}^\star) > \delta$, we have $\sup_{\boldsymbol{x}\in\widehat{\mathcal{X}}_{N_r}^{\text{ER-MA}}} \mathbb{E}_{\mathbb{P}^\star}[c(\boldsymbol{x},\boldsymbol{\xi})] - v^\star \ge \inf_{\boldsymbol{x}\in\mathcal{X}_\delta^\star} \mathbb{E}_{\mathbb{P}^\star}[c(\boldsymbol{x},\boldsymbol{\xi})] - v^\star =: \kappa_\delta > 0$. In other words, we have $\text{Prob}\left(\sup_{\boldsymbol{x}\in\widehat{\mathcal{X}}_{N_r}^{\text{ER-MA}}} \mathbb{E}_{\mathbb{P}^\star}[c(\boldsymbol{x},\boldsymbol{\xi})] - v^\star \ge \kappa_\delta\right) \ge \beta$, which contradicts the result in part (iii) that $\sup_{\boldsymbol{x}\in\widehat{\mathcal{X}}_{N_r}^{\text{ER-MA}}} \mathbb{E}_{\mathbb{P}^\star}[c(\boldsymbol{x},\boldsymbol{\xi})] \xrightarrow{p} v^\star$. This completes the proof of part (ii). □

Appendix A.7. Proof of Theorem 6

Proof. Consider the term A_N in (13). Define the set of functions $\mathcal{G} := \{\widehat{h}_N(\cdot;\boldsymbol{x}) \mid \boldsymbol{x} \in \mathcal{X}\}$. Note that $A_N \le \mathbb{H}\big(\mathcal{P}_N^{\text{SR}}, \{\widehat{\mathbb{P}}_{f,N}\}, d_{\mathcal{G}}\big) \le 2L_\xi \max_{m\in[M]}\{r_{m,N}\}$, where the first inequality follows from

Proposition EC.3 of Tsang and Shehadeh (2025) and the second inequality follows from (A.12). Since $C_N < \infty$ a.s. with $r_{m,N} = o_p(1)$ for all $m \in [M]$, we have $A_N \xrightarrow{p} 0$. Moreover, by Theorem 5, we have $B_N \xrightarrow{p} 0$. It follows from (13) that $\widehat{v}_N^{\mathrm{SR}}(\boldsymbol{z}) \xrightarrow{p} v^\star(\boldsymbol{z})$. Again, given the convergence of the optimal value in probability, parts (ii) and (iii) then follow from the same arguments in the proof of Theorem 5.

□

Appendix A.8. Proof of Theorem 7

Proof. Consider the term A_N in (13). Define the set of functions $\mathcal{G} = \{\widehat{h}_N(\cdot;\boldsymbol{x}) \mid \boldsymbol{x} \in \mathcal{X}\}$. Note that $A_N \le \mathbb{H}\big(\mathcal{P}_N^{\mathrm{W,p}}, \{\widehat{\mathbb{P}}_{f,N}\}, d_{\mathcal{G}}\big) \le 2L_\xi r_N$, where the first inequality follows from Proposition EC.3 of Tsang and Shehadeh (2025) and the second inequality follows from (A.16). Moreover, by Theorem 5, we have $B_N \xrightarrow{p} 0$. It follows from (13) that $\widehat{v}_N^{W,p} \xrightarrow{p} v^\star$. Again, given the convergence of the optimal value in probability, parts (ii) and (iii) then follow from the same arguments in the proof of Theorem 5. □

Appendix A.9. Proof of Theorem 8

Proof. In the following, we simply denote $v^\star(\boldsymbol{z})$ as $v^\star$, $\widehat{v}_N^{\text{ER-MA}}(\boldsymbol{z})$ as $\widehat{v}_N^{\text{ER-MA}}$, $V(\boldsymbol{z})$ as V by suppressing their dependence on $\boldsymbol{z}$ for ease of notation. First, we derive an upper bound on the conditional probability $\mathrm{Prob}(|\widehat{v}_N^{\text{ER-MA}} - v^\star| > \kappa \mid \widehat{\boldsymbol{w}}_N)$. For notational simplicity, we use $\mathrm{Prob}_{\widehat{\boldsymbol{w}}_N}$ to denote the probability conditional on $\widehat{\boldsymbol{w}}_N$. Note that

$$
\begin{aligned}
&\mathrm{Prob}_{\widehat{\boldsymbol{w}}_N}(|\widehat{v}_N^{\text{ER-MA}} - v^\star| > \kappa) \\
&\le \mathrm{Prob}_{\widehat{\boldsymbol{w}}_N}(B_N > \kappa) && \text{(A.22a)} \\
&\le \mathrm{Prob}_{\widehat{\boldsymbol{w}}_N}(T_{1,N} + T_{2,N} + T_{3,N} > \kappa) && \text{(A.22b)} \\
&\le \mathrm{Prob}_{\widehat{\boldsymbol{w}}_N}\left(T_{1,N} > \frac{\kappa}{3}\right) + \mathrm{Prob}_{\widehat{\boldsymbol{w}}_N}\left(T_{2,N} > \frac{\kappa}{3}\right) + \mathrm{Prob}_{\widehat{\boldsymbol{w}}_N}\left(T_{3,N} > \frac{\kappa}{3}\right), && \text{(A.22c)}
\end{aligned}
$$

where (A.22a) follows from (13), (A.22b) follows from (A.18), and (A.22c) follows from Boole's inequality (union bound). In the following, we derive the bounds for the three probabilities in (A.22c). First,

$$
\begin{aligned}
\mathrm{Prob}_{\widehat{\boldsymbol{w}}_N}\left(T_{1,N} > \frac{\kappa}{3}\right) &\le \mathrm{Prob}_{\widehat{\boldsymbol{w}}_N}\left(2L_\xi \sum_{m=1}^M w_m \|\widehat{f}_{m,N} - f_m^\star\|_\infty > \frac{\kappa}{3}\right) && \text{(A.23a)} \\
&\le \inf_{\boldsymbol{\alpha}\in\Delta_M} \sum_{m=1}^M \mathrm{Prob}_{\widehat{\boldsymbol{w}}_N}\left(\|\widehat{f}_{m,N} - f_m^\star\|_\infty > \frac{\kappa\alpha_m}{6L_\xi w_m}\right) \cdot \mathbf{1}(w_m \neq 0) && \text{(A.23b)} \\
&\le \inf_{\boldsymbol{\alpha}\in\Delta_M} \sum_{m=1}^M K_m\left(\frac{\kappa\alpha_m}{6L_\xi w_m}\right) \exp\left\{-\beta_m\left(N, \frac{\kappa\alpha_m}{6L_\xi w_m}\right)\right\} \cdot \mathbf{1}(w_m \neq 0) && \text{(A.23c)}
\end{aligned}
$$

Here, (A.23a) follows from (A.19). (A.23b) follows from the fact that for any random variables $\{X_m\}_{m=1}^M$ and $\delta > 0$, we have $\mathrm{Prob}(\sum_{m=1}^M X_m > \delta) \le \inf_{\boldsymbol{\alpha}\in\Delta_M} \sum_{m=1}^M \mathrm{Prob}(X_m > \alpha_m\delta)$. Indeed, if $\sum_{m=1}^M X_m > \delta$, then there exists $m \in [M]$ such that $X_m > \alpha_m\delta$ for any given $\boldsymbol{\alpha} \in \Delta_M$. Thus, applying the union bound, we have $\mathrm{Prob}(\sum_{m=1}^M X_m > \delta) \le \sum_{m=1}^M \mathrm{Prob}(X_m > \alpha_m\delta)$ for any $\boldsymbol{\alpha} \in \Delta_M$, which yields the desired inequality. (A.23c) follows from Assumption 6. Next,

$$\mathrm{Prob}_{\widehat{\boldsymbol{w}}_N}\left(T_{2,N} > \frac{\kappa}{3}\right) \le \mathrm{Prob}_{\widehat{\boldsymbol{w}}_N}\left(\sum_{m=1}^M w_m \sup_{\boldsymbol{x}\in\mathcal{X}}\left|\widehat{h}_N(f_m^\star;\boldsymbol{x}) - h^\star(f_m^\star;\boldsymbol{x})\right| > \frac{\kappa}{3}\right) \tag{A.24a}$$

$$\le \inf_{\boldsymbol{\alpha}\in\Delta_M} \sum_{m=1}^M \mathrm{Prob}_{\widehat{\boldsymbol{w}}_N}\left(\sup_{\boldsymbol{x}\in\mathcal{X}}\left|\widehat{h}_N(f_m^\star;\boldsymbol{x}) - h^\star(f_m^\star;\boldsymbol{x})\right| > \frac{\kappa\alpha_m}{3w_m}\right)\cdot \mathbf{1}(w_m \neq 0) \tag{A.24b}$$

$$\le \inf_{\boldsymbol{\alpha}\in\Delta_M} \sum_{m=1}^M K_{f_m^\star}\left(\frac{\kappa\alpha_m}{3w_m}, \boldsymbol{z}\right)\exp\left\{-\beta_{f_m^\star}\left(N, \frac{\kappa\alpha_m}{3w_m}, \boldsymbol{z}\right)\right\}\cdot \mathbf{1}(w_m \neq 0) \tag{A.24c}$$

Here, (A.24a) follows from (A.20); (A.24b) follows from the same argument as in (A.23b); (A.24c) follows from Assumption 7. Finally,

$$\mathrm{Prob}_{\widehat{\boldsymbol{w}}_N}\left(T_{3,N} > \frac{\kappa}{3}\right) \le \mathrm{Prob}_{\widehat{\boldsymbol{w}}_N}\left(V\sum_{m\notin\mathcal{M}_0} w_m > \frac{\kappa}{3}\right) = \mathbf{1}\left(\sum_{m\notin\mathcal{M}_0} w_m > \frac{\kappa}{3V}\right), \tag{A.25}$$

where the inequality follows from (A.21). Combining (A.22)–(A.25), we have

$$\begin{aligned}\mathrm{Prob}_{\widehat{\boldsymbol{w}}_N}(|\widehat{v}_N^{\text{ER-MA}} - v^\star| > \kappa) \le{}& \inf_{\boldsymbol{\alpha}\in\Delta_M} \sum_{m=1}^M K_m\left(\frac{\kappa\alpha_m}{6L_\xi w_m}\right)\exp\left\{-\beta_m\left(N, \frac{\kappa\alpha_m}{6L_\xi w_m}\right)\right\}\cdot \mathbf{1}(w_m \neq 0)\\ &+ \inf_{\boldsymbol{\alpha}\in\Delta_M} \sum_{m=1}^M K_{f_m^\star}\left(\frac{\kappa\alpha_m}{3w_m}, \boldsymbol{z}\right)\exp\left\{-\beta_{f_m^\star}\left(N, \frac{\kappa\alpha_m}{3w_m}, \boldsymbol{z}\right)\right\}\cdot \mathbf{1}(w_m \neq 0)\\ &+ \mathbf{1}\left(\sum_{m\notin\mathcal{M}_0} w_m > \frac{\kappa}{3V}\right).\end{aligned} \tag{A.26}$$

Now, denote $\mathbb{E}_{\widehat{\boldsymbol{w}}_N}$ as the expectation over the distribution of $\widehat{\boldsymbol{w}}_N$. The desired inequality (17) follows from

$$\begin{aligned}&\mathrm{Prob}(|\widehat{v}_N^{\text{ER-MA}} - v^\star| > \kappa)\\ &= \mathbb{E}_{\widehat{\boldsymbol{w}}_N}\left[\mathrm{Prob}(|\widehat{v}_N^{\text{ER-MA}} - v^\star| > \kappa \mid \widehat{\boldsymbol{w}}_N)\right] \qquad \text{(A.27a)}\\ &\le \mathbb{E}_{\widehat{\boldsymbol{w}}_N}\left[\inf_{\boldsymbol{\alpha}\in\Delta_M} \sum_{m=1}^M K_m\left(\frac{\kappa\alpha_m}{6L_\xi \widehat{w}_{m,N}}\right)\exp\left\{-\beta_m\left(N, \frac{\kappa\alpha_m}{6L_\xi \widehat{w}_{m,N}}\right)\right\}\cdot \mathbf{1}(\widehat{w}_{m,N} \neq 0)\right]\end{aligned}$$

$$+\mathbb{E}_{\widehat{\boldsymbol{w}}_N}\left[\inf_{\boldsymbol{\alpha}\in\Delta_M}\sum_{m=1}^{M}K_{f_m^\star}\left(\frac{\kappa\alpha_m}{3\widehat{w}_{m,N}},\boldsymbol{z}\right)\exp\left\{-\beta_{f_m^\star}\left(N,\frac{\kappa\alpha_m}{3\widehat{w}_{m,N}},\boldsymbol{z}\right)\right\}\cdot\mathbf{1}(\widehat{w}_{m,N}\neq 0)\right]$$
$$+\operatorname{Prob}\left(\sum_{m\notin\mathcal{M}_0}\widehat{w}_{m,N}>\frac{\kappa}{3V}\right), \tag{A.27b}$$
$$\leq \mathbb{E}_{\widehat{\boldsymbol{w}}_N}\left[\sum_{m=1}^{M}K_m\left(\frac{\kappa}{6L_\xi}\right)\exp\left\{-\beta_m\left(N,\frac{\kappa}{6L_\xi}\right)\right\}\cdot\mathbf{1}(\widehat{w}_{m,N}\neq 0)\right]$$
$$+\mathbb{E}_{\widehat{\boldsymbol{w}}_N}\left[\sum_{m=1}^{M}K_{f_m^\star}\left(\frac{\kappa}{3},\boldsymbol{z}\right)\exp\left\{-\beta_{f_m^\star}\left(N,\frac{\kappa}{3},\boldsymbol{z}\right)\right\}\cdot\mathbf{1}(\widehat{w}_{m,N}\neq 0)\right]$$
$$+\operatorname{Prob}\left(\sum_{m\notin\mathcal{M}_0}\widehat{w}_{m,N}>\frac{\kappa}{3V}\right), \tag{A.27c}$$
$$\leq \sum_{m=1}^{M}K_m\left(\frac{\kappa}{6L_\xi}\right)\exp\left\{-\beta_m\left(N,\frac{\kappa}{6L_\xi}\right)\right\}+\sum_{m=1}^{M}K_{f_m^\star}\left(\frac{\kappa}{3},\boldsymbol{z}\right)\exp\left\{-\beta_{f_m^\star}\left(N,\frac{\kappa}{3},\boldsymbol{z}\right)\right\}$$
$$+\operatorname{Prob}\left(\sum_{m\notin\mathcal{M}_0}\widehat{w}_{m,N}>\frac{\kappa}{3V}\right), \tag{A.27d}$$
$$\leq \overline{K}(\kappa,\boldsymbol{z})\exp\left\{-\overline{\beta}(N,\kappa,\boldsymbol{z})\right\}+\operatorname{Prob}\left(\sum_{m\notin\mathcal{M}_0}\widehat{w}_{m,N}>\frac{\kappa}{3V}\right). \tag{A.27e}$$

Here, (A.27a) follows from the total law of expectation; (A.27b) follows from (A.26); (A.27c) follows from setting $\boldsymbol{\alpha}=\widehat{\boldsymbol{w}}_N$; (A.27d) follows from the fact that $\operatorname{Prob}(\widehat{w}_{m,N}\neq 0)\leq 1$ for all $m\in[M]$; (A.27e) follows from setting $\overline{K}(\kappa,\boldsymbol{z})=\max_{m\in[M]}\left\{\max\{K_m(\kappa/6L_\xi),K_{f_m^\star}(\kappa/3,\boldsymbol{z})\}\right\}$ and $\overline{\beta}(N,\kappa,\boldsymbol{z})=\min_{m\in[M]}\left\{\min\{\beta_m(N,\kappa/6L_\xi),\beta_{f_m^\star}(N,\kappa/3,\boldsymbol{z})\}\right\}$. □

Appendix A.10. Proof of Theorem 9

Proof. In the following, we simply denote $\upsilon^\star(\boldsymbol{z})$ as $\upsilon^\star$, $\widehat{\upsilon}_N^{\mathrm{SR}}(\boldsymbol{z})$ as $\widehat{\upsilon}_N^{\mathrm{SR}}$, and $V(\boldsymbol{z})$ as V by suppressing their dependence on $\boldsymbol{z}$ for ease of notation. For a fixed $\kappa>0$, let $R_N=\operatorname{ess\,sup}\max_{m\in[M]}\{r_{m,N}\}$ and $R_N^c=\kappa-2L_\xi\operatorname{ess\,sup}\max_{m\in[M]}\{r_{m,N}\}$. Note that $2L_\xi R_N+R_N^c=\kappa$. Also, since $r_{m,N}\leq\kappa/2L_\xi-\epsilon$ a.s. for $m\in[M]$, we have $R_N^c>0$. It follows from (13) and the union bound that

$$\operatorname{Prob}(|\widehat{\upsilon}_N^{\mathrm{SR}}-\upsilon^\star|>\kappa)\leq\operatorname{Prob}(A_N>2L_\xi R_N)+\operatorname{Prob}(B_N>R_N^c). \tag{A.28}$$

It follows from the proof of Theorem 6 that $A_N\leq 2L_\xi\max_{m\in[M]}\{r_{m,N}\}\leq 2L_\xi R_N$ a.s. for some constant $K>0$. Therefore, we have

$$\operatorname{Prob}(A_N>2L_\xi R_N)\leq\operatorname{Prob}(2L_\xi R_N>2L_\xi R_N)=0. \tag{A.29}$$

Next, it follows from Theorem 8 that there exist constants $\overline{K}(\kappa, \boldsymbol{z}) > 0$ and $\overline{\beta}(N, \kappa, \boldsymbol{z}) > 0$ with $\lim_{N\to\infty} \overline{\beta}(N, \kappa, \boldsymbol{z}) = \infty$ such that

$$\text{Prob}(B_N > R_N^c) \leq \overline{K}(R_N^c, \boldsymbol{z}) \exp\big\{ -\overline{\beta}(N, R_N^c, \boldsymbol{z})\big\} + \text{Prob}\left(\sum_{m\notin\mathcal{M}_0} \widehat{w}_{m,N} > \frac{R_N^c}{3V}\right). \tag{A.30}$$

The desired inequality (18) follows from (A.28)–(A.30). □

Appendix A.11. Proof of Theorem 10

Proof. In the following, we simply denote $\upsilon^\star(\boldsymbol{z})$ as $\upsilon^\star$, $\widehat{\upsilon}_N^{W,p}(\boldsymbol{z})$ as $\widehat{\upsilon}_N^{W,p}$, and $V(\boldsymbol{z})$ as V by suppressing their dependence on $\boldsymbol{z}$ for ease of notation. For a fixed $\kappa > 0$, let $R_N = \operatorname{ess\,sup} r_N$ and $R_N^c = \kappa - 2L_\xi \operatorname{ess\,sup} r_N$. Note that $2L_\xi R_N + R_N^c = \kappa$. Also, since $r_N \leq \kappa/2L_\xi - \epsilon$ a.s. for $m \in [M]$, we have $R_N^c > 0$. It follows from (13) and the union bound that

$$\text{Prob}(|\widehat{\upsilon}_N^{W,p} - \upsilon^\star| > \kappa) \leq \text{Prob}(A_N > 2L_\xi R_N) + \text{Prob}(B_N > R_N^c). \tag{A.31}$$

From the proof of Theorem 7, we have a.s. $A_N \leq 2L_\xi r_N \leq 2L_\xi R_N$. Therefore, we have

$$\text{Prob}(A_N > 2L_\xi R_N) \leq \text{Prob}(2L_\xi R_N > 2L_\xi R_N) = 0. \tag{A.32}$$

Next, it follows from Theorem 8 that there exist constants $\overline{K}(\kappa, \boldsymbol{z}) > 0$ and $\overline{\beta}(N, \kappa, \boldsymbol{z}) > 0$ with $\lim_{N\to\infty} \overline{\beta}(N, \kappa, \boldsymbol{z}) = \infty$ such that

$$\text{Prob}(B_N > R_N^c) \leq \overline{K}(R_N^c, \boldsymbol{z}) \exp\big\{ -\overline{\beta}(N, R_N^c, \boldsymbol{z})\big\} + \text{Prob}\left(\sum_{m\notin\mathcal{M}_0} \widehat{w}_{m,N} > \frac{R_N^c}{3V}\right). \tag{A.33}$$

The desired inequality (19) follows from (A.31)–(A.33). □

Appendix B. The C&CG Algorithm for Solving the MUR-SAA Models

In this section, we present a C&CG algorithm to solve the reformulations (20) and (22) of the MUR-SAA model formed the sample robust and p-Wasserstein MUR ambiguity sets, respectively. In Appendix B.1, we first outline the general C&CG framework. Subsequently, in Appendix B.2, we derive tractable reformulations of the subproblem solved during each iteration of the C&CG algorithm.

Appendix B.1. The C&CG Framework

Algorithm 1 summarizes the steps of our C&CG algorithm. We initialize the algorithm with a lower bound $L \leftarrow 0$, an upper bound $U \leftarrow \infty$, and scenario set $\widehat{\mathcal{U}}_{m,1} \leftarrow \{\boldsymbol{\Theta}_{m,1}\}$ for all $m \in [M]$. In each iteration $s \in \mathbb{N}$, we first solve a master problem, which is a relaxation. For the sample robust MUR ambiguity set, the master problem is a relaxation of (20) defined on $\widehat{\mathcal{U}}_{m,s} = \{\boldsymbol{\theta}_{m,1}, \ldots, \boldsymbol{\theta}_{m,s}\} \subseteq$

$\mathfrak{T}_m$ for $m \in [M]$:

$$\underset{\boldsymbol{x}\in\mathcal{X},\,\boldsymbol{\eta}}{\text{minimize}} \quad \sum_{m=1}^{M} w_m \eta_m \tag{B.1a}$$

$$\text{subject to} \quad \eta_m \geq \frac{1}{N}\sum_{i=1}^{N} c\Big(\boldsymbol{x}, \operatorname{proj}_{\Xi}\left([f(\boldsymbol{z};\boldsymbol{\theta}_{m,\sigma}) - f(\boldsymbol{z}^i;\boldsymbol{\theta}_{m,\sigma})] + \widehat{\boldsymbol{\xi}}^i\right)\Big), \quad \forall \sigma \in [s],\, m \in [M], \tag{B.1b}$$

For the p-Wasserstein MUR ambiguity set, the master problem is a relaxation of (20) defined on $\widehat{\mathcal{U}}_{m,s} = \{\boldsymbol{\theta}_{m,1}, \ldots, \boldsymbol{\theta}_{m,s}\} \subseteq \mathfrak{T}$ for $m \in [M]$:

$$\underset{\boldsymbol{x}\in\mathcal{X},\,\lambda,\,\boldsymbol{\eta}}{\text{minimize}} \quad r^p\lambda + \sum_{m=1}^{M} w_m \eta_m \tag{B.2a}$$

$$\text{subject to} \quad \eta_m \geq \frac{1}{N}\sum_{i=1}^{N} c\Big(\boldsymbol{x}, \operatorname{proj}_{\Xi}\left([f(\boldsymbol{z};\boldsymbol{\theta}_{m,\sigma}) - f(\boldsymbol{z}^i;\boldsymbol{\theta}_{m,\sigma})] + \widehat{\boldsymbol{\xi}}^i\right)\Big) - \lambda\|\boldsymbol{\theta}_{m,\sigma} - \widehat{\boldsymbol{\theta}}_m\|^p,$$

$$\forall \sigma \in [s],\, m \in [M]. \tag{B.2b}$$

Note that if the cost function $c(\cdot, \boldsymbol{\xi})$ is convex in $\boldsymbol{x}$ for any $\boldsymbol{\xi} \in \Xi$ and the feasible set $\mathcal{X}$ is characterized by linear constraints, both master problems (B.1) and (B.2) are (mixed-integer) convex programs, which can be solved using off-the-shelf optimization solvers. Let $(\boldsymbol{x}_s, \boldsymbol{\eta}_s)$ (resp. $(\boldsymbol{x}_s, \lambda_s, \boldsymbol{\eta}_s)$) and υ_s denote the optimal solution and value of (B.1) (resp. (B.2)) in iteration s. Optimal value of the master problems (B.1) and (B.2) provide a lower bound. Hence, in Step 1.2, we update the lower bound: $L \leftarrow \upsilon_s$.

Next, in Step 2.1, we solve a subproblem at the identified solution $\boldsymbol{x}_s$ to identify potential new $\boldsymbol{\theta}$ to enlarge the sets $\{\widehat{\mathcal{U}}_{m,s}\}_{m=1}^{M}$. For the sample robust MUR ambiguity set, we solve the following subproblem for each $m \in [M]$:

$$\sup_{\boldsymbol{\theta}_m \in \mathfrak{T}_m} \frac{1}{N}\sum_{i=1}^{N} c\Big(\boldsymbol{x}, \operatorname{proj}_{\Xi}\left([f(\boldsymbol{z};\boldsymbol{\theta}_m) - f(\boldsymbol{z}^i;\boldsymbol{\theta}_m)] + \widehat{\boldsymbol{\xi}}^i\right)\Big). \tag{B.3}$$

For the p-Wasserstein MUR ambiguity set, we solve the following subproblem for each $m \in [M]$:

$$\sup_{\boldsymbol{\theta}_m \in \mathfrak{T}} \left\{ \frac{1}{N}\sum_{i=1}^{N} c\Big(\boldsymbol{x}, \operatorname{proj}_{\Xi}\left([f(\boldsymbol{z};\boldsymbol{\theta}_m) - f(\boldsymbol{z}^i;\boldsymbol{\theta}_m)] + \widehat{\boldsymbol{\xi}}^i\right)\Big) - \lambda_s\|\boldsymbol{\theta}_m - \widehat{\boldsymbol{\theta}}_m\|^p \right\}. \tag{B.4}$$

Denote an optimal solution to (B.3) or (B.4) as $\boldsymbol{\theta}_{m,s+1}$ and the optimal value as $V_{m,s+1}$ for all $m \in [M]$. At the end of this step, we have a potential new upper bound. We therefore update the upper bound $U \leftarrow \min\{U, \sum_{m=1}^{M} w_m V_{m,s+1}\}$ for the sample robust MUR ambiguity set or $U \leftarrow \min\{U, r^p\lambda_s + \sum_{m=1}^{M} w_m V_{m,s+1}\}$ for the p-Wasserstein MUR ambiguity set in Step 2.2. Finally, in Step 3.1, we compute the absolute gap $U - L$. If this gap is within the pre-specified tolerance

Algorithm 1: A C&CG method for solving the MUR-SAA model with the sample robust (`SR`) and the p-Wasserstein (`Wass`) MUR ambiguity sets.

Initialization: $s \leftarrow 1$, $\varepsilon > 0$, $L \leftarrow 0$, $U \leftarrow \infty$, $\widehat{\mathcal{U}}_{m,1} \leftarrow \{\boldsymbol{\Theta}_{m,1}\}$ for all $m \in [M]$.

1. Master Problem and Lower Bound.

1.1 Solve the master problem (B.1) if `SR` or (B.2) if `Wass`. Record its optimal value υ_s and the obtained optimal solution $(\boldsymbol{x}_s, \boldsymbol{\eta}_s)$ if `SR` or $(\boldsymbol{x}_s, \lambda_s, \boldsymbol{\eta}_s)$ if `Wass`.

1.2 Update the lower bound $L \leftarrow \upsilon_k$.

2. Subproblems and Upper Bound.

2.1 Solve the subproblem (B.3) if `SR` or (B.4) if `Wass` for $m \in [M]$. Record the optimal value $\{V_{m,s+1}\}_{m=1}^M$ and optimal solution $\{\boldsymbol{\Theta}_{m,s+1}\}_{m=1}^M$.

2.2 Update the upper bound $U \leftarrow \min\{U, \sum_{m=1}^M w_m V_{m,s+1}\}$ if `SR` or $U \leftarrow \min\{U, r^p \lambda_s + \sum_{m=1}^M \widehat{w}_{m,N} V_m^{s+1}\}$ if `Wass`.

3. Optimality Check.

3.1 If $U - L \leq \varepsilon$, terminate. Return the best feasible solution with objective value U.

3.2 Otherwise, enlarge the set $\widehat{\mathcal{U}}_{m,s+1} \leftarrow \widehat{\mathcal{U}}_{m,s} \cup \{\boldsymbol{\theta}_{m,s+1}\}$ for all $m \in [M]$. Update $s \leftarrow s+1$. Return to Step 1.

ε, we terminate and return the best feasible solution with the best objective value U. Otherwise, we proceed to Step 3.2, where we enlarge the set $\widehat{\mathcal{U}}_{m,s+1} \leftarrow \widehat{\mathcal{U}}_{m,s} \cup \{\boldsymbol{\theta}_{m,s+1}\}$ for all $m \in [M]$, and return to step 1 to solve the master problem with these enlarged sets. It follows from the proof of Theorem 4 in Tsang and Shehadeh (2026) that Algorithm 1 converges in a finite number of iterations for any $\varepsilon > 0$ if the set $\mathcal{X}$ is compact and $c\big(\boldsymbol{x}, \mathrm{proj}_{\Xi}\left([f(\boldsymbol{z};\boldsymbol{\theta}) - f(\boldsymbol{z}^i;\boldsymbol{\theta})] + \widehat{\boldsymbol{\xi}}^i\right)\big)$ is uniformly Lipschitz continuous in $\boldsymbol{x}$ over $\boldsymbol{\theta}_m \in \mathfrak{T}$ for all $i \in [N]$.

Appendix B.2. Reformulation of the Subproblem

Let us now examine the subproblems given in (B.3) and (B.4) in more detail. Solving subproblems (B.3) and (B.4) in the presented form are challenging mainly because they are (nonlinear) max-min problems with a possibly nonlinear cost function c combined with the projection term $\mathrm{proj}_{\Xi}\left([f(\boldsymbol{z};\boldsymbol{\theta}_{m,\sigma}) - f(\boldsymbol{z}^i;\boldsymbol{\theta}_{m,\sigma})] + \widehat{\boldsymbol{\xi}}^i\right)\big)$ that depends on decisions $\boldsymbol{\theta}_m$. Observe that, in contrast, the projection term in the master problems' constraints (B.1b) and (B.2b) do not depend on the decision variables of the master problem and can therefore be calculated offline whenever a new $\boldsymbol{\theta}_{m,\sigma}$ is obtained. Thus, in this section, we focus on deriving tractable reformulations of the subproblem (B.3) for the sample robust MUR ambiguity set under some popular choices of function classes and cost function. Using the same technique, one could easily derive tractable reformulations of the subproblem (B.4) for the p-Wasserstein robust MUR ambiguity set with the reformulation the additional term $\|\boldsymbol{\theta}_m - \widehat{\boldsymbol{\theta}}_m\|^p$.

To derive tractable reformulation of the subproblem (B.3), we assume that the cost function takes the form $c(\boldsymbol{x},\boldsymbol{\xi}) = \boldsymbol{d}^\top \boldsymbol{x} + Q(\boldsymbol{x},\boldsymbol{\xi})$, where $Q(\boldsymbol{x},\boldsymbol{\xi}) = \min_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x},\boldsymbol{\xi})}\{\boldsymbol{q}^\top \boldsymbol{y}\}$ and $\mathcal{Y}(\boldsymbol{x},\boldsymbol{\xi}) := \{\boldsymbol{y} \geq \boldsymbol{0} \mid \boldsymbol{T}\boldsymbol{x} + \boldsymbol{W}\boldsymbol{y} + \boldsymbol{C}\boldsymbol{\xi} \geq \boldsymbol{h}\} \subseteq \mathbb{R}^{d_y}$ with $\boldsymbol{h} \in \mathbb{R}^{d_h}$, $\boldsymbol{T} \in \mathbb{R}^{d_h \times d_x}$, $\boldsymbol{W} \in \mathbb{R}^{d_h \times d_y}$, and $\boldsymbol{C} \in \mathbb{R}^{d_h \times d_\xi}$. Here, $\boldsymbol{C}$

can be a function of $\boldsymbol{x}$. This cost function is common in many two-stage stochastic optimization models, where $\boldsymbol{d}^\top \boldsymbol{x}$ and $Q(\boldsymbol{x}, \boldsymbol{\xi})$ represent the first-stage and second-stage costs, respectively. We assume that $Q(\boldsymbol{x}, \boldsymbol{\xi})$ is finite for any $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{\xi} \in \Xi$ to ensure that the second-stage problem is always feasible. In addition, we consider the following the function classes:

- *Linear Regression:* Let $\mathcal{F} = \{f(\boldsymbol{z}) = \boldsymbol{\Theta}\boldsymbol{z} \mid \boldsymbol{\Theta} \in \mathfrak{T}\}$ be the class of linear regression functions, where $\mathfrak{T} \subseteq \mathbb{R}^{d_\xi \times d_z}$ is compact. The covariate vector $\boldsymbol{z}$ may include a constant component to account for the intercept term. This class encompasses a wide range of commonly used linear models, including ordinary least squares, LASSO, and ridge regression. The subproblem (B.3) then takes the form:
$$\sup_{\boldsymbol{\Theta} \in \mathfrak{T}} \frac{1}{N} \sum_{i=1}^{N} Q\Big(\boldsymbol{x}_s, \mathrm{proj}_\Xi \left(\boldsymbol{\Theta}(\boldsymbol{z} - \boldsymbol{z}^i) + \widehat{\boldsymbol{\xi}}^i\right)\Big). \tag{B.5}$$

- *k-Nearest Neighborhood Regression:* Let
$$\mathcal{F} = \left\{ f_{k\mathrm{NN}}(\boldsymbol{z}; k) = \frac{1}{k} \sum_{j \in \mathcal{N}_k(\boldsymbol{z})} \widehat{\boldsymbol{\xi}}^j \;\middle|\; k \in \mathcal{K} \right\}$$
be the class of k-nearest neighborhood (kNN) regression functions, where $\mathcal{K} \subseteq [N]$ and $\mathcal{N}_k(\boldsymbol{z})$ denotes the k-nearest neighborhood of $\boldsymbol{z}$ defined as follows. First, given the data $\mathcal{D}_N = \{(\boldsymbol{z}^i, \widehat{\boldsymbol{\xi}}^i)\}_{i=1}^N$, we define $\{(\boldsymbol{z}^{\sigma(i)}, \widehat{\boldsymbol{\xi}}^{\sigma(i)})\}_{i=1}^N$ as its ordered set such that $\|\boldsymbol{z} - \boldsymbol{z}^{\sigma(i)}\| \leq \|\boldsymbol{z} - \boldsymbol{z}^{\sigma(j)}\|$ for any $\sigma(i) < \sigma(j)$, where $\sigma(\cdot) = \sigma(\cdot; \boldsymbol{z})$ denotes a permutation function on $[N]$. If $\|\boldsymbol{z} - \boldsymbol{z}^i\| = \|\boldsymbol{z} - \boldsymbol{z}^j\|$ for some $\{i, j\} \subseteq [N]$ (i.e., a tie), we set $\sigma(i) < \sigma(j)$ if $i < j$, i.e., we break the tie by their indices in the data $\mathcal{D}_N$ (before ordering). Then, $\mathcal{N}_k(\boldsymbol{z}) := \{\sigma^{-1}(j)\}_{j \in [k]}$ is the set of indices corresponding to the first k smallest entries in this order. Thus, the kNN regression predicts the value of $\boldsymbol{\xi}$ of a given covariate $\boldsymbol{z}$ by averaging the k scenarios with covariates that are closest to $\boldsymbol{z}$. Accordingly, the subproblem reads as follows:
$$\sup_{k \in \mathcal{K}} \frac{1}{N} \sum_{i=1}^{N} Q\left(\boldsymbol{x}_s, \mathrm{proj}_\Xi \left(\frac{1}{k}\left[\sum_{j \in \mathcal{N}_k(\boldsymbol{z})} \widehat{\boldsymbol{\xi}}^j - \sum_{j \in \mathcal{N}_k(\boldsymbol{z}^i)} \widehat{\boldsymbol{\xi}}^j\right] + \widehat{\boldsymbol{\xi}}^i\right)\right). \tag{B.6}$$

- *Kernel Regression:* Let
$$\mathcal{F} = \left\{ f_{\mathrm{kern}}(\boldsymbol{z}; h_{\mathrm{bw}}) = \sum_{j=1}^{N} \frac{K(\|\boldsymbol{z} - \boldsymbol{z}^j\|; h_{\mathrm{bw}})}{\sum_{j'=1}^{N} K(\|\boldsymbol{z} - \boldsymbol{z}^{j'}\|; h_{\mathrm{bw}})} \widehat{\boldsymbol{\xi}}^j \;\middle|\; h_{\mathrm{bw}} \in \mathcal{H} \right\}$$
be the class of kernel regression functions, where K is a univariate kernel function with bandwidth $h_{\mathrm{bw}} \in \mathcal{H} \subseteq \mathbb{R}_{++}$. We assume that the set $\mathcal{H}$ is closed. We focus on the box kernel, one of the common kernel functions employed in the literature: $K(u; h_{\mathrm{bw}}) = \mathbf{1}(u \leq h_{\mathrm{bw}})$ for any $u \geq 0$.

Under these settings, the subproblem reads as follows:

$$\sup_{h_{\mathrm{bw}}\in\mathcal{H}} \frac{1}{N}\sum_{i=1}^{N} Q\left(\boldsymbol{x}_s, \mathrm{proj}_{\Xi}\left(\sum_{j=1}^{N}\left[\frac{\mathbf{1}(\|\boldsymbol{z}-\boldsymbol{z}^j\|\le h_{\mathrm{bw}})}{\sum_{j'=1}^{N}\mathbf{1}(\|\boldsymbol{z}-\boldsymbol{z}^{j'}\|\le h_{\mathrm{bw}})}-\frac{\mathbf{1}(\|\boldsymbol{z}^i-\boldsymbol{z}^j\|\le h_{\mathrm{bw}})}{\sum_{j'=1}^{N}\mathbf{1}(\|\boldsymbol{z}^i-\boldsymbol{z}^{j'}\|\le h_{\mathrm{bw}})}\right]\widehat{\boldsymbol{\xi}}^j+\widehat{\boldsymbol{\xi}}^i\right)\right). \tag{B.7}$$

In Theorems 11–13, we derive equivalent and tractable reformulations of the subproblems (B.5)–(B.7) under these choices of the function class $\mathcal{F}$.

Theorem 11. *There exists $\overline{M}>0$ such that solving subproblem (B.5) is equivalent to solving*

$$\begin{aligned}
\underset{\boldsymbol{\Theta},\boldsymbol{\gamma},\boldsymbol{y},\boldsymbol{\pi},\boldsymbol{a},\boldsymbol{b}}{\text{maximize}}\quad & \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{q}^\top\boldsymbol{y}_i && \text{(B.8a)}\\
\text{subject to}\quad & \boldsymbol{\Theta}\in\mathfrak{T}, && \text{(B.8b)}\\
& \boldsymbol{T}\boldsymbol{x}_s+\boldsymbol{W}\boldsymbol{y}_i+\boldsymbol{C}\boldsymbol{\gamma}_i\ge\boldsymbol{h},\quad \forall i\in[N], && \text{(B.8c)}\\
& \boldsymbol{W}^\top\boldsymbol{\pi}_i\le\boldsymbol{q},\quad \forall i\in[N], && \text{(B.8d)}\\
& \boldsymbol{q}-\boldsymbol{W}^\top\boldsymbol{\pi}_i\le\overline{M}\boldsymbol{a}_i,\quad \boldsymbol{y}_i\le\overline{M}(\mathbf{1}-\boldsymbol{a}_i),\quad \forall i\in[N], && \text{(B.8e)}\\
& \boldsymbol{T}\boldsymbol{x}_s+\boldsymbol{W}\boldsymbol{y}_i+\boldsymbol{C}\boldsymbol{\gamma}_i-\boldsymbol{h}\le\overline{M}\boldsymbol{b}_i,\quad \forall i\in[N], && \text{(B.8f)}\\
& \boldsymbol{\pi}_i\le\overline{M}(\mathbf{1}-\boldsymbol{b}_i),\quad \forall i\in[N], && \text{(B.8g)}\\
& \boldsymbol{\gamma}_i\in\Gamma_i(\boldsymbol{\Theta}),\quad \boldsymbol{y}_i\ge 0,\quad \boldsymbol{\pi}_i\ge 0,\quad \boldsymbol{a}_i\in\{0,1\}^{d_y},\quad \boldsymbol{b}_i\in\{0,1\}^{d_h},\quad \forall i\in[N], && \text{(B.8h)}
\end{aligned}$$

where $\Gamma_i(\boldsymbol{\Theta})=\left\{\boldsymbol{\gamma}\in\mathbb{R}^{d_\xi}\mid\boldsymbol{\gamma}=\mathrm{proj}_{\Xi}\left(\boldsymbol{\Theta}(\boldsymbol{z}-\boldsymbol{z}^i)+\widehat{\boldsymbol{\xi}}^i\right)\right\}$ *for* $i\in[N]$.

Proof. Since $Q(\boldsymbol{x},\boldsymbol{\xi})$ is finite for any $\boldsymbol{x}\in\mathcal{X}$ and $\boldsymbol{\xi}\in\Xi$, by LP duality, we have $Q(\boldsymbol{x},\boldsymbol{\xi})=\max_{\boldsymbol{\pi}\in\Pi}\{(\boldsymbol{h}-\boldsymbol{T}\boldsymbol{x}-\boldsymbol{C}\boldsymbol{\xi})^\top\boldsymbol{\pi}\}$, where $\Pi=\{\boldsymbol{\pi}\ge\mathbf{0}\mid\boldsymbol{W}^\top\boldsymbol{\pi}\le\boldsymbol{q}\}$. For any $i\in[N]$, $(\boldsymbol{y}_i,\boldsymbol{\pi}_i)$ are optimal primal and dual solutions to $Q\left(\boldsymbol{x}_s,\mathrm{proj}_{\Xi}\left(\boldsymbol{\Theta}(\boldsymbol{z}-\boldsymbol{z}^i)+\widehat{\boldsymbol{\xi}}^i\right)\right)$ if and only if $\boldsymbol{y}_i$ is primal feasible, $\boldsymbol{\pi}_i$ is dual feasible, and the following complementary slackness conditions hold:

$$\begin{aligned}
&(\boldsymbol{q}-\boldsymbol{W}^\top\boldsymbol{\pi}_i)\circ\boldsymbol{y}_i=0, && \text{(B.9a)}\\
&\left(\boldsymbol{T}\boldsymbol{x}_s+\boldsymbol{W}\boldsymbol{y}_i+\boldsymbol{C}\left[\mathrm{proj}_{\Xi}\left(\boldsymbol{\Theta}(\boldsymbol{z}-\boldsymbol{z}^i)+\widehat{\boldsymbol{\xi}}^i\right)\right]-\boldsymbol{h}\right)\circ\boldsymbol{\pi}_i=0, && \text{(B.9b)}
\end{aligned}$$

where $\circ$ denotes the element-wise product. Therefore, problem (B.5) is equivalent to

$$\begin{aligned}
\underset{\boldsymbol{\Theta},\boldsymbol{\gamma},\boldsymbol{y},\boldsymbol{\pi}}{\text{maximize}}\quad & \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{q}^\top\boldsymbol{y}_i && \text{(B.10a)}\\
\text{subject to}\quad & \text{(B.8b)–(B.8d)},\ \{\text{(B.9a)–(B.9b)},\ \forall i\in[N]\}, && \text{(B.10b)}\\
& \boldsymbol{\gamma}_i\in\Gamma_i(\boldsymbol{\Theta}),\quad \boldsymbol{y}_i\ge 0,\quad \boldsymbol{\pi}_i\ge 0,\quad \forall i\in[N]. && \text{(B.10c)}
\end{aligned}$$

Let $\Lambda = \{\boldsymbol{\xi} = \text{proj}_{\Xi}\left(\boldsymbol{\Theta}(\boldsymbol{z} - \boldsymbol{z}^i) + \widehat{\boldsymbol{\xi}}^i\right) \mid \boldsymbol{\Theta} \in \mathfrak{T}\}$. Since $\mathfrak{T}$ and Ξ are compact, Λ is also compact. Moreover, because $Q(\boldsymbol{x}, \boldsymbol{\xi})$ is continuous in $\boldsymbol{\xi}$ and Λ is compact, we have $-\infty < \min_{\boldsymbol{\xi}\in\Lambda} Q(\boldsymbol{x}_s, \boldsymbol{\xi}) \leq \max_{\boldsymbol{\xi}\in\Lambda} Q(\boldsymbol{x}_s, \boldsymbol{\xi}) < \infty$. It follows that there exists a compact set $\mathcal{Y}^\star(\Lambda)$ containing, for each $\boldsymbol{\xi} \in \Lambda$, at least one primal optimal solution of $Q(\boldsymbol{x}, \boldsymbol{\xi})$. On the dual side, since $Q(\boldsymbol{x}, \boldsymbol{\xi})$ is finite for every $\boldsymbol{\xi} \in \Lambda$, there exists for each $\boldsymbol{\xi} \in \Lambda$ a dual optimal solution that belongs to $\text{ext}(\Pi)$, where $\text{ext}(\Pi)$ denotes the set of extreme points of the dual feasible region Π. Taking together, the set $\mathcal{Y}^\star(\Lambda) \times \text{ext}(\Pi)$ contains, for every $\boldsymbol{\xi} \in \Lambda$, a primal-dual optimal pair associated with $Q(\boldsymbol{x}_s, \boldsymbol{\xi})$. Consequently, there exists a sufficiently large constant $\overline{M} > 0$ such that the complementary slackness constraints in (B.9) can be linearized by introducing new binary variables $\boldsymbol{a}_i \in \{0,1\}^{d_y}$ and $\boldsymbol{b}_i \in \{0,1\}^{d_h}$ and replacing (B.9a)–(B.9b) by (B.8e)–(B.8g). This shows that problem (B.5) is equivalent to (B.8). □

Theorem 12. *There exists $\overline{M} > 0$ such that the subproblem (B.6) is equivalent to*

$$\underset{k,\boldsymbol{\gamma},\boldsymbol{\phi},\boldsymbol{\rho},\boldsymbol{y},\boldsymbol{\pi},\boldsymbol{a},\boldsymbol{b}}{\text{maximize}} \quad \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{q}^\top \boldsymbol{y}_i \tag{B.11a}$$

$$\text{subject to} \quad k \in \mathcal{K}, \quad \big\{(\boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{\pi}, \boldsymbol{a}, \boldsymbol{b}) \mid \text{(B.8c)–(B.8g)}\big\}, \tag{B.11b}$$

$$\sum_{j=1}^{N} \phi_j = k, \quad \phi_{j-1} \geq \phi_j, \quad \forall j \in \{2, \ldots, N\}, \tag{B.11c}$$

$$\sum_{j=1}^{N} \rho_{i,j} = k, \quad \rho_{i,j-1} \geq \rho_{i,j}, \quad \forall i \in [N],\, j \in \{2, \ldots, N\}, \tag{B.11d}$$

$$\boldsymbol{\gamma}_i \in \Gamma_i(k), \quad \boldsymbol{y}_i \geq 0, \quad \boldsymbol{\pi}_i \geq 0, \quad \boldsymbol{a}_i \in \{0,1\}^{d_y}, \quad \boldsymbol{b}_i \in \{0,1\}^{d_h}, \quad \forall i \in [N], \tag{B.11e}$$

$$\phi_j \in \{0,1\}, \quad \rho_{i,j} \in \{0,1\}, \quad \forall i \in [N],\, j \in [N], \tag{B.11f}$$

where $\Gamma_i(k) = \big\{\boldsymbol{\gamma} \in \mathbb{R}^{d_\xi} \;\big|\; \boldsymbol{\gamma} = \text{proj}_{\Xi}\big(k^{-1}\big[\sum_{j=1}^{N} \phi_j \widehat{\boldsymbol{\xi}}^{\sigma^{-1}(j;\boldsymbol{z})} - \sum_{j=1}^{N} \rho_{i,j} \widehat{\boldsymbol{\xi}}^{\sigma^{-1}(j;\boldsymbol{z}^i)}\big] + \widehat{\boldsymbol{\xi}}^i\big)\big\}$ *for* $i \in [N]$.

Proof. Note that from the proof of Theorem 11, the subproblem (B.6) is equivalent to

$$\underset{k,\boldsymbol{\gamma},\boldsymbol{\phi},\boldsymbol{\rho},\boldsymbol{y},\boldsymbol{\pi},\boldsymbol{a},\boldsymbol{b}}{\text{maximize}} \quad \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{q}^\top \boldsymbol{y}_i \tag{B.12a}$$

$$\text{subject to} \quad k \in \mathcal{K}, \quad \big\{(\boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{\pi}, \boldsymbol{a}, \boldsymbol{b}) \mid \text{(B.8c)–(B.8g)}\big\}, \tag{B.12b}$$

$$\boldsymbol{\gamma}_i = \text{proj}_{\Xi}\left(\frac{1}{k}\left[\sum_{j\in\mathcal{N}_k(\boldsymbol{z})} \widehat{\boldsymbol{\xi}}^j - \sum_{j\in\mathcal{N}_k(\boldsymbol{z}^i)} \widehat{\boldsymbol{\xi}}^j\right] + \widehat{\boldsymbol{\xi}}^i\right), \quad \forall i \in [N], \tag{B.12c}$$

$$\boldsymbol{y}_i \geq 0, \quad \boldsymbol{\pi}_i \geq 0, \quad \boldsymbol{a}_i \in \{0,1\}^{d_y}, \quad \boldsymbol{b}_i \in \{0,1\}^{d_h}, \quad \forall i \in [N]. \tag{B.12d}$$

We now derive an equivalent reformulation of (B.12c). First, consider the nonlinear term $\sum_{j\in\mathcal{N}_k(\boldsymbol{z})} \widehat{\boldsymbol{\xi}}^j$.

We claim that $\sum_{j=1}^N \phi_j \widehat{\boldsymbol{\xi}}^{\sigma^{-1}(j;\boldsymbol{z})}$ together with binary variables $\boldsymbol{\phi}$ and constraints (B.11c) define $\sum_{j\in\mathcal{N}_k(\boldsymbol{z})} \widehat{\boldsymbol{\xi}}^j$. To prove this claim, note that constraints (B.11c) ensure that there are exactly k of the binary variables $\{\phi_j\}_{j=1}^N$ are one and that $\{\phi_j\}_{j=1}^N$ are sorted in descending order based on their indices. Thus, constraints (B.11c) enforce that $\phi_j = 1$ for $j \in [k]$ and $\phi_j = 0$ otherwise. Therefore, we have

$$\sum_{j=1}^N \phi_j \widehat{\boldsymbol{\xi}}^{\sigma^{-1}(j;\boldsymbol{z})} = \sum_{j=1}^k \widehat{\boldsymbol{\xi}}^{\sigma^{-1}(j;\boldsymbol{z})} = \sum_{j\in\mathcal{N}_k(\boldsymbol{z})} \widehat{\boldsymbol{\xi}}^j,$$

where the last inequality follows from the definition of $\mathcal{N}_k(\boldsymbol{z})$. Using the same argument, for all $i \in [N]$, we can show that $\sum_{j=1}^N \rho_{i,j} \widehat{\boldsymbol{\xi}}^{\sigma^{-1}(j;\boldsymbol{z}^i)}$ together with binary variables ρ and constraints (B.11d) define $\sum_{j\in\mathcal{N}_k(\boldsymbol{z}^i)} \widehat{\boldsymbol{\xi}}^j$. Hence, constraints (B.12c) becomes

$$\boldsymbol{\gamma}_i = \mathrm{proj}_{\Xi}\left(\frac{1}{k}\left[\sum_{j=1}^N \phi_j \widehat{\boldsymbol{\xi}}^{\sigma^{-1}(j;\boldsymbol{z})} - \sum_{j=1}^N \rho_{i,j} \widehat{\boldsymbol{\xi}}^{\sigma^{-1}(j;\boldsymbol{z}^i)}\right] + \widehat{\boldsymbol{\xi}}^i\right), \quad \forall i \in [N].$$

This completes the proof. □

Theorem 13. *Let $\mathcal{H} = \bigcup_{\ell=1}^L [h_{\min}^\ell, h_{\max}^\ell]$, where $h_{\max}^\ell$ could be ∞. Define $\mathcal{S} := \big(\bigcup_{i\in[N]}\{\|\boldsymbol{z}-\boldsymbol{z}^i\|\}\big) \cup \big(\bigcup_{i\in[N]}\bigcup_{j\in[N]}\{\|\boldsymbol{z}^i-\boldsymbol{z}^j\|\}\big) \cup \big(\bigcup_{\ell=1}^L\{h_{\min}^\ell, h_{\max}^\ell\}\big)$. Denote by $\mathcal{S}^{\mathrm{d}} := \{s^k\}_{k=1}^{|\mathcal{S}^{\mathrm{d}}|}$ the set of all distinct elements in $\mathcal{S}$ with $s^1 < s^2 < \cdots < s^{|\mathcal{S}^{\mathrm{d}}|}$, and define $\delta = \min_{k=2,\ldots,|\mathcal{S}^{\mathrm{d}}|}\big\{(s^k - s^{k-1})/2\big\}$. There exists $\overline{M} > 0$ such that the subproblem (B.7) is equivalent to*

$$\begin{aligned}
\underset{h_{\mathrm{bw}},\boldsymbol{\gamma},\boldsymbol{\phi},\boldsymbol{\rho},\boldsymbol{y},\boldsymbol{\pi},\boldsymbol{a},\boldsymbol{b}}{\text{maximize}} \quad & \frac{1}{N}\sum_{i=1}^N \boldsymbol{q}^\top \boldsymbol{y}_i && \text{(B.13a)}\\
\text{subject to} \quad & h_{\mathrm{bw}} \in \mathcal{H}, \quad \big\{(\boldsymbol{\gamma},\boldsymbol{y},\boldsymbol{\pi},\boldsymbol{a},\boldsymbol{b}) \mid \text{(B.8c)–(B.8g)}\big\}, && \text{(B.13b)}\\
& h_{\mathrm{bw}} + \delta - \overline{M}\phi_j \le \|\boldsymbol{z} - \boldsymbol{z}^j\| \le h_{\mathrm{bw}} + \overline{M}(1-\phi_j), \quad \forall j \in [N], && \text{(B.13c)}\\
& h_{\mathrm{bw}} + \delta - \overline{M}\rho_{i,j} \le \|\boldsymbol{z}^i - \boldsymbol{z}^j\| \le h_{\mathrm{bw}} + \overline{M}(1-\rho_{i,j}), \quad \forall i \in [N],\ j \in [N], && \text{(B.13d)}\\
& \boldsymbol{\gamma}_i \in \Gamma_i(\boldsymbol{\phi},\boldsymbol{\rho}), \quad \boldsymbol{y}_i \ge 0, \quad \boldsymbol{\pi}_i \ge 0, \quad \boldsymbol{a}_i \in \{0,1\}^{d_y}, \quad \boldsymbol{b}_i \in \{0,1\}^{d_h}, \quad \forall i \in [N], && \text{(B.13e)}\\
& \phi_j \in \{0,1\}, \quad \rho_{i,j} \in \{0,1\}, \quad \forall i \in [N],\ j \in [N], && \text{(B.13f)}
\end{aligned}$$

where $\Gamma_i(\boldsymbol{\phi},\boldsymbol{\rho}) = \big\{\boldsymbol{\gamma} \in \mathbb{R}^{d_\xi} \;\big|\; \boldsymbol{\gamma} = \mathrm{proj}_{\Xi}\big(\sum_{j=1}^N \phi_j \widehat{\boldsymbol{\xi}}_j / \sum_{j'=1}^N \phi_{j'} - \sum_{j=1}^N \rho_{i,j}\widehat{\boldsymbol{\xi}}_j / \sum_{j''=1}^N \rho_{i,j''} + \widehat{\boldsymbol{\xi}}^i\big)\big\}$ for $i \in [N]$.

Proof. First, following the proof of Theorem 11, the subproblem (B.7) is equivalent to

$$\underset{h_{\mathrm{bw}},\boldsymbol{\gamma},\boldsymbol{\phi},\boldsymbol{\rho},\boldsymbol{y},\boldsymbol{\pi},\boldsymbol{a},\boldsymbol{b}}{\text{maximize}} \quad \frac{1}{N}\sum_{i=1}^N \boldsymbol{q}^\top \boldsymbol{y}_i \qquad \text{(B.14a)}$$

$$\text{subject to} \quad h_{\text{bw}} \in \mathcal{H}, \quad \big\{(\boldsymbol{\gamma}, \boldsymbol{y}, \boldsymbol{\pi}, \boldsymbol{a}, \boldsymbol{b}) \mid \text{(B.8c)–(B.8g)}\big\}, \tag{B.14b}$$

$$\boldsymbol{\gamma}_i = \text{proj}_{\Xi}\left(\sum_{j=1}^{N}\left[\frac{\mathbf{1}(\|\boldsymbol{z}-\boldsymbol{z}^j\| \le h_{\text{bw}})}{\sum_{j'=1}^{N}\mathbf{1}(\|\boldsymbol{z}-\boldsymbol{z}^{j'}\| \le h_{\text{bw}})} - \frac{\mathbf{1}(\|\boldsymbol{z}^i-\boldsymbol{z}^j\| \le h_{\text{bw}})}{\sum_{j'=1}^{N}\mathbf{1}(\|\boldsymbol{z}^i-\boldsymbol{z}^{j'}\| \le h_{\text{bw}})}\right]\widehat{\boldsymbol{\xi}}^j + \widehat{\boldsymbol{\xi}}^i\right), \quad \forall i \in [N], \tag{B.14c}$$

$$\boldsymbol{y}_i \ge 0, \quad \boldsymbol{\pi}_i \ge 0, \quad \boldsymbol{a}_i \in \{0,1\}^{d_y}, \quad \boldsymbol{b}_i \in \{0,1\}^{d_h}, \quad \forall i \in [N]. \tag{B.14d}$$

Next, we derive an equivalent reformulation of (B.14c). Let us first consider constraints (B.13c) and the following three cases. (Note that we assume $\overline{M} > 0$ is sufficiently large; in particular, $\overline{M} > s^{|\mathcal{S}^{\text{d}}|} - s^1 + \delta$).

- Suppose that $\|\boldsymbol{z}-\boldsymbol{z}^j\| \le h_{\text{bw}}$. In this case, feasibility of (B.13c) requires that $\phi_j = 1$, and thus $\phi_j = 1 = \mathbf{1}(\|\boldsymbol{z}-\boldsymbol{z}^j\| \le h_{\text{bw}})$.
- Suppose that $\|\boldsymbol{z}-\boldsymbol{z}^j\| \ge h_{\text{bw}} + \delta$. In this case, feasibility of (B.13c) requires $\phi_j = 0$, and thus $\phi_j = 0 = \mathbf{1}(\|\boldsymbol{z}-\boldsymbol{z}^j\| \le h_{\text{bw}})$.
- Suppose that $h_{\text{bw}} < \|\boldsymbol{z}-\boldsymbol{z}^j\| < h_{\text{bw}} + \delta$. It is straightforward to see that (B.13c) is infeasible for any $\phi_j \in \{0,1\}$. We now claim that the values of $\{\boldsymbol{\gamma}_i\}_{i=1}^N$ in (B.14c) at h_{bw} is the same as some $s \in \mathcal{H} \cap \mathcal{S}^{\text{d}}$. In other words, without loss of optimality, we do not need to consider any $h_{\text{bw}} \in \mathcal{H}$ that belongs to the third case. To prove this claim, note that by the definition of $\delta = \min_{k=2,\ldots,|\mathcal{S}^{\text{d}}|}\{(s^k - s^{k-1})/2\}$, the interval $[h_{\text{bw}}, h_{\text{bw}}+\delta]$ does not intersect with any $s \in \mathcal{S}^{\text{d}}$ except for the one that equals $\|\boldsymbol{z}-\boldsymbol{z}^j\|$, say $s^r = \|\boldsymbol{z}-\boldsymbol{z}^j\|$ for some $r > 1$. Since $h_{\text{bw}} \in \mathcal{H}$ and $\mathcal{H}$ is the union of closed intervals, the indicator functions in constraints (B.14c) take the same value at both h_{bw} and $s^{r-1} \in \mathcal{H} \cap \mathcal{S}^{\text{d}}$. Thus, the values of $\{\boldsymbol{\gamma}_i\}_{i=1}^N$ in (B.14c) at h_{bw} is the same as some $s \in \mathcal{H} \cap \mathcal{S}^{\text{d}}$.

Combining all three cases, constraints (B.13c) either give a characterization of the indicator function, i.e., $\phi_j = \mathbf{1}(\|\boldsymbol{z}-\boldsymbol{z}^j\| \le h_{\text{bw}})$, or exclude values of $h \in \mathcal{H}$ that lead to equivalent solutions. Thus, without loss of optimality, we can replace $\mathbf{1}(\|\boldsymbol{z}-\boldsymbol{z}^j\| \le h_{\text{bw}})$ with the binary variable ϕ_j. Applying the same argument, constraints (B.13d) allow us to replace $\mathbf{1}(\|\boldsymbol{z}^i-\boldsymbol{z}^j\| \le h_{\text{bw}})$ with the binary variable $\rho_{i,j}$. Hence, with the new binary variables, constraints (B.14c) become

$$\boldsymbol{\gamma}_i = \text{proj}_{\Xi}\left(\sum_{j=1}^{N}\left(\frac{\phi_j}{\sum_{j'=1}^{N}\phi_{j'}} - \frac{\rho_{i,j}}{\sum_{j''=1}^{N}\rho_{i,j''}}\right)\widehat{\boldsymbol{\xi}}^j + \widehat{\boldsymbol{\xi}}^i\right), \quad \forall i \in [N].$$

This completes the proof. □

Remark 3. Note that it suffices to optimize h_{bw} over the finite set $\mathcal{H} \cap \mathcal{S}^{\text{d}}$. Indeed, for a fixed value of h_{bw}, the optimal value of the subproblem (B.7), i.e., $v(h_{\text{bw}}) = \frac{1}{N}\sum_{i=1}^{N} Q\big(\boldsymbol{x}_s, \text{proj}_{\Xi}\big(f_{\text{kern}}(\boldsymbol{z}; h_{\text{bw}}) + [\widehat{\boldsymbol{\xi}}^i - f_{\text{kern}}(\boldsymbol{z}^i; h_{\text{bw}})]\big)\big)$, is the same on each interval $[s^{k-1}, s^k)$ for $k \in \{2, \ldots, |\mathcal{S}^{\text{d}}|\}$. Hence, if the sample size N is small, and thus the number of elements in $\mathcal{S}^{\text{d}}$ is small, one could alternatively solve

the subproblem (B.7) by enumeration: for each element $h_{\rm bw} \in \mathcal{H} \cap \mathcal{S}^{\rm d}$, we first compute $\upsilon(h_{\rm bw})$. Then, the optimal bandwidth $h^\star_{\rm bw}$ is the one that gives the worst (i.e., largest) value of $\upsilon(h_{\rm bw})$.

Finally, we note that reformulations (B.8), (B.11), and (B.13) remain nonlinear, primarily due to the projection operator embedded within the sets Γ_i. To address this, Proposition 2 derives equivalent mixed-integer reformulations for these sets, formulated under the generic form: $\Gamma_i(\boldsymbol{\theta}) = \left\{\boldsymbol{\gamma} \in \mathbb{R}^{d_\xi} \,\middle|\, \boldsymbol{\gamma} = \text{proj}_\Xi\left([f(\boldsymbol{z};\boldsymbol{\theta}) - f(\boldsymbol{z}^i;\boldsymbol{\theta})] + \widehat{\boldsymbol{\xi}}^i\right)\right\}$ for $i \in [N]$ under three popular choices of the support set Ξ.

Proposition 2. *Let* $\Gamma_i(\boldsymbol{\theta}) = \left\{\boldsymbol{\gamma} \in \mathbb{R}^{d_\xi} \,\middle|\, \boldsymbol{\gamma} = \text{proj}_\Xi\left([f(\boldsymbol{z};\boldsymbol{\theta}) - f(\boldsymbol{z}^i;\boldsymbol{\theta})] + \widehat{\boldsymbol{\xi}}^i\right)\right\}$ *for* $i \in [N]$.

(i) When $\Xi = \mathbb{R}^{d_\xi}$, *it is equivalent to* $\Gamma_i(\boldsymbol{\theta}) = \left\{\boldsymbol{\gamma} \in \mathbb{R}^{d_\xi} \,|\, \boldsymbol{\gamma} = \left[f(\boldsymbol{z};\boldsymbol{\theta}) - f(\boldsymbol{z}^i;\boldsymbol{\theta})\right] + \widehat{\boldsymbol{\xi}}^i\right\}$,

(ii) When $\Xi = \mathbb{R}^{d_\xi}_+$, *it is equivalent to*

$$\Gamma_i(\boldsymbol{\theta}) = \left\{ \boldsymbol{\gamma} \in \mathbb{R}^{d_\xi} \,\middle|\, \exists \boldsymbol{\tau} \in \{0,1\}^{d_\xi} \text{ s.t. } \begin{array}{l} \boldsymbol{\gamma} \le \left[f(\boldsymbol{z};\boldsymbol{\theta}) - f(\boldsymbol{z}^i;\boldsymbol{\theta})\right] + \overline{M}\boldsymbol{\tau} \\ \boldsymbol{\gamma} \le \overline{M}(\mathbf{1} - \boldsymbol{\tau}) \\ \boldsymbol{\gamma} \ge \left[f(\boldsymbol{z};\boldsymbol{\theta}) - f(\boldsymbol{z}^i;\boldsymbol{\theta})\right] + \widehat{\boldsymbol{\xi}}^i \\ \boldsymbol{\gamma} \ge 0 \end{array} \right\}, \text{ and} \tag{B.15}$$

(iii) When $\Xi = \{\boldsymbol{\xi} \in \mathbb{R}^{d_\xi} \,|\, \underline{\boldsymbol{\xi}} \le \boldsymbol{\xi} \le \bar{\boldsymbol{\xi}}\}$ *with* $-\infty < \underline{\boldsymbol{\xi}} < \bar{\boldsymbol{\xi}} < \infty$, *it is equivalent to*

$$\Gamma_i(\boldsymbol{\theta}) = \left\{ \boldsymbol{\gamma} \in \mathbb{R}^{d_\xi} \,\middle|\, \exists \{\boldsymbol{\tau}, \boldsymbol{\psi}\} \subseteq \{0,1\}^{d_\xi} \text{ s.t. } \begin{array}{l} \boldsymbol{\gamma} \le \left[f(\boldsymbol{z};\boldsymbol{\theta}) - f(\boldsymbol{z}^i;\boldsymbol{\theta})\right] + \overline{M}\boldsymbol{\tau} \\ \boldsymbol{\gamma} \ge \left[f(\boldsymbol{z};\boldsymbol{\theta}) - f(\boldsymbol{z}^i;\boldsymbol{\theta})\right] - \overline{M}\boldsymbol{\psi} \\ \boldsymbol{\gamma} \le \underline{\boldsymbol{\xi}} + \overline{M}(\mathbf{1} - \boldsymbol{\tau}) \\ \boldsymbol{\gamma} \ge \bar{\boldsymbol{\xi}} - \overline{M}(\mathbf{1} - \boldsymbol{\psi}) \\ \underline{\boldsymbol{\xi}} \le \boldsymbol{\gamma} \le \bar{\boldsymbol{\xi}} \end{array} \right\}. \tag{B.16}$$

Proof. Next, we derive a characterization of the set $\Gamma_i(\boldsymbol{\theta})$ under the stated choices for the support set Ξ. For part (i), since $\Xi = \mathbb{R}^{d_\xi}$, we have $\Gamma_i(\boldsymbol{\theta}) = \left\{\boldsymbol{\gamma} \in \mathbb{R}^{d_\xi} \,|\, \boldsymbol{\gamma} = \left[f(\boldsymbol{z};\boldsymbol{\theta}) - f(\boldsymbol{z}^i;\boldsymbol{\theta})\right] + \widehat{\boldsymbol{\xi}}^i\right\}$. Now, we prove part (ii). For simplicity, let $\boldsymbol{u}_i = \left[f(\boldsymbol{z};\boldsymbol{\theta}) - f(\boldsymbol{z}^i;\boldsymbol{\theta})\right] + \widehat{\boldsymbol{\xi}}^i$ for $i \in [N]$. It suffices to show that $\boldsymbol{\gamma}_i$ and the new binary variables $\boldsymbol{\tau}_i$ define $\text{proj}_\Xi(\boldsymbol{u}_i)$ through the set of constraints in (B.15). Since $\Xi = \mathbb{R}^{d_\xi}_+$, we have $[\text{proj}_\Xi(\boldsymbol{u}_i)]_l = \max\{u_{il}, 0\}$ for all $l \in [d_\xi]$. The lth constraint in (B.15) reads as follows:

$$\gamma_{il} \le u_{il} + \overline{M}\tau_{il}, \quad \gamma_{il} \le \overline{M}(1 - \tau_{il}), \quad \gamma_{il} \ge u_{il}, \quad \gamma_{il} \ge 0, \quad \tau_{il} \in \{0,1\}. \tag{B.17}$$

Consider the following three cases:

- Suppose $u_{il} > 0$. If $\tau_{il} = 1$, then constraints (B.17) imply $0 < u_{il} \le \gamma_{il} = 0$, a contradiction. Hence, for the set of constraints (B.17) to be feasible, we must have $\tau_{il} = 0$. With $\tau_{il} = 0$, constraints (B.17) reduces to $\gamma_{il} = u_{il} \ge 0$.

- Suppose $u_{il} < 0$. If $\tau_{il} = 0$, then constraints (B.17) imply $u_{il} = \gamma_{il} \geq 0$, a contradiction with $u_{il} < 0$. Hence, for the set of constraints (B.17) to be feasible, we must have $\tau_{il} = 1$. With $\tau_{il} = 1$, constraints (B.17) reduces to $u_{il} \leq \gamma_{il} = 0$.
- Suppose $u_{il} = 0$. It is easy to verify that for any $\tau_{il} \in \{0, 1\}$, constraints (B.17) are feasible and implies $\gamma_{il} = 0$.

Combining the three cases, we have $\gamma_{il} = [\mathrm{proj}_{\Xi}(\boldsymbol{u}_i)]_l$. This shows that $\boldsymbol{\gamma}_i$ and the new binary variables $\boldsymbol{\tau}_i$ define $\gamma_{il} = [\mathrm{proj}_{\Xi}(\boldsymbol{u}_i)]_l$ through the set of constraints in (B.15). This completes the proof of part (ii).

Finally, we prove part (iii). Let $\boldsymbol{u}_i = \big[f(\boldsymbol{z};\boldsymbol{\theta}) - f(\boldsymbol{z}^i;\boldsymbol{\theta})\big] + \widehat{\boldsymbol{\xi}}^i$ for $i \in [N]$. It suffices to show that $\boldsymbol{\gamma}_i$ and the new binary variables $(\boldsymbol{\tau}_i, \boldsymbol{\psi}_i)$ define $\mathrm{proj}_{\Xi}(\boldsymbol{u}_i)$ through the set of constraints in (B.16). Since $\Xi = \{\boldsymbol{\xi} \in \mathbb{R}^{d_\xi} \mid \underline{\boldsymbol{\xi}} \leq \boldsymbol{\xi} \leq \overline{\boldsymbol{\xi}}\}$, we have

$$[\mathrm{proj}_{\Xi}(\boldsymbol{u}_i)]_l = \begin{cases} \underline{\xi}_l & \text{if } u_{ik} < \underline{\xi}_l, \\ u_{il} & \text{if } u_{ik} \in [\underline{\xi}_l, \overline{\xi}_l], \\ \overline{\xi}_l & \text{if } u_{ik} > \overline{\xi}_l, \end{cases}$$

for all $l \in [d_\xi]$. The lth constraint in (B.16) reads as follows:

$$\gamma_{il} \leq u_{il} + \overline{M}\tau_{il}, \quad \gamma_{il} \leq \underline{\xi}_l + \overline{M}(1 - \tau_{il}), \quad \gamma_{il} \geq \underline{\xi}_l, \quad \tau_{il} \in \{0, 1\}, \tag{B.18a}$$

$$\gamma_{il} \geq u_{il} - \overline{M}\psi_{il}, \quad \gamma_{il} \geq \overline{\xi}_l - \overline{M}(1 - \psi_{il}), \quad \gamma_{il} \leq \overline{\xi}_l, \quad \psi_{il} \in \{0, 1\}. \tag{B.18b}$$

Consider the following five cases:

- Suppose $u_{il} < \underline{\xi}_l$. If $\tau_{il} = 0$, the constraints (B.18) implies $u_{il} \geq \gamma_{il} \geq \underline{\xi}_l$, which contradicts with $u_{il} < \underline{\xi}_l$. Hence, for the set of constraints (B.18) to be feasible, we must have $\tau_{il} = 1$. Now, if $\tau_{il} = 1$ and $\psi_{il} = 1$, constraints (B.18) implies $\gamma_{il} = \underline{\xi}_l = \overline{\xi}_l$, which contradicts with $\underline{\xi}_l < \overline{\xi}_l$. It follows that for the set of constraints (B.18) to be feasible, we must have $\tau_{il} = 1$ and $\psi_{il} = 0$. With $\tau_{il} = 1$ and $\psi_{il} = 0$, constraints (B.18) reduces to $u_{il} \leq \gamma_{il} = \underline{\xi}_l$.
- Suppose $u_{il} > \overline{\xi}_l$. If $\psi_{il} = 0$, then constraints (B.18) implies $u_{il} \leq \gamma_{il} \leq \overline{\xi}_l$, which contradicts with $u_{il} > \overline{\xi}_l$. Thus, for the set of constraints (B.18) to be feasible, we must have $\psi_{il} = 1$. Again, if $\psi_{il} = 1$ and $\tau_{il} = 1$, then constraints (B.18) implies $\gamma_{il} = \underline{\xi}_l = \overline{\xi}_l$, which contradicts with $\underline{\xi}_l < \overline{\xi}_l$. It follows that for the set of constraints (B.18) to be feasible, we must have $\tau_{il} = 0$ and $\psi_{il} = 1$. With $\tau_{il} = 0$ and $\psi_{il} = 1$, the set of constraints (B.18) reduces to $u_{il} \geq \gamma_{il} = \overline{\xi}_l$.
- Suppose $u_{il} \in (\underline{\xi}_l, \overline{\xi}_l)$. If $\tau_{il} = 1$ and $\psi_{il} = 0$, then the set of constraints (B.18) implies $u_{il} \leq \gamma_{il} = \underline{\xi}_l$, which contradicts with $u_{il} \in (\underline{\xi}_l, \overline{\xi}_l)$. Again, if $\tau_{il} = 1$ and $\psi_{il} = 1$, then the set of constraints (B.18) implies $\gamma_{il} = \underline{\xi}_l = \overline{\xi}_l$, which contradicts with $\underline{\xi}_l < \overline{\xi}_l$. Hence, for the set of constraints (B.18) to be feasible, we must have $\tau_{il} = 0$. Now, if $\tau_{il} = 0$ and $\psi_{il} = 1$, then constraints (B.18)

implies $u_{il} \geq \gamma_{il} = \overline{\xi}_l$, which contradicts with $u_{il} < \overline{\xi}_l$. It follows that for the set of constraints (B.18) to be feasible, we must have $\tau_{il} = 0$ and $\psi_{il} = 0$. With $\tau_{il} = 0$ and $\psi_{il} = 0$, constraints (B.18) reduces to $u_{il} = \gamma_{il} \in [\underline{\xi}_l, \overline{\xi}_l]$.

- Suppose $u_{il} = \underline{\xi}_l$. If $\psi_{il} = 1$, then constraints (B.18) implies $\underline{\xi}_l \geq \gamma_{il} = \overline{\xi}_l$ when $\tau_{il} = 1$ and $u_{il} \geq \gamma_{il} = \overline{\xi}_l$ when $\tau_{il} = 0$. In either case, we arrive at a contradiction with $\underline{\xi}_l = u_{il} < \overline{\xi}_l$. Hence, for the set of constraints (B.18) to be feasible, we must have $\psi_{il} = 0$. It is easy to verify that for any $\tau_{il} \in \{0, 1\}$, the set of constraints (B.18) reduces to $\gamma_{il} = \underline{\xi}_{il}$.
- Suppose $u_{il} = \overline{\xi}_l$. If $\tau_{il} = 1$, then constraints (B.18) implies $\underline{\xi}_l \geq \gamma_{il} = \overline{\xi}_l$ when $\psi_{il} = 1$ and $u_{il} \leq \gamma_{il} = \underline{\xi}_l$ when $\psi_{il} = 0$. In either case, we arrive at a contradiction with $\underline{\xi}_l < u_{il} = \overline{\xi}_l$. Hence, for the set of constraints (B.18) to be feasible, we must have $\tau_{il} = 0$. It is easy to verify that for any $\psi_{il} \in \{0, 1\}$, the set of constraints (B.18) reduces to $\gamma_{il} = \overline{\xi}_{il}$.

Combining the five cases, we have $\gamma_{il} = [\text{proj}_{\Xi}(\boldsymbol{u}_i)]_l$. This shows that $\boldsymbol{\gamma}_i$ and binary variables $(\boldsymbol{\tau}_i, \boldsymbol{\psi}_i)$ define $\gamma_{il} = [\text{proj}_{\Xi}(\boldsymbol{u}_i)]_l$ through the set of constraints in (B.16). This completes the proof of part (iii) and Theorem 11. □

Proposition 2 shows that the sets Γ_i in reformulations (B.8), (B.11), and (B.13) are mixed-integer representable across three commonly used support sets. When f is linear with respect to $\boldsymbol{\theta}$, as is the case with linear regression models, the constraints characterizing $\boldsymbol{\gamma}$ remain linear in the subproblem decision variables $(\boldsymbol{\gamma}, \boldsymbol{\theta})$. Conversely, if f is nonlinear, these constraints may also become nonlinear, necessitating additional reformulations (e.g., via McCormick inequalities). In the subsequent remarks, we detail the specific settings where f corresponds to kNN and kernel regression models, where we focus on the support set $\Xi = \mathbb{R}^{d_\xi}$ for simplicity.

Remark 4. Consider subproblem (B.6) associated with the class of kNN regression models. Under $\Xi = \mathbb{R}^{d_\xi}$, we can characterize the set $\Gamma_i(k)$ via the following constraint for all $i \in [N]$:

$$k\boldsymbol{\gamma}_i = \sum_{j=1}^{N} \phi_j \widehat{\boldsymbol{\xi}}^{\sigma^{-1}(j;\boldsymbol{z})} - \sum_{j=1}^{N} \rho_{ij} \widehat{\boldsymbol{\xi}}^{\sigma^{-1}(j;\boldsymbol{z}^i)} + k\widehat{\boldsymbol{\xi}}^i.$$

Note that since $\mathcal{K}$ is finite, say $\mathcal{K} = \{k_r\}_{r=1}^{|\mathcal{K}|}$, $\boldsymbol{\gamma}_i$ could only take a finite number of values with varying k, and thus is bounded. Therefore, we can linear $k\boldsymbol{\gamma}_i$ as follows. We first introduce $|\mathcal{K}|$ binary variables $\{\alpha_r\}_{r \in |\mathcal{K}|}$ and write $k = \sum_{r \in |\mathcal{K}|} \alpha_r k_r$ with constraint $\sum_{r \in |\mathcal{K}|} \alpha_r = 1$. Then, we can apply McCormick inequalities to linearize $\alpha_r \boldsymbol{\gamma}_i$, each entry being a product of a binary variable and a bounded continuous variable.

Remark 5. Consider subproblem (B.7) associated with the class of kernel regression models. Under

$\Xi = \mathbb{R}^{d_\xi}$, we can characterize the set $\Gamma_i(\boldsymbol{\phi}, \boldsymbol{\rho})$ via the following constraint for all $i \in [N]$:

$$\sum_{j'=1}^{N}\sum_{j''=1}^{N} \phi_{j'}\rho_{i,j''}\boldsymbol{\gamma}_i = \sum_{j=1}^{N}\left(\phi_j \sum_{j''=1}^{N}\rho_{i,j''} - \rho_{i,j}\sum_{j'=1}^{N}\phi_{j'}\right)\widehat{\boldsymbol{\xi}}^j + \sum_{j'=1}^{N}\sum_{j''=1}^{N}\phi_{j'}\rho_{i,j''}\widehat{\boldsymbol{\xi}}^i.$$

Note that $\phi_j \in \{0,1\}$ and $\rho_{i,j} \in \{0,1\}$ are bounded. Also, $\boldsymbol{\gamma}_i$ could only take a finite number of values with varying $h_{\rm bw}$, and thus is bounded. Therefore, we can linearize the nonlinear terms, e.g., $\phi_j\rho_{i,j''}$ and $\phi_{j'}\rho_{i,j''}\boldsymbol{\gamma}_i$, using McCormick inequalities.

Appendix C. Reformulations of the MUR-SAA Models for the Portfolio Optimization Problem

Appendix C.1. Reformulation of the MUR-SAA model with 1-Wasserstein MUR Ambiguity Set

Applying the duality theory for Wasserstein DRO models as in (22), we can reformulate our MUR-SAA model with the 1-Wasserstein MUR ambiguity set as follows:

$$\begin{aligned}\underset{\boldsymbol{x},\,\tau,\,\lambda\geq 0}{\text{minimize}} \quad & \rho\tau + r\lambda + \sum_{m=1}^{M}\widehat{w}_m \sup_{\boldsymbol{\Theta}\in\mathfrak{T}}\left\{\frac{1}{N}\sum_{i=1}^{N}\left\{-\boldsymbol{x}^\top\left[\boldsymbol{\Theta}(\boldsymbol{z}-\boldsymbol{z}^i)+\widehat{\boldsymbol{\xi}}^i\right]\right.\right. \\ & \left.\left. + \frac{\rho}{1-\beta}\Big(-\boldsymbol{x}^\top\left[\boldsymbol{\Theta}(\boldsymbol{z}-\boldsymbol{z}^i)+\widehat{\boldsymbol{\xi}}^i\right]-\tau\Big)^+\right\} - \lambda\|\boldsymbol{\Theta}-\widehat{\boldsymbol{\Theta}}_m\|\right\} \qquad \text{(C.1a)}\\ \text{subject to} \quad & \mathbf{1}^\top\boldsymbol{x} = 1, \quad \boldsymbol{x}\geq 0, \qquad \text{(C.1b)}\end{aligned}$$

where $\mathfrak{T} = \mathbb{R}^{d_\xi\times d_z}$. Consider the inner supremum problem in the objective (C.1a). Define the inner product of two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ as $\langle \boldsymbol{A}, \boldsymbol{B}\rangle = \mathrm{tr}(\boldsymbol{A}^\top\boldsymbol{B})$, where tr denotes the trace of a matrix. Let $\boldsymbol{A}_1^i = -(\boldsymbol{z}-\boldsymbol{z}^i)\boldsymbol{x}^\top$, $\boldsymbol{A}_2^i = -\big[1+\rho/(1-\beta)\big](\boldsymbol{z}-\boldsymbol{z}^i)\boldsymbol{x}^\top$, $b_1^i = -\boldsymbol{x}^\top\widehat{\boldsymbol{\xi}}^i$, and $b_2^i = -\boldsymbol{x}^\top\widehat{\boldsymbol{\xi}}^i - \rho/(1-\beta)\tau$, where we suppress the dependence of these matrices and values on $(\boldsymbol{x},\tau)$ for notational simplicity. Then, we can rewrite the inner supremum problem in the objective (C.1a) as follows:

$$\begin{aligned} & \sup_{\Theta\in\mathbb{R}^{d_\xi\times d_z}}\left\{\frac{1}{N}\sum_{i=1}^{N}\max\left\{\langle\boldsymbol{A}_1^i,\boldsymbol{\Theta}\rangle+b_1^i,\,\langle\boldsymbol{A}_2^i,\boldsymbol{\Theta}\rangle+b_2^i\right\} - \lambda\|\boldsymbol{\Theta}-\widehat{\boldsymbol{\Theta}}_m\|\right\} \qquad \text{(C.2a)}\\ = & \sup_{\Theta\in\mathbb{R}^{d_\xi\times d_z}}\max_{\boldsymbol{k}\in\{1,2\}^N}\left\{\frac{1}{N}\sum_{i=1}^{N}\left(\langle\boldsymbol{A}_{k_i}^i,\boldsymbol{\Theta}\rangle+b_{k_i}^i\right) - \lambda\|\boldsymbol{\Theta}-\widehat{\boldsymbol{\Theta}}_m\|\right\} \qquad \text{(C.2b)}\\ = & \sup_{\Theta\in\mathbb{R}^{d_\xi\times d_z}}\left\{\max_{\boldsymbol{k}\in\{1,2\}^N}\left\{\langle\overline{\boldsymbol{A}}_{\boldsymbol{k}},\boldsymbol{\Theta}\rangle+\bar{b}_{\boldsymbol{k}}\right\} - \lambda\|\boldsymbol{\Theta}-\widehat{\boldsymbol{\Theta}}_m\|\right\}, \qquad \text{(C.2c)}\end{aligned}$$

where $\overline{\boldsymbol{A}}_{\boldsymbol{k}} = N^{-1}\sum_{i=1}^{N}\boldsymbol{A}_{k_i}^i$ and $\bar{b}_{\boldsymbol{k}} = N^{-1}\sum_{i=1}^{N}b_{k_i}^i$. Here, (C.2b) follows from the fact that the (inner) objective function in (C.2b) is separable in $\boldsymbol{k}$, and (C.2c) follows from the definitions of $\overline{\boldsymbol{A}}_{\boldsymbol{k}}$ and $\bar{b}_{\boldsymbol{k}}$. Note that the function $\max_{\boldsymbol{k}\in\{1,2\}^N}\left\{\langle\overline{\boldsymbol{A}}_{\boldsymbol{k}},\boldsymbol{\Theta}\rangle+\bar{b}_{\boldsymbol{k}}\right\}$ in (C.2c) is a max-affine function in $\boldsymbol{\Theta}$. Applying the same argument as in Theorem 6.3 of Mohajerin Esfahani and Kuhn (2018) (see also

Remark 6.6), the supremum problem (C.2) is equivalent to

$$\begin{cases} \max_{\boldsymbol{k}\in\{1,2\}^N} \left\{ \langle \overline{\boldsymbol{A}}_{\boldsymbol{k}}, \widehat{\boldsymbol{\Theta}}_m \rangle + \bar{b}_{\boldsymbol{k}} \right\}, & \text{if } \lambda \geq \|\overline{\boldsymbol{A}}_{\boldsymbol{k}}\|_* \text{ for all } \boldsymbol{k} \in \{1,2\}^N, \\ \infty & \text{otherwise,} \end{cases} \tag{C.3}$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$. Therefore, using (C.3), we derive the following equivalent reformulation of (C.1).

$$\begin{aligned} \underset{\boldsymbol{x},\,\tau,\,\lambda\geq 0}{\text{minimize}} \quad & \rho\tau + r\lambda + \sum_{m=1}^{M} \widehat{w}_m \sup_{\Theta\in\mathfrak{T}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left\{ -\boldsymbol{x}^\top \left[\widehat{\boldsymbol{\Theta}}_m (\boldsymbol{z} - \boldsymbol{z}^i) + \widehat{\boldsymbol{\xi}}^i \right] \right.\right. \\ & \left.\left. + \frac{\rho}{1-\beta} \Big(-\boldsymbol{x}^\top \left[\widehat{\boldsymbol{\Theta}}_m (\boldsymbol{z} - \boldsymbol{z}^i) + \widehat{\boldsymbol{\xi}}^i \right] - \tau \Big)^+ \right\} \right\} \qquad \text{(C.4a)} \\ \text{subject to} \quad & \mathbf{1}^\top \boldsymbol{x} = 1, \quad \boldsymbol{x} \geq 0, \qquad \text{(C.4b)} \\ & \lambda \geq \left\| \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{A}^i_{k_i} \right\|_*, \quad \forall \boldsymbol{k} \in \{1,2\}^N. \qquad \text{(C.4c)} \end{aligned}$$

Since we employ the vectorized ℓ_1 norm, i.e., $\|\boldsymbol{\Theta}\| = \|\mathrm{vec}(\boldsymbol{\Theta})\|_1$, its dual norms is the vectorized ℓ_∞ norm. The set of constraints (C.4c) is then equivalent to

$$\lambda \geq \left| \frac{1}{N} \sum_{i=1}^{N} A^i_{k_i,j,\ell} \right|, \quad \forall \boldsymbol{k} \in \{1,2\}^N,\ j \in [d_\xi],\ \ell \in [d_z], \tag{C.5}$$

where $A^i_{k_i,j,\ell}$ is the (k,ℓ)-th entry of the matrix $\boldsymbol{A}^i_{k_i}$. Note that for any $i \in [N]$, $j \in [d_\xi]$ and $\ell \in [d_z]$, since $x_j \geq 0$, we have $A^i_{1,j,\ell} = -\big(1 + \rho/(1-\beta)\big)(z_\ell - z^i_\ell)x_j \leq -(z_\ell - z^i_\ell)x_j = A^i_{2,j,\ell} \leq 0$ if $z_\ell - z^i_\ell > 0$, and $A^i_{1,j,\ell} = -\big(1 + \rho/(1-\beta)\big)(z_\ell - z^i_\ell)x_j \geq -(z_\ell - z^i_\ell)x_j = A^i_{2,j,\ell} \geq 0$ if $z_\ell - z^i_\ell < 0$. Let $\mathcal{I}^+_\ell = \big\{ i \in [N] \,\big|\, z_\ell - z^i_\ell > 0 \big\}$ and $\mathcal{I}^-_\ell = \big\{ i \in [N] \,\big|\, z_\ell - z^i_\ell < 0 \big\}$. Then, we have the following upper and lower bounds for any $\boldsymbol{k} \in \{1,2\}^N$:

$$\frac{1}{N} \sum_{i=1}^{N} A^i_{k_i,j,\ell} \leq \frac{1}{N} \left[\sum_{i\in\mathcal{I}^+_\ell} -(z_\ell - z^i_\ell)x_j + \sum_{i\in\mathcal{I}^-_\ell} -\left(1 + \frac{\rho}{1-\beta}\right)(z_\ell - z^i_\ell)x_j \right], \tag{C.6a}$$

$$\frac{1}{N} \sum_{i=1}^{N} A^i_{k_i,j,\ell} \geq \frac{1}{N} \left[\sum_{i\in\mathcal{I}^+_\ell} -\left(1 + \frac{\rho}{1-\beta}\right)(z_\ell - z^i_\ell)x_j + \sum_{i\in\mathcal{I}^-_\ell} -(z_\ell - z^i_\ell)x_j \right]. \tag{C.6b}$$

Therefore, using (C.6), we can reformulate (C.5) as

$$\lambda \geq \left| \frac{1}{N} \left[\sum_{i\in\mathcal{I}^+_\ell} -(z_\ell - z^i_\ell) + \sum_{i\in\mathcal{I}^-_\ell} -\left(1 + \frac{\rho}{1-\beta}\right)(z_\ell - z^i_\ell) \right] x_j \right| =: q_{1,\ell} x_j, \quad \forall j \in [d_\xi],\ \ell \in [d_z], \tag{C.7a}$$

$$\lambda \geq \left| \frac{1}{N} \left[\sum_{i \in \mathcal{I}_\ell^+} -\left(1 + \frac{\rho}{1-\beta}\right)(z_\ell - z_\ell^i) + \sum_{i \in \mathcal{I}_\ell^-} -(z_\ell - z_\ell^i) \right] x_j \right| =: q_{2,\ell} x_j, \quad \forall j \in [d_\xi],\, \ell \in [d_z]. \quad \text{(C.7b)}$$

Let $q_\ell = \max\{q_{1,\ell}, q_{2,\ell}\}$ and $q_{\max} = \max_{\ell \in [d_z]}\{q_\ell\}$. Constraints (C.7) imply that $\lambda \geq q_{\max}\|\boldsymbol{x}\|_\infty$. Thus, (C.1) is equivalent to

$$\begin{aligned}
\underset{\boldsymbol{x},\,\tau,\,\lambda \geq 0}{\text{minimize}} \quad & \rho\tau + r\lambda + \sum_{m=1}^{M} \widehat{w}_m \left\{ \frac{1}{N} \sum_{i=1}^{N} \left\{ -\boldsymbol{x}^\top \left[\widehat{\boldsymbol{\Theta}}_m (\boldsymbol{z} - \boldsymbol{z}^i) + \widehat{\boldsymbol{\xi}}^i \right] \right.\right. \\
& \left.\left. + \frac{\rho}{1-\beta} \left(-\boldsymbol{x}^\top \left[\widehat{\boldsymbol{\Theta}}_m (\boldsymbol{z} - \boldsymbol{z}^i) + \widehat{\boldsymbol{\xi}}^i \right] - \tau \right)^+ \right\} \right. \qquad \text{(C.8a)} \\
\text{subject to} \quad & \mathbf{1}^\top \boldsymbol{x} = 1, \quad \boldsymbol{x} \geq 0, \qquad \text{(C.8b)} \\
& \lambda \geq q_{\max} \|\boldsymbol{x}\|_\infty. \qquad \text{(C.8c)}
\end{aligned}$$

This shows that with a single training regression model (i.e., $M = 1$), our MUR-SAA model with 1-Wasserstein MUR ambiguity set is equivalent to the ER-DRO model with 1-Wasserstein ambiguity set with a scaled radius; see Mohajerin Esfahani and Kuhn (2018) for an equivalent reformulation of the latter model. Thus, we do not consider the latter model in our experiments.

Appendix C.2. Reformulation of the MUR-SAA model with Sample Robust MUR Ambiguity Set

Recall our MUR-SAA model with the sample robust MUR ambiguity set:

$$\begin{aligned}
\underset{\boldsymbol{x},\,\tau}{\text{minimize}} \quad & \rho\tau + \sum_{m=1}^{M} \widehat{w}_m \sup_{\Theta \in \mathfrak{T}_m} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left\{ -\boldsymbol{x}^\top \left[\boldsymbol{\Theta} (\boldsymbol{z} - \boldsymbol{z}^i) + \widehat{\boldsymbol{\xi}}^i \right] \right.\right. \\
& \left.\left. + \frac{\rho}{1-\beta} \left(-\boldsymbol{x}^\top \left[\boldsymbol{\Theta} (\boldsymbol{z} - \boldsymbol{z}^i) + \widehat{\boldsymbol{\xi}}^i \right] - \tau \right)^+ \right\} \right\} \qquad \text{(C.9a)} \\
\text{subject to} \quad & \mathbf{1}^\top \boldsymbol{x} = 1, \quad \boldsymbol{x} \geq 0, \qquad \text{(C.9b)}
\end{aligned}$$

Consider the inner supremum problem in (C.9a). Note that the objective function is convex in $\boldsymbol{\Theta}$. Therefore, there always exists an extreme point of $\mathfrak{T}_m$ that is an optimal solution to the supremum problem. In other words, we can replace the set $\mathfrak{T}_m$ with its set of extreme points. Note that $\mathfrak{T}_m = \{\boldsymbol{\Theta} \mid \|\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}_m\|_1 \leq r_m\}$. Thus, the set of extreme points are $\{\boldsymbol{\Theta} = \widehat{\boldsymbol{\Theta}}_m \pm r_m \boldsymbol{E}_{j\ell} \mid j \in [d_\xi], \ell \in [d_z]\}$, where $\boldsymbol{E}_{j\ell}$ is a matrix with all zero entries except for the (j,ℓ)-th entry being one. If we enumerate the set of extreme points as $\{\boldsymbol{\Theta}_m^k\}_{k=1}^{2\times d_\xi \times d_z}$ for all $m \in [M]$. we can derive the following equivalent reformulation of the MUR-SAA model with the sample robust MUR ambiguity set:

$$\begin{aligned}
\underset{\boldsymbol{x},\,\tau,\,\boldsymbol{\eta},\,\boldsymbol{t}}{\text{minimize}} \quad & \rho\tau + \sum_{m=1}^{M} \widehat{w}_m \eta_m \qquad \text{(C.10a)} \\
\text{subject to} \quad & \mathbf{1}^\top \boldsymbol{x} = 1, \quad \boldsymbol{x} \geq 0, \qquad \text{(C.10b)}
\end{aligned}$$

$$\eta_m \geq \frac{1}{N}\sum_{i=1}^{N}\Big\{ -\boldsymbol{x}^\top\big[\boldsymbol{\Theta}_m^k(\boldsymbol{z}-\boldsymbol{z}^i)+\widehat{\boldsymbol{\xi}}^i\big]+\frac{\rho}{1-\beta}t_{i,m}^k\Big\},$$

$$\forall m\in[M],\, k\in[2\times d_\xi\times d_z], \tag{C.10c}$$

$$t_{i,m}^k \geq -\boldsymbol{x}^\top\big[\boldsymbol{\Theta}_m^k(\boldsymbol{z}-\boldsymbol{z}^i)+\widehat{\boldsymbol{\xi}}^i\big]-\tau,\;\; t_{i,m}^k\geq 0,$$

$$\forall i\in[N],\, m\in[M],\, k\in[2\times d_\xi\times d_z]. \tag{C.10d}$$

References

Babyak, M.A., 2004. What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. Biopsychosocial Science and Medicine 66, 411–421.

Ban, G.Y., Gallien, J., Mersereau, A.J., 2019. Dynamic procurement of new products with covariate information: The residual tree method. Manufacturing & Service Operations Management 21, 798–815.

Ban, G.Y., Rudin, C., 2019. The big data newsvendor: Practical insights from machine learning. Operations Research 67, 90–108.

Bertsimas, D., Kallus, N., 2020. From predictive to prescriptive analytics. Management Science 66, 1025–1044.

Bertsimas, D., Koduri, N., 2022. Data-driven optimization: A reproducing kernel Hilbert space approach. Operations Research 70, 454–471.

Bertsimas, D., Pauphilet, J., Van Parys, B., 2020. Sparse regression: Scalable algorithms and empirical performance. Statistical Science 35, 555–578.

Bertsimas, D., Shtern, S., Sturt, B., 2022. Two-stage sample robust optimization. Operations Research 70, 624–640.

Bertsimas, D., Van Parys, B., 2022. Bootstrap robust prescriptive analytics. Mathematical Programming 195, 39–78.

Chatfield, C., 1995. Model uncertainty, data mining and statistical inference. Journal of the Royal Statistical Society Series A: Statistics in Society 158, 419–444.

Chen, R., Paschalidis, I., 2019. Selecting optimal decisions via distributionally robust nearest-neighbor regression. Advances in Neural Information Processing Systems 32.

Chen, Z., Zhang, J., Xu, W., Yang, Y., 2022. Consistency of BIC model averaging. Statistica Sinica 32, 635–640.

Costa, G., Iyengar, G.N., 2023. Distributionally robust end-to-end portfolio construction. Quantitative Finance 23, 1465–1482.

Deng, Y., Sen, S., 2018. Learning enabled optimization: Towards a fusion of statistical learning and stochastic programming. Available at Optimization Online .

Deng, Y., Sen, S., 2022. Predictive stochastic programming. Computational Management Science 19, 65–98.

Donti, P., Amos, B., Kolter, J.Z., 2017. Task-based end-to-end model learning in stochastic optimization. Advances in Neural Information Processing Systems 30.

Dou, X., Anitescu, M., 2019. Distributionally robust optimization with correlated data from vector autoregressive processes. Operations Research Letters 47, 294–299.

Draper, D., 1995. Assessment and propagation of model uncertainty. Journal of the Royal Statistical Society Series B: Statistical Methodology 57, 45–70.

Duchi, J.C., Glynn, P.W., Namkoong, H., 2021. Statistics of robust optimization: A generalized empirical likelihood approach. Mathematics of Operations Research 46, 946–969.

Elmachtoub, A.N., Grigas, P., 2022. Smart "predict, then optimize". Management Science 68, 9–26.

Fan, J., Lou, Z., Yu, M., 2024. Are latent factor regression and sparse regression adequate? Journal of the American Statistical Association 119, 1076–1088.

Fan, J., Lv, J., 2010. A selective overview of variable selection in high dimensional feature space. Statistica Sinica 20, 101.

Gao, R., 2023. Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality. Operations Research 71, 2291–2306.

Gao, R., Kleywegt, A., 2023. Distributionally robust stochastic optimization with Wasserstein distance. Mathematics of Operations Research 48, 603–655.

Garling, D.J., 2018. Analysis on Polish Spaces and an Introduction to Optimal Transportation. volume 89. Cambridge University Press.

Hanasusanto, G.A., Kuhn, D., Wallace, S.W., Zymler, S., 2015. Distributionally robust multi-item newsvendor problems with multimodal demand distributions. Mathematical Programming 152, 1–32.

Hansen, B.E., 2007. Least squares model averaging. Econometrica 75, 1175–1189.

Hastie, T., Tibshirani, R., Tibshirani, R., 2020. Best subset, forward stepwise or LASSO? Analysis and recommendations based on extensive comparisons. Statistical Science 35, 579–592.

Hurvich, C.M., Tsai, C.L., 1989. Regression and time series model selection in small samples. Biometrika 76, 297–307.

Kannan, R., Bayraksan, G., Luedtke, J.R., 2024. Residuals-based distributionally robust optimization with covariate information. Mathematical Programming 207, 369–425.

Kannan, R., Bayraksan, G., Luedtke, J.R., 2025. Data-driven sample average approximation with covariate information. Operations Research 73, 3245–3259.

Lewandowski, D., Kurowicka, D., Joe, H., 2009. Generating random correlation matrices based on vines and extended onion method. Journal of Multivariate Analysis 100, 1989–2001.

Liao, J., Wan, A.T., He, S., Zou, G., 2022. Optimal model averaging for multivariate regression models. Journal of Multivariate Analysis 189, 104858.

Liu, Y., Xu, H., 2013. Stability analysis of stochastic programs with second order dominance constraints. Mathematical Programming 142, 435–460.

McNeish, D.M., 2015. Using LASSO for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. Multivariate Behavioral Research 50, 471–484.

Mohajerin Esfahani, P., Kuhn, D., 2018. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. Mathematical Programming 171, 115–166.

Muñoz, M.A., Pineda, S., Morales, J.M., 2022. A bilevel framework for decision-making under uncertainty with contextual information. Omega 108, 102575.

Newey, W.K., 1991. Uniform convergence in probability and stochastic equicontinuity. Econometrica: Journal of the Econometric Society , 1161–1167.

Oroojlooyjadid, A., Snyder, L.V., Takáč, M., 2020. Applying deep learning to the newsvendor problem. IISE Transactions 52, 444–463.

Parthasarathy, K.R., 2005. Probability Measures on Metric Spaces. volume 352. American Mathematical Society.

Pflug, G.C., Pichler, A., 2014. Multistage Stochastic Optimization. volume 1104. Springer.

Pichler, A., Xu, H., 2022. Quantitative stability analysis for minimax distributionally robust risk optimization. Mathematical Programming 191, 47–77.

Qi, M., Grigas, P., Shen, Z.J., 2025. Integrated conditional estimation-optimization. Operations Research .

Qi, M., Shen, Z.J., 2022. Integrating prediction/estimation and optimization with applications in operations management, in: Tutorials in Operations Research: Emerging and Impactful Topics in Operations. INFORMS, pp. 36–58.

Rachev, S.T., Römisch, W., 2002. Quantitative stability in stochastic programming: The method of probability metrics. Mathematics of Operations Research 27, 792–818.

Riley, R.D., Snell, K.I., Martin, G.P., Whittle, R., Archer, L., Sperrin, M., Collins, G.S., 2021. Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. Journal of Clinical Epidemiology 132, 88–96.

Römisch, W., Schultz, R., 1991. Stability analysis for stochastic programs. Annals of Operations Research 30, 241–266.

Sadana, U., Chenreddy, A., Delage, E., Forel, A., Frejinger, E., Vidal, T., 2025. A survey of contextual optimization methods for decision-making under uncertainty. European Journal of Operational Research 320, 271–289.

Sarwar, O., Sauk, B., Sahinidis, N.V., 2020. A discussion on practical considerations with sparse regression methodologies. Statistical Science 35, 593–601.

Shapiro, A., Dentcheva, D., Ruszczyński, A., 2014. Lectures on Stochastic Programming: Modeling and Theory. Second ed., SIAM.

Sim, M., Tang, Q., Zhou, M., Zhu, T., 2025. The analytics of robust satisficing: predict, optimize, satisfice, then fortify. Operations Research 73, 2708–2728.

Song, M., Zou, G., Wan, A.T., 2026. Bootstrap model averaging. Journal of Business & Economic Statistics , 1–21.

Subramanian, J., Simon, R., 2013. Overfitting in prediction models–is it a problem only in high dimensions? Contemporary Clinical Trials 36, 636–641.

Sun, H., Xu, H., 2016. Convergence analysis for distributionally robust optimization and equilibrium problems. Mathematics of Operations Research 41, 377–401.

Tsang, M.Y., Shehadeh, K.S., 2025. On the tradeoff between distributional belief and ambiguity: Conservatism, finite-sample guarantees, and asymptotic properties. INFORMS Journal on Optimization 7, 240–263.

Tsang, M.Y., Shehadeh, K.S., 2026. Algorithmic approaches for identifying the trade-off between pessimism and optimism in a stochastic fixed charge facility location problem. INFORMS Journal on Optimization .

Wang, F., Mukherjee, S., Richardson, S., Hill, S.M., 2020. High-dimensional regression in practice: An empirical study of finite-sample prediction, variable selection and ranking. Statistics and Computing 30, 697–719.

White, H., 1981. Consequences and detection of misspecified nonlinear regression models. Journal of the American Statistical Association 76, 419–433.

White, H., 1982. Maximum likelihood estimation of misspecified models. Econometrica: Journal of the Econometric Society 50, 1–25.

Wu, D., Zhu, H., Zhou, E., 2018. A Bayesian risk approach to data-driven stochastic optimization: Formulations and asymptotics. SIAM Journal on Optimization 28, 1588–1612.

Xu, H., Zhang, S., 2023. Quantitative statistical robustness in distributionally robust optimization models. Pacific Journal of Optimization 19, 335–361.

Yu, X., Basciftci, B., 2026. Distributionally robust optimization with multimodal decision-dependent ambiguity sets. Mathematical Programming , 1–51.

Yurdakul, O., Bayraksan, G., 2024. A data-driven methodology for contextual unit commitment using regression residuals. IEEE Transactions on Power Systems 39, 6914–6933.

Zhang, K., Li, W., Han, Y., Geng, Z., Chu, C., 2021. Production capacity identification and analysis using novel multivariate nonlinear regression: Application to resource optimization of industrial processes. Journal of Cleaner Production 282, 124469.

Zhang, X., 2015. Consistency of model averaging estimators. Economics Letters 130, 120–123.

Zhang, X., Liu, C.A., 2023. Model averaging prediction by K-fold cross-validation. Journal of Econometrics 235, 280–301.

Zhu, Q., Yu, X., Bayraksan, G., 2024. Residuals-based contextual distributionally robust optimization with decision-dependent uncertainty. arXiv preprint arXiv:2406.20004 .

Zhu, T., Xie, J., Sim, M., 2022. Joint estimation and robustness optimization. Management Science 68, 1659–1677.