

Data-Driven Sample Average Approximation with Covariate Information*

Rohit Kannan^{†1}, Güzin Bayraksan², and James R. Luedtke³

¹Center for Nonlinear Studies (T-CNLS) and Applied Mathematics & Plasma Physics (T-5),
Los Alamos National Laboratory, Los Alamos, NM, USA. E-mail: rohitk@alum.mit.edu

²Department of Integrated Systems Engineering, The Ohio State University, Columbus, OH, USA.
E-mail: bayraksan.1@osu.edu

³Department of Industrial & Systems Engineering and Wisconsin Institute for Discovery,
University of Wisconsin-Madison, Madison, WI, USA. E-mail: jim.luedtke@wisc.edu

Version 2 (this document): July 25, 2022

Version 1: July 24, 2020

Abstract

We study optimization for data-driven decision-making when we have observations of the uncertain parameters within the optimization model together with concurrent observations of covariates. Given a new covariate observation, the goal is to choose a decision that minimizes the expected cost conditioned on this observation. We investigate three data-driven frameworks that integrate a machine learning prediction model within a stochastic programming sample average approximation (SAA) for approximating the solution to this problem. Two of the SAA frameworks are new and use out-of-sample residuals of leave-one-out prediction models for scenario generation. The frameworks we investigate are flexible and accommodate parametric, nonparametric, and semiparametric regression techniques. We derive conditions on the data generation process, the prediction model, and the stochastic program under which solutions of these data-driven SAAs are consistent and asymptotically optimal, and also derive convergence rates and finite sample guarantees. Computational experiments validate our theoretical results, demonstrate the potential advantages of our data-driven formulations over existing approaches (even when the prediction model is misspecified), and illustrate the benefits of our new data-driven formulations in the limited data regime.

Key words: Data-driven stochastic programming, covariates, regression, sample average approximation, jackknife, large deviations

1 Introduction

We study data-driven decision-making under uncertainty, where the decision-maker (DM) has access to a finite number of observations of uncertain parameters of an optimization model together with concurrent observations of auxiliary features/covariates. Stochastic programming [17, 73] is a popular modeling framework for decision-making under uncertainty in such applications. A standard formulation of a stochastic program is

$$\min_{z \in \mathcal{Z}} \mathbb{E}[c(z, Y)],$$

*An earlier version of this article was available on Optimization Online on July 24, 2020. This version also incorporates analysis from our unpublished technical report arXiv preprint arXiv:2101.03139.

[†]I dedicate this work to the memory of my grandparents (Thathi and Thatha).

where z denotes the decision variables, Y denotes the uncertain model parameters, \mathcal{Z} denotes the feasible region, c is a cost function, and the expectation is computed with respect to the distribution of Y . Data-driven solution methods such as sample average approximation (SAA) traditionally assume access to only samples of the random vector Y [43, 73]. However, in many real-world applications, values of Y (e.g., demand for water and energy) are predicted using available covariate information (e.g., weather).

Motivated by the developments in Ban and Rudin [4], Bertsimas and Kallus [12], and Sen and Deng [72], we study the case in which covariate information is available and can be used to inform the distribution of Y . Specifically, given a new random observation $X = x$ of covariates, the goal of the DM is to solve the conditional stochastic program

$$\min_{z \in \mathcal{Z}} \mathbb{E} [c(z, Y) \mid X = x]. \quad (\text{SP})$$

The aim of this paper is to analyze the SAA framework when a prediction model—obtained by statistical or machine learning—is explicitly integrated into the SAA for (SP) to leverage the covariate observation $X = x$. Here, residuals of the prediction model are scaled and added on to a point prediction of Y at $X = x$ to construct scenarios of Y for use within the SAA. We formally define our data-driven approximations to (SP) in Section 2.

Applications of this framework include (i) the data-driven newsvendor problem [4], where the product’s demand can be predicted using seasonality and location data before making order decisions, (ii) dynamic procurement of a new product [5] whose demand can be predicted using historical data for similar past products, (iii) shipment planning under uncertainty [12], where historical demands, weather forecasts, and web search results can be used to predict products’ demands before making production and inventory decisions, and (iv) grid scheduling under uncertainty [30], where seasonality, weather, and historical demand data can be used to predict the load before creating generator schedules.

Formulation (SP) requires knowledge of the conditional distribution of the random variables given a new realization of the covariates. Since this distribution is typically unknown, we are interested in using an estimate of it to approximately solve (SP) given access to a finite set of joint observations of (X, Y) . In this setting, we would like to construct approximations to (SP) that not only have good statistical properties, but are also practically effective in the limited data regime. At a minimum, we would like a data-driven approach that is asymptotically optimal in the sense that the objective value of its solutions approaches the optimal value of (SP) as the number of samples increases. We would also like to determine the rate at which this convergence occurs.

Our first contribution is to generalize and analyze the approach proposed in Ban et al. [5] and Sen and Deng [72], in which data-driven approximations to (SP) are constructed using explicit models that predict the random vector Y using the covariates X . In this approach, a prediction model is first used to generate a point prediction of Y at the new observation $X = x$. The residuals obtained during the training of the prediction model are then scaled and added on to this point prediction to construct scenarios for use within an SAA framework to approximate the solution to (SP). We refer to this approximation as the *empirical residuals-based SAA* (ER-SAA). We demonstrate asymptotic optimality, rates of convergence, and finite sample guarantees of solutions obtained from the ER-SAA under mild assumptions. Inspired by jackknife-based methods for constructing prediction intervals [6], we also propose two new data-driven SAA frameworks that use *leave-one-out residuals* instead of empirical residuals, and demonstrate how our analysis can be extended to these frameworks. The motivation for these new data-driven SAA formulations is that using leave-one-out residuals might result in a better approximation of the true conditional distribution of Y given $X = x$, particularly when the sample size is small.

The prediction frameworks we analyze are flexible and accommodate parametric, nonparametric, and semiparametric regression techniques [41, 78, 83]. While our results imply that using nonparametric regression techniques within our SAA frameworks results in convergent approximations to (SP) under mild assumptions [cf. 12], the rate at which such approximations converge typically exhibits poor dependence on the dimension of the covariate vector X . Parametric (and semiparametric) regression approaches, on the other hand, presume some knowledge of the functional dependence of Y on X . If the assumed functional dependence is a good approximation of the true dependence, they may yield significantly better solutions

when the number of samples is limited. The tradeoff between employing parametric and nonparametric regression techniques within our framework is evident upon looking at the assumptions under which these approaches are guaranteed to yield convergent approximations to (SP), the rates at which their optimal solutions converge, and numerical experience in Section 4. The generality of our framework enables DMs to choose the modeling approach that works best for their application.

1.1 Relation to existing literature

The papers of Ban et al. [5] and Sen and Deng [72] are most closely related to this work. Motivated by the application of dynamic procurement of a short-life-cycle product in the presence of demand uncertainty, Ban et al. [5] propose a residual tree method for the data-driven solution of multistage stochastic programs (also see references to the operations management literature therein for other data-driven approaches). They propose to use ordinary least squares (OLS) or Lasso regression to generate demand forecasts for a new product using historical demand and covariate data for similar products, and establish asymptotic optimality of their data-driven procurement decisions for their particular application. Sen and Deng [72] also use predictive models to generate scenarios of random variables in stochastic programs with exogenous and endogenous uncertainty when covariate information is available. They propose an empirical additive error method that is similar to the residual tree method of Ban et al. [5]. They also consider estimating distributions of the coefficients and residuals of a linear regression model and propose to subsequently sample from these distributions to generate scenarios of the random variables. They present model validation and model selection strategies for when the DM has access to several candidate prediction models. Kim and Mehrotra [52] use empirical residuals to construct scenarios in a computational study, but conduct no analysis of the approach.

Our work differs from the above in the following respects: we introduce a general framework that applies to a wide range of prediction and optimization models and allows for the covariance matrix of the errors to depend on the covariates; we establish asymptotic optimality of the solutions from the ER-SAA under general conditions; we derive results establishing rates of convergence and finite sample guarantees of the solutions from the ER-SAA; we propose two new frameworks that use leave-one-out residuals and extend the asymptotic optimality and rate of convergence analysis to these frameworks; and we present an empirical study demonstrating the potential advantage of using these frameworks.

Bertsimas and Kallus [12] consider approximating the solution to (SP) by solving a reweighted SAA problem, where the weights are chosen using nonparametric regression methods based on k-nearest neighbors (kNN), kernels, classification and regression trees (CART), or random forests (RF). They pay particular attention to the setting where the joint observations of (X, Y) may not be i.i.d., but arise from a mixing process. They also consider the setting where decisions affect the realization of the uncertainty, and establish asymptotic optimality and consistency of their data-driven solutions. They also consider policy-based empirical risk minimization (ERM) approaches for (SP), and develop out-of-sample guarantees for costs of decisions constructed using such policies. Diao and Sen [29] develop stochastic quasigradient methods for efficiently solving the kNN and kernel-based reweighted SAA formulations of Bertsimas and Kallus [12] without sacrificing theoretical guarantees. Ban and Rudin [4] also propose a policy-based ERM approach and a kernel regression-based nonparametric approach for solving (SP) in the context of the data-driven newsvendor problem. They derive finite sample guarantees on the out-of-sample costs of order decisions, and quantify the gains from using feature information under different demand models. Bazier-Matte and Delage [9] derive out-of-sample performance guarantees for regularized portfolio selection with side information. Bertsimas and McCord [13] extend the analysis of Bertsimas and Kallus [12] to the multistage setting when the covariates evolve according to a Markov process. They establish asymptotic optimality and consistency of their data-driven decisions along with finite sample guarantees for the solutions to the kNN-based approach. Kallus and Mao [49] propose RF-based decision policies for problem (SP) and demonstrate asymptotic optimality of these policies. Finally, Hu et al. [46] show that “estimate and then optimize” methods can have faster regret convergence rates than ERM-based approaches for contextual linear optimization when the estimated parameters appear in the objective coefficients.

Our work differs from the above in the following respects: we propose data-driven approaches to approximate the solution to (SP) that rely on the construction of explicit models to predict the random variables from covariates, allow for both parametric and nonparametric regression models, and derive convergence rates and finite sample guarantees for solutions to our approximations that complement the above analyses.

Another stream of research has been investigating methods that change the training of the prediction model in order to obtain better solutions to (SP) [e.g., see 26, 30, 34, 35]. The philosophy behind these approaches is that, instead of constructing the prediction model purely for high predictive accuracy, the DM should construct a model to predict Y using X such that the resulting optimization decisions provide the lowest cost solution to the true conditional stochastic program (SP). These methods result in harder joint estimation and optimization problems that can only be solved to optimality in special settings. In contrast, we focus on the setting where the prediction framework is independent of the stochastic programming model. This is common in many real-world applications and facilitates easily changing or improving the prediction model.

Several recent works [e.g., see 14, 31, 36, 61] use distributionally robust optimization (DRO) in a bid to construct better approximations to (SP) than SAA in the limited data regime. In follow-up work [50], we study residuals-based DRO formulations that are built around our data-driven SAA formulations, analyze their theoretical guarantees, and illustrate their advantages in the limited data regime through a case study. Our data-driven SAA formulations in this work are flexible and remain tractable under milder assumptions on (SP) compared to such DRO approaches.

A ‘traditional data-driven SAA approach’ for the conditional stochastic program (SP) would involve constructing a model to predict the random variables Y given X , fitting a distribution to the residuals of the prediction model, and using samples from this distribution along with the prediction model to construct scenarios for Y given $X = x$. While it is difficult to pin down a reference that is *the first* to adopt this approach, we point to the works of Schütz et al. [69], Royset and Wets [67], and the references therein for applications-motivated versions. Recent work by Grigas et al. [40] contributes to this approach by attempting to directly estimate the conditional distribution of Y given X while considering the structure of the optimization problem (assuming Y has finite support). Instead of fitting a distribution to the residuals of the prediction model, we propose and analyze methods that directly use empirical residuals within the SAA framework. These methods avoid the need to fit a distribution of the residuals, and hence we expect them to be advantageous when the available data is insufficient to provide a good estimate of the residuals distribution.

1.2 Summary of main contributions

The key contributions of this paper are as follows:

1. We demonstrate asymptotic optimality, rates of convergence, and finite sample guarantees of solutions to the ER-SAA formulation under mild assumptions on the data, the prediction framework, and the stochastic programming formulation.
2. We introduce and analyze two new variants of ER-SAA that use leave-one-out residuals instead of empirical residuals, which may lead to better solutions when data is limited.
3. We verify that the assumptions on the underlying stochastic programming formulation hold for a broad class of two-stage stochastic programs, including two-stage stochastic mixed-integer programming (MIP) with continuous recourse. Additionally, we verify that the assumptions on the prediction step hold for a broad class of M-estimation procedures and nonparametric regression methods, including OLS, Lasso, kNN, and RF regression.
4. Finally, we empirically validate our theoretical results, demonstrate the advantages of our data-driven SAA formulations over existing approaches in the limited data regime, and demonstrate the potential benefit of using a structured prediction model even if it is misspecified.

2 Data-driven SAA frameworks

Recall that our goal is to approximate the solution to the conditional stochastic program (SP):

$$\min_{z \in \mathcal{Z}} \mathbb{E} [c(z, Y) \mid X = x],$$

where $X = x$ is a new random observation of the covariates and the expectation is taken with respect to the conditional distribution of Y given $X = x$. Let P_X and P_Y denote the marginal distributions of the covariates X and the random vector Y , respectively, and $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ denote their supports. We assume that the support \mathcal{Y} is nonempty and convex and $c : \mathcal{Z} \times \mathcal{Y} \rightarrow \mathbb{R}$.

We assume that the ‘true relationship’ between the random vector Y and the random covariates X can be described as

$$Y = f^*(X) + Q^*(X)\varepsilon,$$

where $f^*(x) := \mathbb{E}[Y \mid X = x]$ is the regression function, $Q^*(X)$ is the square root of the conditional covariance matrix of the error term, and the zero-mean random errors ε are independent of the covariates X . Because the error term $Q^*(X)\varepsilon$ is influenced by the covariate X , our model is *heteroscedastic*. When $Q^* \equiv I$, as assumed in Ban et al. [5] and Sen and Deng [72], the error distribution is *homoscedastic*. Heteroscedasticity arises, for instance, when variability of the random vector Y such as the variability of product demands or wind power availability depends on the covariates X like location and seasonality. It can also arise when the DM cannot fully identify all the covariates and the remaining covariates appear in the error term.

We suppose that f^* and Q^* belong to known classes of functions \mathcal{F} and \mathcal{Q} , respectively. The model classes \mathcal{F} and \mathcal{Q} may comprise parametric or nonparametric models. Let $\Xi \subseteq \mathbb{R}^{d_y}$ denote the support of ε and P_ε denote its distribution. We assume $Q^*(\bar{x}) \succ 0$, $\forall \bar{x} \in \mathcal{X}$, where $A \succ 0$ denotes the matrix A is positive definite. We may also assume without loss of generality that the covariance matrix $\mathbb{E}[\varepsilon\varepsilon^\top] = I$ since a general covariance matrix for ε can be handled by suitably redefining Q^* .

Under these structural assumptions, the conditional stochastic program (SP) is equivalent to

$$v^*(x) := \min_{z \in \mathcal{Z}} \{g(z; x) := \mathbb{E}[c(z, f^*(x) + Q^*(x)\varepsilon)]\}, \quad (1)$$

where the expectation is computed with respect to the distribution P_ε of ε . We refer to problem (1) as the *true problem*, and denote its optimal solution set by $S^*(x)$. Throughout, we assume that the feasible set $\mathcal{Z} \subset \mathbb{R}^{d_z}$ is nonempty and compact, $\mathbb{E}[|c(z, f^*(x) + Q^*(x)\varepsilon)|] < +\infty$ for each $z \in \mathcal{Z}$ and almost every (a.e.) $x \in \mathcal{X}$, and the function $g(\cdot; x)$ is lower semicontinuous (lsc) on \mathcal{Z} for a.e. $x \in \mathcal{X}$ (see Theorem 7.42 of Shapiro et al. [73] for conditions that guarantee $g(\cdot; x)$ is lsc). These assumptions ensure problem (1) is well defined and the solution set $S^*(x) \neq \emptyset$ for a.e. $x \in \mathcal{X}$.

Let $\mathcal{D}_n := \{(y^i, x^i)\}_{i=1}^n$ denote joint observations of (Y, X) . If the functions f^* and Q^* are known, then the *full-information SAA* (FI-SAA) counterpart to the true problem (1) using data \mathcal{D}_n is

$$\min_{z \in \mathcal{Z}} \left\{ g_n^*(z; x) := \frac{1}{n} \sum_{i=1}^n c(z, f^*(x) + Q^*(x)\varepsilon^i) \right\}, \quad (2)$$

where $\varepsilon^i := [Q^*(x^i)]^{-1}(y^i - f^*(x^i))$, $\forall i \in \{1, \dots, n\}$ denote the realizations of the errors ε at the given observations. We cannot solve problems (1) or (2) directly because the functions f^* and Q^* are unknown. A practical alternative is to first estimate f^* from the data \mathcal{D}_n , for instance by using an M-estimator [76, 78] of the form

$$\hat{f}_n(\cdot) \in \arg \min_{f(\cdot) \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \ell(y^i, f(x^i)) \quad (3)$$

with some loss function $\ell : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \rightarrow \mathbb{R}_+$. We sometimes assume that the regression model class \mathcal{F} is parameterized by θ (e.g., the parameters of a linear regression model) and let θ^* denote the true value

of θ corresponding to the regression function f^* . In this setting, the aim of the regression step (3) is to estimate θ^* , and we denote the estimate corresponding to \hat{f}_n by $\hat{\theta}_n$.

Given an estimate \hat{f}_n of f^* , we use the fact that $\mathbb{E}[(Y - f^*(X))(Y - f^*(X))^T | X = x] = Q^*(x)Q^*(x)^T$ and plug in \hat{f}_n instead of f^* to determine the best regression estimate \hat{Q}_n of Q^* in the model class \mathcal{Q} [8]. We may then update our estimate \hat{f}_n using an estimate \hat{Q}_n of Q^* , e.g., using weighted least squares regression [66], which could yield an improved estimate of f^* with lower variance. Alternatively, we could estimate f^* and Q^* jointly using M-estimation [27]. In the homoscedastic setting, we simply set $\hat{Q}_n := Q^* \equiv I$. Throughout, we reference equation (3) for both regression steps (i.e., for estimating f^* and Q^*) with the understanding that our regression setup is not restricted to M-estimation.

Given an estimate \hat{f}_n of f^* and a nonsingular estimate \hat{Q}_n of Q^* , residuals $\hat{\varepsilon}_n^i := [\hat{Q}_n(x^i)]^{-1}(y^i - \hat{f}_n(x^i))$, $i \in \{1, \dots, n\}$, of this estimate can be used as proxy for samples of ε from P_ε . Let $\text{proj}_{\mathcal{Y}}(v)$ denote the orthogonal projection of $v \in \mathbb{R}^{d_y}$ onto \mathcal{Y} . Then, the *empirical residuals-based SAA* (ER-SAA) corresponding to problem (1) is defined as

$$\hat{v}_n^{ER}(x) := \min_{z \in \mathcal{Z}} \left\{ \hat{g}_n^{ER}(z; x) := \frac{1}{n} \sum_{i=1}^n c(z, \text{proj}_{\mathcal{Y}}(\hat{f}_n(x) + \hat{Q}_n(x)\hat{\varepsilon}_n^i)) \right\}. \quad (4)$$

While the ER-SAA problem (4) is equally tractable and does not lose any theoretical guarantees if the scenarios $\{\hat{f}_n(x) + \hat{Q}_n(x)\hat{\varepsilon}_n^i\}$ are *not* projected onto the support \mathcal{Y} , this projection step may be helpful in situations where (some of) the ER-SAA scenarios $\{\hat{f}_n(x) + \hat{Q}_n(x)\hat{\varepsilon}_n^i\}$ lie outside the support \mathcal{Y} even though the “true” FI-SAA scenarios $\{f^*(x) + Q^*(x)\varepsilon^i\}$ are contained in \mathcal{Y} . We let $\hat{z}_n^{ER}(x)$ denote an optimal solution to problem (4) and $\hat{S}_n^{ER}(x)$ denote its optimal solution set. We assume throughout that the set $\hat{S}_n^{ER}(x)$ is nonempty for a.e. $x \in \mathcal{X}$, which holds, for example, if the function $c(\cdot, y)$ is lower semicontinuous on \mathcal{Z} for each $y \in \mathcal{Y}$. We stress that problem (4) is different from the following *naive SAA* (N-SAA) problem that directly uses the observations $\{y^i\}_{i=1}^n$ of the random vector Y without using the new covariate observation $X = x$:

$$\hat{v}_n^{\text{NSAA}} := \min_{z \in \mathcal{Z}} \frac{1}{n} \sum_{i=1}^n c(z, y^i). \quad (5)$$

The computational complexity of the ER-SAA problem (4) is similar to that of the N-SAA problem (5) with the only additional computation cost being the cost of estimating f^* and Q^* . Problem (4) also differs from the following point prediction-based deterministic approximation to (1):

$$\hat{v}_n^{\text{PP}}(x) := \min_{z \in \mathcal{Z}} c(z, \text{proj}_{\mathcal{Y}}(\hat{f}_n(x))). \quad (6)$$

Problem (4) is a modification of problem (6) that accounts for the uncertainty in the point estimate.

We also propose two alternatives to the ER-SAA problem (4) that construct scenarios differently. Note that the n observations $\mathcal{D}_n = \{(y^i, x^i)\}_{i=1}^n$ that are used to estimate the regression functions \hat{f}_n and \hat{Q}_n are also used to estimate the errors $\hat{\varepsilon}_n^i$, $i \in \{1, \dots, n\}$. This can cause a bias in the estimation of the residuals, especially when the sample size n is small, yielding suboptimal solutions for some problems. To alleviate this issue, we propose to use jackknife-based variants of the ER-SAA problem. For each $i \in \{1, \dots, n\}$, let \hat{f}_{-i} and \hat{Q}_{-i} denote the estimates of f^* and Q^* obtained by omitting the data point (y^i, x^i) from the training set \mathcal{D}_n while carrying out the regression step (3), and define the residual term $\hat{\varepsilon}_{n,J}^i := [\hat{Q}_{-i}(x^i)]^{-1}(y^i - \hat{f}_{-i}(x^i))$, calculated at the omitted point (y^i, x^i) . The alternatives we propose are

$$\hat{v}_n^J(x) := \min_{z \in \mathcal{Z}} \left\{ \hat{g}_n^J(z; x) := \frac{1}{n} \sum_{i=1}^n c(z, \text{proj}_{\mathcal{Y}}(\hat{f}_n(x) + \hat{Q}_n(x)\hat{\varepsilon}_{n,J}^i)) \right\}, \quad (7)$$

$$\hat{v}_n^{J+}(x) := \min_{z \in \mathcal{Z}} \left\{ \hat{g}_n^{J+}(z; x) := \frac{1}{n} \sum_{i=1}^n c(z, \text{proj}_{\mathcal{Y}}(\hat{f}_{-i}(x) + \hat{Q}_{-i}(x)\hat{\varepsilon}_{n,J}^i)) \right\}. \quad (8)$$

We call problems (7) and (8) *jackknife-based SAA* (J-SAA) and *jackknife+-based SAA* (J+-SAA), respectively [cf. 6]. These data-driven SAAs are well-motivated when the data \mathcal{D}_n is independent, in which case the leave-one-out residual $\hat{\varepsilon}_{n,J}^i$ may be a significantly more accurate estimate of the scaled prediction error at the covariate observation x^i than the empirical residual $\hat{\varepsilon}_n^i$, particularly when n is small relative to the complexity of the regression step (3) due to overfitting [6]. When \mathcal{D}_n is not independently generated, omitting blocks of data (instead of individual observations as in the jackknife-based methods) during the regression steps (3) can yield better-motivated variants of the J-SAA and J+-SAA formulations [54].

Problems (7) and (8) roughly require the construction of n regression models, which may be computationally unattractive in some settings. This extra computational burden can be alleviated in some special settings such as OLS regression by re-using information from one regression model to the next (see page 13 of Barber et al. [6] for other regression setups that can re-use information). We make use of this computational speed-up in our experiments in Section 4.

We use the following two-stage stochastic linear program (LP) as our running example for problem (1). Section E in the Appendix includes a discussion of more general forms of problem (1) that satisfy the assumptions of our framework. The more general classes of problems discussed in the Appendix subsume the problem class presented in Example 1.

Example 1. [Two-stage stochastic LP] The set \mathcal{Z} is a nonempty convex polytope and the function $c(z, Y) := c_z^T z + V(z, Y)$, with $V(z, Y) := \min_{v \in \mathbb{R}_+^{d_y}} \{c_v^T v : Wv = Y - Tz\}$. The quantities c_z , c_v , W , and T have commensurate dimensions. We assume that $V(z, y) < +\infty$ for each $z \in \mathcal{Z}$ and $y \in \mathbb{R}^{d_y}$, the matrix W has full row rank, and the dual feasible set $\{\lambda : \lambda^T W \leq c_v^T\}$ is nonempty.

We also use OLS regression with a structured parametric model for heteroscedasticity as our running example for the regression step (3). Section F in the Appendix includes a detailed discussion of how other prediction models fit within our framework.

Example 2. [OLS regression] The model class $\mathcal{F} := \{f(\cdot) : f(X) = \theta X \text{ for some } \theta \in \mathbb{R}^{d_y \times d_x}\}$, with $X_1 \equiv 1$ and loss function is $\ell(y, \hat{y}) := \|y - \hat{y}\|^2$. We assume the regression function is $f^*(X) = \theta^* X$ for some $\theta^* \in \mathbb{R}^{d_y \times d_x}$, and estimate θ^* by $\hat{\theta}_n \in \arg \min_{\theta \in \mathbb{R}^{d_y \times d_x}} \frac{1}{n} \sum_{i=1}^n \|y^i - \theta x^i\|^2$. In the homoscedastic setting, the model class is $\mathcal{Q} := \{Q : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_y \times d_y} : Q \equiv I\}$ and we simply set $\hat{Q}_n := Q^* \equiv I$. In the heteroscedastic setting, following Romano and Wolf [66], the model class is a set of diagonal matrices $\mathcal{Q} := \{Q : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_y \times d_y} : Q(X) = \text{diag}(q_1(X), q_2(X), \dots, q_{d_y}(X))\}$, where $q_j^2(X) := \exp(\sum_{i=1}^{d_x} \pi_i^j \log|X_i|)$ for some $\pi^j \in \mathbb{R}^{d_x}$. We assume that $Q^* \in \mathcal{Q}$ with parameters $\{\pi^{j*}\}_{j=1}^{d_y}$ and $\mathbb{P}_X\{X = 0\} = 0$. For $j \in \{1, \dots, d_y\}$, we use $\mathbb{E}[(Y_j - f_j^*(X))^2 | X = x] = q_j^2(x)$ to estimate π^{j*} by

$$\hat{\pi}_n^j \in \arg \min_{\pi^j \in \mathbb{R}^{d_x}} \frac{1}{n} \sum_{i=1}^n \left(\log(\max\{\delta_n, y_j^i - (\hat{\theta}_n x^i)_j\}^2) - \sum_{k=1}^{d_x} \pi_k^j \log|X_k| \right)^2.$$

The tolerance sequence $\{\delta_n\} \downarrow 0$ is chosen to avoid ill-conditioning [66].

There is an inherent tradeoff between using parametric and nonparametric regression techniques for estimating the functions f^* and Q^* . If the function classes \mathcal{F} and \mathcal{Q} are correctly specified, then parametric regression approaches may yield much faster rates of convergence of the data-driven SAA estimators relative to nonparametric approaches (see Section 3.3). On the other hand, misspecification of the prediction model can result in our data-driven solutions being asymptotically inconsistent and suboptimal. Empirical evidence in Section 4 indicates that it *may* still be beneficial to use a misspecified prediction model when we do not have access to an abundance of data.

Remark 1. Although we assume that the functions $f^* \in \mathcal{F}$ and $Q^* \in \mathcal{Q}$ to establish our theoretical guarantees, our ER-SAA formulation (4) is well defined even when $f^* \notin \mathcal{F}$ and $Q^* \notin \mathcal{Q}$, i.e., when the regression models are misspecified. In this misspecified setting, the regression estimates satisfy $\hat{f}_n \xrightarrow{P} \bar{f} \in \mathcal{F}$ and $\hat{Q}_n \xrightarrow{P} \bar{Q} \in \mathcal{Q}$, where \xrightarrow{P} denotes convergence in probability, and \bar{f} and \bar{Q} are the best (in terms of

prediction error) approximations to f^* and Q^* in \mathcal{F} and \mathcal{Q} , respectively, under mild assumptions. The optimal value and solutions of the ER-SAA formulation then converge in probability to the optimal value and solutions of

$$\min_{z \in \mathcal{Z}} \frac{1}{n} \sum_{i=1}^n c(z, \text{proj}_{\mathcal{Y}}(\bar{f}(x) + \bar{Q}(x)\bar{\varepsilon}^i)),$$

under mild assumptions, where $\bar{\varepsilon}^i := [\bar{Q}(x^i)]^{-1}(f^*(x^i) - \bar{f}(x^i) + Q^*(x^i)\varepsilon^i)$. Although the ER-SAA estimators may no longer be consistent, their asymptotic and finite sample properties in this setting may be characterized by replacing f^* , Q^* , $\{\varepsilon^i\}$ by \bar{f} , \bar{Q} , and $\{\bar{\varepsilon}^i\}$ in our assumptions and results.

Notation. Let $[n] := \{1, \dots, n\}$, $|S|$ denote the cardinality of a finite set S , $\|\cdot\|$ denote the Euclidean norm or its induced matrix norm, $\|\cdot\|_0$ denote the ℓ_0 “norm”, $\mathcal{B}_\delta(v)$ denote a Euclidean ball of radius $\delta > 0$ around a point v , $M_{[j]}$ denote the j^{th} row of a matrix M and $M \succ 0$ denote that it is positive definite. For sets $A, B \subseteq \mathbb{R}^{d_z}$, let $\mathbb{D}(A, B) := \sup_{v \in A} \text{dist}(v, B)$ denote the deviation of A from B , where $\text{dist}(v, B) := \inf_{w \in B} \|v - w\|$. A random vector V is said to be sub-Gaussian with variance proxy σ^2 if $\mathbb{E}[V] = 0$ and $\mathbb{E}[\exp(su^T V)] \leq \exp(0.5\sigma^2 s^2)$, $\forall s \in \mathbb{R}$ and $\|u\| = 1$. The abbreviations ‘a.e.’, ‘LLN’, and ‘r.h.s.’ are shorthand for ‘almost everywhere’, ‘law of large numbers’, and ‘right-hand side’. By ‘a.e. X ’ and ‘a.e. Y ’, we mean P_X -a.e. $x \in \mathcal{X}$ and P_Y -a.e. Y . Throughout, ‘a.s.’ is written to mean almost surely with respect to the probability measure by which the data $\{(x^i, \varepsilon^i)\}$ is generated. The symbols \xrightarrow{p} , $\xrightarrow{a.s.}$, and \xrightarrow{d} are used to denote convergence in probability, almost surely, and in distribution with respect to this probability measure. For sequences of random variables $\{V_n\}$ and $\{W_n\}$, $V_n = o_p(W_n)$ and $V_n = O_p(W_n)$ convey that $V_n = R_n W_n$ with $\{R_n\}$ converging in probability to zero ($R_n \xrightarrow{p} 0$), or being bounded in probability, respectively (see Chapter 2 of van der Vaart [78] for basic theory). We write $O(1)$ to denote generic constants. We assume throughout this work that all functions, sets and selections are measurable (see van der Vaart and Wellner [79] and Shapiro et al. [73] for detailed consideration of these issues).

3 Analysis of the empirical residuals-based SAA

We first analyze the theoretical properties of solutions to the ER-SAA problem (4). In particular, after establishing some preliminary results that are useful for the remaining analysis (Section 3.1) we investigate conditions under which solutions to problem (4) are asymptotically optimal and consistent (Section 3.2) and develop finite sample guarantees for solutions to problem (4) using large deviations theory (Section 3.3). We outline the modifications required to analyze the J-SAA and J+-SAA methods in Section 3.4. Omitted proofs are provided in Appendix A. We also present a number of complementary results in the Appendix. In Section B of the Appendix, we briefly discuss alternative assumptions under which our theoretical guarantees hold. We analyze the rate of convergence of the optimal value of problem (4) to that of problem (1) in Section C. In Section D, we provide more details about the adaptation of our analysis to the J-SAA and J+-SAA methods. Finally, in Sections E and F, we verify that a variety of stochastic optimization and regression setups satisfy the assumptions made in our analysis.

3.1 Preliminary results

The difference between the objective functions of the ER-SAA problem (4) and the true problem (1) can be bounded uniformly over the decision variables $z \in \mathcal{Z}$ as follows:

$$\sup_{z \in \mathcal{Z}} |\hat{g}_n^{ER}(z; x) - g(z; x)| \leq \sup_{z \in \mathcal{Z}} |\hat{g}_n^{ER}(z; x) - g_n^*(z; x)| + \sup_{z \in \mathcal{Z}} |g_n^*(z; x) - g(z; x)|. \quad (9)$$

The second term on the r.h.s. of inequality (9) corresponds to the maximum deviation between the FI-SAA objective function (2) and the objective function of the true problem on \mathcal{Z} . This term converges to zero in probability whenever a uniform weak LLN result holds [73]. The first term on the r.h.s. of inequality (9) corresponds to the maximum deviation between the ER-SAA and FI-SAA objective functions on \mathcal{Z} . Let

$\tilde{\varepsilon}_n^i(x)$ denote the difference between the i th ER-SAA scenario ($\hat{f}_n(x) + \hat{Q}_n(x)\tilde{\varepsilon}_n^i$) and the i th FI-SAA scenario ($f^*(x) + Q^*(x)\varepsilon^i$), i.e.,

$$\tilde{\varepsilon}_n^i(x) := (\hat{f}_n(x) + \hat{Q}_n(x)\tilde{\varepsilon}_n^i) - (f^*(x) + Q^*(x)\varepsilon^i), \quad \forall i \in [n].$$

Then, the first term on the r.h.s. of inequality (9) can be bounded using the *deviation sequence* $\{\tilde{\varepsilon}_n^i(x)\}_{i=1}^n$ whenever the following Lipschitz assumption holds.

Assumption 1. For each $z \in \mathcal{Z}$, the function c in problem (1) satisfies the Lipschitz condition

$$|c(z, \bar{y}) - c(z, y)| \leq L(z)\|\bar{y} - y\|, \quad \forall y, \bar{y} \in \mathcal{Y},$$

with Lipschitz constant L satisfying $\sup_{z \in \mathcal{Z}} L(z) < +\infty$.

Assumption 1 requires the function $c(z, \cdot)$ to be Lipschitz continuous on \mathcal{Y} for each $z \in \mathcal{Z}$. We show in Section E in the Appendix that it is satisfied by Example 1. When Assumption 1 holds, the first term on the r.h.s. of inequality (9) can be bounded as follows.

Lemma 1. Suppose Assumption 1 holds. Then

$$\sup_{z \in \mathcal{Z}} |\hat{g}_n^{ER}(z; x) - g_n^*(z; x)| \leq \left(\sup_{z \in \mathcal{Z}} L(z) \right) \left(\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\| \right).$$

Therefore, when Assumption 1 holds, convergence of the mean deviation term $\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|$ to zero in probability readily translates to convergence to zero in probability of the first term on the r.h.s. of inequality (9). Next, we focus on bounding the mean deviation term $\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|$ using the arguments in Section 3.1 of Kannan et al. [51].

Lemma 2. Given regression estimates \hat{f}_n of f^* and \hat{Q}_n of Q^* with $[\hat{Q}_n(\bar{x})]^{-1} \succ 0$ for each $\bar{x} \in \mathcal{X}$:

1. In the homoscedastic setting (i.e., $\hat{Q}_n := Q^* \equiv I$), we have

$$\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\| \leq \|\hat{f}_n(x) - f^*(x)\| + \frac{1}{n} \sum_{i=1}^n \|\hat{f}_n(x^i) - f^*(x^i)\|.$$

2. In the heteroscedastic setting, we have

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\| &\leq \|\hat{f}_n(x) - f^*(x)\| + \|\hat{Q}_n(x) - Q^*(x)\| \left(\frac{1}{n} \sum_{i=1}^n \|\varepsilon^i\| \right) + \\ &\|\hat{Q}_n(x)\| \left(\frac{1}{n} \sum_{i=1}^n \left\| [\hat{Q}_n(x^i)]^{-1} - [Q^*(x^i)]^{-1} \right\|^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n \|Q^*(x^i)\|^4 \right)^{1/4} \left(\frac{1}{n} \sum_{i=1}^n \|\varepsilon^i\|^4 \right)^{1/4} + \\ &\|\hat{Q}_n(x)\| \left(\frac{1}{n} \sum_{i=1}^n \left\| [\hat{Q}_n(x^i)]^{-1} \right\|^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_n(x^i)\|^2 \right)^{1/2}. \end{aligned} \quad (10)$$

Proof. We begin by noting that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\| &:= \frac{1}{n} \sum_{i=1}^n \|(\hat{f}_n(x) + \hat{Q}_n(x)\tilde{\varepsilon}_n^i) - (f^*(x) + Q^*(x)\varepsilon^i)\| \\ &\leq \|\hat{f}_n(x) - f^*(x)\| + \frac{1}{n} \sum_{i=1}^n \|\hat{Q}_n(x)\tilde{\varepsilon}_n^i - Q^*(x)\varepsilon^i\|. \end{aligned} \quad (11)$$

In the homoscedastic case, we can bound the r.h.s. of inequality (11) further as

$$\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\| \leq \|\hat{f}_n(x) - f^*(x)\| + \frac{1}{n} \sum_{i=1}^n \|\hat{\varepsilon}_n^i - \varepsilon^i\| = \|\hat{f}_n(x) - f^*(x)\| + \frac{1}{n} \sum_{i=1}^n \|\hat{f}_n(x^i) - f^*(x^i)\|.$$

We now focus on the heteroscedastic setting by bounding the second term on the r.h.s. of (11).

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \|\hat{Q}_n(x) \hat{\varepsilon}_n^i - Q^*(x) \varepsilon^i\| \\ &= \frac{1}{n} \sum_{i=1}^n \|\hat{Q}_n(x) [\hat{Q}_n(x^i)]^{-1} (y^i - \hat{f}_n(x^i)) - Q^*(x) \varepsilon^i\| \\ &= \frac{1}{n} \sum_{i=1}^n \|\hat{Q}_n(x) [\hat{Q}_n(x^i)]^{-1} (f^*(x^i) - \hat{f}_n(x^i) + Q^*(x^i) \varepsilon^i) - Q^*(x) \varepsilon^i\| \\ &\leq \frac{1}{n} \sum_{i=1}^n \|\hat{Q}_n(x) [\hat{Q}_n(x^i)]^{-1} (f^*(x^i) - \hat{f}_n(x^i))\| + \frac{1}{n} \sum_{i=1}^n \|(\hat{Q}_n(x) [\hat{Q}_n(x^i)]^{-1} Q^*(x^i) - Q^*(x)) \varepsilon^i\|. \end{aligned} \quad (12)$$

We have for each $i \in [n]$

$$\begin{aligned} \hat{Q}_n(x) [\hat{Q}_n(x^i)]^{-1} Q^*(x^i) - Q^*(x) &= (\hat{Q}_n(x) [\hat{Q}_n(x^i)]^{-1} - Q^*(x) [Q^*(x^i)]^{-1}) Q^*(x^i) \\ &= \hat{Q}_n(x) ([\hat{Q}_n(x^i)]^{-1} - [Q^*(x^i)]^{-1}) Q^*(x^i) + [\hat{Q}_n(x) - Q^*(x)]. \end{aligned}$$

Plugging the above equality into inequality (12), we get

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \|\hat{Q}_n(x) \hat{\varepsilon}_n^i - Q^*(x) \varepsilon^i\| \\ &\leq \frac{1}{n} \sum_{i=1}^n \left\| \left(\hat{Q}_n(x) ([\hat{Q}_n(x^i)]^{-1} - [Q^*(x^i)]^{-1}) Q^*(x^i) + (\hat{Q}_n(x) - Q^*(x)) \right) \varepsilon^i \right\| + \\ & \quad \frac{1}{n} \sum_{i=1}^n \|\hat{Q}_n(x) [\hat{Q}_n(x^i)]^{-1} (f^*(x^i) - \hat{f}_n(x^i))\| \\ &\leq \frac{1}{n} \sum_{i=1}^n \|\hat{Q}_n(x) ([\hat{Q}_n(x^i)]^{-1} - [Q^*(x^i)]^{-1}) Q^*(x^i) \varepsilon^i\| + \frac{1}{n} \sum_{i=1}^n \|(\hat{Q}_n(x) - Q^*(x)) \varepsilon^i\| + \\ & \quad \frac{1}{n} \sum_{i=1}^n \|\hat{Q}_n(x) [\hat{Q}_n(x^i)]^{-1} (f^*(x^i) - \hat{f}_n(x^i))\| \\ &\leq \|\hat{Q}_n(x)\| \left(\frac{1}{n} \sum_{i=1}^n \|[\hat{Q}_n(x^i)]^{-1} - [Q^*(x^i)]^{-1}\|^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n \|Q^*(x^i)\|^4 \right)^{1/4} \left(\frac{1}{n} \sum_{i=1}^n \|\varepsilon^i\|^4 \right)^{1/4} + \\ & \quad \|\hat{Q}_n(x) - Q^*(x)\| \left(\frac{1}{n} \sum_{i=1}^n \|\varepsilon^i\| \right) + \|\hat{Q}_n(x)\| \left(\frac{1}{n} \sum_{i=1}^n \|[\hat{Q}_n(x^i)]^{-1}\|^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_n(x^i)\|^2 \right)^{1/2}, \end{aligned} \quad (13)$$

where the last step above follows by the Cauchy-Schwarz inequality. Finally, using inequality (13) in inequality (11), we get the stated result. \square

In the homoscedastic setting, the bound on the mean deviation term $\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|$ can be interpreted as the sum of the prediction error $\|\hat{f}_n(x) - f^*(x)\|$ at the new covariate realization $x \in \mathcal{X}$ and the average estimation error $\frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_n(x^i)\|$ at the training data points $\{x^i\}$.

We remark that Section B of the Appendix presents alternative assumptions under which the theoretical guarantees studied in the paper continue to hold. These alternative assumptions relax the uniform Lipschitz continuity of Assumption 1 to a weaker *local* Lipschitz continuity condition but require more stringent conditions on the regression step. Conditions and examples under which these alternative assumptions hold are discussed in the Appendix.

3.2 Consistency and asymptotic optimality

In this section, we investigate conditions under which the optimal value and optimal solutions to the ER-SAA problem (4) asymptotically converge to those of the true problem (1). We begin by making the following assumption on the uniform convergence of the sequence of objective functions of the FI-SAA problem (2) to the objective function of the true problem (1) on \mathcal{Z} .

Assumption 2. For a.e. $x \in \mathcal{X}$, the sequence of sample average functions $\{g_n^*(\cdot; x)\}$ defined in (2) converges in probability to the true function $g(\cdot; x)$ defined in (1) uniformly on the set \mathcal{Z} .

Assumption 2 is a uniform weak LLN result that is guaranteed to hold if $c(\cdot, y)$ is continuous for a.e. $y \in \mathcal{Y}$, $c(\cdot, y)$ is dominated by an integrable function for a.e. $y \in \mathcal{Y}$, and the errors $\{\varepsilon^i\}$ are i.i.d. [see Theorem 7.48 of 73]. Using pointwise LLN results in Walk [84] and White [85], we can show that Assumption 2 also holds for some mixing/stationary processes by noting the proof of Theorem 7.48 of Shapiro et al. [73] also extends to these settings. Proposition 20 in the Appendix shows that Assumption 2 holds for our running example (Example 1) of two-stage stochastic LP whenever $\mathbb{E}[\|Y\|] < +\infty$ and the errors $\{\varepsilon^i\}$ are i.i.d.

Next, we need the following weak LLN assumptions on the function Q^* and the errors ε . These assumptions hold, for instance, when the samples $\{(x^i, \varepsilon^i)\}$ are i.i.d. and the quantities $\mathbb{E}[\|Q^*(X)\|^4]$, $\mathbb{E}[\|[Q^*(X)]^{-1}\|^2]$, and $\mathbb{E}[\|\varepsilon\|^4]$ are finite. In particular, Assumption 3 holds for Example 2 if $\{x^i\}$ is i.i.d. and $\mathbb{E}[(\exp(\sum_{k=1}^{d_x} \pi_k^{j*} \log(|X_k|)))^2] < +\infty$ and $\mathbb{E}[\exp(-\sum_{k=1}^{d_x} \pi_k^{j*} \log(|X_k|))] < +\infty$ for each $j \in \{1, \dots, d_y\}$. Assumptions 3 and 4 also hold for non-i.i.d. data arising from mixing/stationary processes that satisfy suitable assumptions (see the discussion above).

Assumption 3. The function Q^* and the covariate samples $\{x^i\}_{i=1}^n$ satisfy the weak LLNs

$$\frac{1}{n} \sum_{i=1}^n \|Q^*(x^i)\|^4 \xrightarrow{P} \mathbb{E}[\|Q^*(X)\|^4] \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^n \|[Q^*(x^i)]^{-1}\|^2 \xrightarrow{P} \mathbb{E}[\|[Q^*(X)]^{-1}\|^2].$$

Assumption 4. The error samples $\{\varepsilon^i\}_{i=1}^n$ satisfy the weak LLN $\frac{1}{n} \sum_{i=1}^n \|\varepsilon^i\|^4 \xrightarrow{P} \mathbb{E}[\|\varepsilon\|^4]$.

Finally, we need the assumption below on the consistency of the regression estimates \hat{f}_n and \hat{Q}_n .

Assumption 5. The regression estimates \hat{f}_n and \hat{Q}_n possess the following consistency properties:

$$\begin{aligned} (5a) \quad \hat{f}_n(x) &\xrightarrow{P} f^*(x) \text{ for a.e. } x \in \mathcal{X}, & (5c) \quad \hat{Q}_n(x) &\xrightarrow{P} Q^*(x) \text{ for a.e. } x \in \mathcal{X}, \\ (5b) \quad \frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_n(x^i)\|^2 &\xrightarrow{P} 0, & (5d) \quad \frac{1}{n} \sum_{i=1}^n \|[\hat{Q}_n(x^i)]^{-1} - [Q^*(x^i)]^{-1}\|^2 &\xrightarrow{P} 0. \end{aligned}$$

Assumption (5a) holds for our running example of OLS regression (Example 2) if the parameter estimate $\hat{\theta}_n$ is weakly consistent (i.e., $\hat{\theta}_n \xrightarrow{P} \theta^*$), and Assumption (5b) holds if, in addition, the weak LLN $\frac{1}{n} \sum_{i=1}^n \|x^i\|^2 \xrightarrow{P} \mathbb{E}[\|X\|^2]$ is satisfied (see Chapter 3 of White [85] for various assumptions on the data \mathcal{D}_n and the distributions P_X and P_ε under which these conditions hold). The quantity $\frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_n(x^i)\|^2$ is called the empirical L^2 semi-norm in the empirical process theory literature [76]. Assumption (5c) holds for our running example of OLS regression with structured heteroscedasticity if the parameter estimate $\hat{\pi}_n$ is weakly consistent (i.e., $\hat{\pi}_n \xrightarrow{P} \pi^*$, see Appendix B.2 of Romano and Wolf [66] for assumptions under

which this holds). Assumption (5d) holds for our running example if, e.g., we additionally have the support \mathcal{X} to be compact and bounded away from the origin (i.e., for each $x \in \mathcal{X}$, $b \leq \|x\| \leq B$ for constants $b, B > 0$) and assume that the estimates $\{\hat{\pi}_n\}$ lie in a compact set a.s. for n large enough. Assumption 5 is implied by the stronger assumption of uniform convergence of the estimates \hat{f}_n and \hat{Q}_n to the functions f^* and Q^* , respectively, on the support \mathcal{X} of the covariates, i.e., when $\sup_{x \in \mathcal{X}} \|f^*(x) - \hat{f}_n(x)\| \xrightarrow{P} 0$, $\sup_{x \in \mathcal{X}} \|Q^*(x) - \hat{Q}_n(x)\| \xrightarrow{P} 0$, and $\sup_{x \in \mathcal{X}} \|[Q^*(x)]^{-1} - [\hat{Q}_n(x)]^{-1}\| \xrightarrow{P} 0$. Section F in the Appendix expands on the above arguments and shows that Assumption 5 also holds when f^* is estimated using Lasso, kNN, and RF regression under certain conditions.

The following result will prove useful in our analysis of ER-SAA in the heteroscedastic setting.

Lemma 3. We have

$$\left(\frac{1}{n} \sum_{i=1}^n \|[\hat{Q}_n(x^i)]^{-1}\|^2 \right)^{1/2} \leq \left(\frac{1}{n} \sum_{i=1}^n \|[\hat{Q}_n(x^i)]^{-1} - [Q^*(x^i)]^{-1}\|^2 \right)^{1/2} + \left(\frac{1}{n} \sum_{i=1}^n \|[Q^*(x^i)]^{-1}\|^2 \right)^{1/2}.$$

Proof. The triangle inequality for the operator norm implies

$$\|[\hat{Q}_n(x^i)]^{-1}\| \leq \|[\hat{Q}_n(x^i)]^{-1} - [Q^*(x^i)]^{-1}\| + \|[Q^*(x^i)]^{-1}\|, \quad \forall i \in [n].$$

Therefore, the following component-wise inequality holds:

$$0 \leq \begin{pmatrix} \|[\hat{Q}_n(x^1)]^{-1}\| \\ \vdots \\ \|[\hat{Q}_n(x^n)]^{-1}\| \end{pmatrix} \leq \begin{pmatrix} \|[\hat{Q}_n(x^1)]^{-1} - [Q^*(x^1)]^{-1}\| \\ \vdots \\ \|[\hat{Q}_n(x^n)]^{-1} - [Q^*(x^n)]^{-1}\| \end{pmatrix} + \begin{pmatrix} \|[Q^*(x^1)]^{-1}\| \\ \vdots \\ \|[Q^*(x^n)]^{-1}\| \end{pmatrix}.$$

The stated result then follows as a consequence of the triangle inequality for the ℓ_2 -norm. \square

Our next result uses Lemmas 1 and 2 to establish conditions under which the sequence of objective functions of the ER-SAA problem (4) converges uniformly to the objective function of the true problem (1) on the feasible region \mathcal{Z} .

Proposition 4. Suppose Assumptions 1 through 5 hold. Then, for a.e. $x \in \mathcal{X}$, the sequence of objective functions of the ER-SAA problem (4) converges in probability to the objective function of the true problem (1) uniformly on the feasible region \mathcal{Z} .

Proof. We wish to show that $\sup_{z \in \mathcal{Z}} |\hat{g}_n^{ER}(z; x) - g(z; x)| \xrightarrow{P} 0$ for a.e. $x \in \mathcal{X}$. Equation (9) yields

$$|\hat{g}_n^{ER}(z; x) - g(z; x)| \leq \sup_{z \in \mathcal{Z}} |\hat{g}_n^{ER}(z; x) - g_n^*(z; x)| + \sup_{z \in \mathcal{Z}} |g_n^*(z; x) - g(z; x)|.$$

The second term on the r.h.s. of the above inequality vanishes in the limit in probability under Assumption 2. If the first term also vanishes in the limit in probability, by $o_p(1) + o_p(1) = o_p(1)$, we obtain the desired result. We now show that the first term vanishes in the limit in probability.

Since Assumption 1 holds, Lemma 1 implies the first term converges to zero in probability whenever the mean deviation term $\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\| \xrightarrow{P} 0$. Therefore, the desired result holds if each term on the r.h.s. of inequality (10) converges to zero in probability. The first term on the r.h.s. of (10) converges to zero in probability by Assumption (5a), and the second term converges to zero in probability by Assumptions 4 and (5c). The third term on the r.h.s. of (10) converges to zero in probability by Assumptions 3, 4, (5c), and (5d). Finally, the last term on the r.h.s. of inequality (10) converges to zero in probability by Assumptions 3, (5b), (5c), and (5d) and Lemma 3. \square

Fewer assumptions are needed to establish Proposition 4 in the homoscedastic case. In that setting, Assumptions 3, 4, (5c), and (5d) are not required and Assumption (5b) may be weakened to the assumption $\frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_n(x^i)\| \xrightarrow{P} 0$ on account of Lemma 2. Proposition 4 provides the foundation for the following result, which demonstrates that the optimal value and solutions of the ER-SAA problem (4) converge to those of the true problem (1).

Theorem 5. Suppose Assumptions 1 to 5 hold. Then, we have $\hat{v}_n^{ER}(x) \xrightarrow{P} v^*(x)$, $\mathbb{D}(\hat{S}_n^{ER}(x), S^*(x)) \xrightarrow{P} 0$, and $\sup_{z \in \hat{S}_n^{ER}(x)} g(z; x) \xrightarrow{P} v^*(x)$ for a.e. $x \in \mathcal{X}$.

The proof of Theorem 5 follows a similar outline as the proof of Theorem 5.3 of Shapiro et al. [73], except that we consider convergence in probability rather than almost sure convergence. Under an *inf-compactness* condition on the ER-SAA problem (4), the conclusions of Theorem 5 hold even if the set \mathcal{Z} is unbounded [see the discussion following Theorem 5.3 of 73]. While we consider convergence in probability instead of almost sure convergence (because the statistics literature is typically concerned with conditions under which Assumption 5 holds rather than its almost sure counterpart), note that our results until this point can be naturally extended to the latter setting by suitably strengthening Assumptions 2 to 5.

Next, we identify conditions under which the optimal value of the ER-SAA problem (4) converges to the optimal value of the true problem (1) on average over the space of the covariates. This can be important in situations where the new covariate observations are random and the DM needs to make decisions facing different observations of the covariates. Given $q \in [1, +\infty]$, we write $\|F\|_{L^q}$ to denote the L^q -norm of a measurable function $F : \mathcal{X} \rightarrow \mathbb{R}^{d_F}$, i.e., $\|F\|_{L^q} := (\int_S \|F\|^q dP_X)^{1/q}$. We require the following adaptation of Assumptions 2 and 5.

Assumption 6. The sequence of sample average functions $\{g_n^*(\cdot; x)\}$ defined in (2) satisfies

$$\left\| \sup_{z \in \mathcal{Z}} |g_n^*(z; X) - g(z; X)| \right\|_{L^q} \xrightarrow{P} 0.$$

Assumption 7. The regression estimates \hat{f}_n and \hat{Q}_n possess the following consistency properties:

$$\begin{aligned} (7a) \quad & \|\hat{f}_n(X) - f^*(X)\|_{L^q} \xrightarrow{P} 0, & (7c) \quad & \|\hat{Q}_n(X) - Q^*(X)\|_{L^q} \xrightarrow{P} 0, \\ (7b) \quad & \frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_n(x^i)\|^2 \xrightarrow{P} 0, & (7d) \quad & \frac{1}{n} \sum_{i=1}^n \|[\hat{Q}_n(x^i)]^{-1} - [Q^*(x^i)]^{-1}\|^2 \xrightarrow{P} 0. \end{aligned}$$

Assumption 6 is implied by the uniform convergence in probability of the FI-SAA objective function g_n^* to g on $\mathcal{Z} \times \mathcal{X}$, i.e., $\sup_{(z,x) \in \mathcal{Z} \times \mathcal{X}} |g_n^*(z; x) - g(z; x)| \xrightarrow{P} 0$. We show in Section E that this assumption holds for our running Examples 1 and 2 whenever $\mathbb{E}[\|\varepsilon\|] < \infty$ and the support \mathcal{X} is compact and bounded away from the origin. Assumptions (7a) and (7c) hold for our running Example 2 if the estimates $\hat{\theta}_n$ and $\hat{\pi}_n$ are weakly consistent, $\|X\|_{L^q} < +\infty$, and if, e.g., we additionally have the support \mathcal{X} to be compact and bounded away from the origin and assume that the estimates $\{\hat{\pi}_n\}$ lie in a compact set a.s. for n large enough (see Section F in the Appendix for details). Unlike Assumptions (5a) and (5c) that require \hat{f}_n and \hat{Q}_n to be pointwise consistent, Assumptions (7a) and (7c) only require \hat{f}_n and \hat{Q}_n to be consistent on average over the covariates X . We have the following result.

Theorem 6. Suppose Assumptions 1, 3, 4, 6, and 7 hold and $\|Q^*(X)\|_{L^q} < +\infty$ for some constant $q \in [1, +\infty]$. Then, $\|\hat{v}_n^{ER}(X) - v^*(X)\|_{L^q} \xrightarrow{P} 0$ and $\|g(\hat{z}_n^{ER}(X); X) - v^*(X)\|_{L^q} \xrightarrow{P} 0$.

3.3 Finite sample guarantees

We now establish a lower bound on the probability that solutions to the ER-SAA problem (4) are nearly optimal to the true problem (1). Section C in the Appendix investigates the rate of convergence of ER-SAA estimators under weaker assumptions. Our next assumption is motivated by the analysis in Section 2 of Homem-de-Mello [42] and Section 7.2.9 of Shapiro et al. [73].

Assumption 8. The full-information SAA problem (2) possesses the following uniform exponential bound property: for any constant $\kappa > 0$ and a.e. $x \in \mathcal{X}$, there exist positive constants $K(\kappa, x)$ and $\beta(\kappa, x)$ such that $\mathbb{P}\left\{ \sup_{z \in \mathcal{Z}} |g_n^*(z; x) - g(z; x)| > \kappa \right\} \leq K(\kappa, x) \exp(-n\beta(\kappa, x))$, $\forall n \in \mathbb{N}$.

Lemma 2.4 of Homem-de-Mello [42] provides conditions under which Assumption 8 holds [also see Section 7.2.9 of 73]. In particular, Homem-de-Mello [42] shows that Assumption 8 holds whenever the function $c(\cdot, y)$ is Lipschitz continuous on \mathcal{Z} for a.e. $y \in \mathcal{Y}$ with an integrable Lipschitz constant and some pointwise exponential bounds hold. When the errors $\{\varepsilon^i\}$ are i.i.d., Section 7.2.9 of Shapiro et al. [73] presents conditions under which these pointwise exponential bound conditions are satisfied via Cramér's large deviation theorem. Bryc and Dembo [18] present mixing conditions on the errors $\{\varepsilon^i\}$ under which these assumptions are also satisfied (also see the references therein). The Gärtner-Ellis Theorem [see Section 2.3 of 28] provides an alternative avenue for verifying Assumption 8 for non-i.i.d. errors $\{\varepsilon^i\}$ [24]. If we also assume that the random variable $c(z, f^*(x) + Q^*(x)\varepsilon) - \mathbb{E}[c(z, f^*(x) + Q^*(x)\varepsilon)]$ is sub-Gaussian for each $z \in \mathcal{Z}$ and a.e. $x \in \mathcal{X}$, then we can characterize the dependence of $\beta(\kappa, x)$ on κ , see Assumption (C4) on page 396 and Theorem 7.67 of Shapiro et al. [73].

Proposition 20 in the Appendix shows that Assumption 8 holds for Example 1 whenever the errors $\{\varepsilon^i\}$ are i.i.d. and sub-Gaussian (which includes zero-mean Gaussian). Unlike Assumption 2, Assumption 8 may not hold when the distribution of the errors ε is heavy-tailed (heavy-tailed error distributions such as the Pareto and Weibull distributions occur in finance, weather forecasting, and reliability engineering applications).

Next, we require the following strengthening of Assumptions 3 and 4.

Assumption 9. For any constant $\kappa > 0$ and $n \in \mathbb{N}$, there exist positive constants $J_Q(\kappa)$, $\gamma_Q(n, \kappa)$, $\bar{J}_Q(\kappa)$, and $\bar{\gamma}_Q(n, \kappa)$, with $\lim_{n \rightarrow \infty} \gamma_Q(n, \kappa) = \infty$ and $\lim_{n \rightarrow \infty} \bar{\gamma}_Q(n, \kappa) = \infty$ for each $\kappa > 0$, such that

$$\begin{aligned} \mathbb{P}\left\{\left(\frac{1}{n} \sum_{i=1}^n \|[Q^*(x^i)]^{-1}\|^2\right)^{1/2} > \left(\mathbb{E}\left[\|[Q^*(X)]^{-1}\|^2\right]\right)^{1/2} + \kappa\right\} &\leq J_Q(\kappa) \exp(-\gamma_Q(n, \kappa)), \\ \mathbb{P}\left\{\left(\frac{1}{n} \sum_{i=1}^n \|Q^*(x^i)\|^4\right)^{1/4} > \left(\mathbb{E}\left[\|Q^*(X)\|^4\right]\right)^{1/4} + \kappa\right\} &\leq \bar{J}_Q(\kappa) \exp(-\bar{\gamma}_Q(n, \kappa)). \end{aligned}$$

Assumption 10. For any constant $\kappa > 0$ and $n \in \mathbb{N}$, there exist positive constants $J_\varepsilon(\kappa)$, $\gamma_\varepsilon(n, \kappa)$, $\bar{J}_\varepsilon(\kappa)$, and $\bar{\gamma}_\varepsilon(n, \kappa)$, with $\lim_{n \rightarrow \infty} \gamma_\varepsilon(n, \kappa) = \infty$ and $\lim_{n \rightarrow \infty} \bar{\gamma}_\varepsilon(n, \kappa) = \infty$ for each $\kappa > 0$, such that

$$\begin{aligned} \mathbb{P}\left\{\frac{1}{n} \sum_{i=1}^n \|\varepsilon^i\| > \mathbb{E}\left[\|\varepsilon\|\right] + \kappa\right\} &\leq J_\varepsilon(\kappa) \exp(-\gamma_\varepsilon(n, \kappa)), \\ \mathbb{P}\left\{\left(\frac{1}{n} \sum_{i=1}^n \|\varepsilon^i\|^4\right)^{1/4} > \left(\mathbb{E}\left[\|\varepsilon\|^4\right]\right)^{1/4} + \kappa\right\} &\leq \bar{J}_\varepsilon(\kappa) \exp(-\bar{\gamma}_\varepsilon(n, \kappa)). \end{aligned}$$

The first part of Assumption 9 holds, e.g., if for each $\kappa > 0$ and $n \in \mathbb{N}$, constants $J_Q(\kappa) > 0$ and $\gamma_Q(n, \kappa) > 0$ exist such that

$$\mathbb{P}\left\{\frac{1}{n} \sum_{i=1}^n \|[Q^*(x^i)]^{-1}\|^2 > \mathbb{E}\left[\|[Q^*(X)]^{-1}\|^2\right] + \kappa^2\right\} \leq J_Q(\kappa) \exp(-\gamma_Q(n, \kappa)).$$

The function γ_Q in the inequality above is related to the so-called rate function in large deviations theory (see Dembo and Zeitouni [28] and Section 7.2.8 of [73]). Similar conclusions hold for the probability inequalities involving the terms $\frac{1}{n} \sum_{i=1}^n \|Q^*(x^i)\|^4$ and $\frac{1}{n} \sum_{i=1}^n \|\varepsilon^i\|^4$ in Assumptions 9 and 10. Using large deviations theory, we can show that the constants $\gamma_Q(n, \kappa)$ and $\bar{\gamma}_Q(n, \kappa)$ in Assumption 9 increase linearly with the sample size n (i.e., $\gamma_Q(n, \kappa) = n\gamma_{Q,1}(\kappa)$ and $\bar{\gamma}_Q(n, \kappa) = n\bar{\gamma}_{Q,1}(\kappa)$) for our running Example 2 with i.i.d. data $\{(x^i, \varepsilon^i)\}$ whenever the support \mathcal{X} of the covariates is compact and bounded away from the origin. The constant $\gamma_\varepsilon(n, \kappa)$ in Assumption 10 also increases linearly with n whenever the errors ε are sub-Gaussian (see Chapter 3 of Vershynin [81]). The discussion following Assumption 8 provides avenues for verifying Assumptions 9 and 10 for non-i.i.d. data $\{(x^i, \varepsilon^i)\}$.

Next, we make the following large deviation assumption on the regression procedure (3) that is similar in spirit to Assumption 8.

Assumption 11. The regression estimates \hat{f}_n and \hat{Q}_n possess the following finite sample properties: for any constant $\kappa > 0$ and $n \in \mathbb{N}$, there exist positive constants $K_f(\kappa, x)$, $\bar{K}_f(\kappa)$, $\beta_f(n, \kappa, x)$, $\bar{\beta}_f(n, \kappa)$, $K_Q(\kappa, x)$, $\bar{K}_Q(\kappa)$, $\beta_Q(n, \kappa, x)$, and $\bar{\beta}_Q(n, \kappa)$, with $\lim_{n \rightarrow \infty} \beta_f(n, \kappa, x) = \infty$, $\lim_{n \rightarrow \infty} \bar{\beta}_f(n, \kappa) = \infty$, $\lim_{n \rightarrow \infty} \beta_Q(n, \kappa, x) = \infty$, and $\lim_{n \rightarrow \infty} \bar{\beta}_Q(n, \kappa) = \infty$ for each $\kappa > 0$ and a.e. $x \in \mathcal{X}$, such that

$$(11a) \quad \mathbb{P}\{\|f^*(x) - \hat{f}_n(x)\| > \kappa\} \leq K_f(\kappa, x) \exp(-\beta_f(n, \kappa, x)) \text{ for a.e. } x \in \mathcal{X},$$

$$(11b) \quad \mathbb{P}\left\{\frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_n(x^i)\|^2 > \kappa^2\right\} \leq \bar{K}_f(\kappa) \exp(-\bar{\beta}_f(n, \kappa)),$$

$$(11c) \quad \mathbb{P}\{\|Q^*(x) - \hat{Q}_n(x)\| > \kappa\} \leq K_Q(\kappa, x) \exp(-\beta_Q(n, \kappa, x)) \text{ for a.e. } x \in \mathcal{X},$$

$$(11d) \quad \mathbb{P}\left\{\frac{1}{n} \sum_{i=1}^n \|[\hat{Q}_n(x^i)]^{-1} - [Q^*(x^i)]^{-1}\|^2 > \kappa^2\right\} \leq \bar{K}_Q(\kappa) \exp(-\bar{\beta}_Q(n, \kappa)).$$

We verify in Section F of the Appendix that Assumptions (11a) and (11b) hold for OLS regression and the Lasso with constants $\beta_f(n, \kappa, x)$ and $\bar{\beta}_f(n, \kappa)$ scaling as $O(n\kappa^2)$ under sub-Gaussian assumptions on the errors ε . The finite sample guarantees on the estimate \hat{Q}_n in Assumptions (11c) and (11d) are typically harder to verify.

The next result provides conditions under which the maximum deviation of the ER-SAA objective from the full-information SAA objective on the feasible set \mathcal{Z} satisfies a qualitatively similar large deviations bound as that in Assumption 8.

Lemma 7. Suppose Assumptions 1, 8 to 11 hold. Then for any constant $\kappa > 0$, $n \in \mathbb{N}$, and a.e. $x \in \mathcal{X}$, there exist positive constants $\bar{K}(\kappa, x)$ and $\bar{\beta}(n, \kappa, x)$, with $\lim_{n \rightarrow \infty} \bar{\beta}(n, \kappa, x) = \infty$ for each $\kappa > 0$ and a.e. $x \in \mathcal{X}$, such that $\mathbb{P}\left\{\sup_{z \in \mathcal{Z}} |\hat{g}_n^{ER}(z; x) - g_n^*(z; x)| > \kappa\right\} \leq \bar{K}(\kappa, x) \exp(-\bar{\beta}(n, \kappa, x))$.

We are now ready to present the main result of this section. It extends finite sample results that are known for traditional SAA estimators, see, e.g., Theorem 2.3 of Homem-de-Mello [42] and Section 5.3 of Shapiro et al. [73], to the ER-SAA setting.

Theorem 8. Suppose Assumptions 1, 8 to 11 hold. Then, for each $n \in \mathbb{N}$ and a.e. $x \in \mathcal{X}$, given $\eta > 0$, there exist constants $Q(\eta, x) > 0$ and $\gamma(n, \eta, x) > 0$, with $\lim_{n \rightarrow \infty} \gamma(n, \eta, x) = \infty$ for each $\eta > 0$ and a.e. $x \in \mathcal{X}$, such that $\mathbb{P}\{\text{dist}(\hat{z}_n^{ER}(x), S^*(x)) \geq \eta\} \leq Q(\eta, x) \exp(-\gamma(n, \eta, x))$.

Assumptions 9, 10, (11c), and (11d) are not required to establish Lemma 7 and Theorem 8 in the homoscedastic case ($\hat{Q}_n := Q^* \equiv I$). Additionally, Assumption (11b) can be weakened in this setting to $\mathbb{P}\left\{\frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_n(x^i)\| > \kappa\right\} \leq \bar{K}_f(\kappa) \exp(-\bar{\beta}_f(n, \kappa))$ on account of Lemma 2.

To give an example of how Theorem 8 can be used to provide a qualitative estimate of sample size required to obtain a desired accuracy with high probability, we now specialize the results in this section to the homoscedastic setting when the ER-SAA formulation is applied to two-stage stochastic LP with OLS, Lasso, or kNN regression for estimating f^* . Given $\kappa > 0$, let $S^\kappa(x) := \{z \in \mathcal{Z} : g(z; x) \leq v^*(x) + \kappa\}$ denote the set of κ -optimal solutions to the true problem (1). Given an unreliability level $\delta \in (0, 1)$, our goal is to estimate the sample size n required for every solution to the ER-SAA problem (4) to be κ -optimal to the true problem (1) with probability at least $1 - \delta$, i.e., to estimate n such that $\mathbb{P}\{\hat{S}_n^{ER}(x) \subseteq S^\kappa(x)\} \geq 1 - \delta$. In the following, we make stronger than necessary assumptions on the regression setups for readability.

Our sample size estimate for ER-SAA proceeds by estimating the sample size required for the full-information SAA problem (2) to be ‘close to’ the true problem (1) and for the ER-SAA problem (4) to be ‘close to’ the FI-SAA problem (2); see also (9). From Section 5.3 of Shapiro et al. [73], we have the following sample size estimate for every solution to the FI-SAA problem (2) to be κ -optimal to the true problem (1) with probability at least $1 - \delta$:

$$n \geq n^* := \frac{O(1)\sigma_c^2(x)}{\kappa^2} \left[d_z \log \left(\frac{O(1)D}{\kappa} \right) + \log \left(\frac{O(1)}{\delta} \right) \right].$$

Learning of the regression function f^* introduces additional terms in the estimate for ER-SAA that depend on the dimensions d_y and d_x of the random vector Y and the random covariates X .

Proposition 9. Consider Example 1 and assume $\hat{Q}_n = Q^* \equiv I$. Suppose \mathcal{Z} is compact with diameter D and for each $z \in \mathcal{Z}$ and a.e. $x \in \mathcal{X}$, the random variable $c(z, f^*(x) + \varepsilon) - \mathbb{E}[c(z, f^*(x) + \varepsilon)]$ is sub-Gaussian with variance proxy $\sigma_c^2(x)$. Let $\{\varepsilon^i\}_{i \in [n]}$ be i.i.d. sub-Gaussian random vectors with variance proxy σ^2 , $\kappa > 0$ be the target optimality gap, and $\delta \in (0, 1)$ be the desired unreliability level.

1. Suppose the regression function f^* is linear, the regression step (3) is OLS regression, the covariance matrix Σ_X of the covariates is positive definite, and the random vector $\Sigma_X^{-\frac{1}{2}}X$ is sub-Gaussian. Then, we have $\mathbb{P}\{\hat{S}_n^{ER}(x) \subseteq S^\kappa(x)\} \geq 1 - \delta$ for sample size n satisfying

$$n \geq n^* + \frac{O(1)\sigma^2 d_y}{\kappa^2} \left[\log\left(\frac{O(1)}{\delta}\right) + d_x \right].$$

2. Suppose the regression function f^* is linear with $\|\theta_{[j]}^*\|_0 \leq s$, $\forall j \in [d_y]$, the regression step (3) is Lasso regression, the support \mathcal{X} of the covariates X is compact, $\mathbb{E}[|X_j|^2] > 0$, $\forall j \in [d_x]$, and the matrix $\mathbb{E}[XX^T] - \tau \text{diag}(\mathbb{E}[XX^T])$ is positive semidefinite for some constant $\tau \in (0, 1]$. Then, we have $\mathbb{P}\{\hat{S}_n^{ER}(x) \subseteq S^\kappa(x)\} \geq 1 - \delta$ for sample size n satisfying

$$n \geq n^* + \frac{O(1)\sigma^2 s d_y}{\kappa^2} \left[\log\left(\frac{O(1)}{\delta}\right) + \log(d_x) \right].$$

3. Suppose the regression function f^* is Lipschitz continuous, the regression step (3) is kNN regression with parameter $k = \lceil O(1)n^\gamma \rceil$ for some constant $\gamma \in (0, 1)$, the support \mathcal{X} of the covariates X is compact, and there exists a constant $\tau > 0$ such that $\mathbb{P}\{X \in \mathcal{B}_\kappa(x)\} \geq \tau \kappa^{d_x}$, $\forall x \in \mathcal{X}$ and $\kappa > 0$.

Then, we have $\mathbb{P}\{\hat{S}_n^{ER}(x) \subseteq S^\kappa(x)\} \geq 1 - \delta$ for sample size n satisfying $n \geq O(1) \left(\frac{O(1)}{\kappa}\right)^{\frac{d_x}{1-\gamma}}$,

$$\frac{n^\gamma}{\log(n)} \geq \frac{O(1)d_x d_y \sigma^2}{\kappa^2}, \text{ and}$$

$$n \geq n^* + \left(\frac{O(1)\sigma^2 d_y}{\kappa^2}\right)^{\frac{1}{\gamma}} \left[d_x \log\left(\frac{O(1)}{d_x}\right) + \log\left(\frac{O(1)}{\delta}\right) \right]^{\frac{1}{\gamma}} + \left(\frac{O(1)d_y}{\kappa^2}\right)^{d_x} \left[\frac{d_x}{2} \log\left(\frac{O(1)d_x d_y}{\kappa^2}\right) + \log\left(\frac{O(1)}{\delta}\right) \right].$$

Proposition 9 illustrates the tradeoff between using parametric and nonparametric regression approaches within the ER-SAA framework. The sample size estimates in Proposition 9 involve the sum of two contributions, the FI-SAA contribution n^* and additional regression-related terms introduced by the use of estimates of f^* within the ER-SAA. Assuming that the regression function f^* satisfies the necessary structural properties, using OLS regression or the Lasso for the regression step (3) can yield sample size estimates that depend modestly on the accuracy κ and the dimensions d_x and d_y compared to kNN regression. On the other hand, unlike OLS and Lasso regression, the sample size estimates for kNN regression are valid under mild assumptions on the regression function f^* . Nevertheless, we empirically demonstrate in Section 4 that it may be beneficial to use a structured but misspecified prediction model when we do not have an abundance of data. Note that the OLS estimate includes a term that depends linearly on the dimension d_x of the covariates, whereas the corresponding term in the Lasso estimate only depends logarithmically on d_x .

3.4 Outline of analysis for the jackknife-based estimators

The results thus far carry over to the J-SAA and J+-SAA estimators if the assumptions that ensure that the ER-SAA mean deviation term $\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\| \xrightarrow{P} 0$ at a certain rate are adapted to ensure that the J-SAA

and J+-SAA mean deviation terms $\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^{i,J}(x)\|$ and $\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^{i,J^+}(x)\|$ converge to zero in probability at a certain rate, where

$$\begin{aligned}\tilde{\varepsilon}_n^{i,J}(x) &:= \left(\hat{f}_n(x) + \hat{Q}_n(x) \hat{\varepsilon}_{n,J}^i \right) - \left(f^*(x) + Q^*(x) \varepsilon^i \right), \quad \forall i \in [n], \\ \tilde{\varepsilon}_n^{i,J^+}(x) &:= \left(\hat{f}_{-i}(x) + \hat{Q}_{-i}(x) \hat{\varepsilon}_{n,J}^i \right) - \left(f^*(x) + Q^*(x) \varepsilon^i \right), \quad \forall i \in [n].\end{aligned}$$

Lemma 14 in Section D of the Appendix presents the analogue of Lemma 2 for the jackknife-based mean deviation terms. It also provides guidance for how the assumptions on the quantities appearing in inequality (10) could be replaced with assumptions on the quantities appearing in the bounds of the jackknife-based mean deviation terms to derive similar results for the J-SAA and J+-SAA estimators as the ER-SAA estimator. Since the formal statements of the assumptions and results for the J-SAA and J+-SAA estimators closely mirror those for the ER-SAA given in Sections 3.2 and 3.3, we present these details in Section D of the Appendix.

4 Computational experiments

We consider instances of the following resource allocation model adapted from Luedtke [56]:

$$\min_{z \in \mathbb{R}_+^{|\mathcal{I}|}} c_z^\top z + \mathbb{E}[Q(z, Y)],$$

where the second-stage function is defined as

$$Q(z, y) := \min_{v \in \mathbb{R}_+^{|\mathcal{I}| \times |\mathcal{J}|}, w \in \mathbb{R}_+^{|\mathcal{J}|}} \left\{ q_w^\top w : \sum_{j \in \mathcal{J}} v_{ij} \leq \rho_i z_i, \quad \forall i \in \mathcal{I}, \quad \sum_{i \in \mathcal{I}} \mu_{ij} v_{ij} + w_j \geq y_j, \quad \forall j \in \mathcal{J} \right\}.$$

The first-stage variables z_i denote the order quantities of resources $i \in \mathcal{I}$, and the second-stage variables v_{ij} and w_j denote the amount of resource $i \in \mathcal{I}$ allocated to customer type $j \in \mathcal{J}$ and the unmet demand of customer type j , respectively. We consider instances with $|\mathcal{I}| = 20$ and $|\mathcal{J}| = 30$. The yield and service rate parameters ρ and μ and the cost coefficients c_z and q_w are assumed to be deterministic. Parameters c_z , ρ , and μ are set using the procedure described in Luedtke [56], and the coefficients q_w are determined by $q_w := \tau \|c_z\|_\infty$, where each component of the vector τ is drawn independently from a lognormal $\mathcal{LN}(0.5, 0.05)$ distribution.

The demands y_j , $j \in \mathcal{J}$, of the customer types are considered to be stochastic with $\mathcal{Y} = \mathbb{R}_+^{|\mathcal{J}|}$. We assume that some of the variability in the demands can be explained with knowledge of covariates X_l , $l \in \mathcal{L}$, where $|\mathcal{L}| = d_x$. The demands Y are assumed to be related to the covariates through

$$Y_j = \varphi_j^* + \sum_{l \in \mathcal{L}^*} \zeta_{jl}^* (X_l)^p + q_j^*(X) \varepsilon_j, \quad \forall j \in \mathcal{J},$$

where $p \in \{0.5, 1, 2\}$ is a fixed parameter that determines the model class, $q_j^* : \mathcal{X} \rightarrow \mathbb{R}_+$ is defined as $q_j^*(X) = (s_j \exp(\sum_{l \in \mathcal{L}^*} \pi_{jl}^* \log|1 + X_l|))^{1/2}$ for parameters $\{\pi_{jl}^*\}_{l \in \mathcal{L}^*}$ [66], errors $\varepsilon_j \sim \mathcal{N}(0, \sigma_j^2)$ are independent of X , φ^* , ζ^* and σ_j are additional model parameters, and $\mathcal{L}^* \subseteq \mathcal{L}$ contains the indices of a subset of covariates with predictive power (note that \mathcal{L}^* does not depend on $j \in \mathcal{J}$). In the definition of $q_j^*(X)$, the scaling factor s_j is a numerical approximation of the median of $\exp(\sum_{l \in \mathcal{L}^*} \pi_{jl}^* \log|1 + X_l|)$ so that $\mathbb{P}_X(q_j^*(X) > 1) \approx 0.5$. The form of the heteroscedasticity functions q_j^* is chosen to simulate increasing error variance with increasing magnitude of the covariates [66]. Throughout, we assume $|\mathcal{L}^*| = 3$, i.e., the demands truly depend only on three covariates. We simulate i.i.d. data $\{(x^i, \varepsilon^i)\}$ with φ^* and ζ^* randomly generated, $\sigma_j = \sigma = 5$, $\forall j \in \mathcal{J}$, unless otherwise specified, and draw covariate samples $\{x^i\}_{i=1}^n$ from a multivariate folded normal distribution (see Section G in the Appendix for details). We vary the heteroscedasticity level $\omega \in \{1, 2, 3\}$, where $\omega = 1, 2$, and 3 correspond to zero, moderate, and severe heteroscedasticity, respectively, and sample $\{\pi_{jl}^*\}_{l \in \mathcal{L}^*}$

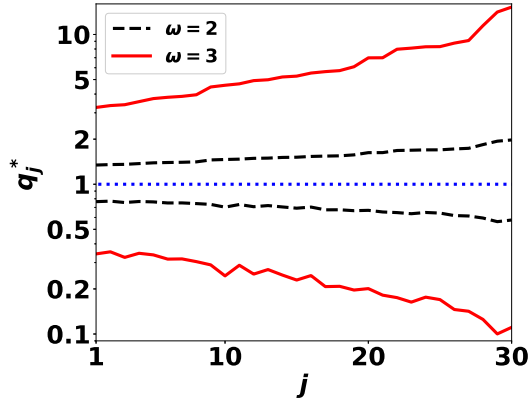


Figure 1: Median (dotted line) and the 10th and 90th quantiles (dashed lines for $\omega = 2$ [moderate heteroscedasticity] and solid lines for $\omega = 3$ [severe heteroscedasticity]) of the heteroscedasticity function q_j^* over 1100 independent realizations of the covariates. The y -axis is in logarithmic scale and the indices j are reordered to generate approximately monotonic plots.

independently from the uniform distribution $U(0, 2(\omega - 1)^2)$. Figure 1 plots statistics of the functions q_j^* for moderate and severe heteroscedasticity.

Given data \mathcal{D}_n on the demands and covariates, we estimate the coefficients of the linear model

$$Y_j = \varphi_j + \sum_{l \in \mathcal{L}} \zeta_{jl} X_l + \eta_j, \quad \forall j \in \mathcal{J},$$

where η_j are zero-mean errors, using OLS or Lasso regression and use this prediction model within the ER-SAA, J-SAA, and J+-SAA frameworks. We use this linear prediction model even when the degree $p \neq 1$, in which case the prediction model is misspecified. We also evaluate the performance of the ER-SAA approach when kNN regression is used to predict Y from X (the parameter k in kNN regression is chosen from $[\lfloor n^{0.1} \rfloor, \lceil n^{0.9} \rceil]$ to minimize the 5-fold CV test error). We compare these ER-SAA and J-SAA estimators with the point prediction-based (PP) deterministic approximation (6) that uses OLS or Lasso regression (assuming a linear model). While we do not estimate the heteroscedasticity functions $\{q_j^*\}$ by default in our experiments (equivalent to setting $\hat{Q}_n \equiv I$), we also compare the performance of ER-SAA formulations that ignore heteroscedasticity with heteroscedasticity-aware ER-SAA formulations where each q_j^* is estimated using Lasso regression (cf. Example 2). When the function f^* is estimated using OLS regression, we use the estimate of Q^* to update the estimate \hat{f}_n of f^* using weighted least squares [66]. While estimating Q^* , we assume that its parametric form is known; model misspecification for f^* also simulates model misspecification for heteroscedasticity in our two-step approach for estimating Q^* .

We compare our data-driven SAA estimators with the kNN-based reweighted SAA (kNN-SAA) approach of Bertsimas and Kallus [12] on a few test instances by varying the dimensions of the covariates d_x , the sample size n , the degree p , the standard deviation σ of the errors ε , and the degree of heteroscedasticity ω . While our case studies illustrate the potential advantages of employing parametric regression models (such as OLS and the Lasso) within our data-driven formulations, we do not claim that this advantage holds for arbitrary model instances. We choose the kNN-SAA approach to compare against because it is easy to implement and tune this approach, and the empirical results of Bertsimas and Kallus [12] and Bertsimas and McCord [13] show that this approach is one of the better performing reweighted SAA approaches. The parameter k in kNN-SAA is once again chosen from $[\lfloor n^{0.1} \rfloor, \lceil n^{0.9} \rceil]$ to minimize the 5-fold CV test error.

Solutions from the different approaches are compared by estimating a normalized version of the upper bound of a 99% confidence interval (UCB) on their optimality gaps using the multiple replication procedure

of Mak et al. [57] (see Section G for details). We choose to compare 99% UCBs of the different estimators as it provides a conservative estimate of their suboptimality and mitigates the variability of a single random evaluation sample. Because the data-driven solutions depend on the realization of samples \mathcal{D}_n , we perform 100 replications per test instance and report our results in the form of box plots of these UCBs (the boxes denote the 25th, 50th, and 75th percentiles of the 99% UCBs, and the whiskers denote their 5th and 95th percentiles over the 100 replicates).

Source code and data for the test instances are available at <https://github.com/rohitkannan/DD-SAA>. Our codes are written in Julia 0.6.4 [15], use Gurobi 8.1.0 to solve LPs through the JuMP 0.18.5 interface [33], and use `glmnet` 0.3.0 [39] for Lasso regression. All computational tests were conducted through the UW-Madison high throughput computing software HTCondor (<http://chtc.cs.wisc.edu/>).

Effect of varying covariate dimension. Figure 2 compares the performance of the kNN-SAA, ER-SAA+kNN, ER-SAA+OLS, and PP+OLS approaches for the homoscedastic case ($\omega = 1$) by varying the model degree p , the covariate dimension among $d_x \in \{3, 10, 100\}$, and the sample size among $n \in \{5(d_x + 1), 20(d_x + 1), 100(d_x + 1)\}$. Note that OLS regression estimates $d_x + 1$ parameters for each $j \in \mathcal{J}$. The ER-SAA approaches in this case study do not estimate $Q^* \equiv I$, but directly assume $\hat{Q}_n \equiv I$. The kNN-SAA and ER-SAA+kNN approaches exhibit similar performance overall. When the prediction model is correctly specified (i.e., $p = 1$), the ER-SAA+OLS approach unsurprisingly dominates the ER-SAA+kNN and kNN-SAA approaches. When $p \neq 1$, as anticipated, the ER-SAA+OLS approach does not yield a consistent estimator, whereas the ER-SAA+kNN and kNN-SAA approaches yields consistent estimators, albeit with a slow rate of convergence (cf. Proposition 9). However, the ER-SAA+OLS approach consistently outperforms the ER-SAA+kNN and kNN-SAA approaches when $p = 0.5$ even for the largest sample size of $n = 100(d_x + 1)$. When the degree $p = 2$, the ER-SAA+kNN and kNN-SAA approaches fare better than the ER-SAA+OLS approach only for a sample size of $n \geq 80$ when the covariate dimension is small ($d_x = 3$), and lose this advantage in the larger covariate dimensions. Although the PP+OLS does not yield a consistent estimator even when $p = 1$, it performs better than the ER-SAA+kNN and kNN-SAA approaches in many cases even for large sample sizes n . However, the ER-SAA+OLS approach outperforms the PP+OLS approach across all cases. While we do not show results, we mention that the N-SAA estimator is not asymptotically optimal for all three model instances with the median values of the 99% UCBs of its percentage optimality gaps being about 11%, 5%, and 26% for the $p = 1$, $p = 0.5$, and $p = 2$ instances, respectively, for $n = 10100$. This indicates that using covariate information can be advantageous in these instances.

Impact of the jackknife-based formulations. Figure 3 compares the performance of the ER-SAA and J-SAA approaches with OLS regression by varying the model degree p , the covariate dimension among $d_x \in \{10, 100\}$, and the sample size among $n \in \{1.3(d_x + 1), 1.5(d_x + 1), 2(d_x + 1)\}$ for $\omega = 1$. Once again, we directly assume $\hat{Q}_n \equiv I$. We employ smaller sample sizes in these experiments to see if the jackknife-based SAAs perform better in the limited data regime. We observe that the solutions obtained from the J-SAA formulation typically have smaller 75th and 95th percentiles of the 99% UCBs than those from the ER-SAA formulation, particularly when the sample size n is small. Performance gains are more pronounced for larger sample sizes when the covariate dimension is larger ($d_x = 100$), possibly because the OLS estimators overfit more. Note that, as expected, the J-SAA results converge to the ER-SAA results when the sample size increases. We do not plot the results for the J+-SAA formulation because they are similar to those of the J-SAA formulation.

Impact of the prediction setup. Figure 3 also compares the performance of the ER-SAA+OLS, ER-SAA+Lasso, and PP+Lasso approaches by varying the model degree p , the covariate dimension among $d_x \in \{10, 100\}$, and the sample size among $n \in \{1.3(d_x + 1), 1.5(d_x + 1), 2(d_x + 1)\}$ for $\omega = 1$. We again assume $\hat{Q}_n \equiv I$ in the ER-SAA approaches. We observe that the ER-SAA+Lasso formulation yields better estimators than the ER-SAA+OLS formulation when the sample size n is small relative to the covariate dimension d_x . This effect is accentuated when the covariate dimension is larger ($d_x = 100$), in which case the

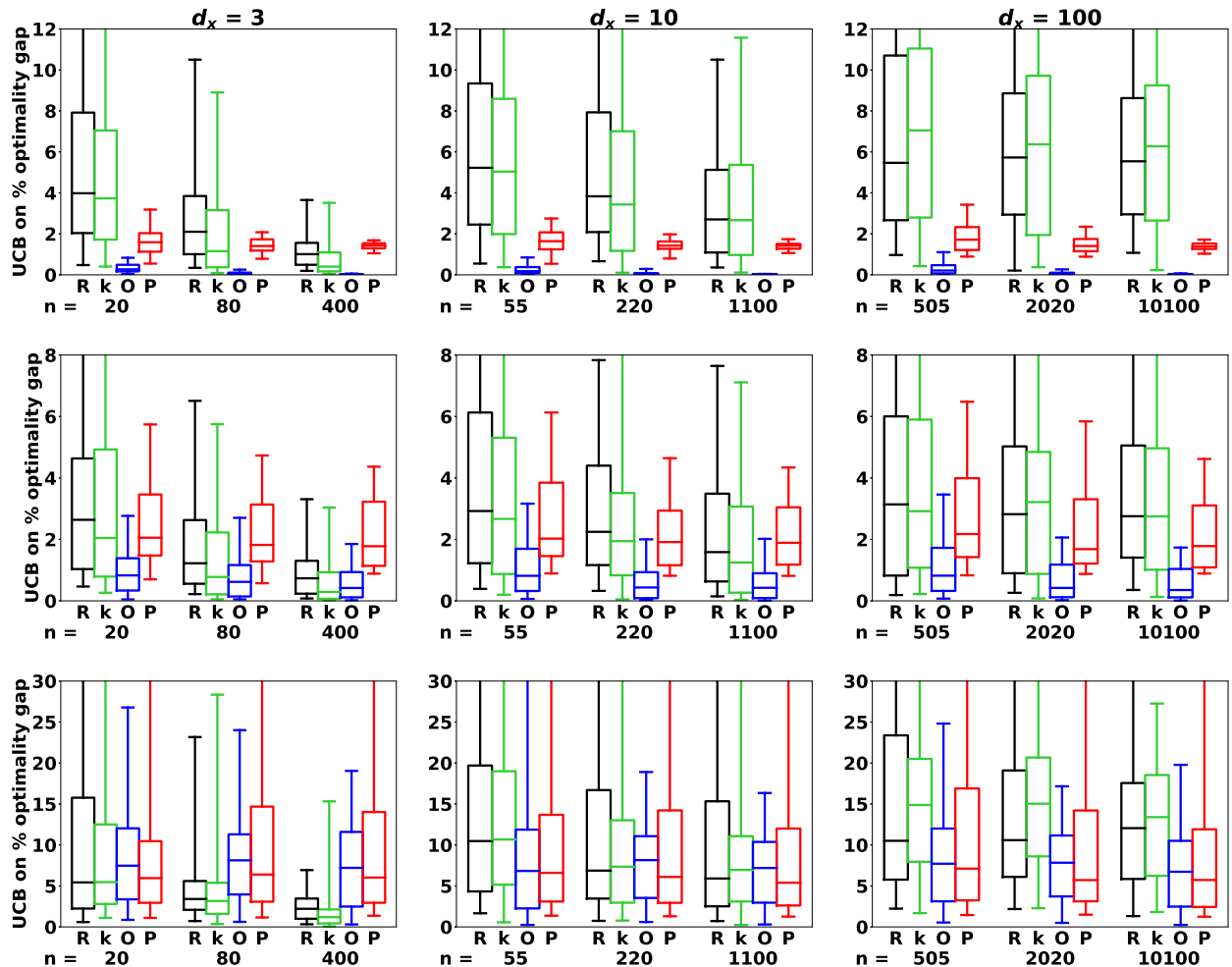


Figure 2: Comparison of kNN-SAA (R), ER-SAA+kNN (k), ER-SAA+OLS (O), and PP+OLS (P) approaches for $\omega = 1$. Top row: $p = 1$. Middle row: $p = 0.5$. Bottom row: $p = 2$. Left column: $d_x = 3$. Middle column: $d_x = 10$. Right column: $d_x = 100$.

OLS-based estimators overfit more and there is increased benefit in using the Lasso to fit a sparser model. The advantage of the Lasso-based estimators shrinks as the sample size increases. The ER-SAA+Lasso approach outperforms the PP+Lasso approach with increased gains for larger sample sizes and covariate dimensions.

Impact of the error variance. Figure 4 and Figure 7 in Section G of the Appendix compare the performance of the kNN-SAA, ER-SAA+kNN, ER-SAA+OLS, and PP+OLS approaches by varying the standard deviation of the errors ε among $\sigma \in \{5, 10, 20\}$ (the case studies thus far used $\sigma = 5$), the model degree p , and the sample size among $n \in \{5(d_x + 1), 20(d_x + 1), 100(d_x + 1)\}$ for $d_x = 10$ and $\omega = 1$. We observe that the ER-SAA+OLS formulation needs a larger sample size to yield a similar certificate of optimality as the standard deviation σ increases. On the other hand, the performance of the ER-SAA+kNN and kNN-SAA approaches appear to be unaffected (and even slightly improve!) with increasing error variance. The performance of the PP+OLS approach deteriorates significantly with increasing σ , especially for $p \neq 2$, and it no longer dominates the ER-SAA+kNN and kNN-SAA approaches for $\sigma = 20$.

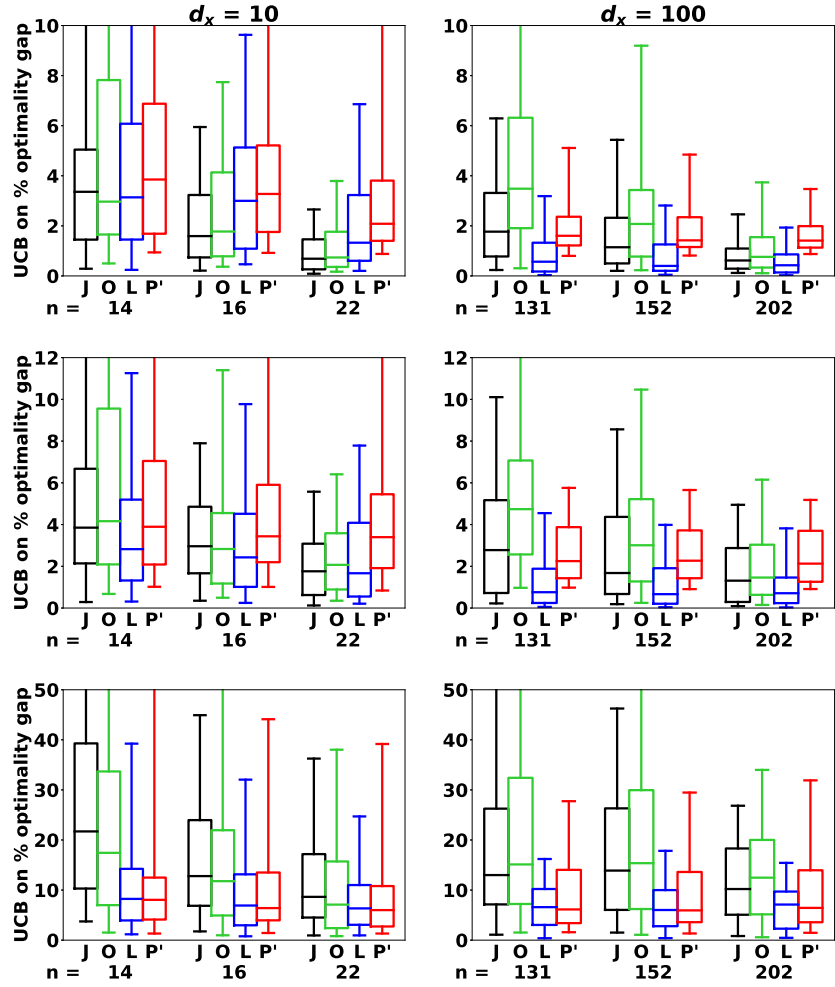


Figure 3: Comparison of J-SAA+OLS (J), ER-SAA+OLS (O), ER-SAA+Lasso (L), and PP+Lasso (P') approaches for $\omega = 1$. Top row: $p = 1$. Middle row: $p = 0.5$. Bottom row: $p = 2$. Left column: $d_x = 10$. Right column: $d_x = 100$.

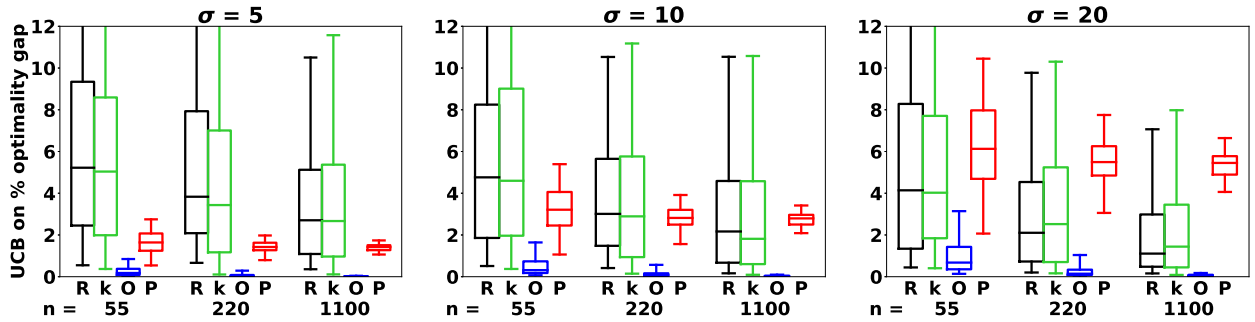


Figure 4: Effect of increasing σ on kNN-SAA (R), ER-SAA+kNN (k), ER-SAA+OLS (O), and PP+OLS (P) approaches when $d_x = 10$, $p = 1$, and $\omega = 1$.

Impact of heteroscedasticity. Figures 5 and 6 compare the performance of the kNN-SAA approach, ER-SAA+kNN approaches with and without estimation of Q^* , ER-SAA+OLS approaches with and without estimation of Q^* , and the PP+OLS approach by varying the heteroscedasticity level $\omega \in \{1, 2, 3\}$, the model degree p , and the sample size among $n \in \{5(d_x+1), 20(d_x+1), 100(d_x+1)\}$ for $d_x = 10$ and $\sigma = 5$. We assume $\hat{Q}_n \equiv I$ in the ER-SAA approaches that do not estimate Q^* . The kNN-SAA approach and the ER-SAA+kNN approach without heteroscedasticity estimation again exhibit similar performance across the different test instances, with the former performing slightly better on instances with severe heteroscedasticity ($\omega = 3$) and the latter performing slightly better on instances with zero and moderate heteroscedasticity ($\omega \in \{1, 2\}$). The ER-SAA+kNN approach with heteroscedasticity estimation outperforms the kNN-SAA approach and the ER-SAA+kNN approach without heteroscedasticity estimation, particularly for model degree $p = 2$ and for severe heteroscedasticity ($\omega = 3$). The ER-SAA+OLS approach without heteroscedasticity estimation outperforms the kNN-SAA approach and the ER-SAA+kNN approaches for $\omega = 1$ and $\omega = 2$ irrespective of the model degree p . However, the kNN-SAA approach outperforms the ER-SAA+OLS approach without heteroscedasticity estimation for $\omega = 3$, $p \in \{0.5, 1\}$, and small sample sizes, with the reverse holding true for large sample sizes. Similar to the observation for kNN regression, the ER-SAA+OLS approach with estimation of Q^* outperforms the ER-SAA+OLS approach without estimation of Q^* in several instances, particularly for large sample sizes and for severe heteroscedasticity ($\omega = 3$). Finally, the ER-SAA+OLS approaches demonstrate significant gains over the PP+OLS approach, especially when the degree of model misspecification is low ($p = 1$ or $p = 0.5$). The results for the PP+OLS and ER-SAA+OLS approaches for $\omega = 3$ may be partly explained by an increase in the variance of the error terms (cf. Figure 1). Note that the median values of the 99% UCBs of the percentage optimality gaps of the N-SAA estimator (for $n = 10100$) for both $\omega = 2$ and $\omega = 3$ are roughly 11%, 5%, and 25% for $p = 1, 0.5,$ and 2 , respectively, which underscores the benefit of using covariate information.

5 Conclusion and future work

We propose three data-driven SAA frameworks for approximating the solution to two-stage stochastic programs when the DM has access to a finite number of samples of random variables and concurrently observed covariates. These formulations fit a model to predict the random variables given covariate values, and use the prediction model and its (out-of-sample) residuals on the given data to construct scenarios for the original stochastic program at a new covariate realization. We provide conditions on the prediction and optimization frameworks and the data generation process under which these data-driven estimators are asymptotically optimal, possess a certain rate of convergence, and possess finite sample guarantees. In particular, we show that our assumptions hold for two-stage stochastic LP in conjunction with popular regression setups such as OLS, Lasso, kNN, and RF regression under various assumptions on the data generation process. Numerical experiments demonstrate the benefits of our data-driven SAA frameworks, in particular, those of our new data-driven formulations in the limited data regime.

Verifying the assumptions on the prediction setup for other frameworks of interest is an important task to be undertaken by the DM. Ongoing work includes analysis of an extension of the ER-SAA approach to multistage stochastic programming [cf. 5, 13]. Designing asymptotically optimal estimators for problems with stochastic constraints [44] and problems with decision-dependent uncertainty [12] are interesting avenues for future work. Modifications of the ER-SAA that achieve better performance guarantees for a given covariate $x \in \mathcal{X}$ would also be interesting to explore.

Acknowledgments

This research is supported by the Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program under Contract Number DE-AC02-06CH11357, and was performed using the computing resources of the UW-Madison Center For High Throughput Computing (CHTC) in the Dept. of Computer Sciences. The CHTC is supported by UW-Madison, the Advanced

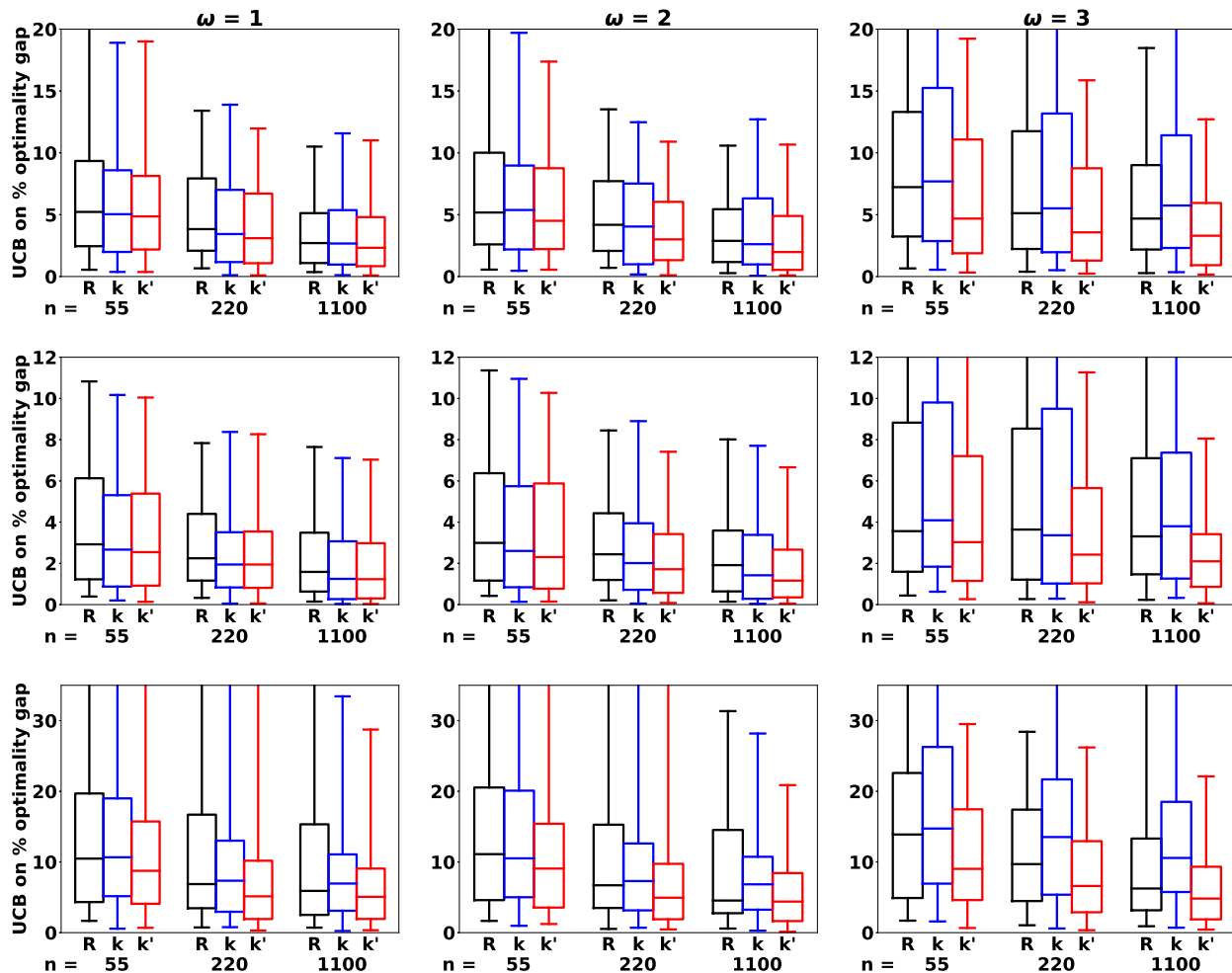


Figure 5: Comparison of kNN-SAA (R), ER-SAA+kNN without heteroscedasticity estimation (k) and ER-SAA+kNN with heteroscedasticity estimation (k') approaches for $d_x = 10$. Top row: $p = 1$. Middle row: $p = 0.5$. Bottom row: $p = 2$. Left column: $\omega = 1$. Middle column: $\omega = 2$. Right column: $\omega = 3$.

Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institutes for Discovery, and the National Science Foundation. R.K. also gratefully acknowledges the support of the U.S. Department of Energy through the LANL/LDRD Program and the Center for Nonlinear Studies. The authors thank the three anonymous reviewers, the associate editor, and Prof. Erick Delage for suggestions that helped improve this paper. R.K. also thanks Dr. Rui Chen and Prof. Garvesh Raskutti for helpful discussions.

References

- [1] D. W. Andrews and D. Pollard. An introduction to functional central limit theorems for dependent stochastic processes. *International Statistical Review*, 62(1):119–132, 1994.
- [2] M. A. Arcones. Limit theorems for nonlinear functionals of a stationary Gaussian sequence of vectors. *The Annals of Probability*, 22(4):2242–2274, 1994.
- [3] M. A. Arcones and B. Yu. Central limit theorems for empirical and U-processes of stationary mixing sequences. *Journal of Theoretical Probability*, 7(1):47–71, 1994.
- [4] G.-Y. Ban and C. Rudin. The big data newsvendor: Practical insights from machine learning. *Operations Research*, 67(1):90–108, 2018.

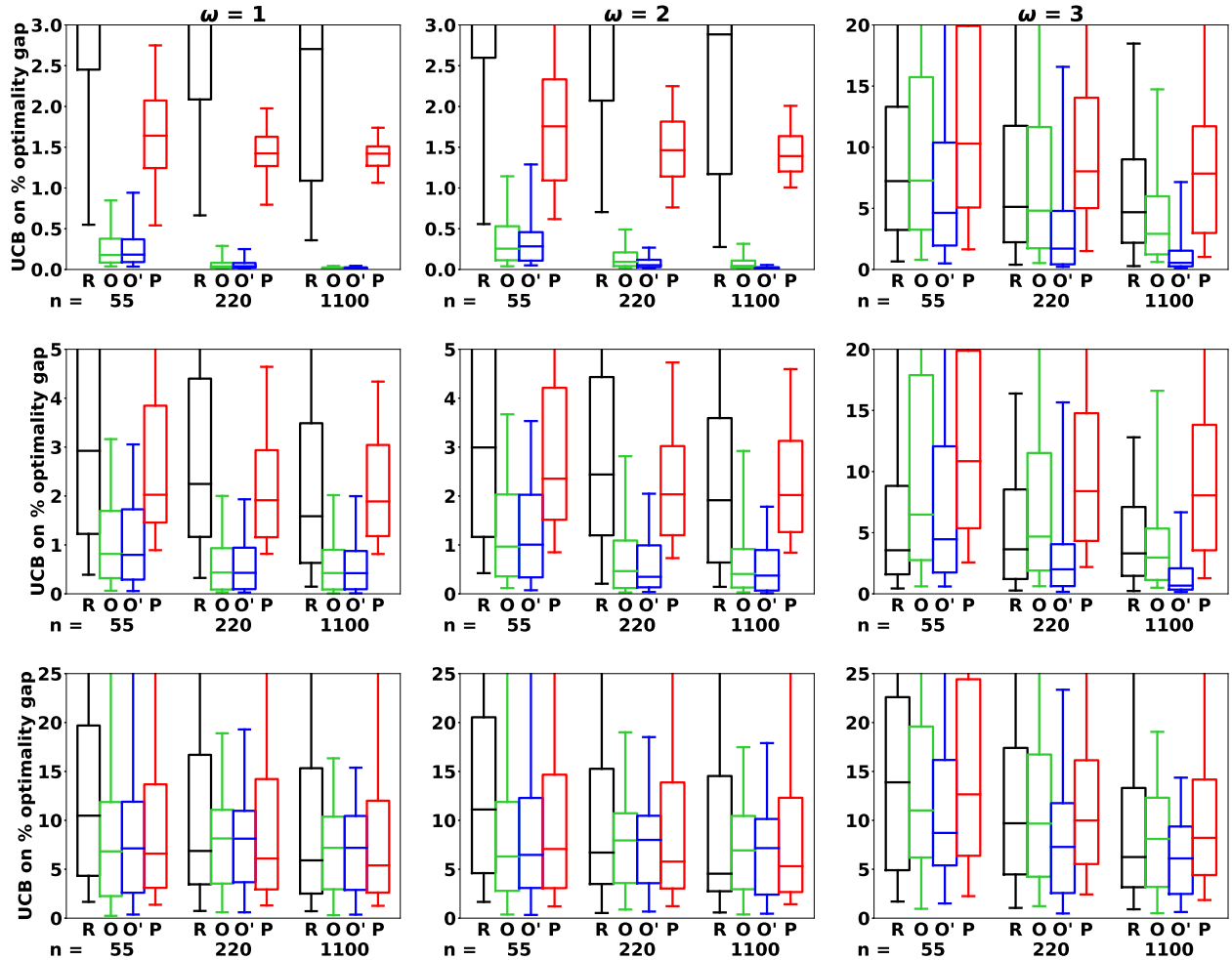


Figure 6: Comparison of kNN-SAA (R), ER-SAA+OLS without heteroscedasticity estimation (O), ER-SAA+OLS with heteroscedasticity estimation (O') and PP+OLS (P) approaches for $d_x = 10$. Top row: $p = 1$. Middle row: $p = 0.5$. Bottom row: $p = 2$. Left column: $\omega = 1$. Middle column: $\omega = 2$. Right column: $\omega = 3$.

- [5] G.-Y. Ban, J. Gallien, and A. J. Mersereau. Dynamic procurement of new products with covariate information: The residual tree method. *Manufacturing & Service Operations Management*, 21:798–815, 2019.
- [6] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49:486–507, 2021.
- [7] S. Basu and G. Michailidis. Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, 43(4):1535–1567, 2015.
- [8] L. Bauwens, S. Laurent, and J. V. Rombouts. Multivariate GARCH models: a survey. *Journal of applied econometrics*, 21(1):79–109, 2006.
- [9] T. Bazier-Matte and E. Delage. Generalization bounds for regularized portfolio selection with market side information. *INFOR: Information Systems and Operational Research*, 58(2):374–401, 2020.
- [10] A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012.
- [11] A. Belloni, V. Chernozhukov, and L. Wang. Pivotal estimation via square-root lasso in nonparametric regression. *The Annals of Statistics*, 42(2):757–788, 2014.
- [12] D. Bertsimas and N. Kallus. From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044, 2020.

- [13] D. Bertsimas and C. McCord. From predictions to prescriptions in multistage optimization problems. *arXiv preprint arXiv:1904.11637*, pages 1–38, 2019.
- [14] D. Bertsimas, C. McCord, and B. Sturt. Dynamic optimization with side information. *European Journal of Operational Research*, 2022.
- [15] J. Bezanson, A. Edelman, S. Karpinski, and V. Shah. Julia: a fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017.
- [16] G. Biau and L. Devroye. *Lectures on the nearest neighbor method*, volume 246. Springer, 2015.
- [17] J. R. Birge and F. Louveaux. *Introduction to stochastic programming*. Springer Science & Business Media, 2011.
- [18] W. Bryc and A. Dembo. Large deviations and strong mixing. In *Annales de l’IHP Probabilités et statistiques*, volume 32, pages 549–569, 1996.
- [19] F. Bunea, A. Tsybakov, M. Wegkamp, et al. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.
- [20] R. J. Carroll and D. Ruppert. Robust estimation in heteroscedastic linear models. *The Annals of Statistics*, pages 429–441, 1982.
- [21] S. Chatterjee. Assumptionless consistency of the Lasso. *arXiv preprint arXiv:1303.5817*, 2013.
- [22] G. H. Chen and D. Shah. Explaining the success of nearest neighbor methods in prediction. *Foundations and Trends in Machine Learning*, 10(5-6):337–588, 2018.
- [23] C. Chesneau, S. El Kolei, J. Kou, and F. Navarro. Nonparametric estimation in a regression model with additive and multiplicative noise. *Journal of Computational and Applied Mathematics*, page 112971, 2020.
- [24] L. Dai, C. Chen, and J. Birge. Convergence properties of two-stage stochastic programming. *Journal of Optimization Theory and Applications*, 106(3):489–509, 2000.
- [25] A. Dalalyan, M. Hebiri, K. Meziani, and J. Salmon. Learning heteroscedastic models by convex programming under group sparsity. In *Proceedings of the 30th International Conference on Machine Learning*, pages 379–387, 2013.
- [26] D. Davarnia, B. Kocuk, and G. Cornuéjols. Computational aspects of Bayesian solution estimators in stochastic optimization. *INFORMS Journal on Optimization*, 2:256–272, 2020.
- [27] M. Davidian and R. J. Carroll. Variance function estimation. *Journal of the American statistical association*, 82(400):1079–1091, 1987.
- [28] A. Dembo and O. Zeitouni. *Large deviations techniques and applications*, volume 38 of *Stochastic Modelling and Applied Probability*. Springer, 2nd edition, 2010.
- [29] S. Diao and S. Sen. Distribution-free algorithms for learning enabled optimization with non-parametric estimation. Optimization Online. URL: http://www.optimization-online.org/DB_HTML/2020/03/7661.html, 2020.
- [30] P. Donti, B. Amos, and J. Z. Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, pages 5484–5494, 2017.
- [31] X. Dou and M. Anitescu. Distributionally robust optimization with correlated data from vector autoregressive processes. *Operations Research Letters*, 47(4):294–299, 2019.
- [32] P. Doukhan, P. Massart, and E. Rio. Invariance principles for absolutely regular empirical processes. In *Annales de l’IHP Probabilités et statistiques*, volume 31, pages 393–427, 1995.
- [33] I. Dunning, J. Huchette, and M. Lubin. JuMP: A modeling language for mathematical optimization. *SIAM Review*, 59(2):295–320, 2017.
- [34] O. El Balghiti, A. N. Elmachtoub, P. Grigas, and A. Tewari. Generalization bounds in the predict-then-optimize framework. *Advances in neural information processing systems*, 32, 2019.
- [35] A. N. Elmachtoub and P. Grigas. Smart “predict, then optimize”. *Management Science*, 68(1):9–26, 2022.
- [36] A. Esteban-Pérez and J. M. Morales. Distributionally robust stochastic programs with side information based on trimmings. *Mathematical Programming*, pages 1–37, 2021.
- [37] J. Fan and Q. Yao. *Nonlinear time series: nonparametric and parametric methods*. Springer Science & Business Media, 2008.
- [38] J. Fan, L. Qi, and D. Xiu. Quasi-maximum likelihood estimation of GARCH models with heavy-tailed likelihoods. *Journal of Business & Economic Statistics*, 32(2):178–191, 2014.
- [39] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.
- [40] P. Grigas, M. Qi, and Z.-J. Shen. Integrated conditional estimation-optimization, 2021.
- [41] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.
- [42] T. Homem-de-Mello. On rates of convergence for stochastic optimization problems under non-independent and identically distributed sampling. *SIAM Journal on Optimization*, 19(2):524–551, 2008.

- [43] T. Homem-de Mello and G. Bayraksan. Monte Carlo sampling-based methods for stochastic optimization. *Surveys in Operations Research and Management Science*, 19(1):56–85, 2014.
- [44] T. Homem-de-Mello and G. Bayraksan. Stochastic constraints and variance reduction techniques. In M. C. Fu, editor, *Handbook of Simulation Optimization*, pages 245–276. Springer New York, 2015.
- [45] D. Hsu, S. M. Kakade, and T. Zhang. Random design analysis of ridge regression. In *Conference on learning theory*, pages 9–1, 2012.
- [46] Y. Hu, N. Kallus, and X. Mao. Fast rates for contextual linear optimization. *Management Science*, 68:4236–4245, 2022.
- [47] H. Jiang. Non-asymptotic uniform rates of consistency for k-NN regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3999–4006, 2019.
- [48] S. Jin, L. Su, and Z. Xiao. Adaptive nonparametric regression with conditional heteroskedasticity. *Econometric Theory*, 31(6):1153, 2015.
- [49] N. Kallus and X. Mao. Stochastic optimization forests. *Management Science*, 2022. Articles in Advance.
- [50] R. Kannan, G. Bayraksan, and J. R. Luedtke. Residuals-based distributionally robust optimization with covariate information. *arXiv preprint arXiv:2012.01088*, 2020.
- [51] R. Kannan, G. Bayraksan, and J. Luedtke. Heteroscedasticity-aware residuals-based contextual stochastic optimization. *arXiv preprint arXiv:2101.03139*, 2021.
- [52] K. Kim and S. Mehrotra. A two-stage stochastic integer programming approach to integrated staffing and scheduling with application to nurse management. *Operations Research*, 63(6):1431–1451, 2015.
- [53] V. Koltchinskii. The Dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15(3):799–828, 2009.
- [54] S. N. Lahiri. *Resampling methods for dependent data*. Springer Science & Business Media, 2013.
- [55] D. Lewandowski, D. Kurowicka, and H. Joe. Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis*, 100(9):1989–2001, 2009.
- [56] J. Luedtke. A branch-and-cut decomposition algorithm for solving chance-constrained mathematical programs with finite support. *Mathematical Programming*, 146(1-2):219–244, 2014.
- [57] W.-K. Mak, D. P. Morton, and R. K. Wood. Monte Carlo bounding techniques for determining solution quality in stochastic programs. *Operations Research Letters*, 24(1-2):47–56, 1999.
- [58] R. Mazumder, A. Choudhury, G. Iyengar, and B. Sen. A computational framework for multivariate convex regression and its variants. *Journal of the American Statistical Association*, 114(525):318–331, 2019.
- [59] M. C. Medeiros and E. F. Mendes. ℓ_1 -regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic errors. *Journal of Econometrics*, 191(1):255–271, 2016.
- [60] S. N. Negahban, P. Ravikumar, M. J. Wainwright, B. Yu, et al. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.
- [61] V. A. Nguyen, F. Zhang, J. Blanchet, E. Delage, and Y. Ye. Distributionally robust local non-parametric conditional estimation. *Advances in Neural Information Processing Systems*, 33:15232–15242, 2020.
- [62] J. L. Powell. Models, testing, and correction of heteroskedasticity. Lecture notes, Department of Economics, University of California, Berkeley. URL: https://eml.berkeley.edu/~powell/e240b_sp10/hetnotes.pdf, 2010.
- [63] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13(Feb):389–427, 2012.
- [64] P. Rigollet and J.-C. Hütter. High dimensional statistics. Lecture notes for MIT’s 18.657 course, 2017. URL <http://www-math.mit.edu/~rigollet/PDFs/RigNotes17.pdf>.
- [65] P. M. Robinson. Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica: Journal of the Econometric Society*, pages 875–891, 1987.
- [66] J. P. Romano and M. Wolf. Resurrecting weighted least squares. *Journal of Econometrics*, 197(1):1–19, 2017.
- [67] J. O. Royset and R. J. Wets. From data to assessments and decisions: Epi-spline technology. In *Bridging data and decisions*, pages 27–53. INFORMS, 2014.
- [68] D. Ruppert, M. P. Wand, U. Holst, and O. Hössjer. Local polynomial variance-function estimation. *Technometrics*, 39(3): 262–273, 1997.
- [69] P. Schütz, A. Tomasgard, and S. Ahmed. Supply chain design under uncertainty using sample average approximation and dual decomposition. *European Journal of Operational Research*, 199(2):409–419, 2009.
- [70] G. A. Seber and A. J. Lee. *Linear regression analysis*. John Wiley & Sons, 2003.
- [71] E. Seijo and B. Sen. Nonparametric least squares estimation of a multivariate convex regression function. *The Annals of Statistics*, 39(3):1633–1657, 2011.
- [72] S. Sen and Y. Deng. Predictive stochastic programming. *Computational Management Science*, 19:1–45, 2022.
- [73] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.
- [74] Q. Sun, W.-X. Zhou, and J. Fan. Adaptive Huber regression. *Journal of the American Statistical Association*, 115(529):

254–265, 2020.

- [75] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.
- [76] S. A. van de Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.
- [77] S. A. van de Geer. High-dimensional generalized linear models and the Lasso. *The Annals of Statistics*, 36(2):614–645, 2008.
- [78] A. W. van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 1998.
- [79] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer, 1996.
- [80] I. Van Keilegom and L. Wang. Semiparametric modeling and estimation of heteroscedasticity in regression analysis of cross-sectional data. *Electronic Journal of Statistics*, 4:133–160, 2010.
- [81] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.
- [82] S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.
- [83] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.
- [84] H. Walk. Strong laws of large numbers and nonparametric estimation. In *Recent Developments in Applied Probability and Statistics*, pages 183–214. Springer, 2010.
- [85] H. White. *Asymptotic theory for econometricians*. Academic press, 2014.
- [86] W.-X. Zhou, K. Bose, J. Fan, and H. Liu. A new perspective on robust M-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *Annals of Statistics*, 46(5):1904, 2018.
- [87] F. Ziel. Iteratively reweighted adaptive Lasso for conditional heteroscedastic time series with applications to AR-ARCH type processes. *Computational Statistics & Data Analysis*, 100:773–793, 2016.

Appendix

Section A provides omitted proofs for results in Section 3. We continue in Section B with a discussion of alternative assumptions under which the results of Section 3 hold and derive rates of convergence for the ER-SAA in Section C. We then outline the analysis for the jackknife-based SAA variants in Section D. Then, in Section E, we present a class of two-stage stochastic programs that satisfy our assumptions. Section F lists several prediction setups (including M-estimators, OLS, Lasso, kNN, CART, and RF regression) that satisfy the assumptions in our analysis. Finally, we end with Section G by providing omitted details for the computational experiments.

A Omitted proofs

A.1 Proof of Lemma 1

We have

$$\begin{aligned}
 \sup_{z \in \mathcal{Z}} |\hat{g}_n^{ER}(z; x) - g_n^*(z; x)| &= \sup_{z \in \mathcal{Z}} \left| \frac{1}{n} \sum_{i=1}^n c(z, \text{proj}_{\mathcal{Y}}(\hat{f}_n(x) + \hat{Q}_n(x)\hat{\varepsilon}_n^i)) - \frac{1}{n} \sum_{i=1}^n c(z, f^*(x) + Q^*(x)\varepsilon^i) \right| \\
 &\leq \sup_{z \in \mathcal{Z}} \frac{1}{n} \sum_{i=1}^n |c(z, \text{proj}_{\mathcal{Y}}(\hat{f}_n(x) + \hat{Q}_n(x)\hat{\varepsilon}_n^i)) - c(z, f^*(x) + Q^*(x)\varepsilon^i)| \\
 &\leq \sup_{z \in \mathcal{Z}} \frac{1}{n} \sum_{i=1}^n L(z) \|\text{proj}_{\mathcal{Y}}(\hat{f}_n(x) + \hat{Q}_n(x)\hat{\varepsilon}_n^i) - (f^*(x) + Q^*(x)\varepsilon^i)\| \\
 &\leq \left(\sup_{z \in \mathcal{Z}} L(z) \right) \left(\frac{1}{n} \sum_{i=1}^n \|\hat{\varepsilon}_n^i(x)\| \right),
 \end{aligned}$$

where the second inequality above follows by Assumption 1 and the final inequality follows by the Lipschitz continuity of orthogonal projections. \square

A.2 Proof of Theorem 5

Before we prove Theorem 5, we present the following lemma that is needed in its proof.

Lemma 10. Let $W \subset \mathbb{R}^{d_w}$ be a nonempty and compact set and $h : W \rightarrow \mathbb{R}$ be a lower semicontinuous function. Define $W^* := \arg \min_{w \in W} h(w)$. Suppose there exists $\delta > 0$ and $\bar{w} \in W$ such that $\text{dist}(\bar{w}, W^*) \geq \delta$. Then, there exists $\kappa > 0$ such that $h(\bar{w}) \geq \min_{w \in W} h(w) + \kappa$.

Proof. Let $W_\delta := \{w \in W : \text{dist}(w, W^*) \geq \delta\}$, and note that $\bar{w} \in W_\delta$. Since h is lower semicontinuous and W_δ is nonempty and compact, $\inf_{w \in W_\delta} h(w)$ is attained. Furthermore, we readily have $\min_{w \in W_\delta} h(w) > \min_{w \in W} h(w)$. Setting $\kappa = \min_{w \in W_\delta} h(w) - \min_{w \in W} h(w)$ yields the desired result. \square

Proof of Theorem 5. Let $z^*(x) \in S^*(x)$ and $\hat{z}_n^{ER}(x) \in \hat{S}_n^{ER}(x)$. Consider any constant $\delta > 0$. From Proposition 4, we have for a.e. $x \in \mathcal{X}$:

$$\begin{aligned} \mathbb{P}\{|\hat{g}_n^{ER}(z^*(x); x) - v^*(x)| > \delta\} &\rightarrow 0 \implies \mathbb{P}\{\hat{v}_n^{ER}(x) > v^*(x) + \delta\} \rightarrow 0 \quad \text{and} \\ \mathbb{P}\{|\hat{v}_n^{ER}(x) - g(\hat{z}_n^{ER}(x); x)| > \delta\} &\rightarrow 0 \implies \mathbb{P}\{v^*(x) > \hat{v}_n^{ER}(x) + \delta\} \rightarrow 0. \\ \therefore \mathbb{P}\{|\hat{v}_n^{ER}(x) - v^*(x)| > \delta\} &\rightarrow 0 \implies \hat{v}_n^{ER}(x) \xrightarrow{P} v^*(x) \text{ for a.e. } x \in \mathcal{X}. \end{aligned}$$

Suppose for contradiction that $\mathbb{D}(\hat{S}_n^{ER}(x), S^*(x)) \not\xrightarrow{P} 0, \forall x \in \bar{\mathcal{X}}$, where $\bar{\mathcal{X}} \subseteq \mathcal{X}$ with $P_X(\bar{\mathcal{X}}) > 0$. This implies for any $\bar{x} \in \bar{\mathcal{X}}$, there exist constants $\delta > 0$ and $\beta > 0$ and a subsequence $\{n_q\}$ of \mathbb{N} such that $\mathbb{P}\{\mathbb{D}(\hat{S}_{n_q}^{ER}(\bar{x}), S^*(\bar{x})) \geq \delta\} \geq \beta, \forall q \in \mathbb{N}$. Lemma 10 then implies that for a.e. $\bar{x} \in \bar{\mathcal{X}}$, there exists $\kappa(\bar{x}) > 0$ such that

$$\mathbb{P}\left\{\sup_{z \in \hat{S}_{n_q}^{ER}(\bar{x})} g(z; \bar{x}) > v^*(\bar{x}) + \kappa(\bar{x})\right\} \geq \beta, \quad \forall q \in \mathbb{N}. \quad (14)$$

From Proposition 4, we have for a.e. $\bar{x} \in \bar{\mathcal{X}}$:

$$\begin{aligned} \mathbb{P}\left\{\sup_{z \in \hat{S}_{n_q}^{ER}(\bar{x})} |\hat{g}_n^{ER}(z; \bar{x}) - g(z; \bar{x})| \leq 0.5\kappa(\bar{x})\right\} &\rightarrow 1 \implies \mathbb{P}\left\{\sup_{z \in \hat{S}_{n_q}^{ER}(\bar{x})} g(z; \bar{x}) \leq \hat{v}_n^{ER}(\bar{x}) + 0.5\kappa(\bar{x})\right\} \rightarrow 1, \\ \mathbb{P}\left\{\sup_{z \in S^*(\bar{x})} |\hat{g}_n^{ER}(z; \bar{x}) - g(z; \bar{x})| \leq 0.5\kappa(\bar{x})\right\} &\rightarrow 1 \implies \mathbb{P}\left\{\sup_{z \in S^*(\bar{x})} \hat{g}_n^{ER}(z; \bar{x}) \leq v^*(\bar{x}) + 0.5\kappa(\bar{x})\right\} \rightarrow 1. \end{aligned}$$

Since $\hat{v}_n^{ER}(\bar{x}) \leq \sup_{z \in S^*(\bar{x})} \hat{g}_n^{ER}(z; \bar{x})$ by definition, the above inequalities in turn imply

$$\mathbb{P}\left\{\sup_{z \in \hat{S}_{n_q}^{ER}(\bar{x})} g(z; \bar{x}) \leq v^*(\bar{x}) + \kappa(\bar{x})\right\} \rightarrow 1,$$

which contradicts the inequality (14). The above arguments also readily imply that the ER-SAA estimators are asymptotically optimal, i.e., $\sup_{z \in \hat{S}_n^{ER}(x)} g(z; x) \xrightarrow{P} v^*(x)$ for a.e. $x \in \mathcal{X}$. \square

A.3 Proof of Theorem 6

Since

$$\begin{aligned} \|\hat{v}_n^{ER}(X) - v^*(X)\|_{L^q} &= \left\| \min_{z \in \hat{\mathcal{Z}}} \hat{g}_n^{ER}(z; X) - \min_{z \in \mathcal{Z}} g(z; X) \right\|_{L^q} \leq \left\| \sup_{z \in \hat{\mathcal{Z}}} |\hat{g}_n^{ER}(z; X) - g(z; X)| \right\|_{L^q}, \\ \|g(\hat{z}_n^{ER}(X); X) - v^*(X)\|_{L^q} &\leq \|g(\hat{z}_n^{ER}(X); X) - \hat{v}_n^{ER}(X)\|_{L^q} + \|\hat{v}_n^{ER}(X) - v^*(X)\|_{L^q}, \end{aligned}$$

we focus on establishing convergence of $\hat{g}_n^{ER}(\cdot; X)$ to $g(\cdot; X)$ with respect to the L^q -norm on \mathcal{X} . From inequality (9) and the triangle inequality for the L^q -norm, we have

$$\left\| \sup_{z \in \mathcal{Z}} |\hat{g}_n^{ER}(z; X) - g(z; X)| \right\|_{L^q} \leq \left\| \sup_{z \in \mathcal{Z}} |\hat{g}_n^{ER}(z; X) - g_n^*(z; X)| \right\|_{L^q} + \left\| \sup_{z \in \mathcal{Z}} |g_n^*(z; X) - g(z; X)| \right\|_{L^q}.$$

Assumption 6 implies that the second term on the r.h.s. of the above inequality converges to zero in probability. We now show that the first term also converges to zero in probability. We have

$$\begin{aligned} & \left\| \sup_{z \in \mathcal{Z}} |\hat{g}_n^{ER}(z; X) - g_n^*(z; X)| \right\|_{L^q} \\ & \leq \left(\sup_{z \in \mathcal{Z}} L(z) \right) \left\| \frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(X)\| \right\|_{L^q} \\ & \leq O(1) \left\| \hat{f}_n(X) - f^*(X) \right\|_{L^q} + O(1) \left\| \hat{Q}_n(X) - Q^*(X) \right\|_{L^q} \left(\frac{1}{n} \sum_{i=1}^n \|\varepsilon^i\| \right) + \\ & \quad O(1) \left\| \hat{Q}_n(X) \right\|_{L^q} \left(\frac{1}{n} \sum_{i=1}^n \left\| [\hat{Q}_n(x^i)]^{-1} - [Q^*(x^i)]^{-1} \right\|^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n \|Q^*(x^i)\|^4 \right)^{1/4} \left(\frac{1}{n} \sum_{i=1}^n \|\varepsilon^i\|^4 \right)^{1/4} + \\ & \quad O(1) \left\| \hat{Q}_n(X) \right\|_{L^q} \left(\frac{1}{n} \sum_{i=1}^n \left\| [\hat{Q}_n(x^i)]^{-1} \right\|^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_n(x^i)\|^2 \right)^{1/2}, \end{aligned}$$

where the first inequality follows from Lemma 1 and the second inequality follows from Lemma 2 and the triangle inequality. The rest of the arguments for $\left\| \sup_{z \in \mathcal{Z}} |\hat{g}_n^{ER}(z; X) - g_n^*(z; X)| \right\|_{L^q} \xrightarrow{P} 0$ follow by similar arguments as in the proof of Proposition 4 upon noting that

$$\left\| \hat{Q}_n(X) \right\|_{L^q} \leq \left\| \hat{Q}_n(X) - Q^*(X) \right\|_{L^q} + \left\| Q^*(X) \right\|_{L^q}$$

and by replacing Assumptions 2 and 5 with Assumptions 6 and 7. \square

A.4 Proofs of Lemma 7, Theorem 8, and Proposition 9

Proof of Lemma 7. Assumption 1 and Lemma 1 imply

$$\begin{aligned} \mathbb{P} \left\{ \sup_{z \in \mathcal{Z}} |\hat{g}_n^{ER}(z; x) - g_n^*(z; x)| > \kappa \right\} & \leq \mathbb{P} \left\{ \left(\sup_{z \in \mathcal{Z}} L(z) \right) \left(\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\| \right) > \kappa \right\} \\ & \leq \mathbb{P} \left\{ \left(\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\| \right) > \frac{\kappa}{\left(\sup_{z \in \mathcal{Z}} L(z) \right)} \right\}. \end{aligned} \quad (15)$$

Therefore, it suffices to bound $\mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\| > \kappa \right\}$ for any $\kappa > 0$.

From Lemma 2 and the inequality $\mathbb{P}\{V + W > c_1 + c_2\} \leq \mathbb{P}\{V > c_1\} + \mathbb{P}\{W > c_2\}$ for any random variables V, W and constants c_1, c_2 , we have:

$$\begin{aligned} & \mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\| > \kappa \right\} \\ & \leq \mathbb{P} \left\{ \|\hat{f}_n(x) - f^*(x)\| > \frac{\kappa}{4} \right\} + \mathbb{P} \left\{ \left(\frac{1}{n} \sum_{i=1}^n \|\varepsilon^i\| \right) \|\hat{Q}_n(x) - Q^*(x)\| > \frac{\kappa}{4} \right\} + \\ & \quad \mathbb{P} \left\{ \|\hat{Q}_n(x)\| \left(\frac{1}{n} \sum_{i=1}^n \left\| [\hat{Q}_n(x^i)]^{-1} - [Q^*(x^i)]^{-1} \right\|^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n \|Q^*(x^i)\|^4 \right)^{1/4} \left(\frac{1}{n} \sum_{i=1}^n \|\varepsilon^i\|^4 \right)^{1/4} > \frac{\kappa}{4} \right\} + \\ & \quad \mathbb{P} \left\{ \|\hat{Q}_n(x)\| \left(\frac{1}{n} \sum_{i=1}^n \left\| [\hat{Q}_n(x^i)]^{-1} \right\|^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_n(x^i)\|^2 \right)^{1/2} > \frac{\kappa}{4} \right\}. \end{aligned} \quad (16)$$

For a.e. $x \in \mathcal{X}$, the first term on the r.h.s. of (16) can be bounded using Assumption (11a) as

$$\mathbb{P}\left\{\|\hat{f}_n(x) - f^*(x)\| > \frac{\kappa}{4}\right\} \leq K_f(\frac{\kappa}{4}, x) \exp(-\beta_f(n, \frac{\kappa}{4}, x)).$$

Next, consider the second term on the r.h.s. of inequality (16). We have for a.e. $x \in \mathcal{X}$

$$\begin{aligned} & \mathbb{P}\left\{\left(\frac{1}{n} \sum_{i=1}^n \|\varepsilon^i\|\right) \|\hat{Q}_n(x) - Q^*(x)\| > \frac{\kappa}{4}\right\} \\ & \leq \mathbb{P}\left\{\frac{1}{n} \sum_{i=1}^n \|\varepsilon^i\| > \mathbb{E}[\|\varepsilon\|] + \kappa\right\} + \mathbb{P}\left\{(\mathbb{E}[\|\varepsilon\|] + \kappa) \|\hat{Q}_n(x) - Q^*(x)\| > \frac{\kappa}{4}\right\} \\ & \leq J_\varepsilon(\kappa) \exp(-\gamma_\varepsilon(n, \kappa)) + \mathbb{P}\left\{\|\hat{Q}_n(x) - Q^*(x)\| > \frac{\kappa}{4(\mathbb{E}[\|\varepsilon\|] + \kappa)}\right\} \\ & \leq J_\varepsilon(\kappa) \exp(-\gamma_\varepsilon(n, \kappa)) + K_Q\left(\frac{\kappa}{4(\mathbb{E}[\|\varepsilon\|] + \kappa)}, x\right) \exp(-\beta_Q(n, \frac{\kappa}{4(\mathbb{E}[\|\varepsilon\|] + \kappa)}, x)), \end{aligned}$$

where the second inequality follows by Assumption 10 and the last step by Assumption (11c).

The third term on the r.h.s. of inequality (16) can be bounded for a.e. $x \in \mathcal{X}$ as

$$\begin{aligned} & \mathbb{P}\left\{\|\hat{Q}_n(x)\| \left(\frac{1}{n} \sum_{i=1}^n \|\hat{Q}_n(x^i)\|^{-1} - \|Q^*(x^i)\|^{-1}\right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n \|Q^*(x^i)\|^4\right)^{1/4} \left(\frac{1}{n} \sum_{i=1}^n \|\varepsilon^i\|^4\right)^{1/4} > \frac{\kappa}{4}\right\} \\ & \leq \mathbb{P}\left\{\|\hat{Q}_n(x)\| > \|Q^*(x)\| + \kappa\right\} + \mathbb{P}\left\{\left(\frac{1}{n} \sum_{i=1}^n \|Q^*(x^i)\|^4\right)^{1/4} > (\mathbb{E}[\|Q^*(X)\|^4])^{1/4} + \kappa\right\} + \\ & \quad \mathbb{P}\left\{\left(\frac{1}{n} \sum_{i=1}^n \|\varepsilon^i\|^4\right)^{1/4} > (\mathbb{E}[\|\varepsilon\|^4])^{1/4} + \kappa\right\} + \\ & \quad \mathbb{P}\left\{(\|Q^*(x)\| + \kappa) ((\mathbb{E}[\|Q^*(X)\|^4])^{1/4} + \kappa) ((\mathbb{E}[\|\varepsilon\|^4])^{1/4} + \kappa) \left(\frac{1}{n} \sum_{i=1}^n \|\hat{Q}_n(x^i)\|^{-1} - \|Q^*(x^i)\|^{-1}\right)^{1/2} > \frac{\kappa}{4}\right\} \\ & \leq K_Q(\kappa, x) \exp(-\beta_Q(n, \kappa, x)) + \bar{J}_Q(\kappa) \exp(-\bar{\gamma}_Q(n, \kappa)) + \bar{J}_\varepsilon(\kappa) \exp(-\bar{\gamma}_\varepsilon(n, \kappa)) + \\ & \quad \mathbb{P}\left\{\left(\frac{1}{n} \sum_{i=1}^n \|\hat{Q}_n(x^i)\|^{-1} - \|Q^*(x^i)\|^{-1}\right)^{1/2} > h_1(\kappa, x)\right\} \\ & \leq K_Q(\kappa, x) \exp(-\beta_Q(n, \kappa, x)) + \bar{J}_Q(\kappa) \exp(-\bar{\gamma}_Q(n, \kappa)) + \bar{J}_\varepsilon(\kappa) \exp(-\bar{\gamma}_\varepsilon(n, \kappa)) + \\ & \quad \bar{K}_Q(h_1(\kappa, x)) \exp(-\bar{\beta}_Q(n, h_1(\kappa, x))), \end{aligned}$$

where the second inequality follows from Assumptions 9, 10, and (11c), the final inequality follows from Assumption (11d), and

$$h_1(\kappa, x) := \frac{\kappa}{4(\|Q^*(x)\| + \kappa) ((\mathbb{E}[\|Q^*(X)\|^4])^{1/4} + \kappa) ((\mathbb{E}[\|\varepsilon\|^4])^{1/4} + \kappa)}.$$

Finally, the fourth term on the r.h.s. of inequality (16) can be bounded for a.e. $x \in \mathcal{X}$ as

$$\begin{aligned}
& \mathbb{P}\left\{\|\hat{Q}_n(x)\|\left(\frac{1}{n}\sum_{i=1}^n\|[\hat{Q}_n(x^i)]^{-1}\|^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^n\|f^*(x^i)-\hat{f}_n(x^i)\|^2\right)^{1/2}>\frac{\kappa}{4}\right\} \\
& \leq \mathbb{P}\left\{\|\hat{Q}_n(x)\|>\|Q^*(x)\|+\kappa\right\}+\mathbb{P}\left\{\left(\frac{1}{n}\sum_{i=1}^n\|[\hat{Q}_n(x^i)]^{-1}\|^2\right)^{1/2}>\left(\mathbb{E}\|[\|Q^*(X)\|^{-1}\|^2]\right)^{1/2}+2\kappa\right\}+ \\
& \quad \mathbb{P}\left\{\left(\|Q^*(x)\|+\kappa\right)\left(\left(\mathbb{E}\|[\|Q^*(X)\|^{-1}\|^2]\right)^{1/2}+2\kappa\right)\left(\frac{1}{n}\sum_{i=1}^n\|f^*(x^i)-\hat{f}_n(x^i)\|^2\right)^{1/2}>\frac{\kappa}{4}\right\} \\
& \leq K_Q(\kappa,x)\exp(-\beta_Q(n,\kappa,x))+\mathbb{P}\left\{\left(\frac{1}{n}\sum_{i=1}^n\|f^*(x^i)-\hat{f}_n(x^i)\|^2\right)^{1/2}>h_2(\kappa,x)\right\}+ \\
& \quad \mathbb{P}\left\{\left(\frac{1}{n}\sum_{i=1}^n\|[\hat{Q}_n(x^i)]^{-1}-[Q^*(x^i)]^{-1}\|^2\right)^{1/2}+\left(\frac{1}{n}\sum_{i=1}^n\|[\|Q^*(x^i)\|^{-1}\|^2]\right)^{1/2}>\left(\mathbb{E}\|[\|Q^*(X)\|^{-1}\|^2]\right)^{1/2}+2\kappa\right\} \\
& \leq K_Q(\kappa,x)\exp(-\beta_Q(n,\kappa,x))+\bar{K}_f(h_2(\kappa,x))\exp(-\bar{\beta}_f(n,h_2(\kappa,x)))+\bar{K}_Q(\kappa)\exp(-\bar{\beta}_Q(n,\kappa))+ \\
& \quad J_Q(\kappa)\exp(-\gamma_Q(n,\kappa)),
\end{aligned}$$

where the second inequality follows by Assumption 11 and Lemma 3, the final inequality follows by Assumptions 11 and 9 and the probability inequality stated at the beginning of this proof, and

$$h_2(\kappa,x):=\frac{\kappa}{4\left(\|Q^*(x)\|+\kappa\right)\left(\left(\mathbb{E}\|[\|Q^*(X)\|^{-1}\|^2]\right)^{1/2}+2\kappa\right)}.$$

Putting the above bounds together in inequality (16), we have for a.e. $x \in \mathcal{X}$:

$$\begin{aligned}
\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n\|\hat{\varepsilon}_n^i(x)\|>\kappa\right\}&\leq J_\varepsilon(\kappa)\exp(-\gamma_\varepsilon(n,\kappa))+\bar{J}_\varepsilon(\kappa)\exp(-\bar{\gamma}_\varepsilon(n,\kappa))+ \\
& \quad J_Q(\kappa)\exp(-\gamma_Q(n,\kappa))+\bar{J}_Q(\kappa)\exp(-\bar{\gamma}_Q(n,\kappa))+ \\
& \quad K_f\left(\frac{\kappa}{4},x\right)\exp(-\beta_f(n,\frac{\kappa}{4},x))+\bar{K}_f(h_2(\kappa,x))\exp(-\bar{\beta}_f(n,h_2(\kappa,x)))+ \\
& \quad K_Q\left(\frac{\kappa}{4\left(\mathbb{E}\|[\|\hat{\varepsilon}_n\|]+\kappa\right)},x\right)\exp(-\beta_Q(n,\frac{\kappa}{4\left(\mathbb{E}\|[\|\hat{\varepsilon}_n\|]+\kappa\right)},x))+2K_Q(\kappa,x)\exp(-\beta_Q(n,\kappa,x))+ \\
& \quad \bar{K}_Q(\kappa)\exp(-\bar{\beta}_Q(n,\kappa))+\bar{K}_Q(h_1(\kappa,x))\exp(-\bar{\beta}_Q(n,h_1(\kappa,x))),
\end{aligned} \tag{17}$$

which along with inequality (15) implies the desired result. \square

Proof of Theorem 8. Note that for any $\kappa > 0$:

$$\mathbb{P}\left\{\sup_{z \in \mathcal{Z}}|\hat{g}_n^{ER}(z;x)-g(z;x)|>\kappa\right\}\leq\mathbb{P}\left\{\sup_{z \in \mathcal{Z}}|\hat{g}_n^{ER}(z;x)-g_n^*(z;x)|>\frac{\kappa}{2}\right\}+\mathbb{P}\left\{\sup_{z \in \mathcal{Z}}|g_n^*(z;x)-g(z;x)|>\frac{\kappa}{2}\right\}.$$

Bounding the two terms on the r.h.s. of the above inequality using Assumption 8 and Lemma 7 yields for a.e. $x \in \mathcal{X}$:

$$\mathbb{P}\left\{\sup_{z \in \mathcal{Z}}|\hat{g}_n^{ER}(z;x)-g(z;x)|>\kappa\right\}\leq\tilde{K}(\kappa,x)\exp(-\tilde{\beta}(n,\kappa,x)), \tag{18}$$

where $\tilde{K}(\kappa,x):=2\max\{K(0.5\kappa,x),\bar{K}(0.5\kappa,x)\}$ and $\tilde{\beta}(n,\kappa,x):=\min\{n\beta(0.5\kappa,x),\bar{\beta}(n,0.5\kappa,x)\}$. Inequal-

ity (18) implies for $n \in \mathbb{N}$, a.e. $x \in \mathcal{X}$, and any $\kappa := \kappa(x) > 0$:

$$\begin{aligned} & \mathbb{P}\left\{\sup_{z \in \hat{S}_n^{ER}(x)} |\hat{g}_n^{ER}(z; x) - g(z; x)| \leq 0.5\kappa\right\} \geq 1 - \tilde{K}(0.5\kappa, x) \exp(-\tilde{\beta}(n, 0.5\kappa, x)) \\ \implies & \mathbb{P}\left\{\sup_{z \in \hat{S}_n^{ER}(x)} g(z; x) \leq \hat{v}_n^{ER}(x) + 0.5\kappa\right\} \geq 1 - \tilde{K}(0.5\kappa, x) \exp(-\tilde{\beta}(n, 0.5\kappa, x)), \\ \text{and } & \mathbb{P}\left\{\sup_{z \in S^*(x)} |\hat{g}_n^{ER}(z; x) - g(z; x)| \leq 0.5\kappa\right\} \geq 1 - \tilde{K}(0.5\kappa, x) \exp(-\tilde{\beta}(n, 0.5\kappa, x)) \\ \implies & \mathbb{P}\left\{\sup_{z \in S^*(x)} \hat{g}_n^{ER}(z; x) \leq v^*(x) + 0.5\kappa\right\} \geq 1 - \tilde{K}(0.5\kappa, x) \exp(-\tilde{\beta}(n, 0.5\kappa, x)). \end{aligned}$$

Since $\hat{v}_n^{ER}(x) \leq \sup_{z \in S^*(x)} \hat{g}_n^{ER}(z; x)$ by the definition of $\hat{v}_n^{ER}(x)$, the above two inequalities imply

$$\begin{aligned} & \mathbb{P}\left\{\sup_{z \in \hat{S}_n^{ER}(x)} g(z; x) \leq v^*(x) + \kappa\right\} \geq 1 - 2\tilde{K}(0.5\kappa, x) \exp(-\tilde{\beta}(n, 0.5\kappa, x)), \quad (19) \\ \implies & \mathbb{P}\left\{g(\hat{z}_n^{ER}(x); x) \leq v^*(x) + \kappa(x)\right\} \geq 1 - 2\tilde{K}(0.5\kappa, x) \exp(-\tilde{\beta}(n, 0.5\kappa, x)). \end{aligned}$$

Suppose $\text{dist}(\hat{z}_n^{ER}(x), S^*(x)) \geq \eta$ for some $x \in \mathcal{X}$ and some sample path. Since $g(\cdot; x)$ is lsc on the compact set \mathcal{Z} for a.e. $x \in \mathcal{X}$, Lemma 10 implies that there exists $\kappa(\eta, x) > 0$ such that $g(\hat{z}_n^{ER}(x); x) > v^*(x) + \kappa(\eta, x)$ on that path (except for some paths of measure zero). We now provide a bound on the probability of this event. By the above arguments, we have for a.e. $x \in \mathcal{X}$:

$$\begin{aligned} \mathbb{P}\left\{\text{dist}(\hat{z}_n^{ER}(x), S^*(x)) \geq \eta\right\} & \leq \mathbb{P}\left\{g(\hat{z}_n^{ER}(x); x) > v^*(x) + \kappa(\eta, x)\right\} \\ & \leq 2\tilde{K}(0.5\kappa(\eta, x), x) \exp(-\tilde{\beta}(n, 0.5\kappa(\eta, x), x)). \quad \square \end{aligned}$$

Proof of Proposition 9. We show that for each $n \in \mathbb{N}$ and a.e. $x \in \mathcal{X}$, there exist positive constants $\tilde{K}(\kappa, x)$ and $\tilde{\beta}(n, \kappa, x)$ such that inequality (18) holds. Following the arguments in the proof of Theorem 8, inequality (19) then implies

$$\mathbb{P}\left\{\sup_{z \in \hat{S}_n^{ER}(x)} g(z; x) \leq v^*(x) + \kappa\right\} \geq 1 - 2\tilde{K}(0.5\kappa, x) \exp(-\tilde{\beta}(n, 0.5\kappa, x)),$$

which along with the definition of $S^\kappa(x)$ in turn implies that

$$\mathbb{P}\left\{\hat{S}_n^{ER}(x) \subseteq S^\kappa(x)\right\} \geq 1 - 2\tilde{K}(0.5\kappa, x) \exp(-\tilde{\beta}(n, 0.5\kappa, x)).$$

We now state results that can be used to bound the constants $\tilde{K}(\kappa, x)$ and $\tilde{\beta}(\kappa, x)$ in inequality (18); we ignore their dependence on x to keep the exposition simple. Theorems 7.66 and 7.67 of Shapiro et al. [73] imply for our setting of two-stage stochastic LP the bound

$$\mathbb{P}\left\{\sup_{z \in \mathcal{Z}} |g_n^*(z; x) - g(z; x)| > \kappa\right\} \leq O(1) \left(\frac{O(1)D}{\kappa}\right)^{d_z} \exp\left(-\frac{n\kappa^2}{O(1)\sigma_c^2(x)}\right) \quad (20)$$

for a.e. $x \in \mathcal{X}$. The following large deviation inequalities for our three different regression setups (see Section F) can be used to specialize the bound afforded by Lemma 7:

1. OLS regression: $\mathbb{P}\left\{\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|^2 > \kappa^2\right\} \leq \exp(d_x) \exp\left(-\frac{n\kappa^2}{O(1)\sigma^2 d_y}\right)$, which follows from Remark 12 of Hsu et al. [45], Theorem 2.2 and Remark 2.3 of Rigollet and Hütter [64].

2. Lasso regression: $\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n\|\tilde{\varepsilon}_n^i(x)\|^2 > \kappa^2\right\} \leq 2d_x \exp\left(-\frac{n\kappa^2}{O(1)\sigma^2sd_y}\right)$, which follows from Theorem 2.1 and Corollary 1 of Bunea et al. [19].

3. kNN regression: Whenever $n \geq O(1)\left(\frac{O(1)}{\kappa}\right)^{\frac{d_x}{1-\gamma}}$ and $\frac{n^\gamma}{\log(n)} \geq \frac{O(1)d_xd_y\sigma^2}{\kappa^2}$, we have

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n\|\tilde{\varepsilon}_n^i(x)\|^2 > \kappa^2d_y\right\} \leq \left(\frac{O(1)\sqrt{d_x}}{\kappa}\right)^{d_x} \exp\left(-O(1)n(O(1)\kappa)^{2d_x}\right) + O(1)n^{2d_x}\left(\frac{O(1)}{d_x}\right)^{d_x} \exp\left(-\frac{n^\gamma\kappa^2}{O(1)\sigma^2}\right)$$

from Lemma 10 of Bertsimas and McCord [13].

Suppose the regression step (3) is Lasso regression. We have from Lemma 7 that

$$\mathbb{P}\left\{\sup_{z \in \mathcal{Z}}|\hat{g}_n^{ER}(z; x) - g_n^*(z; x)| > \kappa\right\} \leq O(1)d_x \exp\left(-\frac{n\kappa^2}{O(1)\sigma^2sd_y}\right).$$

Along with the uniform exponential bound inequality (20), this yields for a.e. $x \in \mathcal{X}$:

$$\mathbb{P}\left\{\sup_{z \in \mathcal{Z}}|\hat{g}_n^{ER}(z; x) - g(z; x)| > \kappa\right\} \leq O(1)\left(\frac{O(1)D}{\kappa}\right)^{d_z} \exp\left(-\frac{n\kappa^2}{O(1)\sigma_\varepsilon^2(x)}\right) + O(1)d_x \exp\left(-\frac{n\kappa^2}{O(1)\sigma^2sd_y}\right).$$

Requiring each term in the r.h.s. of the above inequality to be $\leq 0.5\delta$ and using the union bound yields the stated conservative sample size estimates. Sample complexities for OLS and kNN regression can be similarly derived. \square

B Alternative assumptions for analysis of the ER-SAA

In this section, we present alternative sets of assumptions under which theoretical guarantees similar to those in Section 3 hold for the ER-SAA problem. The following assumption provides an alternative to the Lipschitz continuity Assumption 1, and can be used to derive asymptotic guarantees for the ER-SAA.

Assumption 12. Problem (1), the regression step (3), and the data \mathcal{D}_n satisfy:

(12a) there exists a function $\delta : \mathcal{X} \rightarrow \mathbb{R}_+$ such that for a.e. $x \in \mathcal{X}$, $\|\tilde{\varepsilon}_n^i(x)\| \leq \delta(x)$, $\forall i \in [n]$, a.s. for n large enough,

(12b) for a.e. $x \in \mathcal{X}$ and for each $(z, \varepsilon) \in \mathcal{Z} \times \Xi$, the function c in problem (1) satisfies the following local Lipschitz inequality for each $\bar{y} \in \mathcal{B}_{\delta(x)}(f^*(x) + Q^*(x)\varepsilon) \cap \mathcal{Y}$:

$$|c(z, \bar{y}) - c(z, f^*(x) + Q^*(x)\varepsilon)| \leq L_{\delta(x)}(z, f^*(x) + Q^*(x)\varepsilon)\|\bar{y} - (f^*(x) + Q^*(x)\varepsilon)\|,$$

with the ‘local Lipschitz constant’ $L_{\delta(x)}(z, f^*(x) + Q^*(x)\varepsilon)$ satisfying

$$\sup_{z \in \mathcal{Z}} \frac{1}{n} \sum_{i=1}^n L_{\delta(x)}^2(z, f^*(x) + Q^*(x)\varepsilon^i) = O_p(1).$$

Unlike Assumption 1, Assumption (12b) only requires the function $c(z, \cdot)$ to be locally Lipschitz continuous for each $z \in \mathcal{Z}$ with the local Lipschitz constant satisfying a uniform stochastic boundedness condition, but using this weaker assumption necessitates the stronger Assumption (12a) on the regression step (3). Consider the homoscedastic setting (i.e., $\hat{Q}_n := Q^* \equiv I$). Because the deviation terms satisfy $\|\tilde{\varepsilon}_n^i(x)\| \leq \|f^*(x) - \hat{f}_n(x)\| + \|f^*(x^i) - \hat{f}_n(x^i)\|$, $\forall i \in [n]$, Assumption (12a) is satisfied in this setting for our running example of OLS regression, e.g., if the support \mathcal{X} is compact, the population regression problem has a unique solution θ^* , the strong pointwise LLN holds for the objective function (i.e., the empirical loss)

of the regression problem (3), and $\mathbb{E}[\|\varepsilon\|^2] < +\infty$ (see Theorem 5.4 of Shapiro et al. [73] for details). The support \mathcal{X} of the covariates being compact may not be an overly restrictive assumption because in many applications the covariates (e.g., temperature, precipitation, wind, or location) can be assumed to be bounded for all practical purposes. For the heteroscedastic setting in Example 2, following the proof of Lemma 2, we can show that Assumption (12a) holds if in addition the supports \mathcal{X} and Ξ of the covariates X and the errors ε are compact and a.s. for n large enough the parameter estimates $\hat{\pi}_n$ of π^* lie in a compact set. We present conditions under which Assumption (12b) holds in Section E of the Appendix (note that it readily holds for our running example of two-stage stochastic LP).

Replacing Assumption 1 with Assumption 12 yields the following analogue of Lemma 1.

Lemma 11. Let Assumption 12 hold. Then, for a.e. $x \in \mathcal{X}$, we a.s. have for n large enough:

$$\sup_{z \in \mathcal{Z}} |\hat{g}_n^{ER}(z; x) - g_n^*(z; x)| \leq \sup_{z \in \mathcal{Z}} \left(\frac{1}{n} \sum_{i=1}^n L_{\delta(x)}^2(z, f^*(x) + Q^*(x)\varepsilon^i) \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|^2 \right)^{1/2}.$$

Proof. We have

$$\begin{aligned} & \sup_{z \in \mathcal{Z}} |\hat{g}_n^{ER}(z; x) - g_n^*(z; x)| \\ &= \sup_{z \in \mathcal{Z}} \left| \frac{1}{n} \sum_{i=1}^n c(z, \text{proj}_{\mathcal{Y}}(\hat{f}_n(x) + \hat{Q}_n(x)\tilde{\varepsilon}_n^i)) - \frac{1}{n} \sum_{i=1}^n c(z, f^*(x) + Q^*(x)\varepsilon^i) \right| \\ &\leq \sup_{z \in \mathcal{Z}} \frac{1}{n} \sum_{i=1}^n |c(z, \text{proj}_{\mathcal{Y}}(\hat{f}_n(x) + \hat{Q}_n(x)\tilde{\varepsilon}_n^i)) - c(z, f^*(x) + Q^*(x)\varepsilon^i)| \\ &\leq \sup_{z \in \mathcal{Z}} \frac{1}{n} \sum_{i=1}^n L_{\delta(x)}(z, f^*(x) + Q^*(x)\varepsilon^i) \|\text{proj}_{\mathcal{Y}}(\hat{f}_n(x) + \hat{Q}_n(x)\tilde{\varepsilon}_n^i) - (f^*(x) + Q^*(x)\varepsilon^i)\| \\ &\leq \sup_{z \in \mathcal{Z}} \left(\frac{1}{n} \sum_{i=1}^n L_{\delta(x)}^2(z, f^*(x) + Q^*(x)\varepsilon^i) \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|^2 \right)^{1/2}, \end{aligned}$$

where the second inequality follows by Assumption 12 and the final inequality follows by the Cauchy-Schwarz inequality and the Lipschitz continuity of orthogonal projections. \square

Comparing with Lemma 1, Proposition 4, and Theorem 5, we see that convergence in probability of the root mean square deviation term $(\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|^2)^{1/2} \xrightarrow{P} 0$ directly translates to the asymptotic guarantees for ER-SAA in Section 3.2 when Assumption 1 is replaced with Assumption 12. The following analogue of Lemma 2, which we state without proof, bounds the root mean square deviation term $(\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|^2)^{1/2}$.

Lemma 12. Given estimates \hat{f}_n of f^* and \hat{Q}_n of Q^* with $[\hat{Q}_n(\bar{x})]^{-1} \succ 0$ for each $\bar{x} \in \mathcal{X}$:

1. In the homoscedastic setting (i.e., $\hat{Q}_n = Q^* \equiv I$), we have

$$\left(\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|^2 \right)^{1/2} \leq \sqrt{2} \|\hat{f}_n(x) - f^*(x)\| + \left(\frac{2}{n} \sum_{i=1}^n \|\hat{f}_n(x^i) - f^*(x^i)\|^2 \right)^{1/2}.$$

2. In the heteroscedastic setting, we have

$$\begin{aligned}
& \left(\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|^2 \right)^{1/2} \\
& \leq O(1) \|\hat{f}_n(x) - f^*(x)\| + O(1) \|\hat{Q}_n(x) - Q^*(x)\| \left(\frac{1}{n} \sum_{i=1}^n \|\varepsilon^i\|^2 \right)^{1/2} + \\
& \quad O(1) \|\hat{Q}_n(x)\| \left(\frac{1}{n} \sum_{i=1}^n \|\hat{Q}_n(x^i)^{-1} - [Q^*(x^i)]^{-1}\|^4 \right)^{1/4} \left(\frac{1}{n} \sum_{i=1}^n \|Q^*(x^i)\|^8 \right)^{1/8} \left(\frac{1}{n} \sum_{i=1}^n \|\varepsilon^i\|^8 \right)^{1/8} + \\
& \quad O(1) \|\hat{Q}_n(x)\| \left(\frac{1}{n} \sum_{i=1}^n \|\hat{Q}_n(x^i)^{-1}\|^4 \right)^{1/4} \left(\frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_n(x^i)\|^4 \right)^{1/4}.
\end{aligned}$$

Proof. Use the AM-QM inequality and follow the arguments in the proof of Lemma 2. \square

Lemma 12 informs how Assumptions 3, 4, and 5 may to be adapted to derive similar theoretical guarantees for ER-SAA as in Section 3.2. Next, we consider the following strengthening of Assumption 12 under which similar finite guarantees as in Section 3.3 hold for the ER-SAA.

Assumption 13. There exists a function $\delta : \mathcal{X} \rightarrow \mathbb{R}_+$ such that for a.e. $x \in \mathcal{X}$, the regression step and the data \mathcal{D}_n satisfy $\|\tilde{\varepsilon}_n^i(x)\| \leq \delta(x)$, $\forall i \in [n]$ and $n \in \mathbb{N}$. Furthermore, Assumption (12b) holds with the local Lipschitz constant satisfying for a.e. $x \in \mathcal{X}$, $\kappa > 0$, and $n \in \mathbb{N}$:

$$\mathbb{P} \left\{ \left(\frac{1}{n} \sum_{i=1}^n \sup_{z \in \mathcal{Z}} L_{\delta(x)}^2(z, f^*(x) + Q^*(x)\varepsilon^i) \right)^{1/2} > \left(\mathbb{E} \left[\sup_{z \in \mathcal{Z}} L_{\delta(x)}^2(z, f^*(x) + Q^*(x)\varepsilon) \right] \right)^{1/2} + \kappa \right\} \leq J_L(\kappa) \exp(-\gamma_L(n, \kappa; x)),$$

where $J_L(\kappa)$, $\gamma_L(n, \kappa; x) > 0$ with $\lim_{n \rightarrow \infty} \gamma_L(n, \kappa; x) = \infty$ for each $\kappa > 0$ and a.e. $x \in \mathcal{X}$.

The discussion following Assumption 12 provides conditions under which $\|\tilde{\varepsilon}_n^i(x)\| \leq \delta(x)$, $\forall i \in [n]$, $n \in \mathbb{N}$, if we further assume the estimates $\hat{\theta}_n$ and $\hat{\pi}_n$ lie in compact sets. Section E identifies conditions under which the uniform Lipschitz condition in Assumption 13 holds (again, it readily holds for our running example of two-stage stochastic LP).

When Assumption 13 is used in place of Assumption 1, we can derive the following analogue of inequality (15) in the proof of Lemma 7:

$$\begin{aligned}
& \mathbb{P} \left\{ \sup_{z \in \mathcal{Z}} |\hat{g}_n^{ER}(z; x) - g_n^*(z; x)| > \kappa \right\} \\
& \leq \mathbb{P} \left\{ \left(\frac{1}{n} \sum_{i=1}^n \sup_{z \in \mathcal{Z}} L_{\delta(x)}^2(z, f^*(x) + Q^*(x)\varepsilon^i) \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|^2 \right)^{1/2} > \kappa \right\} \\
& \leq \mathbb{P} \left\{ \left(\frac{1}{n} \sum_{i=1}^n \sup_{z \in \mathcal{Z}} L_{\delta(x)}^2(z, f^*(x) + Q^*(x)\varepsilon^i) \right)^{1/2} > \left(\mathbb{E} \left[\sup_{z \in \mathcal{Z}} L_{\delta(x)}^2(z, f^*(x) + Q^*(x)\varepsilon) \right] \right)^{1/2} + \kappa \right\} + \\
& \quad \mathbb{P} \left\{ \left[\left(\mathbb{E} \left[\sup_{z \in \mathcal{Z}} L_{\delta(x)}^2(z, f^*(x) + Q^*(x)\varepsilon) \right] \right)^{1/2} + \kappa \right] \left(\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|^2 \right)^{1/2} > \kappa \right\} \\
& \leq J_L(\kappa) \exp(-\gamma_L(n, \kappa; x)) + \mathbb{P} \left\{ \left(\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|^2 \right)^{1/2} > \frac{\kappa}{B_L(\kappa; x)} \right\},
\end{aligned}$$

where $B_L(\kappa; x) := \left(\mathbb{E} \left[\sup_{z \in \mathcal{Z}} L_{\delta(x)}^2(z, f^*(x) + Q^*(x)\varepsilon) \right] \right)^{1/2} + \kappa$ and the last inequality follows by Assumption 13. Therefore, finite sample guarantees for the root mean square deviation term $\left(\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|^2 \right)^{1/2}$ directly translate to finite sample guarantees for ER-SAA of the form in Lemma 7 and Theorem 8. Once again, Lemma 12 provides guidance for how Assumptions 9, 10, and 11 may to be adapted to derive similar theoretical guarantees for ER-SAA as in Section 3.3.

C Rate of convergence of the ER-SAA estimator

We investigate the rate of convergence of the optimal objective value of the sequence of ER-SAA problems (4) to that of the true problem (1). This analysis requires the following additional assumptions on the true problem (1) and the regression step (3).

Assumption 14. The function c in problem (1) and the errors $\{\varepsilon^i\}$ satisfy the following functional central limit theorem (CLT) for the full-information SAA objective (2):

$$\sqrt{n}(g_n^*(\cdot; x) - g(\cdot; x)) \xrightarrow{d} V(\cdot; x), \quad \text{for a.e. } x \in \mathcal{X},$$

where $g_n^*(\cdot; x)$, $g(\cdot; x)$, and $V(\cdot; x)$ are (random) elements of $\mathcal{C}(\mathcal{Z})$, the Banach space of real-valued continuous functions on \mathcal{Z} equipped with the supremum norm.

Assumption 14 holds, for instance, when the errors $\{\varepsilon^i\}$ are i.i.d., the function $c(\cdot, y)$ is Lipschitz continuous on \mathcal{Z} for a.e. $y \in \mathcal{Y}$ with an $L^2(\mathcal{Y})$ Lipschitz constant, and, for a.e. $x \in \mathcal{X}$, there exists $\tilde{z} \in \mathcal{Z}$ such that $\mathbb{E}[(c(\tilde{z}, f^*(x) + Q^*(x)\varepsilon))^2] < +\infty$ (see page 164 of Shapiro et al. [73] for details). Theorem 1 of Doukhan et al. [32], Theorem 2.1 of Arcones and Yu [3], Theorem 9 of Arcones [2], and Corollary 2.3 of Andrews and Pollard [1] provide conditions under which the functional CLT holds under mixing assumptions on $\{\varepsilon^i\}$. Theorems 1.5.4 and 1.5.6 of van der Vaart and Wellner [79] present a general set of conditions under which the functional CLT holds.

The next assumption, which strengthens Assumption 5, ensures that the deviation of the ER-SAA problem (4) from the full-information SAA problem (2) converges at a certain rate.

Assumption 15. There is a constant $0 < \alpha \leq 1$ (that is independent of the number of samples n , but could depend on the dimension d_x of the covariates X) such that the regression estimates \hat{f}_n and \hat{Q}_n satisfy the following convergence rate criteria:

$$(15a) \quad \|\hat{f}_n(x) - f^*(x)\| = O_p(n^{-\alpha/2}) \text{ for a.e. } x \in \mathcal{X},$$

$$(15b) \quad \frac{1}{n} \sum_{i=1}^n \|\hat{f}_n(x^i) - f^*(x^i)\|^2 = O_p(n^{-\alpha}),$$

$$(15c) \quad \|\hat{Q}_n(x) - Q^*(x)\| = O_p(n^{-\alpha/2}) \text{ for a.e. } x \in \mathcal{X},$$

$$(15d) \quad \frac{1}{n} \sum_{i=1}^n \|[\hat{Q}_n(x^i)]^{-1} - [Q^*(x^i)]^{-1}\|^2 = O_p(n^{-\alpha}).$$

Note that the $O_p(\cdot)$ terms in Assumption 15 hide factors proportional to the dimension d_y of the random vector Y . For our running example of OLS regression, Assumptions (15a) and (15b) hold with $\alpha = 1$ under mild assumptions on the data \mathcal{D}_n and the distribution P_X of the covariates [see Chapter 5 of 85]. A similar rate holds for Lasso, best subset selection, and many other parametric regression procedures under mild assumptions. Nonparametric regression procedures such as kNN and RF regression, on the other hand, typically only satisfy these assumptions with constant $\alpha = \frac{O(1)}{d_x}$. This rate cannot be improved upon in general, and is commonly referred to as the curse of dimensionality. Structured nonparametric regression methods such as sparse additive models [63] can hope to break the curse of dimensionality and achieve rates with $\alpha = 1$. Assumptions (15c) and (15d) also hold for our running Example 2 if the parameter estimates $\hat{\pi}_n$ satisfy $\|\hat{\pi}_n - \pi^*\| = O_p(n^{-\alpha/2})$. Section F of the Appendix verifies that Assumption 15 holds for these prediction setups with the stated constants α .

Our main result of this section extends Theorem 5.7 of Shapiro et al. [73] to establish a rate at which the optimal objective value of the ER-SAA problem (4) converges to that of the true problem (1). We hide the dependence of the convergence rate on the dimensions d_x and d_y of the covariates X and random vector Y . We discuss how these dimensions affect the rate of convergence via a non-asymptotic/finite sample analysis in Section 3.3. Note that the convergence rate analysis in Theorem 5.7 of Shapiro et al. [73] for

the full-information SAA problem (2) is sharper than Theorem 13 in the sense that it also characterizes the asymptotic distribution of the optimal objective value, see equations (5.25) and (5.26) therein. Deriving the asymptotic distribution of the optimal value of the ER-SAA is an interesting direction for future work.

Theorem 13. Suppose Assumptions 1, 3, 4, 14 and 15 hold. Then, we have $|\hat{v}_n^{ER}(x) - v^*(x)| = O_p(n^{-\frac{\alpha}{2}})$ and $|g(\hat{z}_n^{ER}(x); x) - \hat{v}_n^{ER}(x)| = O_p(n^{-\frac{\alpha}{2}})$ for a.e. $x \in \mathcal{X}$.

Proof. We begin by showing that $\sup_{z \in \mathcal{Z}} |\hat{g}_n^{ER}(z; x) - g(z; x)| = O_p(n^{-\alpha/2})$ for a.e. $x \in \mathcal{X}$.

Assumption 14 implies $\sqrt{n} \sup_{z \in \mathcal{Z}} |g_n^*(z; x) - g(z; x)| = O_p(1)$ for a.e. $x \in \mathcal{X}$, which in turn implies $\sup_{z \in \mathcal{Z}} |g_n^*(z; x) - g(z; x)| = O_p(n^{-1/2})$ for a.e. $x \in \mathcal{X}$.

We now bound $\sup_{z \in \mathcal{Z}} |\hat{g}_n^{ER}(z; x) - g(z; x)|$ from above using Lemmas 1 and 2. Assumption 1 and Lemma 1 imply

$$\sup_{z \in \mathcal{Z}} |\hat{g}_n^{ER}(z; x) - g_n^*(z; x)| \leq \left(\sup_{z \in \mathcal{Z}} L(z) \right) \left(\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\| \right) = O(1) \left(\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\| \right).$$

Lemma 2 along with Assumptions 3, 4, and 15 and the continuous mapping theorem imply $\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\| = O_p(n^{-\alpha/2})$ for a.e. $x \in \mathcal{X}$. Consequently, $\sup_{z \in \mathcal{Z}} |\hat{g}_n^{ER}(z; x) - g_n^*(z; x)| = O_p(n^{-\alpha/2})$ for a.e. $x \in \mathcal{X}$. Since $\alpha \in (0, 1]$, the above two probability inequalities yield

$$\sup_{z \in \mathcal{Z}} |\hat{g}_n^{ER}(z; x) - g(z; x)| = O_p(n^{-\alpha/2}), \quad \text{for a.e. } x \in \mathcal{X}.$$

This implies that for a.e. $x \in \mathcal{X}$ and any $\beta > 0$, there exists $M_\beta > 0$ such that

$$\mathbb{P} \left\{ \sup_{z \in \mathcal{Z}} |\hat{g}_n^{ER}(z; x) - g(z; x)| > M_\beta n^{-\alpha/2} \right\} < \beta.$$

Consequently, we have for a.e. $x \in \mathcal{X}$:

$$\begin{aligned} \mathbb{P} \left\{ \hat{v}_n^{ER}(x) > v^*(x) + M_\beta n^{-\alpha/2} \right\} &\leq \mathbb{P} \left\{ \hat{g}_n^{ER}(z^*(x); x) > v^*(x) + M_\beta n^{-\alpha/2} \right\} \\ &\leq \mathbb{P} \left\{ |\hat{g}_n^{ER}(z^*(x); x) - v^*(x)| > M_\beta n^{-\alpha/2} \right\} \leq \beta, \\ \mathbb{P} \left\{ v^*(x) > \hat{v}_n^{ER}(x) + M_\beta n^{-\alpha/2} \right\} &\leq \mathbb{P} \left\{ g(\hat{z}_n^{ER}(x); x) > \hat{v}_n^{ER}(x) + M_\beta n^{-\alpha/2} \right\} \\ &\leq \mathbb{P} \left\{ |\hat{v}_n^{ER}(x) - g(\hat{z}_n^{ER}(x); x)| > M_\beta n^{-\alpha/2} \right\} \leq \beta. \end{aligned}$$

Therefore, both $|\hat{v}_n^{ER}(x) - v^*(x)|$ and $|g(\hat{z}_n^{ER}(x); x) - v^*(x)|$ are $O_p(n^{-\alpha/2})$. \square

Note that Assumptions 3, 4, (15c), and (15d) are not required to establish Theorem 13 in the homoscedastic case ($\hat{Q}_n := Q^* \equiv I$). Additionally, Assumption (15b) may be weakened in this setting to $\frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_n(x^i)\| = O_p(n^{-\alpha/2})$ on account of Lemma 2. When Assumption 1 is replaced with Assumption 12, rates of convergence for the root mean square deviation term $(\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|^2)^{1/2} \xrightarrow{p} 0$ directly translate to rates of convergence of the ER-SAA estimators; similar to Theorem 13 (cf. Section B of the Appendix).

D Analysis for the jackknife and jackknife+ estimators

In this section, we analyze the consistency, rate of convergence, and finite sample guarantees of the J-SAA and J+-SAA estimators obtained by solving problems (7) and (8), respectively, under certain assumptions on the true problem (1) and the prediction step (3). We omit proofs because they are similar to the proofs of

results in Section 3. In place of the sequence of deviation terms $\{\tilde{\varepsilon}_n^i(x)\}$ considered in Section 3, we consider the following deviation sequences $\{\tilde{\varepsilon}_n^{i,J}(x)\}$ and $\{\tilde{\varepsilon}_n^{i,J^+}(x)\}$:

$$\begin{aligned}\tilde{\varepsilon}_n^{i,J}(x) &:= \left(\hat{f}_n(x) + \hat{Q}_n(x)\tilde{\varepsilon}_{n,J}^i\right) - (f^*(x) + Q^*(x)\varepsilon^i), \quad \forall i \in [n], \\ \tilde{\varepsilon}_n^{i,J^+}(x) &:= \left(\hat{f}_{-i}(x) + \hat{Q}_{-i}(x)\tilde{\varepsilon}_{n,J}^i\right) - (f^*(x) + Q^*(x)\varepsilon^i), \quad \forall i \in [n].\end{aligned}$$

We let $\hat{z}_n^J(x)$ and $\hat{z}_n^{J^+}(x)$ denote an optimal solution to problem (7) and (8), respectively, and $\hat{S}_n^J(x)$ and $\hat{S}_n^{J^+}(x)$ denote the corresponding sets of optimal solutions. We assume throughout that the sets $\hat{S}_n^J(x)$ and $\hat{S}_n^{J^+}(x)$ are nonempty for a.e. $x \in \mathcal{X}$.

We have the following analogue of Lemma 2 for the jackknife-based mean deviation terms.

Lemma 14. Given regression estimates $\hat{f}_n, \{\hat{f}_{-i}\}$ of f^* and $\hat{Q}_n, \{\hat{Q}_{-i}\}$ of Q^* with $[\hat{Q}_n(\bar{x})]^{-1} \succ 0$ and $[\hat{Q}_{-i}(\bar{x})]^{-1} \succ 0$ for each $i \in [n]$ and $\bar{x} \in \mathcal{X}$:

1. In the homoscedastic setting (i.e., $\hat{Q}_n = \hat{Q}_{-i} = Q^* \equiv I$), we have

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^{i,J}(x)\| &\leq \|\hat{f}_n(x) - f^*(x)\| + \frac{1}{n} \sum_{i=1}^n \|\hat{f}_{-i}(x^i) - f^*(x^i)\|, \\ \frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^{i,J^+}(x)\| &\leq \frac{1}{n} \sum_{i=1}^n \|\hat{f}_{-i}(x) - f^*(x)\| + \frac{1}{n} \sum_{i=1}^n \|\hat{f}_{-i}(x^i) - f^*(x^i)\|.\end{aligned}$$

2. In the heteroscedastic setting, we have

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^{i,J}(x)\| &\leq \\ &\|\hat{f}_n(x) - f^*(x)\| + \|\hat{Q}_n(x) - Q^*(x)\| \left(\frac{1}{n} \sum_{i=1}^n \|\varepsilon^i\|\right) + \\ &\|\hat{Q}_n(x)\| \left(\frac{1}{n} \sum_{i=1}^n \|[\hat{Q}_{-i}(x^i)]^{-1} - [Q^*(x^i)]^{-1}\|^2\right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n \|Q^*(x^i)\|^4\right)^{1/4} \left(\frac{1}{n} \sum_{i=1}^n \|\varepsilon^i\|^4\right)^{1/4} + \\ &\|\hat{Q}_n(x)\| \left(\frac{1}{n} \sum_{i=1}^n \|[\hat{Q}_{-i}(x^i)]^{-1}\|^2\right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_{-i}(x^i)\|^2\right)^{1/2}\end{aligned}$$

and

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n \|\tilde{\varepsilon}_n^{i,J^+}(x)\| &\leq \\ &\frac{1}{n} \sum_{i=1}^n \|\hat{f}_{-i}(x) - f^*(x)\| + \left(\frac{1}{n} \sum_{i=1}^n \|\hat{Q}_{-i}(x) - Q^*(x)\|^2\right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n \|\varepsilon^i\|^2\right)^{1/2} + \\ &\left(\frac{1}{n} \sum_{i=1}^n \|\hat{Q}_{-i}(x)\|^2\right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n \|[\hat{Q}_{-i}(x^i)]^{-1} - [Q^*(x^i)]^{-1}\|^4\right)^{1/4} \left(\frac{1}{n} \sum_{i=1}^n \|Q^*(x^i)\|^8\right)^{1/8} \left(\frac{1}{n} \sum_{i=1}^n \|\varepsilon^i\|^8\right)^{1/8} + \\ &\left(\frac{1}{n} \sum_{i=1}^n \|\hat{Q}_{-i}(x)\|^2\right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n \|[\hat{Q}_{-i}(x^i)]^{-1}\|^4\right)^{1/4} \left(\frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_{-i}(x^i)\|^4\right)^{1/4}.\end{aligned}$$

Therefore, assumptions on the quantities appearing in Lemma 2 for the mean deviation term of the ER-SAA may be replaced with assumptions on the quantities appearing in the above inequalities to derive similar results for the jackknife-based estimators as the ER-SAA estimator. While we focus on analyzing the jackknife-based SAAs when Assumption 1 holds, note that a similar analysis can be carried out when Assumptions 12 and 13 are adapted for the jackknife-based SAAs by following the arguments in Section B. We omit these details for brevity.

D.1 Consistency and asymptotic optimality

We present conditions under which the optimal value and optimal solutions to the J-SAA and J+-SAA problems (7) and (8) asymptotically converge to those of the true problem (1). We make the following assumptions on the consistency of the (leave-one-out version of the) regression procedure (3) that adapts Assumption 5 for the J-SAA and J+-SAA approaches.

Assumption 5J. *The regression estimates \hat{f}_n and $\{\hat{f}_{-i}\}$ of f^* and the regression estimates \hat{Q}_n and $\{\hat{Q}_{-i}\}$ of Q^* satisfy the following consistency properties:*

$$(5Ja) \quad \hat{f}_n(x) \xrightarrow{P} f^*(x) \text{ for a.e. } x \in \mathcal{X}, \quad (5Jc) \quad \hat{Q}_n(x) \xrightarrow{P} Q^*(x) \text{ for a.e. } x \in \mathcal{X},$$

$$(5Jb) \quad \frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_{-i}(x^i)\|^2 \xrightarrow{P} 0, \quad (5Jd) \quad \frac{1}{n} \sum_{i=1}^n \left\| [\hat{Q}_{-i}(x^i)]^{-1} - [Q^*(x^i)]^{-1} \right\|^2 \xrightarrow{P} 0.$$

Assumption 5J+. *The regression estimates \hat{f}_n and $\{\hat{f}_{-i}\}$ of f^* and the regression estimates \hat{Q}_n and $\{\hat{Q}_{-i}\}$ of Q^* satisfy the following consistency properties:*

$$(5J+a) \quad \frac{1}{n} \sum_{i=1}^n \|f^*(x) - \hat{f}_{-i}(x)\| \xrightarrow{P} 0 \text{ for a.e. } x \in \mathcal{X}, \quad (5J+c) \quad \frac{1}{n} \sum_{i=1}^n \|Q^*(x) - \hat{Q}_{-i}(x)\|^2 \xrightarrow{P} 0 \text{ for a.e. } x \in \mathcal{X},$$

$$(5J+b) \quad \frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_{-i}(x^i)\|^4 \xrightarrow{P} 0, \quad (5J+d) \quad \frac{1}{n} \sum_{i=1}^n \left\| [\hat{Q}_{-i}(x^i)]^{-1} - [Q^*(x^i)]^{-1} \right\|^4 \xrightarrow{P} 0.$$

Section F identifies conditions under which Assumptions 5J and 5J+ hold for OLS, Lasso, kNN, and RF regression with i.i.d. data $\{(x^i, \varepsilon^i)\}$. We also require the following strengthening of Assumptions 3 and 4 for the J+-SAA problem.

Assumption 3J+. *The function Q^* and the data $\{x^i\}$ satisfy the weak LLNs*

$$\frac{1}{n} \sum_{i=1}^n \|Q^*(x^i)\|^8 \xrightarrow{P} \mathbb{E}[\|Q^*(X)\|^8] \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^n \left\| [Q^*(x^i)]^{-1} \right\|^4 \xrightarrow{P} \mathbb{E}[\| [Q^*(X)]^{-1} \|^4].$$

Assumption 4J+. *The error samples $\{\varepsilon^i\}_{i=1}^n$ satisfy the weak LLN $\frac{1}{n} \sum_{i=1}^n \|\varepsilon^i\|^8 \xrightarrow{P} \mathbb{E}[\|\varepsilon\|^8]$.*

We now state conditions under which the sequence of objective functions of problems (7) and (8) converge uniformly to the objective function of the true problem (1) on the set \mathcal{Z} . We group the results for the J-SAA and J+-SAA problems for brevity (the individual results are apparent).

Proposition 15. *Suppose Assumptions 1 through 4 and Assumptions 3J+, 4J+, 5J, and 5J+ hold. Then, for a.e. $x \in \mathcal{X}$, the sequences of objective functions of J-SAA and J+-SAA problems (7) and (8) converge uniformly in probability to the objective function of the true problem (1) on the feasible region \mathcal{Z} .*

Proposition 15 helps us establish conditions under which the optimal objective values and solutions of the J-SAA and J+-SAA problems (7) and (8) converge to those of the true problem (1).

Theorem 16. *Suppose Assumptions 1 through 4 and Assumptions 3J+, 4J+, 5J, and 5J+ hold. Then, we have $\hat{v}_n^J(x) \xrightarrow{P} v^*(x)$, $\hat{v}_n^{J+}(x) \xrightarrow{P} v^*(x)$, $\mathbb{D}(\hat{S}_n^J(x), S^*(x)) \xrightarrow{P} 0$, $\mathbb{D}(\hat{S}_n^{J+}(x), S^*(x)) \xrightarrow{P} 0$, $\sup_{z \in \hat{S}_n^J(x)} g(z; x) \xrightarrow{P} v^*(x)$, and $\sup_{z \in \hat{S}_n^{J+}(x)} g(z; x) \xrightarrow{P} v^*(x)$ for a.e. $x \in \mathcal{X}$.*

Assumptions 3, 4, 3J+, 4J+, (5Jc), (5J+c), (5Jd), and (5J+d) are not required to establish Proposition 15 and Theorem 16 in the homoscedastic case (assumptions that only involve \hat{Q}_n and $\{\hat{Q}_{-i}\}$ may be omitted in this setting). Additionally, Assumptions (5Jb) and (5J+b) may be weakened in this setting to $\frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_{-i}(x^i)\| \xrightarrow{P} 0$ on account of Lemma 14.

D.2 Rates of convergence

We derive rates of convergence of the optimal objective value of the sequence of J-SAA and J+-SAA problems (7) and (8) to the optimal objective value of the true problem (1). To enable this, we make the following assumptions on the regression procedure (3) that adapt Assumption 15 to strengthen Assumptions 5J and 5J+. Assumptions 15J and 15J+ ensure that the deviations of the J-SAA and J+-SAA problems from the FI-SAA problem (2) converge at a certain rate.

Assumption 15J. *There is a constant $0 < \alpha \leq 1$ (that is independent of the number of samples n , but could depend on the dimension d_x of the covariates X) such that the regression procedure (3) satisfies the following asymptotic convergence rate criterion:*

$$(15Ja) \quad \|\hat{f}_n(x) - f^*(x)\| = O_p(n^{-\alpha/2}) \text{ for a.e. } x \in \mathcal{X},$$

$$(15Jb) \quad \frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_{-i}(x^i)\|^2 = O_p(n^{-\alpha}),$$

$$(15Jc) \quad \|\hat{Q}_n(x) - Q^*(x)\| = O_p(n^{-\alpha/2}) \text{ for a.e. } x \in \mathcal{X},$$

$$(15Jd) \quad \frac{1}{n} \sum_{i=1}^n \|\hat{Q}_{-i}(x^i)^{-1} - [Q^*(x^i)]^{-1}\|^2 = O_p(n^{-\alpha}).$$

Assumption 15J+. *There is a constant $0 < \alpha \leq 1$ (that is independent of the number of samples n , but could depend on the dimension d_x of the covariates X) such that the regression procedure (3) satisfies the following asymptotic convergence rate criterion:*

$$(15J+a) \quad \frac{1}{n} \sum_{i=1}^n \|f^*(x) - \hat{f}_{-i}(x)\| = O_p(n^{-\alpha/2}) \text{ for a.e. } x \in \mathcal{X},$$

$$(15J+b) \quad \frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_{-i}(x^i)\|^4 = O_p(n^{-2\alpha}),$$

$$(15J+c) \quad \frac{1}{n} \sum_{i=1}^n \|Q^*(x) - \hat{Q}_{-i}(x)\|^2 = O_p(n^{-\alpha}) \text{ for a.e. } x \in \mathcal{X},$$

$$(15J+d) \quad \frac{1}{n} \sum_{i=1}^n \|\hat{Q}_{-i}(x^i)^{-1} - [Q^*(x^i)]^{-1}\|^4 = O_p(n^{-2\alpha}).$$

Section F demonstrates that Assumptions 15J and 15J+ hold with rates similar to those in Assumption 15 when the data $\{(x^i, \varepsilon^i)\}$ is i.i.d. Along with Lemma 14, Assumptions 15J and 15J+ imply that the mean deviation terms for the J-SAA and J+-SAA approaches can be bounded as $\frac{1}{n} \sum_{i=1}^n \|\hat{\varepsilon}_n^{i,J}(x)\| = O_p(n^{-\alpha})$ and $\frac{1}{n} \sum_{i=1}^n \|\hat{\varepsilon}_n^{i,J+}(x)\| = O_p(n^{-\alpha})$ for a.e. $x \in \mathcal{X}$.

We now establish rates at which the optimal objective value of the J-SAA and J+-SAA problems converge to the optimal objective value of the true problem (1). We hide the dependence of the convergence rate on the dimensions d_x and d_y of the covariates X and random vector Y . The analysis in the next section can account for how these dimensions affect the rate of convergence.

Theorem 17. Suppose Assumptions 1, 3, 4, 3J+, 4J+, 15J, and 15J+ hold. Then, for a.e. $x \in \mathcal{X}$, we have $|\hat{v}_n^J(x) - v^*(x)| = O_p(n^{-\frac{\alpha}{2}})$, $|\hat{v}_n^{J+}(x) - v^*(x)| = O_p(n^{-\frac{\alpha}{2}})$, $|g(\hat{z}_n^J(x); x) - v^*(x)| = O_p(n^{-\frac{\alpha}{2}})$, and $|g(\hat{z}_n^{J+}(x); x) - v^*(x)| = O_p(n^{-\frac{\alpha}{2}})$.

Assumptions 3, 4, 3J+, 4J+, (15Jc), (15J+c), (15Jd), and (15J+d) are not required to establish Theorem 17 in the homoscedastic case ($\hat{Q}_n = \hat{Q}_{-i} = Q^* \equiv I$). Additionally, Assumptions (15Jb) and (15J+b) may be weakened in this setting to $\frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_{-i}(x^i)\| = O_p(n^{-\alpha/2})$ on account of Lemma 14.

D.3 Finite sample guarantees

We now establish exponential convergence of solutions to the J-SAA and J+-SAA problems to solutions to the true problem (1) under additional assumptions. We begin by adapting Assumption 11 to assume that the regression procedure (3) satisfies the following large deviation properties.

Assumption 11J. *The regression estimates \hat{f}_n , $\{\hat{f}_{-i}\}$, \hat{Q}_n , and $\{\hat{Q}_{-i}\}$ possess the following large deviation properties: for any constant $\kappa > 0$ and $n \in \mathbb{N}$, there exist positive constants $K_f^J(\kappa, x)$, $\bar{K}_f^J(\kappa)$, $\beta_f^J(n, \kappa, x)$, $\bar{\beta}_f^J(n, \kappa)$, $K_Q^J(\kappa, x)$, $\bar{K}_Q^J(\kappa)$, $\beta_Q^J(n, \kappa, x)$, and $\bar{\beta}_Q^J(n, \kappa)$, with $\lim_{n \rightarrow \infty} \beta_f^J(n, \kappa, x) = \infty$, $\lim_{n \rightarrow \infty} \bar{\beta}_f^J(n, \kappa) = \infty$, $\lim_{n \rightarrow \infty} \beta_Q^J(n, \kappa, x) = \infty$, and $\lim_{n \rightarrow \infty} \bar{\beta}_Q^J(n, \kappa) = \infty$ for each $\kappa > 0$ and a.e. $x \in \mathcal{X}$, satisfying*

$$(11Ja) \quad \mathbb{P}\{\|f^*(x) - \hat{f}_n(x)\| > \kappa\} \leq K_f^J(\kappa, x) \exp(-\beta_f^J(n, \kappa, x)) \text{ for a.e. } x \in \mathcal{X},$$

$$(11Jb) \quad \mathbb{P}\left\{\frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_{-i}(x^i)\|^2 > \kappa^2\right\} \leq \bar{K}_f^J(\kappa) \exp(-\bar{\beta}_f^J(n, \kappa)),$$

$$(11Jc) \quad \mathbb{P}\{\|Q^*(x) - \hat{Q}_n(x)\| > \kappa\} \leq K_Q^J(\kappa, x) \exp(-\beta_Q^J(n, \kappa, x)) \text{ for a.e. } x \in \mathcal{X},$$

$$(11Jd) \quad \mathbb{P}\left\{\frac{1}{n} \sum_{i=1}^n \|[\hat{Q}_{-i}(x^i)]^{-1} - [Q^*(x^i)]^{-1}\|^2 > \kappa^2\right\} \leq \bar{K}_Q^J(\kappa) \exp(-\bar{\beta}_Q^J(n, \kappa)).$$

Assumption 11J+. *The regression estimates \hat{f}_n , $\{\hat{f}_{-i}\}$, \hat{Q}_n , and $\{\hat{Q}_{-i}\}$ possess the following large deviation properties: for any constant $\kappa > 0$ and $n \in \mathbb{N}$, there exist positive constants $K_f^{J+}(\kappa, x)$, $\bar{K}_f^{J+}(\kappa)$, $\beta_f^{J+}(n, \kappa, x)$, $\bar{\beta}_f^{J+}(n, \kappa)$, $K_Q^{J+}(\kappa, x)$, $\bar{K}_Q^{J+}(\kappa)$, $\beta_Q^{J+}(n, \kappa, x)$, and $\bar{\beta}_Q^{J+}(n, \kappa)$, with $\lim_{n \rightarrow \infty} \beta_f^{J+}(n, \kappa, x) = \infty$, $\lim_{n \rightarrow \infty} \bar{\beta}_f^{J+}(n, \kappa) = \infty$, $\lim_{n \rightarrow \infty} \beta_Q^{J+}(n, \kappa, x) = \infty$, and $\lim_{n \rightarrow \infty} \bar{\beta}_Q^{J+}(n, \kappa) = \infty$ for each $\kappa > 0$ and a.e. $x \in \mathcal{X}$, satisfying*

$$(11J+a) \quad \mathbb{P}\left\{\frac{1}{n} \sum_{i=1}^n \|f^*(x) - \hat{f}_{-i}(x)\| > \kappa\right\} \leq K_f^{J+}(\kappa, x) \exp(-\beta_f^{J+}(n, \kappa, x)) \text{ for a.e. } x \in \mathcal{X},$$

$$(11J+b) \quad \mathbb{P}\left\{\frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_{-i}(x^i)\|^4 > \kappa^4\right\} \leq \bar{K}_f^{J+}(\kappa) \exp(-\bar{\beta}_f^{J+}(n, \kappa)),$$

$$(11J+c) \quad \mathbb{P}\left\{\frac{1}{n} \sum_{i=1}^n \|Q^*(x) - \hat{Q}_{-i}(x)\|^2 > \kappa^2\right\} \leq K_Q^{J+}(\kappa, x) \exp(-\beta_Q^{J+}(n, \kappa, x)) \text{ for a.e. } x \in \mathcal{X},$$

$$(11J+d) \quad \mathbb{P}\left\{\frac{1}{n} \sum_{i=1}^n \|[\hat{Q}_{-i}(x^i)]^{-1} - [Q^*(x^i)]^{-1}\|^4 > \kappa^4\right\} \leq \bar{K}_Q^{J+}(\kappa) \exp(-\bar{\beta}_Q^{J+}(n, \kappa)).$$

Assumptions 11J and 11J+ strengthen Assumptions 15J and 15J+ by imposing restrictions on the tails of the regression estimators. Please see the discussion in Section F for when these strengthened assumptions are satisfied with i.i.d. data $\{(x^i, \varepsilon^i)\}$. We also require the following strengthening of Assumptions 9 and 10 for the J+-SAA problem.

Assumption 9J+. *For any $\kappa > 0$ and $n \in \mathbb{N}$, there exist positive constants $L_Q^{J+}(\kappa)$, $\gamma_Q^{J+}(n, \kappa)$, $\bar{L}_Q^{J+}(\kappa)$, and $\bar{\gamma}_Q^{J+}(n, \kappa)$, with $\lim_{n \rightarrow \infty} \gamma_Q^{J+}(n, \kappa) = \infty$ and $\lim_{n \rightarrow \infty} \bar{\gamma}_Q^{J+}(n, \kappa) = \infty$ for each $\kappa > 0$, such that*

$$\begin{aligned} \mathbb{P}\left\{\left(\frac{1}{n} \sum_{i=1}^n \|[Q^*(x^i)]^{-1}\|^4\right)^{1/4} > \left(\mathbb{E} \left[\|[Q^*(X)]^{-1}\|^4\right]\right)^{1/4} + \kappa\right\} &\leq L_Q^{J+}(\kappa) \exp(-\gamma_Q^{J+}(n, \kappa)), \\ \mathbb{P}\left\{\left(\frac{1}{n} \sum_{i=1}^n \|Q^*(x^i)\|^8\right)^{1/8} > \left(\mathbb{E} \left[\|Q^*(X)\|^8\right]\right)^{1/8} + \kappa\right\} &\leq \bar{L}_Q^{J+}(\kappa) \exp(-\bar{\gamma}_Q^{J+}(n, \kappa)). \end{aligned}$$

Assumption 10J+. For any $\kappa > 0$ and $n \in \mathbb{N}$, there exist positive constants $L_\varepsilon^{J+}(\kappa)$, $\gamma_\varepsilon^{J+}(n, \kappa)$, $\bar{L}_\varepsilon^{J+}(\kappa)$, and $\bar{\gamma}_\varepsilon^{J+}(n, \kappa)$, with $\lim_{n \rightarrow \infty} \gamma_\varepsilon^{J+}(n, \kappa) = \infty$ and $\lim_{n \rightarrow \infty} \bar{\gamma}_\varepsilon^{J+}(n, \kappa) = \infty$ for each $\kappa > 0$, such that

$$\begin{aligned} \mathbb{P} \left\{ \left(\frac{1}{n} \sum_{i=1}^n \|\varepsilon^i\|^2 \right)^{1/2} > \left(\mathbb{E} [\|\varepsilon\|^2] \right)^{1/2} + \kappa \right\} &\leq L_\varepsilon^{J+}(\kappa) \exp(-\gamma_\varepsilon^{J+}(n, \kappa)), \\ \mathbb{P} \left\{ \left(\frac{1}{n} \sum_{i=1}^n \|\varepsilon^i\|^8 \right)^{1/8} > \left(\mathbb{E} [\|\varepsilon\|^8] \right)^{1/8} + \kappa \right\} &\leq \bar{L}_\varepsilon^{J+}(\kappa) \exp(-\bar{\gamma}_\varepsilon^{J+}(n, \kappa)). \end{aligned}$$

The next result presents conditions under which the maximum deviations of the J-SAA and J+-SAA objectives from the FI-SAA objective satisfy qualitatively similar large deviations bounds as that in Assumption 8.

Lemma 18. Suppose Assumptions, 1, 8, 9, 10, 9J+, 10J+, 11J and 11J+ hold. Then for any constant $\kappa > 0$, $n \in \mathbb{N}$, and a.e. $x \in \mathcal{X}$, there exist positive constants $\bar{K}^J(\kappa, x)$, $\bar{K}^{J+}(\kappa, x)$, $\bar{\beta}^J(n, \kappa, x)$, and $\bar{\beta}^{J+}(n, \kappa, x)$, with $\lim_{n \rightarrow \infty} \bar{\beta}^J(n, \kappa, x) = \infty$ and $\lim_{n \rightarrow \infty} \bar{\beta}^{J+}(n, \kappa, x) = \infty$ for each $\kappa > 0$ and a.e. $x \in \mathcal{X}$, satisfying

$$\begin{aligned} \mathbb{P} \left\{ \sup_{z \in \mathcal{Z}} |\hat{g}_n^J(z; x) - g_n^*(z; x)| > \kappa \right\} &\leq \bar{K}^J(\kappa, x) \exp(-\bar{\beta}^J(n, \kappa, x)), \quad \text{and} \\ \mathbb{P} \left\{ \sup_{z \in \mathcal{Z}} |\hat{g}_n^{J+}(z; x) - g_n^*(z; x)| > \kappa \right\} &\leq \bar{K}^{J+}(\kappa, x) \exp(-\bar{\beta}^{J+}(n, \kappa, x)). \end{aligned}$$

We now finite sample guarantees for the distances between solutions to the J-SAA and J+-SAA problems (7) and (8) and the set of optimal solutions to the true problem (1).

Theorem 19. Suppose Assumptions, 1, 8, 9, 10, 9J+, 10J+, 11J and 11J+ hold. Then, for each $n \in \mathbb{N}$ and a.e. $x \in \mathcal{X}$, given $\eta > 0$, there exist positive constants $Q^J(\eta, x)$, $Q^{J+}(\eta, x)$, $\gamma^J(n, \eta, x)$, and $\gamma^{J+}(n, \eta, x)$, with $\lim_{n \rightarrow \infty} \gamma^J(n, \eta, x) = \infty$ and $\lim_{n \rightarrow \infty} \gamma^{J+}(n, \eta, x) = \infty$ for each $\eta > 0$ and a.e. $x \in \mathcal{X}$, such that

$$\begin{aligned} \mathbb{P} \left\{ \text{dist}(\hat{z}_n^J(x), S^*(x)) \geq \eta \right\} &\leq Q^J(\eta, x) \exp(-\gamma^J(n, \eta, x)), \quad \text{and} \\ \mathbb{P} \left\{ \text{dist}(\hat{z}_n^{J+}(x), S^*(x)) \geq \eta \right\} &\leq Q^{J+}(\eta, x) \exp(-\gamma^{J+}(n, \eta, x)). \end{aligned}$$

Assumptions 9, 10, 9J+, 10J+, (11Jc), (11J+c), (11Jd), and (11J+d) are not required to establish Theorem 19 in the homoscedastic case. Additionally, Assumptions (11Jb) and (11J+b) may be weakened in this setting to $\mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_{-i}(x^i)\| > \kappa \right\} \leq \bar{K}_f^J(\kappa) \exp(-\bar{\beta}_f^J(n, \kappa))$ and $\mathbb{P} \left\{ \frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_{-i}(x^i)\| > \kappa \right\} \leq \bar{K}_f^{J+}(\kappa) \exp(-\bar{\beta}_f^{J+}(n, \kappa))$ on account of Lemma 14.

E Application to two-stage stochastic programming problems

We present a class of stochastic programs that satisfy Assumptions 1, 2, 6, 8, 12, 13, and 14. We first consider a class of two-stage stochastic programs with continuous recourse decisions that subsumes Example 1, our running example of two-stage stochastic LP. We then briefly outline the verification of these assumptions for a broader class of stochastic programs.

Consider first the two-stage stochastic program

$$\min_{z \in \mathcal{Z}} \mathbb{E} [c(z, Y)] := p(z) + \mathbb{E} [V(z, Y)], \quad (21)$$

where the second-stage function V is defined by the optimal value of the following LP:

$$V(z, y) := \min_{v \in \mathbb{R}_+^{d_v}} \left\{ c_v^\top v : Wv = h(y) - T(y, z) \right\}.$$

We make the following assumptions on problem (21).

Assumption 16. The set \mathcal{Z} is nonempty and compact, the matrix W has full row rank, the set $\Lambda := \{\lambda : \lambda^\top W \leq c_v^\top\}$ is nonempty, and the value function $V(z, y) < +\infty$ for each $(z, y) \in \mathcal{Z} \times \mathcal{Y}$.

Assumption 17. The functions h and $T(\cdot, z)$ are Lipschitz continuous on \mathcal{Y} for each $z \in \mathcal{Z}$ with Lipschitz constants L_h and $L_{T,y}(z)$, and the functions p and $T(y, \cdot)$ are Lipschitz continuous on \mathcal{Z} for each $y \in \mathcal{Y}$ with Lipschitz constants L_p and $L_{T,z}(y)$. Additionally, the Lipschitz constants for the function T satisfy $\sup_{z \in \mathcal{Z}} L_{T,y}(z) < +\infty$ and $\sup_{y \in \mathcal{Y}} L_{T,z}(y) < +\infty$.

Assumption 18. The support \mathcal{X} of the covariates X is compact and the functions f^* and Q^* are continuous on \mathcal{X} . Additionally, $\mathbb{E} \left[\sup_{(z,x) \in \mathcal{Z} \times \mathcal{X}} \|h(f^*(x) + Q^*(x)\varepsilon) - T(f^*(x) + Q^*(x)\varepsilon, z)\| \right] < +\infty$.

Let $\text{vert}(\Lambda)$ denote the finite set of extreme points of the dual feasible region Λ , and define $\mathcal{V}(Y, \lambda, z) := \lambda^\top (h(Y) - T(Y, z))$ for each $Y \in \mathcal{Y}$, $\lambda \in \text{vert}(\Lambda)$, and $z \in \mathcal{Z}$.

Assumption 19. We have $\mathbb{E} \left[\sup_{z \in \mathcal{Z}} \|h(Y) - T(Y, z)\|^2 \right] < +\infty$. Additionally, the random variable $\mathcal{V}(f^*(x) + Q^*(x)\varepsilon, \lambda, z) - \mathbb{E}_{\varepsilon \sim P_\varepsilon} [\mathcal{V}(f^*(x) + Q^*(x)\varepsilon, \lambda, z)]$ is sub-Gaussian with variance proxy $\sigma_c^2(x)$ for each $\lambda \in \text{vert}(\Lambda)$, $z \in \mathcal{Z}$ and a.e. $x \in \mathcal{X}$.

Note that the first-stage feasible set \mathcal{Z} can include integrality constraints. Our running example of two-stage stochastic LP with OLS regression fits within the above setup and readily satisfies Assumptions 16 and 17. It also satisfies Assumption 18 when $\mathbb{E} \|\varepsilon\| < +\infty$ and Assumption 19 when the error ε is sub-Gaussian. Additionally, under Assumption 16, we have by LP duality that for each $y \in \mathcal{Y}$:

$$V(z, y) = \max_{\lambda \in \Lambda} \lambda^\top (h(y) - T(y, z)) = \max_{\lambda \in \text{vert}(\Lambda)} \lambda^\top (h(y) - T(y, z)). \quad (22)$$

Proposition 20. Suppose Assumptions 16, 17, 18, and 19 hold and the data $\{\varepsilon^i\}$ is i.i.d. Then, problem (21) satisfies Assumptions 1, 2, 6, 8, and 14.

Proof. We have by Assumptions 16 and 17 that for any $y, \bar{y} \in \mathcal{Y}$ and $z \in \mathcal{Z}$:

$$\begin{aligned} |c(z, y) - c(z, \bar{y})| &= \left| \max_{\lambda \in \text{vert}(\Lambda)} \lambda^\top (h(y) - T(y, z)) - \max_{\lambda \in \text{vert}(\Lambda)} \lambda^\top (h(\bar{y}) - T(\bar{y}, z)) \right| \\ &\leq \max_{\lambda \in \text{vert}(\Lambda)} \left| \lambda^\top (h(y) - h(\bar{y})) + \lambda^\top (T(\bar{y}, z) - T(y, z)) \right| \\ &\leq \max_{\lambda \in \text{vert}(\Lambda)} \|\lambda\| \|h(y) - h(\bar{y})\| + \max_{\lambda \in \text{vert}(\Lambda)} \|\lambda\| \|T(\bar{y}, z) - T(y, z)\| \\ &\leq [L_h + L_{T,y}(z)] \left(\max_{\lambda \in \text{vert}(\Lambda)} \|\lambda\| \right) \|y - \bar{y}\|. \end{aligned}$$

Therefore, the Lipschitz continuity Assumption 1 holds since $\sup_{z \in \mathcal{Z}} L_{T,y}(z) < +\infty$ and $\max_{\lambda \in \text{vert}(\Lambda)} \|\lambda\| < +\infty$.

The function $c(\cdot, y)$ is continuous on \mathcal{Z} for each $y \in \mathcal{Y}$ by virtue of Assumptions 16 and 17 since p is Lipschitz continuous on \mathcal{Z} and equation (22) implies that $V(\cdot, y)$ is a finite maximum of continuous functions for each $y \in \mathcal{Y}$. Furthermore, for any $\bar{z} \in \mathcal{Z}$:

$$\begin{aligned} |c(\bar{z}, Y)| &= \left| p(\bar{z}) + \max_{\lambda \in \text{vert}(\Lambda)} \lambda^\top (h(Y) - T(Y, \bar{z})) \right| \\ &\leq \max_{z \in \mathcal{Z}} |p(z)| + \left(\max_{\lambda \in \text{vert}(\Lambda)} \|\lambda\| \right) \left(\max_{z \in \mathcal{Z}} \|h(Y) - T(Y, z)\| \right). \end{aligned}$$

The uniform weak LLN Assumption 2 then holds by virtue of Assumption 16, the first part of Assumption 19, and Theorem 7.48 of Shapiro et al. [73], which also implies that the objective function of the true problem (1) is continuous on \mathcal{Z} .

The function $c(\cdot, f^*(\cdot) + Q^*(\cdot)\varepsilon)$ is continuous on $\mathcal{Z} \times \mathcal{X}$ for each $\varepsilon \in \Xi$ by virtue of Assumptions 16, 17, and 18 since p is Lipschitz continuous on \mathcal{Z} , $c(\cdot, y)$ is Lipschitz continuous on \mathcal{Z} with a Lipschitz constant independent of $y \in \mathcal{Y}$ (see below), and equation (22) implies that $V(\cdot, f^*(\cdot) + Q^*(\cdot)\varepsilon)$ is a finite maximum of continuous functions for each $\varepsilon \in \Xi$. Additionally, for any $(\bar{z}, \bar{x}) \in \mathcal{Z} \times \mathcal{X}$:

$$\begin{aligned} & |c(\bar{z}, f^*(\bar{x}) + Q^*(\bar{x})\varepsilon)| \\ &= \left| p(\bar{z}) + \max_{\lambda \in \text{vert}(\Lambda)} \lambda^\top (h(f^*(\bar{x}) + Q^*(\bar{x})\varepsilon) - T(f^*(\bar{x}) + Q^*(\bar{x})\varepsilon, \bar{z})) \right| \\ &\leq \max_{z \in \mathcal{Z}} |p(z)| + \left(\max_{\lambda \in \text{vert}(\Lambda)} \|\lambda\| \right) \left(\max_{(z,x) \in \mathcal{Z} \times \mathcal{X}} \|h(f^*(x) + Q^*(x)\varepsilon) - T(f^*(x) + Q^*(x)\varepsilon, z)\| \right). \end{aligned}$$

Therefore, Assumptions 16, 17, and 18 along with Theorem 7.48 of Shapiro et al. [73] together implies that $\sup_{(z,x) \in \mathcal{Z} \times \mathcal{X}} |g_n^*(z; x) - g(z; x)| \xrightarrow{P} 0$, which in turn implies Assumption 6.

Next, note that for any $z, \bar{z} \in \mathcal{Z}$:

$$\begin{aligned} |c(z, Y) - c(\bar{z}, Y)| &= \left| p(z) + \max_{\lambda \in \text{vert}(\Lambda)} \lambda^\top (h(Y) - T(Y, z)) - p(\bar{z}) - \max_{\lambda \in \text{vert}(\Lambda)} \lambda^\top (h(Y) - T(Y, \bar{z})) \right| \\ &\leq |p(z) - p(\bar{z})| + \max_{\lambda \in \text{vert}(\Lambda)} |\lambda^\top (T(Y, \bar{z}) - T(Y, z))| \\ &\leq \left[L_p + L_{T,z}(Y) \left(\max_{\lambda \in \text{vert}(\Lambda)} \|\lambda\| \right) \right] \|z - \bar{z}\|. \end{aligned}$$

Additionally, for any $\bar{z} \in \mathcal{Z}$:

$$\begin{aligned} \mathbb{E} [c(\bar{z}, Y)^2] &= \mathbb{E} \left[\left(p(\bar{z}) + \max_{\lambda \in \text{vert}(\Lambda)} \lambda^\top (h(Y) - T(Y, \bar{z})) \right)^2 \right] \\ &\leq 2(p(\bar{z}))^2 + 2 \left(\max_{\lambda \in \text{vert}(\Lambda)} \|\lambda\|^2 \right) \mathbb{E} [\|h(Y) - T(Y, \bar{z})\|^2] \\ &\leq 2(p(\bar{z}))^2 + 2 \left(\max_{\lambda \in \text{vert}(\Lambda)} \|\lambda\|^2 \right) \mathbb{E} \left[\sup_{z \in \mathcal{Z}} \|h(Y) - T(Y, z)\|^2 \right]. \end{aligned}$$

Consequently, the functional CLT Assumption 14 holds by virtue of Assumptions 16 and 17, and the first part of Assumption 19, see page 164 of Shapiro et al. [73] for details.

Finally, note that for any $z \in \mathcal{Z}$ and a.e. $x \in \mathcal{X}$:

$$\begin{aligned} & c(z, f^*(x) + Q^*(x)\varepsilon) - g(z; x) \\ &= V(z, f^*(x) + Q^*(x)\varepsilon) - \mathbb{E}_{\bar{\varepsilon} \sim P_\varepsilon} [V(z, f^*(x) + Q^*(x)\bar{\varepsilon})] \\ &= \max_{\lambda \in \text{vert}(\Lambda)} \mathcal{V}(f^*(x) + Q^*(x)\varepsilon, \lambda, z) - \mathbb{E}_{\bar{\varepsilon} \sim P_\varepsilon} \left[\max_{\lambda \in \text{vert}(\Lambda)} \mathcal{V}(f^*(x) + Q^*(x)\bar{\varepsilon}, \lambda, z) \right] \\ &\leq \max_{\lambda \in \text{vert}(\Lambda)} \{ \mathcal{V}(f^*(x) + Q^*(x)\varepsilon, \lambda, z) - \mathbb{E}_{\bar{\varepsilon} \sim P_\varepsilon} [\mathcal{V}(f^*(x) + Q^*(x)\bar{\varepsilon}, \lambda, z)] \} \end{aligned}$$

Consequently, for any $z \in \mathcal{Z}$ and a.e. $x \in \mathcal{X}$:

$$\begin{aligned} & \mathbb{E} \left[\exp \left(t(c(z, f^*(x) + Q^*(x)\varepsilon) - g(z; x)) \right) \right] \\ &\leq \mathbb{E} \left[\exp \left(t \max_{\lambda \in \text{vert}(\Lambda)} \{ \mathcal{V}(f^*(x) + Q^*(x)\varepsilon, \lambda, z) - \mathbb{E}_{\bar{\varepsilon} \sim P_\varepsilon} [\mathcal{V}(f^*(x) + Q^*(x)\bar{\varepsilon}, \lambda, z)] \} \right) \right] \\ &\leq \mathbb{E} \left[\max_{\lambda \in \text{vert}(\Lambda)} \exp \left(t(\mathcal{V}(f^*(x) + Q^*(x)\varepsilon, \lambda, z) - \mathbb{E}_{\bar{\varepsilon} \sim P_\varepsilon} [\mathcal{V}(f^*(x) + Q^*(x)\bar{\varepsilon}, \lambda, z)]) \right) \right] \\ &\leq \sum_{\lambda \in \text{vert}(\Lambda)} \mathbb{E} \left[\exp \left(t(\mathcal{V}(f^*(x) + Q^*(x)\varepsilon, \lambda, z) - \mathbb{E}_{\bar{\varepsilon} \sim P_\varepsilon} [\mathcal{V}(f^*(x) + Q^*(x)\bar{\varepsilon}, \lambda, z)]) \right) \right] \\ &\leq |\text{vert}(\Lambda)| \exp \left(\frac{\sigma_\varepsilon^2(x)t^2}{2} \right) \end{aligned}$$

where the last inequality follows by Assumption 19. Therefore, the large deviation Assumption 8 follows by Assumptions 16, 17, and 19 and Theorem 7.65 of Shapiro et al. [73]. \square

The assumption $\sup_{y \in \mathcal{Y}} L_{T,z}(y) < +\infty$ can be relaxed to assume that the moment generating function (mgf) of $L_{T,z}(Y)$ is finite valued in a neighborhood of zero, see Assumption (C3) and Theorem 7.65 in Section 7.2.9 of Shapiro et al. [73]. The discussion in Section 3 following Assumptions 2 and 8 and the discussion following Assumption 14 provide avenues for relaxing the i.i.d. assumption on the errors $\{\varepsilon^i\}$. The conclusions of Proposition 20 can also be established for the case of objective uncertainty (i.e., only the objective coefficients q depend on Y) if Assumptions 16, 17, 18, and 19 are suitably modified. Note that the second parts of Assumptions 12 and 13 also readily hold for problem (21) whenever Assumption 1 holds.

Generalization to a broader class of stochastic programs. Suppose the feasible region \mathcal{Z} and the support \mathcal{X} are nonempty and compact, functions f^* and Q^* are continuous on \mathcal{X} , and the function c in problem (1) is of the form:

$$c(z, Y) := \sum_{i=1}^{|\mathcal{I}|} c_i(z) \phi_i(Y)$$

for functions $\{c_i : \mathcal{Z} \rightarrow \mathbb{R}\}$ and $\{\phi_i : \mathcal{Y} \rightarrow \mathbb{R}\}$, where $|\mathcal{I}| < +\infty$. Suppose for each $i \in \mathcal{I}$ the function c_i is Lipschitz continuous on \mathcal{Z} with Lipschitz constant $L_{c,i}$ and the function ϕ_i is continuously differentiable on \mathcal{Y} . Additionally, assume for each $i \in \mathcal{I}$ that $\mathbb{E}[(\phi_i(Y))^2] < +\infty$, $\mathbb{E}[\max_{x \in \mathcal{X}} |\phi_i(f^*(x) + Q^*(x)\varepsilon)|] < +\infty$, and $\mathbb{E}[\max_{v \in \mathcal{B}_{\delta(x)}(0)} \|\nabla \phi_i(f^*(x) + Q^*(x)\varepsilon + v)\|^2] < +\infty$ for a.e. $x \in \mathcal{X}$, where δ is defined in Assumption 12. The above conditions hold, e.g., if each function ϕ_i is polynomial and the errors ε are sub-exponential. We argue that the above class of stochastic programs satisfies Assumptions 2, 6, 12, and 14 whenever the errors $\{\varepsilon^i\}$ are i.i.d. Additionally, we argue that if finite sample guarantees of the form

$$\begin{aligned} & \mathbb{P}\left\{\left(\frac{1}{n} \sum_{j=1}^n \max_{v \in \mathcal{B}_{\delta(x)}(0)} \|\nabla \phi_i(f^*(x) + Q^*(x)\varepsilon^j + v)\|^2\right)^{1/2} > \left(\mathbb{E}\left[\max_{v \in \mathcal{B}_{\delta(x)}(0)} \|\nabla \phi_i(f^*(x) + Q^*(x)\varepsilon + v)\|^2\right]\right)^{1/2} + \kappa\right\} \\ & \leq J_{\phi,i}(\kappa; x) \exp(-\gamma_{\phi,i}(n, \kappa; x)) \end{aligned}$$

hold for each $\kappa > 0$, $i \in \mathcal{I}$, $n \in \mathbb{N}$, and a.e. $x \in \mathcal{X}$, then the second part of Assumption 13 holds. Finally, we argue that if the moment generating functions of $\sum_{i=1}^{|\mathcal{I}|} c_i(z) [\phi_i(Y) - \mathbb{E}[\phi_i(Y)]]$ and $\sum_{i=1}^{|\mathcal{I}|} L_{c,i} [|\phi_i(Y)| - \mathbb{E}[|\phi_i(Y)|]]$ are finite-valued for all $z \in \mathcal{Z}$ in a neighborhood of the origin [73, cf. Section 7.2.9], then Assumption 8 also holds. The above two conditions hold, e.g., if each function ϕ_i is polynomial and the distribution of the errors ε is sufficiently light-tail.

The function $c(\cdot, y)$ is continuous on \mathcal{Z} for each $y \in \mathcal{Y}$ since each c_i is assumed to be Lipschitz continuous on \mathcal{Z} . Furthermore, for any $\bar{z} \in \mathcal{Z}$:

$$|c(\bar{z}, Y)| = \left| \sum_{i=1}^{|\mathcal{I}|} c_i(\bar{z}) \phi_i(Y) \right| \leq \sum_{i=1}^{|\mathcal{I}|} \left(\sup_{z \in \mathcal{Z}} |c_i(z)| \right) |\phi_i(Y)|.$$

Therefore, the fact that $\mathbb{E}[(\phi_i(Y))^2] < +\infty$, $\forall i \in \mathcal{I}$, along with Theorem 7.48 of Shapiro et al. [73] implies that the uniform weak LLN Assumption 2 holds.

The function $c(\cdot, f^*(\bar{x}) + Q^*(\bar{x})\varepsilon)$ is continuous on $\mathcal{Z} \times \mathcal{X}$ for each $\varepsilon \in \Xi$ since each c_i and ϕ_i are assumed to be Lipschitz continuous and f^* and Q^* are assumed to be continuous. Furthermore, for any $(\bar{z}, \bar{x}) \in \mathcal{Z} \times \mathcal{X}$:

$$|c(\bar{z}, f^*(\bar{x}) + Q^*(\bar{x})\varepsilon)| = \left| \sum_{i=1}^{|\mathcal{I}|} c_i(\bar{z}) \phi_i(f^*(\bar{x}) + Q^*(\bar{x})\varepsilon) \right| \leq \sum_{i=1}^{|\mathcal{I}|} \left(\sup_{z \in \mathcal{Z}} |c_i(z)| \right) \left(\sup_{x \in \mathcal{X}} |\phi_i(f^*(x) + Q^*(x)\varepsilon)| \right).$$

Therefore, the fact that $\mathbb{E}[\sup_{x \in \mathcal{X}} |\phi_i(f^*(x) + Q^*(x)\varepsilon)|] < +\infty$, $\forall i \in \mathcal{I}$, along with Theorem 7.48 of Shapiro et al. [73] implies that $\sup_{(z,x) \in \mathcal{Z} \times \mathcal{X}} |g_n^*(z; x) - g(z; x)| \xrightarrow{P} 0$, which in turn implies Assumption 6.

Next, note that for any $z, \bar{z} \in \mathcal{Z}$:

$$|c(z, Y) - c(\bar{z}, Y)| = \left| \sum_{i=1}^{|\mathcal{I}|} [c_i(z) - c_i(\bar{z})] \phi_i(Y) \right| \leq \sum_{i=1}^{|\mathcal{I}|} |c_i(z) - c_i(\bar{z})| |\phi_i(Y)| \leq \left(\sum_{i=1}^{|\mathcal{I}|} L_{c,i} |\phi_i(Y)| \right) \|z - \bar{z}\|.$$

with $\mathbb{E}[(\sum_{i=1}^{|\mathcal{I}|} L_{c,i} |\phi_i(Y)|)^2] \leq O(1) \sum_{i=1}^{|\mathcal{I}|} \mathbb{E}[(\phi_i(Y))^2] < +\infty$. Additionally, for any $\bar{z} \in \mathcal{Z}$:

$$\mathbb{E}[(c(\bar{z}, Y))^2] = \mathbb{E}\left[\left(\sum_{i=1}^{|\mathcal{I}|} c_i(\bar{z}) \phi_i(Y)\right)^2\right] \leq \mathbb{E}\left[|\mathcal{I}| \sum_{i=1}^{|\mathcal{I}|} (c_i(\bar{z}))^2 (\phi_i(Y))^2\right] \leq |\mathcal{I}| \sum_{i=1}^{|\mathcal{I}|} (c_i(\bar{z}))^2 \mathbb{E}[(\phi_i(Y))^2] < +\infty.$$

Consequently, the CLT Assumption 14 holds by arguments in page 164 of Shapiro et al. [73]. Since $\sum_{i=1}^{|\mathcal{I}|} L_{c,i} |\phi_i(Y)|$ is a Lipschitz constant for $c(\cdot, Y)$ on \mathcal{Z} , Assumption 8 holds if the stated conditions on the mgfs of $\sum_{i=1}^{|\mathcal{I}|} c_i(z) [\phi_i(Y) - \mathbb{E}[\phi_i(Y)]]$ and $\sum_{i=1}^{|\mathcal{I}|} L_{c,i} [|\phi_i(Y)| - \mathbb{E}[|\phi_i(Y)|]]$ hold.

We focus on verifying the second parts of Assumptions 12 and 13 next. We have for any $z \in \mathcal{Z}$, $y \in \mathcal{Y}$, and $\bar{y} \in \mathcal{B}_{\delta(x)}(y) \cap \mathcal{Y}$:

$$\begin{aligned} |c(z, \bar{y}) - c(z, y)| &= \left| \sum_{i=1}^{|\mathcal{I}|} c_i(z) [\phi_i(\bar{y}) - \phi_i(y)] \right| \leq \sum_{i=1}^{|\mathcal{I}|} |c_i(z)| |\phi_i(\bar{y}) - \phi_i(y)| \\ &\leq \left(\sum_{i=1}^{|\mathcal{I}|} |c_i(z)| \max_{v \in \mathcal{B}_{\delta(x)}(y)} \|\nabla \phi_i(v)\| \right) \|\bar{y} - y\|, \end{aligned}$$

where the last inequality follows by the mean-value theorem. Therefore, we can choose

$$L_{\delta(x)}(z, y) := \left(\sum_{i=1}^{|\mathcal{I}|} |c_i(z)| \max_{v \in \mathcal{B}_{\delta(x)}(y)} \|\nabla \phi_i(v)\| \right)$$

in Assumptions 12 and 13. Noting that

$$\begin{aligned} \sup_{z \in \mathcal{Z}} \frac{1}{n} \sum_{j=1}^n L_{\delta(x)}^2(z, f^*(x) + Q^*(x)\varepsilon^j) &\leq \frac{1}{n} \sum_{j=1}^n \sup_{z \in \mathcal{Z}} \left(\sum_{i=1}^{|\mathcal{I}|} |c_i(z)| \max_{v \in \mathcal{B}_{\delta(x)}(0)} \|\nabla \phi_i(f^*(x) + Q^*(x)\varepsilon^j + v)\| \right)^2 \\ &\leq O(1) \sum_{i=1}^{|\mathcal{I}|} \left(\sup_{z \in \mathcal{Z}} |c_i(z)|^2 \right) \frac{1}{n} \sum_{j=1}^n \max_{v \in \mathcal{B}_{\delta(x)}(0)} \|\nabla \phi_i(f^*(x) + Q^*(x)\varepsilon^j + v)\|^2, \end{aligned}$$

we see that $\sup_{z \in \mathcal{Z}} \frac{1}{n} \sum_{j=1}^n L_{\delta(x)}^2(z, f^*(x) + Q^*(x)\varepsilon^j) = O_p(1)$ by the weak LLN, which implies that Assumption (12b) holds. Additionally, we have for each $\varepsilon \in \Xi$ and a.e. $x \in \mathcal{X}$:

$$\sup_{z \in \mathcal{Z}} L_{\delta(x)}^2(z, f^*(x) + Q^*(x)\varepsilon) \leq |\mathcal{I}| \sum_{i=1}^{|\mathcal{I}|} \left(\sup_{z \in \mathcal{Z}} |c_i(z)|^2 \right) \max_{v \in \mathcal{B}_{\delta(x)}(0)} \|\nabla \phi_i(f^*(x) + Q^*(x)\varepsilon + v)\|^2.$$

Therefore, finite sample guarantees for $\max_{v \in \mathcal{B}_{\delta(x)}(0)} \|\nabla \phi_i(f^*(x) + Q^*(x)\varepsilon + v)\|^2$, $i \in \mathcal{I}$, directly translate to finite sample guarantees of the form in Assumption 13.

F Some prediction setups that satisfy our assumptions

We verify that Assumptions 5, 11, and 15 and the corresponding assumptions for the J-SAA and J+-SAA problems hold for specific regression procedures, and point to resources within the literature for verifying

these assumptions more broadly. We do not attempt to be exhaustive and, for the most part, restrict our attention to M-estimators [76, 78], which encapsulate a rich class of prediction techniques. We often also consider the special case where the true model can be written as $Y = f(X; \theta^*) + Q(X; \pi^*)\varepsilon$ and the goal of the regression procedure (3) is to estimate the finite-dimensional parameters $\theta^* \in \Theta$ and $\pi^* \in \Pi$ using the data \mathcal{D}_n . To summarize, we largely consider the regression setup (possibly with a regularization term)

$$\hat{\theta}_n \in \arg \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \ell(y^i, f(x^i; \theta))$$

for estimating θ^* with a particular emphasis on the squared loss $\ell(y, \hat{y}) := \|y - \hat{y}\|^2$. We call the optimization problem

$$\min_{\theta \in \Theta} \mathbb{E}_{(Y, X)}[\ell(Y, f(X; \theta))]$$

the population regression problem, where the above expectation is taken with respect to the joint distribution of (Y, X) . We mostly assume that the solution set of the population regression problem is the singleton $\{\theta^*\}$, and assume throughout this section that $X_1 \equiv 1$.

We consider two distinct approaches for estimating the heteroscedasticity function Q^* . The first approach uses the fact that $\mathbb{E}[(Y - f^*(X))(Y - f^*(X))^T | X = x] = Q^*(x)Q^*(x)^T$ and plugs in \hat{f}_n instead of f^* to determine the best regression estimate \hat{Q}_n of Q^* in the model class \mathcal{Q} . For the parametric case, once we obtain an estimate $\hat{\theta}_n$ of θ^* , we can estimate π^* using

$$\hat{\pi}_n \in \arg \min_{\pi \in \Pi} \frac{1}{n} \sum_{i=1}^n \bar{\ell}((y^i - f(x^i; \hat{\theta}_n))(y^i - f(x^i; \hat{\theta}_n))^T, Q(x^i; \pi)),$$

where $\bar{\ell}(V, \hat{V}) := \|V - \hat{V}\|_F^2$ is the squared Frobenius norm. In special settings (cf. our running Example 2), the above problem for estimating π^* can be transformed into a linear regression problem. An alternative is to estimate f^* and Q^* jointly using M-estimation:

$$(\hat{\theta}_n, \hat{\pi}_n) \in \arg \min_{(\theta, \pi) \in \Theta \times \Pi} \frac{1}{n} \sum_{i=1}^n \bar{\ell}((y^i - f(x^i; \theta))(y^i - f(x^i; \theta))^T, Q(x^i; \pi))$$

for some loss function $\bar{\ell}$ [27].

Finally, we emphasize that we only deal with the random design case (where the covariates X are considered to be random) in this work. Much of the statistics literature presents results for the fixed design setting in which the covariate observations $\{x^i\}_{i=1}^n$ are deterministic and designed by the DM. These results readily carry over to the random design setting if $Q^* \equiv I$, the errors ε are independent of X , and no restriction is made on the design points $\{x^i\}_{i=1}^n$.

F.1 Parametric regression techniques for estimating f^*

We verify that the parts of Assumptions 5, 11, and 15 relating to the estimate \hat{f}_n of f^* hold for OLS regression, the Lasso and generalized linear regression models under suitable assumptions. We also verify their counterparts for the J-SAA and J+-SAA problems in Section D. Note that Assumption (7a) holds for the regression techniques considered in this section whenever $\hat{\theta}_n \xrightarrow{p} \theta^*$ and $\|X\|_{L^q} < +\infty$. Theorem 2.6 and Corollary 2.8 of Rigollet and Hütter [64] present conditions under which these assumptions hold for best subset selection regression in the homoscedastic setting, and Theorem 2.14 therein presents similar guarantees for the Bayes Information Criterion estimator. Koltchinskii [53] verifies these assumptions for the Dantzig selector under certain conditions (including homoscedasticity of the error distribution). Hsu et al. [45] verifies these conditions for ridge regression. Negahban et al. [60] provides results for regularized M-estimators in the high-dimensional setting.

F.1.1 Ordinary least squares regression

We present sufficient conditions from White [85], Hsu et al. [45], and Rigollet and Hütter [64] under which the parts of Assumptions 5, 11, and 15 relating to the estimate \hat{f}_n of f^* hold. Note that Theorems 2.31 and 4.25 of White [85] present a general set of sufficient conditions for $\hat{\theta}_n \xrightarrow{P} \theta^*$ and for $\sqrt{n}(\hat{\theta}_n - \theta^*)$ to be asymptotically normally distributed. Chapters 3 to 5 of White [85] also present analyses that can handle *instrumental variables*, which can be used to verify Assumptions 5, 15, and 11 when the errors ε are correlated with the features X . We have the following result:

Proposition 21. Suppose $f^*(X) = \theta^* X$ and we use OLS regression to estimate θ^* . Define $\bar{\varepsilon} := Q^*(X)\varepsilon$.

1. Suppose for some $\gamma > 0$, $\mathbb{E}[|X_i \bar{\varepsilon}_j|^{1+\gamma}] < +\infty$, $\forall i \in [d_x]$ and $j \in [d_y]$, $\mathbb{E}[\|X\|^{2(1+\gamma)}] < +\infty$, and $\mathbb{E}[XX^T]$ is positive definite. If $\{(x^i, \varepsilon^i)\}_{i=1}^n$ is i.i.d., then $\hat{\theta}_n$ a.s. exists for n large enough and $\hat{\theta}_n \xrightarrow{a.s.} \theta^*$. Consequently, Assumptions (5a) and (5b) hold.
2. Suppose for some $\gamma > 0$, $\mathbb{E}[|X_i \bar{\varepsilon}_j|^{2+\gamma}] < +\infty$, $\forall i \in [d_x]$ and $j \in [d_y]$, $\mathbb{E}[\|X\|^{2(1+\gamma)}] < +\infty$, $\mathbb{E}[XX^T]$ is positive definite, the covariance matrix of the random variable $\sum_{j=1}^{d_y} X \bar{\varepsilon}_j$ is positive definite, and $\{(x^i, \varepsilon^i)\}_{i=1}^n$ is i.i.d. Then Assumptions (15a) and (15b) hold with $\alpha = 1$.
3. Suppose $\{(x^i, \varepsilon^i)\}_{i=1}^n$ is i.i.d., $\bar{\varepsilon}$ is uniformly sub-Gaussian with variance proxy σ^2 , i.e.,

$$\mathbb{E}[\exp(su^T \bar{\varepsilon}) \mid X = x] \leq \exp(0.5s^2 \sigma^2), \quad \forall s \in \mathbb{R}, \quad \|u\| = 1, \quad \text{a.e. } x \in \mathcal{X},$$

the covariance matrix Σ_X of the covariates is positive definite, and the random vector $\Sigma_X^{-\frac{1}{2}} X$ is sub-Gaussian. Then Assumptions (11a) and (11b) hold with constants $K_f(\kappa, x) = O(\exp(d_x))$, $\beta_f(\kappa, x) = O\left(\frac{\kappa^2}{\sigma^2 d_y \|x\|^2}\right)$, $\bar{K}_f(\kappa) = O(\exp(d_x))$, and $\bar{\beta}_f(\kappa) = O\left(\frac{\kappa^2}{\sigma^2 d_y}\right)$.

Proof. Part 1 follows from Theorems 3.15 of White [85]. Part 2 follows from Theorem 5.13 of White [85]. Part 3 follows from Remark 12 of Hsu et al. [45]. For the homoscedastic case, part 3 also follows from Theorem 2.2 and Remark 2.3 of Rigollet and Hütter [64]. Although Rigollet and Hütter [64] consider the fixed design case, their proof readily extends to the random design setting since no restrictions were placed on the design. \square

For the homoscedastic setting, part 1 of Proposition 21 follows from Theorems 3.5 and 3.37 of White [85] and part 2 follows from Theorem 5.3 of White [85]. Theorems 3.49 and 3.78 of White [85] present sufficient conditions under which Assumptions (5a) and (5b) hold under mixing and martingale conditions on the data $\{(x^i, \varepsilon^i)\}$. Theorem 5.17 and Exercise 5.21 of White [85] present sufficient conditions under which Assumptions (15a) and (15b) hold with $\alpha = 1$ for ergodic and mixing data $\{(x^i, \varepsilon^i)\}$, respectively. Note that results in Bryc and Dembo [18] and [28] can be used to establish Assumptions (11a) and (11b) for the non-i.i.d. setting. Note that once we have an estimate \hat{Q}_n of Q^* , we can re-estimate θ^* using feasible weighted least squares regression [66]. This yields an estimate $\hat{\theta}_n$ that is asymptotically more efficient than the OLS estimator of θ^* whenever \hat{Q}_n is a consistent estimate of Q^* [65, 66]. Even if the estimate \hat{Q}_n of Q^* is inconsistent, the weighted least squares estimator $\hat{\theta}_n$ remains consistent but may no longer be asymptotically efficient (see, e.g., Section 3.3 of [66]).

The above results can be used in conjunction with the techniques in Section F.2 to verify Assumptions 5J, 5J+, 15J, 15J+, 11J, and 11J+ for i.i.d. data $\{(x^i, \varepsilon^i)\}$. In the remainder of this section, we specialize the verification of these assumptions for OLS regression in the homoscedastic setting when problem (1) is a two-stage stochastic LP (see Example 1). We assume that $d_y = 1$ for ease of exposition.

Following Lemma 14 in Section D, it suffices to establish rates and finite sample guarantees for the terms $\frac{1}{n} \sum_{i=1}^n \|\hat{f}_n(x^i) - \hat{f}_{-i}(x^i)\|$ and $\frac{1}{n} \sum_{i=1}^n \|\hat{f}_n(x) - \hat{f}_{-i}(x)\|$ when the assumptions for the ER-SAA problem hold. Let \bar{X} denote the $\mathbb{R}^{n \times d_x}$ design matrix with $\bar{X}_{[i]} = (x^i)^T$, $h^i := (\bar{X}(\bar{X}^T \bar{X})^{-1} \bar{X}^T)_{ii}$ denote the i th leverage

score, and $e^i := y^i - \hat{\theta}_n^T x^i$ denote the residual of the model \hat{f}_n at the i th data point. From Section 10.6.3 of Seber and Lee [70], we have

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \|\hat{f}_n(x^i) - \hat{f}_{-i}(x^i)\| &= \frac{1}{n} \sum_{i=1}^n \frac{h^i |e^i|}{1-h^i} \leq \left(\frac{1}{n} \sum_{i=1}^n (h^i)^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n \left(\frac{e^i}{1-h^i} \right)^2 \right)^{1/2} \leq \frac{d_x}{\sqrt{n}} \left(\frac{1}{n} \sum_{i=1}^n \left(\frac{e^i}{1-h^i} \right)^2 \right)^{1/2}, \\ \frac{1}{n} \sum_{i=1}^n \|\hat{f}_n(x) - \hat{f}_{-i}(x)\| &\leq \frac{\|x\|}{n} \sum_{i=1}^n \left\| \frac{(\bar{X}^T \bar{X})^{-1} x^i e^i}{1-h^i} \right\| \leq \|x\| \left(\frac{1}{n} \sum_{i=1}^n \left\| (\bar{X}^T \bar{X})^{-1} x^i \right\|^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n \left(\frac{e^i}{1-h^i} \right)^2 \right)^{1/2} \\ &\leq \|x\| \left(\frac{1}{n} \text{Tr}((\bar{X}^T \bar{X})^{-1}) \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n \left(\frac{e^i}{1-h^i} \right)^2 \right)^{1/2}, \end{aligned}$$

where Tr denotes the trace operator. The quantity $\left(\frac{1}{n} \sum_{i=1}^n \left(\frac{e^i}{1-h^i} \right)^2 \right)^{1/2}$ is called the *prediction residual sum of squares* statistic and is bounded in probability under mild assumptions. The above inequalities can be used to verify the assumptions for the jackknife-based estimators for Example 1.

F.1.2 The Lasso and high-dimensional generalized linear models

Following van de Geer [77] and Bunea et al. [19], we consider generalized linear models with an ℓ_1 -penalty. We assume $d_y = 1$ for ease of exposition. The setup is: the model class $\mathcal{F} := \{f : f(\cdot; \theta) := \sum_{k=1}^m \theta_k \psi_k(\cdot), \theta \in \Theta\}$, where $\{\psi_k(\cdot)\}_{k=1}^m$ is a sequence of real-valued basis functions with domain \mathcal{X} , the data $\{(x^i, \varepsilon^i)\}$ is assumed to be i.i.d., the number of basis functions m grows subexponentially with the number of data samples n , the set Θ is convex, the loss function ℓ satisfies some Lipschitz assumptions [see Assumption L and Example 4 of 77], and the estimate $\hat{\theta}_n$ of θ^* is obtained as

$$\hat{\theta}_n \in \arg \min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \ell(y^i, f(x^i; \theta)) + \lambda_n \sum_{k=1}^m \left(\frac{1}{n} \sum_{i=1}^n \psi_k^2(x^i) \right)^{\frac{1}{2}} |\theta_k|$$

for some penalty parameter $\lambda_n = O\left(\sqrt{\frac{\log m}{n}}\right)$ that is chosen large enough. The above setup captures both parametric and nonparametric regression models. Theorem 2.2 of van de Geer [77] and Theorems 2.1, 2.2, and 2.3 of Bunea et al. [19] present conditions under which Assumptions 5, 11, and 15 hold for the above setting.

In the remainder of this section, we specialize the results of Bunea et al. [19] to the traditional Lasso setup [75]. In this setup, $m = d_x$, $\psi_k(x) = x_k$, $\Theta = \mathbb{R}^{d_x}$, and $\ell(y, \hat{y}) = \|y - \hat{y}\|^2$. Once again, we define $\bar{\varepsilon} := Q^*(X)\varepsilon$.

Proposition 22. Suppose $f^*(X) = \theta^* X$ with $\|\theta_{[j]}^*\|_0 \leq s$, $\forall j \in [d_y]$, the sequences $\{x^i\}_{i=1}^n$ and $\{\varepsilon^i\}_{i=1}^n$ are i.i.d., and the error $\bar{\varepsilon}$ is uniformly sub-Gaussian with variance proxy σ^2 , i.e.,

$$\mathbb{E} \left[\exp(su^T \bar{\varepsilon}) \mid X = x \right] \leq \exp(0.5s^2 \sigma^2), \quad \forall s \in \mathbb{R}, \quad \|u\| = 1, \quad \text{a.e. } x \in \mathcal{X}.$$

Additionally, suppose the support \mathcal{X} of the covariates X is compact, $\mathbb{E}[|X_j|^2] > 0$, $\forall j \in [d_x]$, and the matrix $\mathbb{E}[XX^T] - \tau \text{diag}(\mathbb{E}[XX^T])$ is positive semidefinite for some constant $\tau \in (0, 1]$. If we use the Lasso to estimate θ^* , then Assumptions (5a) and (5b) hold, Assumptions (15a) and (15b) hold with $\alpha = 1$, and Assumption (11a) and (11b) hold with $\bar{K}_f(\kappa) = K_f(\kappa, x) = O(d_x)$, $\bar{\beta}_f(\kappa) = O\left(\frac{\kappa^2}{\sigma^2 s d_y}\right)$, and $\beta_f(\kappa, x) = O\left(\frac{\kappa^2}{\sigma^2 s d_y \|x\|^2}\right)$.

Proof. Follows from Theorem 2.1 and Corollary 1 of Bunea et al. [19]. \square

Chatterjee [21] establishes consistency of the Lasso in the homoscedastic setting under the following weaker assumptions: the data $\{(x^i, \varepsilon^i)\}$ is i.i.d., the error ε is sub-Gaussian with variance proxy σ^2 and is independent of the covariates X , the support \mathcal{X} of the covariates is compact, and the covariance matrix of the covariates is positive definite. Theorems 1 and 2 therein present conditions under which Assumption 15 holds at a slower rate with $\alpha = 0.5$. Theorem 2.15 of Rigollet and Hütter [64] can then be used to show that Assumptions (11a) and (11b) hold in the homoscedastic setting with $\bar{K}_f(\kappa) = K_f(\kappa, x) = O(d_x)$, $\bar{\beta}_f(\kappa) = O\left(\frac{\kappa^4}{\sigma^2 s^2 d_y^2}\right)$, and $\beta_f(\kappa, x) = O\left(\frac{\kappa^4}{\sigma^2 s^2 d_y^2 \|x\|^2}\right)$. Basu and Michailidis [7] present conditions under which Assumptions 5, 11, and 15 can be verified for time series data $\{(x^i, \varepsilon^i)\}$ in the homoscedastic case. The above results can be used in conjunction with the discussion in Section F.2 to derive rates of convergence and finite sample guarantees for the jackknife-based estimators for i.i.d. data $\{(x^i, \varepsilon^i)\}$.

Theorem 1 of Belloni et al. [10] outlines conditions under which Assumptions 5 and 15 hold for the heteroscedasticity-adapted Lasso with $\alpha = 1$. Medeiros and Mendes [59] and Ziel [87] present asymptotic analyses of the adaptive Lasso for time series data, including GARCH-type processes. Theorems 2 and 3 of [59] and Theorem 1 of [87] present conditions under which Assumptions 5 and 15 hold with $\alpha = 1$. Belloni et al. [11] study asymptotic and finite sample guarantees for the heteroscedasticity-adapted square-root Lasso. Finally, Theorem 5.2 of Dalalyan et al. [25] introduces a scaled heteroscedastic Dantzig selector and presents large deviation bounds for both \hat{f}_n and \hat{Q}_n under certain sparsity assumptions.

F.2 General M-estimation procedures for estimating f^*

We use results from Chapter 5 of van der Vaart [78], Chapter 3 of van der Vaart and Wellner [79], and Shapiro et al. [73] to verify the parts of Assumptions 5, 11, and 15 relating to the estimate \hat{f}_n of f^* for general M-estimators. To begin, we suppose that the regression function $f(x; \cdot)$ is Lipschitz continuous at θ^* for a.e. $x \in \mathcal{X}$ with Lipschitz constant $L_f(x)$, i.e., we a.s. have $\|f(x; \theta^*) - f(x; \hat{\theta}_n)\| \leq L_f(x) \|\theta^* - \hat{\theta}_n\|$. To establish Assumptions (5a), (5b), (15a), and (15b), it suffices to assume that the function $f(x, \cdot)$ is locally Lipschitz continuous at θ^* and a.s. for n large enough, the estimates $\hat{\theta}_n$ of θ^* lie in some compact subset of Θ . Note that

$$\frac{1}{n} \sum_{i=1}^n \|f(x^i; \theta^*) - f(x^i; \hat{\theta}_n)\|^2 \leq \left(\frac{1}{n} \sum_{i=1}^n L_f^2(x^i) \right) \|\theta^* - \hat{\theta}_n\|^2,$$

with the first term in the r.h.s. of the above inequality bounded in probability under a suitable weak LLN assumption. Therefore, our main focus is presenting rates at which $\|\theta^* - \hat{\theta}_n\| \xrightarrow{P} 0$. Note that Assumption (7a) also holds whenever $\hat{\theta}_n \xrightarrow{P} \theta^*$ and $\|L_f(X)\|_{L^q} < +\infty$.

Verifying Assumption 5. Theorem 5.7 of van der Vaart [78] presents conditions under which $\hat{\theta}_n \xrightarrow{P} \theta^*$ for i.i.d. data $\{(x^i, \varepsilon^i)\}$ [cf. Theorems 5.3 and 5.4 of 73]. Similar to the discussion following Assumption 2, this result also holds when \mathcal{D}_n satisfies certain mixing/stationarity assumptions. Section 5.2 of van der Vaart [78] also presents alternative conditions for $\hat{\theta}_n \xrightarrow{P} \theta^*$.

Verifying Assumption 15. We discuss conditions under which $\|\hat{\theta}_n - \theta^*\| \xrightarrow{P} 0$ at certain rates. Theorem 5.23 of van der Vaart [78] presents regularity conditions under which this convergence holds at the conventional $n^{-1/2}$ rate, in which case Assumption 15 holds with $\alpha = 1$ [cf. Theorem 5.8 of 73]. Once again, the above conclusion holds when the observations \mathcal{D}_n satisfy certain mixing/stationarity assumptions. Chapter 5 of van der Vaart [78] and Chapter 3.2 of van der Vaart and Wellner [79] provide some examples of M-estimators that possess this rate of convergence. Theorem 5.52 and Chapter 25 of van der Vaart [78] present conditions under which Assumption 15 holds with constant $\alpha < 1$ (including for semiparametric regression). Similar to the special case of OLS regression, vanilla M-estimators that do not account for heteroscedasticity may no longer be efficient. Feasible weighted M-estimation may provide an asymptotically efficient alternative in the heteroscedastic setting.

Verifying Assumption 11. We verify this assumption by establishing finite sample guarantees for $\hat{\theta}_n$ when the M-estimation problem satisfies uniform exponential bounds similar to Assumption 8. Specifically, suppose for any constant $\kappa > 0$, there exist positive constants $\hat{K}(\kappa)$ and $\hat{\beta}(\kappa)$ such that

$$\mathbb{P}\left\{\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^n \ell(y^i, f(x^i; \theta)) - \mathbb{E}_{(Y, X)}[\ell(Y, f(X; \theta))] \right| > \kappa \right\} \leq \hat{K}(\kappa) \exp(-n\hat{\beta}(\kappa)),$$

see the discussion surrounding Assumption 8 for conditions under which such a uniform exponential bound holds [the main restriction there is that Θ is compact, but this can be relaxed by assuming that the estimates $\hat{\theta}_n$ lie in a compact subset of Θ , see the discussion following Theorem 5.3 of 73]. Theorem 2.3 of Homem-de-Mello [42] then implies that Assumption 11 holds whenever the sample average term $\frac{1}{n} \sum_{i=1}^n L_f^2(x^i)$ satisfies a large deviation property (i.e., it is concentrated around $\mathbb{E}[L_f^2(X)]$). We note that results in Bryc and Dembo [18], Dembo and Zeitouni [28] can be used to establish such uniform exponential bounds for mixing data $\{(x^i, \varepsilon^i)\}$ by adapting Lemma 2.4 of Homem-de-Mello [42]. Theorems 1, 3, and 5 of Sun et al. [74] and Theorem 2.1 of Zhou et al. [86] present large deviation results in the form of Assumptions (11a) and (11b) for adaptive Huber regression when f^* is linear.

Verifying the assumptions for the jackknife-based methods. We now present techniques for verifying the parts of Assumptions 5J, 5J+, 15J, 15J+, 11J, and 11J+ relating to the estimate \hat{f}_n of f^* when the data $\{(x^i, \varepsilon^i)\}$ is i.i.d. Noting from Markov's inequality that

$$\begin{aligned} \mathbb{P}\left\{\frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_{-i}(x^i)\|^4 > \kappa^4\right\} &\leq \frac{1}{n\kappa^4} \sum_{i=1}^n \mathbb{E}_{\mathcal{D}_n} \left[\|f^*(x^i) - \hat{f}_{-i}(x^i)\|^4 \right] = \frac{1}{\kappa^4} \mathbb{E}_{\mathcal{D}_{n-1}, X} \left[\|f^*(X) - \hat{f}_{n-1}(X)\|^4 \right], \\ \mathbb{P}\left\{\frac{1}{n} \sum_{i=1}^n \|f^*(x) - \hat{f}_{-i}(x)\| > \kappa\right\} &\leq \frac{1}{n\kappa} \sum_{i=1}^n \mathbb{E}_{\mathcal{D}_n} \left[\|f^*(x) - \hat{f}_{-i}(x)\| \right] = \frac{1}{\kappa} \mathbb{E}_{\mathcal{D}_{n-1}} \left[\|f^*(x) - \hat{f}_{n-1}(x)\| \right], \end{aligned}$$

when the data \mathcal{D}_n is i.i.d., we have that Assumptions 15J and 15J+ on the jackknife-based methods hold if, for a.e. $x \in \mathcal{X}$, the expectations $\mathbb{E}_{\mathcal{D}_{n-1}, X} \left[\|f^*(X) - \hat{f}_{n-1}(X)\|^4 \right]$ and $\mathbb{E}_{\mathcal{D}_{n-1}} \left[\|f^*(x) - \hat{f}_{n-1}(x)\| \right]$ converge to zero at suitable rates. Under the aforementioned Lipschitz continuity assumption on the function $f(x; \cdot)$ at θ^* and the assumption that $\mathbb{E} \left[L_f^4(X) \right] < +\infty$, it suffices to establish rates of convergence for the expectation term $\mathbb{E}_{\mathcal{D}_{n-1}} \left[\|\theta^* - \hat{\theta}_{n-1}\|^4 \right]$. These results can be readily obtained under assumptions on the curvature of the loss function of the M-estimation problem (e.g., restricted strong convexity) around the true parameter θ^* , see Negahban et al. [60] for instance. Chapter 14 of Biau and Devroye [16] provides similar rate results for kNN regression. Alternatively, we can also bound the terms appearing in the assumptions for the jackknife-based formulations as

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \|f(x; \theta^*) - f(x; \hat{\theta}_{-i})\| &\leq \frac{1}{n} \sum_{i=1}^n L_f(x) \|\theta^* - \hat{\theta}_{-i}\|, \\ \frac{1}{n} \sum_{i=1}^n \|f(x^i; \theta^*) - f(x^i; \hat{\theta}_{-i})\|^4 &\leq \frac{1}{n} \sum_{i=1}^n L_f^4(x^i) \|\theta^* - \hat{\theta}_{-i}\|^4 \leq \left(\frac{1}{n} \sum_{i=1}^n L_f^8(x^i) \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n \|\theta^* - \hat{\theta}_{-i}\|^8 \right)^{1/2}, \end{aligned}$$

with the first term in the r.h.s. of the last inequality bounded under appropriate LLN assumptions. Therefore, an alternative is to establish rates and finite sample guarantees for the two terms $\frac{1}{n} \sum_{i=1}^n \|\theta^* - \hat{\theta}_{-i}\|^8$ and $\frac{1}{n} \sum_{i=1}^n L_f^8(x^i)$. A third direct approach is to use the weaker bounds

$$\begin{aligned} \mathbb{P}\left\{\frac{1}{n} \sum_{i=1}^n \|f^*(x^i) - \hat{f}_{-i}(x^i)\|^4 > \kappa^4\right\} &\leq \sum_{i=1}^n \mathbb{P}\left\{\|f^*(x^i) - \hat{f}_{-i}(x^i)\|^4 > \kappa^4\right\} = n\mathbb{P}\left\{\|f^*(X) - \hat{f}_{n-1}(X)\|^4 > \kappa^4\right\}, \\ \mathbb{P}\left\{\frac{1}{n} \sum_{i=1}^n \|f^*(x) - \hat{f}_{-i}(x)\| > \kappa\right\} &\leq \sum_{i=1}^n \mathbb{P}\left\{\|f^*(x) - \hat{f}_{-i}(x)\| > \kappa\right\} = n\mathbb{P}\left\{\|f^*(x) - \hat{f}_{n-1}(x)\| > \kappa\right\}. \end{aligned}$$

Finally, note that it is sufficient to establish rates and finite sample guarantees for $\frac{1}{n} \sum_{i=1}^n \|\hat{f}_n(x^i) - \hat{f}_{-i}(x^i)\|^4$ and $\frac{1}{n} \sum_{i=1}^n \|\hat{f}_n(x) - \hat{f}_{-i}(x)\|$ when Assumptions 5, 15, and 11 hold.

F.3 Nonparametric regression techniques for estimating f^*

We verify that Assumptions (5a), (5b), (15a), and (15b) hold for kNN regression, CART, and RF regression, and state a large deviation result similar to Assumptions (11a) and (11b) for kNN regression. The discussion in Section F.2 then provides an avenue for verifying the corresponding assumptions for the J-SAA and J+SAA problems for i.i.d. data $\{(x^i, \varepsilon^i)\}$. Theorem 14.5 in Biau and Devroye [16] presents conditions under which Assumption (7a) holds for $q \leq 2$. Note that results in Walk [84], Györfi et al. [41], and Chen and Shah [22] can be used to verify some of these assumptions for kernel regression and semi-recursive Devroye-Wagner estimates for mixing data \mathcal{D}_n , results in Raskutti et al. [63] can be used to verify these assumptions for sparse additive nonparametric regression, Chapter 13 of Wainwright [83] can be used to verify these assumptions for (regularized) nonparametric least squares regression, and results in Seijo and Sen [71] and Mazumder et al. [58] can be used to verify these assumptions for convex regression. In what follows, we only consider the setting where the data $\{(x^i, \varepsilon^i)\}$ is i.i.d.

We assume that the kNN regression estimate is computed as follows: given parameter $k \in [n]$ and $x \in \mathcal{X}$, define $\hat{f}_n(x) := \frac{1}{k} \sum_{i=1}^k y^{(i)}(x)$, where $\{(y^{(i)}(x), x^{(i)}(x))\}_{i=1}^k$ is a reordering of the data $\{(y^i, x^i)\}_{i=1}^n$ such that $\|x^{(j)}(x) - x\| \leq \|x^{(k)}(x) - x\|$ whenever $j \leq k$ (if $\|x^{(j)}(x) - x\| = \|x^{(k)}(x) - x\|$ for some $j < k$, then we assume that $(y^{(j)}(x), x^{(j)}(x))$ appears first in the reordering).

Proposition 23. Suppose the data $\{(x^i, \varepsilon^i)\}$ is i.i.d. and the support \mathcal{X} of the covariates is compact. Define $\bar{\varepsilon} := Q^*(X)\varepsilon$, and suppose the distribution of the errors $\bar{\varepsilon}$ satisfies $\sup_{x \in \mathbb{R}^{d_x}} \mathbb{E}[\exp(\lambda|\bar{\varepsilon}_j|) \mid X = x] < +\infty$ for each $j \in [d_y]$ and some $\lambda > 0$. Consider the setting where we use kNN regression to estimate the regression function f^* with the parameter ‘ k ’ satisfying $\lim_{n \rightarrow \infty} \frac{k}{\log(n)} = \infty$ and $k = o(n)$.

1. Suppose the function f^* is continuous on \mathcal{X} . Then Assumptions (5a) and (5b) hold.
2. Suppose the function f^* is twice continuously differentiable on \mathcal{X} and the random vector X has a density that is twice continuously differentiable. Then, there exists a choice of the parameter ‘ k ’ such that Assumptions (15a) and (15b) hold with $\alpha = \frac{O(1)}{d_x}$.
3. Suppose the function f^* is Lipschitz continuous on \mathcal{X} and there exists a constant $\tau > 0$ such that the distribution P_X of the covariates satisfies $\mathbb{P}\{X \in \mathcal{B}_\kappa(x)\} \geq \tau\kappa^{d_x}, \forall x \in \mathcal{X}$ and $\kappa > 0$. Then, for sample size n satisfying $n \geq O(1)k \left(\frac{O(1)}{\kappa}\right)^{d_x}$ and $\frac{n^\gamma}{\log(n)} \geq \frac{O(1)d_x d_y \sigma^2}{\kappa^2}$, we have

$$\mathbb{P}\left\{\sup_{x \in \mathcal{X}} \|f^*(x) - \hat{f}_n(x)\| > \kappa \sqrt{d_y}\right\} \leq \left(\frac{O(1)\sqrt{d_x}}{\kappa}\right)^{d_x} \exp\left(-O(1)n(O(1)\kappa)^{2d_x}\right) + O(1)n^{2d_x} \left(\frac{O(1)}{d_x}\right)^{d_x} \exp\left(-\frac{k\kappa^2}{O(1)\sigma^2}\right).$$

Proof. The first part follows from Theorem 12.1 of Biau and Devroye [16]. The second part follows from Theorems 14.3 and 14.5 of Biau and Devroye [16] and Markov’s inequality. The last part follows from Lemma 10 of Bertsimas and McCord [13]. \square

Jiang [47] presents improved rates of convergence and finite sample guarantees for kNN regression in the homoscedastic setting when the data \mathcal{D}_n lies on a low-dimensional manifold. Lemma 7 of Bertsimas and McCord [13] presents conditions under which CART regression satisfies Assumption 5. Along with Theorem 8 of Wager and Athey [82], the above result can be used to show that Assumption 15 holds for CART regression with $\alpha = \frac{O(1)}{d_x}$. Lemma 9 of Bertsimas and McCord [13] presents conditions under which RF regression satisfies Assumption 5. Once again, we can use this result along with Theorem 8 of Wager and Athey [82] to show that Assumption 15 holds for RF regression with $\alpha = \frac{O(1)}{d_x}$.

F.4 Verifying assumptions on the estimation of Q^*

We verify that the parts of Assumptions 5, 11, and 15 relating to the estimate \hat{Q}_n of Q^* hold for some setups with structured heteroscedasticity. The techniques in Section F.2 may then be used to verify the assumptions for the jackknife-based estimators in Section D when the data $\{(x^i, \varepsilon^i)\}$ is i.i.d. These assumptions for \hat{Q}_n —in particular, Assumption 11—are not as well-studied in the literature as those for \hat{f}_n and are typically harder to verify. Because deriving theoretical properties of estimators in the heteroscedastic setting and deriving finite sample properties of estimators in general are areas of topical interest, we envision that future research will enable easier verification of these assumptions. For simplicity, we only consider function classes \mathcal{Q} that comprise diagonal covariance matrices [86], i.e.,

$$\mathcal{Q} := \{Q : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_y \times d_y} : Q(X) = \text{diag}(q_1(X), q_2(X), \dots, q_{d_y}(X))\},$$

although the ER-SAA approach is more generally applicable. Bauwens et al. [8] review some model classes \mathcal{Q} with non-diagonal covariance matrices that are popular in time series modeling.

F.4.1 Parametric models for heteroscedasticity

We assume that $Q^*(X) \equiv Q(X; \pi^*)$ for some finite-dimensional parameter $\pi^* \in \Pi$ and the goal is to estimate π^* . Forms of the functions q_j of interest include [62, 66]:

- i. $(q_j(X; \pi))^2 = ((\pi^j)^\top X)^2$,
- ii. $(q_j(X; \pi))^2 = \exp((\pi^j)^\top X)$,
- iii. $(q_j(X; \pi))^2 = \exp((\pi^j)^\top \log(|X|))$, where $X_1 \equiv 2$ (instead of $X_1 \equiv 1$).

The above setup can also accommodate cases where the parameters of the function Q^* include some of the parameters of the function f^* .

Let $\hat{\pi}_n$ denote the estimate of π^* corresponding to the regression estimate \hat{Q}_n . Suppose for a.e. realization $x \in \mathcal{X}$, the function $Q(x; \cdot)$ is Lipschitz continuous with Lipschitz constant $L_Q(x)$ and its inverse $[Q(x; \cdot)]^{-1}$ is also Lipschitz continuous with Lipschitz constant $\bar{L}_Q(x)$. These assumptions hold for the above parametric models if the parameters π therein are restricted to lie in a compact sets (similar to Section F.2, it suffices to assume that the above Lipschitz continuity assumptions hold locally for the asymptotic results). Because

$$\|\hat{Q}_n(x) - Q^*(x)\| \leq L_Q(x) \|\hat{\pi}_n - \pi^*\|, \quad \frac{1}{n} \sum_{i=1}^n \|[\hat{Q}_n(x^i)]^{-1} - [Q^*(x^i)]^{-1}\|^2 \leq \left(\frac{1}{n} \sum_{i=1}^n \bar{L}_Q^2(x^i) \right) \|\hat{\pi}_n - \pi^*\|^2,$$

asymptotic and finite sample guarantees on the estimator $\hat{\pi}_n$ of π^* directly translate to the asymptotic and finite sample guarantees on the estimate \hat{Q}_n in Assumptions (5c), (5d), (15c), (15d), (11c), and (11d). When the functions f^* and Q^* are jointly estimated using an M-estimation procedure, the results in Section F.2 provide conditions under which the estimator $\hat{\pi}_n$ of π^* is consistent and Assumptions (5c), (5d), (15c), and (15d) hold with $\alpha = 1$. Section F.2 also presents a hard-to-verify uniform exponential bound condition under which $\hat{\pi}_n$ possesses a finite sample guarantee. Carroll and Ruppert [20] consider robust M-estimators for π^* that possess a similar rate of convergence when f^* is linear. Dalalyan et al. [25] present asymptotic and finite sample guarantees for a scaled Dantzig estimator of π^* under some sparsity assumptions. Finally, Fan et al. [38] present a quasi-maximum likelihood approach for estimating the parameters of GARCH models and investigate their asymptotic properties.

In the remainder of this section, we specialize the verification of Assumptions (5c), (5d), (15c), and (15d), to Example 2. We are unable to verify Assumptions (11c) and (11d) because the literature lacks suitable finite-sample guarantees for heteroscedasticity estimation.

Verifying Assumptions 5 and 7. We verify Assumptions (5c), (5d), and (7c) for Example 2. Since

$$\begin{aligned} \|\hat{Q}_n(x) - Q^*(x)\| &\leq \max_{j \in [d_y]} |\hat{q}_{jn}(x) - q_j^*(x)| \\ &= \max_{j \in [d_y]} \left| \left(\exp \left(\sum_{k=1}^{d_x} \hat{\pi}_{kn}^j \log(|x_k|) \right) \right)^{1/2} - \left(\exp \left(\sum_{k=1}^{d_x} \pi_k^{j*} \log(|x_k|) \right) \right)^{1/2} \right|, \end{aligned}$$

we have that Assumption (5c) holds whenever $\hat{\pi}_n \xrightarrow{P} \pi^*$ on account of Slutsky's lemma and the continuous mapping theorem. Appendix B.2 of Romano and Wolf [66] identifies conditions under which $\hat{\pi}_n \xrightarrow{P} \pi^*$ for Example 2. To verify Assumption (5d), assume for simplicity that the support \mathcal{X} is compact and bounded away from the origin and the estimates $\{\hat{\pi}_n\}$ lie in a compact set a.s. for n large enough. Since

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \|\hat{Q}_n(x^i)^{-1} - [Q^*(x^i)]^{-1}\|^2 &\leq \frac{1}{n} \sum_{i=1}^n \max_{j \in [d_y]} |(\hat{q}_{jn}(x^i))^{-1} - (q_j^*(x^i))^{-1}|^2 \\ &\leq \frac{1}{n} \sum_{i=1}^n \max_{j \in [d_y]} |(\hat{q}_{jn}(x^i))^{-2} - (q_j^*(x^i))^{-2}| \\ &= \frac{1}{n} \sum_{i=1}^n \max_{j \in [d_y]} \left| \exp \left(- \sum_{k=1}^{d_x} \hat{\pi}_{kn}^j \log(|x_k^i|) \right) - \exp \left(- \sum_{k=1}^{d_x} \pi_k^{j*} \log(|x_k^i|) \right) \right| \\ &\leq \frac{1}{n} \sum_{i=1}^n O(1) \max_{j \in [d_y]} \left| \sum_{k=1}^{d_x} (\pi_k^{j*} - \hat{\pi}_{kn}^j) \log(|x_k^i|) \right| \\ &\leq O(1) \max_{j \in [d_y]} \sum_{k=1}^{d_x} |\pi_k^{j*} - \hat{\pi}_{kn}^j|, \end{aligned}$$

where the last two steps follow by the mean-value theorem, the assumption that \mathcal{X} is compact and bounded away from the origin, and the assumption that the sequence $\{\hat{\pi}_n\}$ lies in a compact set a.s. for n large enough (note that the compactness assumption on \mathcal{X} can be relaxed, e.g., if the sequence $\{\hat{\pi}_n\}$ lies in the nonnegative orthant and $|\bar{x}| \geq 1, \forall \bar{x} \in \mathcal{X}$). Consequently, Assumption (5d) also holds whenever $\hat{\pi}_n \xrightarrow{P} \pi^*$. The above arguments also imply $\sup_{x \in \mathcal{X}} \|[Q^*(x)]^{-1} - [\hat{Q}_n(x)]^{-1}\| \xrightarrow{P} 0$. These conditions also guarantee that Assumption (7c) holds since

$$\begin{aligned} \|\hat{Q}_n(X) - Q^*(X)\|_{L^q} &\leq \left(\mathbb{E} \left[\sum_{j=1}^{d_y} |\hat{q}_{jn}(X) - q_j^*(X)|^q \right] \right)^{1/q} \\ &\leq \left(\mathbb{E} \left[\sum_{j=1}^{d_y} \left| \left(\exp \left(\sum_{k=1}^{d_x} \hat{\pi}_{kn}^j \log(|X_k|) \right) \right)^{1/2} - \left(\exp \left(\sum_{k=1}^{d_x} \pi_k^{j*} \log(|X_k|) \right) \right)^{1/2} \right|^q \right] \right)^{1/q} \\ &\leq \left(\sum_{j=1}^{d_y} \mathbb{E} \left[O(1) \left| \sum_{k=1}^{d_x} (\hat{\pi}_{kn}^j - \pi_k^{j*}) \log(|X_k|) \right|^q \right] \right)^{1/q} \\ &\leq \left(\sum_{j=1}^{d_y} O(1) \left| \sum_{k=1}^{d_x} |\hat{\pi}_{kn}^j - \pi_k^{j*}| \right|^q \right)^{1/q} \\ &\leq \sum_{j=1}^{d_y} O(1) \sum_{k=1}^{d_x} |\hat{\pi}_{kn}^j - \pi_k^{j*}|, \end{aligned}$$

where the third and fourth steps again follow by the mean-value theorem, the compactness of \mathcal{X} and the fact that it is bounded away from the origin, and the a.s. compactness of the sequence $\{\hat{\pi}_n\}$ for n large enough. The above arguments also imply $\sup_{x \in \mathcal{X}} \|Q^*(x) - \hat{Q}_n(x)\| \xrightarrow{P} 0$.

Verifying Assumption 15. We show that Assumptions (15c) and (15d) hold whenever $\|\hat{\pi}_n - \pi^*\| = O_p(n^{-\alpha/2})$, the support \mathcal{X} is compact and bounded away from the origin, and the estimates $\{\hat{\pi}_n\}$ lie in a compact set a.s. for n large enough. Note that

$$\begin{aligned} \sup_{x \in \mathcal{X}} \|Q^*(x) - \hat{Q}_n(x)\| &\leq \max_{j \in [d_y]} \sup_{x \in \mathcal{X}} |\hat{q}_{jn}(x) - q_j^*(x)| \\ &= \max_{j \in [d_y]} \sup_{x \in \mathcal{X}} \left| \left(\exp \left(\sum_{k=1}^{d_x} \hat{\pi}_{kn}^j \log(|x_k|) \right) \right)^{1/2} - \left(\exp \left(\sum_{k=1}^{d_x} \pi_k^{j*} \log(|x_k|) \right) \right)^{1/2} \right| \\ &\leq \max_{j \in [d_y]} O(1) \sum_{k=1}^{d_x} |\pi_k^{j*} - \hat{\pi}_{kn}^j|, \end{aligned}$$

where the last step follows by arguments similar to the derivation of Assumption (7c) using the mean-value theorem, the compactness of \mathcal{X} and the fact that it is bounded away from the origin, and the a.s. compactness of the sequence $\{\hat{\pi}_n\}$ for n large enough. Therefore, Assumption (15c) holds whenever $\|\hat{\pi}_n - \pi^*\| = O_p(n^{-\alpha/2})$. Similarly, we have

$$\begin{aligned} \sup_{x \in \mathcal{X}} \|[Q^*(x)]^{-1} - [\hat{Q}_n(x)]^{-1}\| &\leq \max_{j \in [d_y]} \sup_{x \in \mathcal{X}} |(\hat{q}_{jn}(x))^{-1} - (q_j^*(x))^{-1}| \\ &\leq \max_{j \in [d_y]} \sup_{x \in \mathcal{X}} \left| \left(\exp \left(- \sum_{k=1}^{d_x} \hat{\pi}_{kn}^j \log(|x_k|) \right) \right)^{1/2} - \left(\exp \left(- \sum_{k=1}^{d_x} \pi_k^{j*} \log(|x_k|) \right) \right)^{1/2} \right| \\ &\leq \max_{j \in [d_y]} O(1) \sum_{k=1}^{d_x} |\pi_k^{j*} - \hat{\pi}_{kn}^j|, \end{aligned}$$

by arguments similar to the derivation of Assumption (5d). Therefore, Assumption (15d) holds whenever $\|\hat{\pi}_n - \pi^*\| = O_p(n^{-\alpha/2})$. Appendix B.2 of Romano and Wolf [66] identifies conditions under which $\hat{\pi}_n \xrightarrow{p} \pi^*$ for Example 2 at a certain rate.

F.4.2 Nonparametric models for heteroscedasticity

We assume that each function $q_j : \mathbb{R}^{d_x} \rightarrow \mathbb{R}_+$ is ‘sufficiently smooth’. Chapter 8 of Fan and Yao [37] presents some popular models for the functions q_j in a time series context.

Suppose the function Q^* and its regression estimate \hat{Q}_n are (asymptotically) a.s. uniformly invertible, i.e., $\sup_{\bar{x} \in \mathcal{X}} \|[Q^*(\bar{x})]^{-1}\| < +\infty$ and $\sup_{\bar{x} \in \mathcal{X}} \|[\hat{Q}_n(\bar{x})]^{-1}\| < +\infty$. We have

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \|[\hat{Q}_n(x^i)]^{-1} - [Q^*(x^i)]^{-1}\|^2 &\leq \frac{1}{n} \sum_{i=1}^n \|[Q^*(x^i)]^{-1}\|^2 \|[\hat{Q}_n(x^i)]^{-1}\|^2 \|\hat{Q}_n(x^i) - Q^*(x^i)\|^2 \\ &\leq \left(\sup_{\bar{x} \in \mathcal{X}} \|[Q^*(\bar{x})]^{-1}\|^2 \right) \left(\sup_{\bar{x} \in \mathcal{X}} \|[\hat{Q}_n(\bar{x})]^{-1}\|^2 \right) \left(\frac{1}{n} \sum_{i=1}^n \|\hat{Q}_n(x^i) - Q^*(x^i)\|^2 \right) \end{aligned}$$

Therefore, asymptotic and finite sample guarantees for $\|\hat{Q}_n(x) - Q^*(x)\|$ and $\frac{1}{n} \sum_{i=1}^n \|\hat{Q}_n(x^i) - Q^*(x^i)\|^2$ are sufficient for verifying Assumptions (5c), (5d), (15c), (15d), (11c), and (11d). Theorem 8.5 of Fan and Yao [37] can be used to identify conditions under which these asymptotic guarantees hold for local linear estimators on time series data when the dimension of the covariates $d_x = 1$. They also mention approaches for estimating Q^* when $d_x > 1$. Theorem 2 of Ruppert et al. [68] can be used to verify Assumptions 5 and 15 for local polynomial smoothers. Proposition 2.1 and Theorem 3.1 of Jin et al. [48] identify conditions under which Assumptions 5 and 15 hold for a local likelihood estimator. Van Keilegom and Wang [80] consider semiparametric models for both f^* and Q^* . Theorems 3.1 and 3.2 therein can be used to verify Assumptions 5 and 15 for the estimates \hat{Q}_n . Section 3 of Zhou et al. [86] presents robust estimators of Q^* when f^* is linear and notes that these estimators \hat{Q}_n possess asymptotic and finite sample guarantees in the form of Assumptions 5, 15, and 11. Finally, Theorem 3.1 of Chesneau et al. [23] can be used to derive asymptotic guarantees for wavelet estimators of Q^* .

G Omitted details for the computational experiments

The parameters φ^* and ζ^* in the true demand model are specified as:

$$\varphi_j^* = 50 + 5\delta_{j0}, \quad \zeta_{j1}^* = 10 + \delta_{j1}, \quad \zeta_{j2}^* = 5 + \delta_{j2}, \quad \text{and} \quad \zeta_{j3}^* = 2 + \delta_{j3}, \quad \forall j \in \mathcal{J},$$

where $\{\delta_{j0}\}_{j \in \mathcal{J}}$ are i.i.d. samples from the standard normal distribution $\mathcal{N}(0, 1)$, and $\{\delta_{j1}\}_{j \in \mathcal{J}}$, $\{\delta_{j2}\}_{j \in \mathcal{J}}$, and $\{\delta_{j3}\}_{j \in \mathcal{J}}$ are i.i.d. samples from the uniform distribution $U(-4, 4)$. We generate i.i.d. samples of the covariates X from a *multivariate folded/half-normal distribution*. We specify the underlying normal distribution to have mean $\mu_X = 0$ and set its covariance matrix Σ_X to be a random correlation matrix that is generated using the ‘vine method’ of Lewandowski et al. [55] (each partial correlation is sampled from the Beta(2, 2) distribution and rescaled to $[-1, 1]$). Finally, Algorithm 1 describes our procedure for estimating the normalized 99% UCB on the optimality gap of our data-driven solutions using the multiple replication procedure [57].

Algorithm 1 Estimating the normalized 99% UCB on the optimality gap of a given solution.

- 1: **Input:** Covariate realization $X = x$ and data-driven solution $\hat{z}_n(x)$ for a particular realization of the data \mathcal{D}_n .
- 2: **Output:** $\hat{B}_{99}(x)$, which is a normalized estimate of the 99% UCB on the optimality gap of $\hat{z}_n(x)$.
- 3: **for** $k = 1, \dots, 30$ **do**
- 4: Draw 1000 i.i.d. samples $\bar{\mathcal{D}}^k := \{\bar{\varepsilon}^{k,i}\}_{i=1}^{1000}$ of ε according to the distribution P_ε .
- 5: Estimate the optimal value $v^*(x)$ by solving the full-information SAA problem (2) using the data $\bar{\mathcal{D}}^k$:

$$\bar{v}^k(x) := \min_{z \in \mathcal{Z}} \frac{1}{1000} \sum_{i=1}^{1000} c(z, f^*(x) + Q^*(x)\bar{\varepsilon}^{k,i}).$$

- 6: Estimate the out-of-sample cost of the solution $\hat{z}_n(x)$ using the data $\bar{\mathcal{D}}^k$:

$$\hat{v}^k(x) := \frac{1}{1000} \sum_{i=1}^{1000} c(\hat{z}_n(x), f^*(x) + Q^*(x)\bar{\varepsilon}^{k,i}).$$

- 7: Estimate the optimality gap of the solution $\hat{z}_n(x)$ as $\hat{G}^k(x) = \hat{v}^k(x) - \bar{v}^k(x)$.
- 8: **end for**
- 9: Construct the normalized estimate of the 99% UCB on the optimality gap of $\hat{z}_n(x)$ as

$$\hat{B}_{99}(x) := \frac{100}{|\bar{v}(x)|} \left(\frac{1}{30} \sum_{k=1}^{30} \hat{G}^k(x) + 2.462 \sqrt{\frac{\text{var}(\{\hat{G}^k(x)\})}{30}} \right),$$

where $\bar{v}(x) := \frac{1}{30} \sum_{k=1}^{30} \bar{v}^k(x)$ and $\text{var}(\{\hat{G}^k(x)\})$ denotes the variance of the gaps $\{\hat{G}^k(x)\}_{k=1}^{30}$.

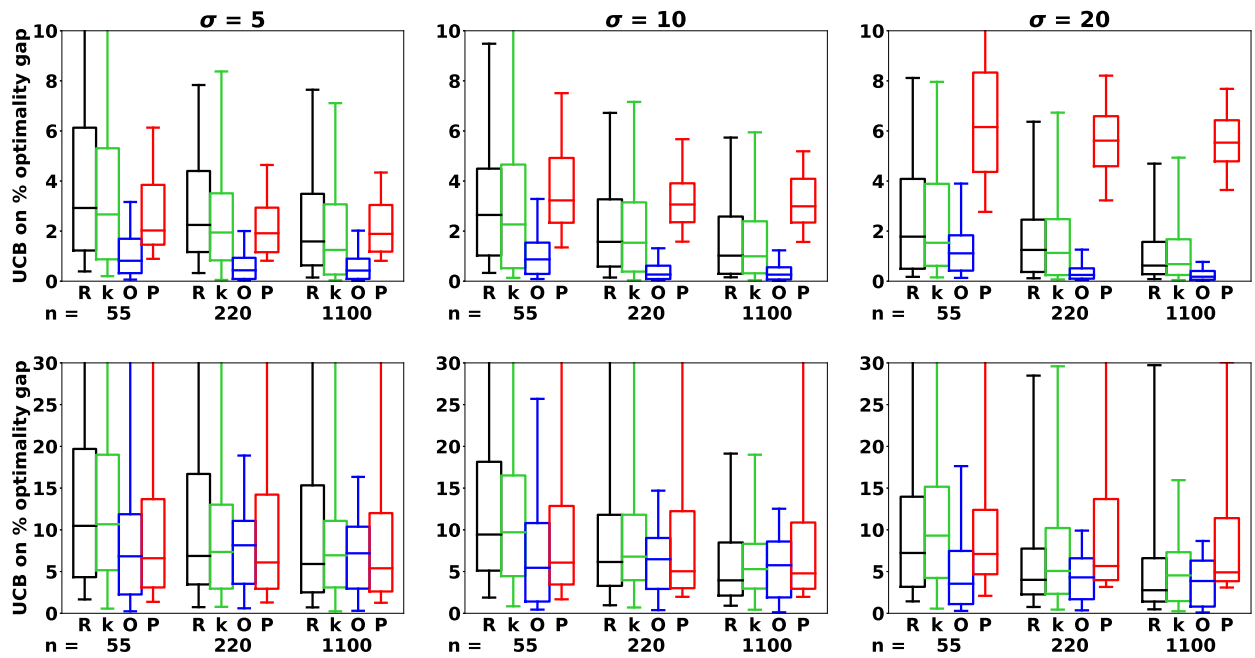


Figure 7: Effect of increasing σ on re-weighted kNN-SAA (R), ER-SAA+kNN (k), ER-SAA+OLS (O), and PP+OLS (P) approaches when $\omega = 1$. Top row: $p = 0.5$. Bottom row: $p = 2$. Left column: $\sigma = 5$. Middle column: $\sigma = 10$. Right column: $\sigma = 20$.