Submitted to
manuscript (Please, provide the manuscript number!)

Data-Driven Sample Average Approximation with Covariate Information

Rohit Kannan*
Wisconsin Institute for Discovery, University of Wisconsin-Madison, Madison, WI, USA. rohit.kannan@wisc.edu

Güzin Bayraksan
Department of Integrated Systems Engineering, The Ohio State University, Columbus, OH, USA. bayraksan.1@osu.edu

James R. Luedtke
Department of Industrial & Systems Engineering and Wisconsin Institute for Discovery, University of Wisconsin-Madison, Madison, WI, USA. jim.luedtke@wisc.edu

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, convergence rate, large deviations

History:

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 (Shapiro

*I dedicate this work to my grandfather (Thatha) and to the memory of my grandmother (Thathi).

et al. 2009, Birge and Louveaux 2011) is a popular modeling framework for decision-making under uncertainty in such applications. A standard formulation of a stochastic program is

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

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 (Shapiro et al. 2009, Homem-de Mello and Bayraksan 2014). 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 (2018), Bertsimas and Kallus (2019), and Sen and Deng (2018), 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

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

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 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 (Ban and Rudin 2018), where the product's demand can be predicted using seasonality and location data before making order decisions, (ii) dynamic procurement of a new product (Ban et al. 2018) whose demand can be predicted using historical data for similar past products, (iii) shipment planning under uncertainty (Bertsimas and Kallus 2019), 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 (Donti et al. 2017), 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. (2018) and Sen and Deng (2018), in which data-driven approximations to (SP) are constructed using explicit models to 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 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 (Barber et al. 2019), 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 (van der Vaart 1998, Györfi et al. 2006, Wainwright 2019). While our results imply that using nonparametric regression techniques within our SAA frameworks results in convergent approximations to (SP) under mild assumptions (cf. Bertsimas and Kallus 2019), 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.

Besides a few exceptions (e.g., Homem-de-Mello 2008), much of the existing SAA theory (Shapiro et al. 2009) focuses on the case when we have independent and identically distributed (i.i.d.) samples of Y. In this work, we establish our results using an abstract set of assumptions, and verify that our assumptions hold for i.i.d. data and for some regression setups with dependent data satisfying widely-used mixing/stationarity assumptions.

1.1. Relation to existing literature

The papers of Ban et al. (2018) and Sen and Deng (2018) 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. (2018) propose a residual tree method for the data-driven solution of multi-stage 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 (2018) 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. (2018). 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 (2015) 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 for a wide range of prediction and optimization models; we establish asymptotic optimality of the solutions from 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, even when the prediction model is misspecified.

Bertsimas and Kallus (2019) 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 (2020) develop stochastic quasigradient methods for efficiently solving the kNN and kernel-based reweighted SAA formulations of Bertsimas and Kallus (2019) without sacrificing theoretical guarantees. Ban and Rudin (2018) 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. Bertsimas and McCord (2019) extend the analysis of Bertsimas and Kallus (2019) to the multi-stage setting when the covariates evolve according to a Markov process. They establish asymptotic optimality and consistency of their data-driven decisions, and also establish finite sample guarantees for the solutions to the kNN-based approach.

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 Donti et al. 2017, Elmachtoub and Grigas 2017, Davarnia et al. 2018). 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.

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. (2009), Royset and Wets (2014), and the references therein for applications-motivated versions. 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 the ER-SAA formulation 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):

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

where $X = x$ is a new random observation of the covariates, the expectation is taken with respect to the conditional distribution of Y given $X = x$, and c is an extended real-valued function defined on $\mathbb{R}^{d_z} \times \mathbb{R}^{d_y}$. 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 'true relationship' between the random vector Y and the covariates X is described by the additive error model $Y = f^*(X) + \varepsilon$, where $f^*(x) := \mathbb{E}\left[Y \mid X = x\right]$ is the regression function and the random variable ε is the associated regression error. Following previous work (Ban et al. 2018, Sen and Deng 2018), we assume that the errors ε are independent of the covariates X and that $\mathbb{E}\left[\varepsilon\right] = 0$. We also assume that f^* belongs to a known class of functions $\mathcal{F}$ with domain a subset of $\mathbb{R}^{d_x}$ and codomain a subset of $\mathbb{R}^{d_y}$. The model class $\mathcal{F}$ can be infinite dimensional and depend on the number of data samples. Let Ξ denote the support of ε and P_ε denote its distribution.

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

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

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, $\mathbb{E}\left[|c(z, f^*(x) + \varepsilon)|\right] < +\infty$ for each $z \in \mathcal{Z}$ and almost every (a.e.) $x \in \mathcal{X}$, and the function $g(\cdot; x)$ is lower semicontinuous on $\mathcal{Z}$ for a.e. $x \in \mathcal{X}$ (see Theorem 7.42 of Shapiro et al. (2009) for conditions that guarantee lower semicontinuity of $g(\cdot; x)$). These assumptions ensure that problem (1) is well defined and the solution set $S^*(x)$ is nonempty for a.e. $x \in \mathcal{X}$.

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

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

where $\{\varepsilon^i\}_{i=1}^n$ denote the realizations of the errors at the given observations, i.e., $\varepsilon^i := y^i - f^*(x^i)$, $\forall i \in \{1, \cdots, n\}$. We cannot solve problems (1) or (2) directly because the regression function f^* is unknown. A practical alternative is to estimate f^* from the data $\mathcal{D}_n$, for instance by using an M-estimator (van der Vaart 1998, van de Geer 2000) of the form

$$\hat{f}_n(\cdot) \in \underset{f(\cdot) \in \mathcal{F}}{\arg\min} \frac{1}{n} \sum_{i=1}^n \ell\left(y^i, f(x^i)\right) \tag{3}$$

with some loss function $\ell : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \to \mathbb{R}_+$. We sometimes assume that the regression model class $\mathcal{F}$ is parameterized by θ (e.g., the parameters of an OLS 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 $\hat{\theta}_n$. Throughout, we will reference equation (3) for the regression step with the understanding that our prediction framework is not restricted to M-estimation.

Given an estimate $\hat{f}_n$ of f^*, the residuals $\hat{\varepsilon}_n^i := y^i - \hat{f}_n(x^i)$, $i \in \{1, \cdots, n\}$, of this estimate can be used as a proxy for samples of ε from the distribution P_ε. The *empirical residuals-based SAA* (ER-SAA) to problem (1) is defined as (cf. Ban et al. 2018, Sen and Deng 2018)

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

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 $y \in \mathbb{R}^{d_y}$. 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 observation $X = x$:

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

The computational complexity of solving the ER-SAA problem (4) is similar to that of solving the N-SAA problem (5) with the only additional computation cost being the cost of estimating the function f^*.

We also propose two alternatives to the ER-SAA problem (4) that construct the scenarios differently. For each $i \in \{1,\dots,n\}$, let $\hat{f}_{-i}$ denote the estimate of f^* 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 $\hat{\varepsilon}^i_{n,J} := y^i - \hat{f}_{-i}(x^i)$. The alternatives we propose are

$$\hat{v}^J_n(x) := \min_{z\in\mathcal{Z}} \left\{ \hat{g}^J_n(z;x) := \frac{1}{n}\sum_{i=1}^{n} c\left(z, \hat{f}_n(x) + \hat{\varepsilon}^i_{n,J}\right) \right\}, \tag{6}$$

$$\hat{v}^{J+}_n(x) := \min_{z\in\mathcal{Z}} \left\{ \hat{g}^{J+}_n(z;x) := \frac{1}{n}\sum_{i=1}^{n} c\left(z, \hat{f}_{-i}(x) + \hat{\varepsilon}^i_{n,J}\right) \right\}. \tag{7}$$

We call problems (6) and (7) *Jackknife-based SAA* (J-SAA) and *Jackknife+-based SAA* (J+-SAA), respectively (cf. Barber et al. 2019). These data-driven SAAs are well-motivated when the data $\mathcal{D}_n$ is i.i.d., in which case the leave-one-out residual $\hat{\varepsilon}^i_{n,J}$ may be a significantly more accurate estimate of the prediction error at the covariate observation x^i than the empirical residual $\hat{\varepsilon}^i_n$, particularly when n is small. When $\mathcal{D}_n$ is not i.i.d., 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 (Lahiri 2013).

Problems (6) and (7) 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. (2019) for other regression setups that can re-use information). We make use of this computational speed-up in our experiments in Section 4. A simple but less data efficient alternative to the ER-SAA and Jackknife-based SAA frameworks is sample-splitting, which leaves out a fraction of the data $\mathcal{D}_n$ from the regression step and uses the out-of-sample residuals on the held-out data along with the regression model to construct scenarios. Another option to mitigate the computational cost of the regression steps for large n is to use K-fold cross-validation (CV) variants of the J-SAA and J+-SAA formulations with $K \ll n$.

We use the following two-stage stochastic linear program (LP) as our running example for problem (1). See Appendix EC.2 (in the electronic companion) for discussion of more general forms of problem (1) that satisfy the assumptions of our framework.

Example 1 (Two-stage stochastic LP). The set $\mathcal{Z}$ is a nonempty convex polytope and the function $c(z,Y) := c_z^{\mathrm{T}} z + Q(z,Y)$, with $Q(z,Y) := \min_{v\in\mathbb{R}^{d_v}_+} \{q_v^{\mathrm{T}} v : Wv = Y - Tz\}$. The quantities c_z, q_v, W, and T have commensurate dimensions. We assume that $Q(z,y) < +\infty$ 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 $\{\lambda : \lambda^{\mathrm{T}} W \le q_v^{\mathrm{T}}\}$ is nonempty.

We also use OLS regression as our running example for the regression step (3). See Appendix EC.3 (in the electronic companion) for a detailed discussion of how other prediction models fit within our framework.

Example 2 (OLS regression). The model class is $\mathcal{F} := \big\{ f(\cdot) : f(X) = \theta X \text{ for some } \theta \in \mathbb{R}^{d_y \times d_x} \big\}$, where the constant term is included as a covariate, and the loss function is $\ell(y, \hat{y}) := \|y - \hat{y}\|^2$. We assume that the regression function is $f^*(X) = \theta^* X$ for some $\theta^* \in \mathbb{R}^{d_y \times d_x}$, and estimate θ^* by $\hat{\theta}_n \in \underset{\theta \in \mathbb{R}^{d_y \times d_x}}{\arg\min} \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \|y^i - \theta x^i\|^2$.

There is an inherent tradeoff between using parametric and nonparametric regression techniques for estimating the function f^*. If the function class $\mathcal{F}$ is correctly specified, then the use of parametric regression approaches may yield much faster rates of convergence of the data-driven SAA estimators relative to the use of nonparametric approaches (see Section 3.2). 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 don't have access to an abundance of data. Note that even if the model class $\mathcal{F}$ is incorrectly specified, the sequence of regression estimates $\{\hat{f}_n\}$ will converge to the best approximation of f^* in $\mathcal{F}$ under mild assumptions. The analysis in Section 3 can then be used to characterize the asymptotic properties of our data-driven SAA estimators even though they are not guaranteed to be consistent in this case.

Notation. Let $[n]$ denote the set $\{1, \dots, n\}$, $|S|$ denote the cardinality of a finite set S, $\|\cdot\|$ denote the Euclidean norm, $\|\cdot\|_0$ denote the ℓ_0 "norm", $\mathcal{B}_\delta(v)$ denote a Euclidean ball of radius $\delta > 0$ around a point v, and $M_{[j]}$ denote the j^{th} row of a matrix M. For sets $A, B \subseteq \mathbb{R}^{d_z}$, let $\mathbb{D}(A, B) := \sup_{v \in A} \operatorname{dist}(v, B)$ denote the deviation of A from B, where $\operatorname{dist}(v, B) := \inf_{w \in B} \|v - w\|$. A random vector V is said to be sub-Gaussian with variance proxy σ^2 if $\mathbb{E}[V] = 0$ and $\mathbb{E}[\exp(s u^{\mathrm{T}} V)] \leq \exp(0.5\sigma^2 s^2)$, $\forall s \in \mathbb{R}$ and $\|u\| = 1$. The abbreviations 'a.e.' (defined earlier), '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 $\mathcal{D}_n$ 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 $\{W_n\}$, the notation $V_n = o_p(W_n)$ and $V_n = O_p(W_n)$ convey that $V_n = R_n W_n$ with the sequence $\{R_n\}$ converging in probability to zero ($R_n \xrightarrow{p} 0$), or being bounded in probability, respectively (see Chapter 2 of van der Vaart (1998) for basic theory). We write $\tilde{o}$ to hide polylogarithmic factors in n, and $O(1)$ to denote generic constants. We ignore measurability-related issues throughout this work (see van der Vaart and Wellner (1996) and Shapiro et al. (2009) 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, we investigate conditions under which solutions to problem (4) are asymptotically optimal and consistent, analyze the rate of convergence of the optimal value of problem (4) to that of problem (1), and develop finite sample guarantees for solutions to problem (4) using large deviations theory. Omitted proofs are provided in Appendix A. We outline the modifications required to analyze the J-SAA and J+-SAA methods in Section 3.4 (see also Appendix EC.1 in the electronic companion). In Appendices EC.2 and EC.3, we verify that a variety of stochastic optimization and prediction setups satisfy the assumptions made in this section.

We establish our results by bounding the 'deviation' of the ER-SAA problem (4) from the full-information SAA problem (2). To facilitate this analysis, denote by $\tilde{\varepsilon}^i_n(x)$ the difference between the ith ER-SAA scenario $(\hat{f}_n(x) + \hat{\varepsilon}^i_n)$ and the corresponding 'true realization' $(f^*(x) + \varepsilon^i)$ of Y given $X = x$, i.e.,

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

The interpretation of $\tilde{\varepsilon}^i_n(x)$ as the sum of the *prediction error* at the point $x \in \mathcal{X}$ and the *estimation error* at the training point $x^i \in \mathcal{X}$ motivates our assumptions in this section.

3.1. Consistency and asymptotic optimality

We begin by investigating conditions under which the optimal value and optimal solutions to the ER-SAA problem (4) asymptotically converge to those of the true problem (1) as the number of data samples n tends to infinity. We make either one of the below assumptions to establish uniform convergence of the sequence of objective functions of the ER-SAA problem (4) to the objective function of the true problem (1) on the feasible region $\mathcal{Z}$.

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

$$|c(z, \bar{y}) - c(z, y)| \leq L(z)\|\bar{y} - y\|, \quad \forall y, \bar{y} \in \mathbb{R}^{d_y},$$

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

Assumption 2. *Problem* (1)*, the regression step* (3)*, and the data $\mathcal{D}_n$ satisfy for a.e. $x \in \mathcal{X}$:*

(2a) there exists a function $\delta : \mathcal{X} \to \mathbb{R}_+$ such that $\|\tilde{\varepsilon}^i_n(x)\| \leq \delta(x)$, $\forall i \in [n]$, a.s. for n large enough,

(2b) for each $z \in \mathcal{Z}$, the function c in problem (1) *satisfies for a.e. $\varepsilon \in \Xi$:*

$$|c(z, \bar{y}) - c(z, f^*(x) + \varepsilon)| \leq L_{\delta(x)}(z, f^*(x) + \varepsilon)\|\bar{y} - (f^*(x) + \varepsilon)\|, \quad \forall \bar{y} \in \mathcal{B}_{\delta(x)}(f^*(x) + \varepsilon),$$

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

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

Assumption 1 requires the function $c(z,\cdot)$ to be globally Lipschitz continuous for each $z \in \mathcal{Z}$. We show in Appendix EC.2 that it is readily satisfied by Example 1. Assumption (2b), on the other hand, 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 (2a) on the regression step (3) in the forthcoming results. Because the deviation terms satisfy $\|\tilde{\varepsilon}^i_n(x)\| \leq \|f^*(x) - \hat{f}_n(x)\| + \|f^*(x^i) - \hat{f}_n(x^i)\|$, $\forall i \in [n]$, Assumption (2a) is satisfied 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 $\mathbb{E}\left[\|\varepsilon\|^2\right] < +\infty$ (see Theorem 5.4 of Shapiro et al. (2009) for details). We present conditions under which Assumption (2b) holds in Appendix EC.2.

The next assumption is also needed to establish uniform convergence of the sequence of objective functions of the ER-SAA problem (4) to the objective function of the true problem (1) on $\mathcal{Z}$.

Assumption 3. *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 3 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 observations $\mathcal{D}_n$ are i.i.d. (see Theorem 7.48 of Shapiro et al. 2009). Using pointwise LLN results in Walk (2010) and White (2014), we can show that Assumption 3 also holds for some mixing/stationary processes by noting that the proof of Theorem 7.48 of Shapiro et al. (2009) also extends to these settings.

Finally, we also need the following assumption on the consistency of the regression step (3).

Assumption 4. *The regression procedure* (3) *satisfies the following consistency properties:*

(4a) Pointwise error consistency: $\hat{f}_n(x) \xrightarrow{p} f^*(x)$ for a.e. $x \in \mathcal{X}$,

(4b) Mean-squared estimation error consistency: $\displaystyle\frac{1}{n}\sum_{i=1}^{n}\|f^*(x^i) - \hat{f}_n(x^i)\|^2 \xrightarrow{p} 0$.

Assumption (4a) holds for our running example of OLS regression if the parameter estimates $\hat{\theta}_n$ are weakly consistent (i.e., $\hat{\theta}_n \xrightarrow{p} \theta^*$), and Assumption (4b) holds if, in addition, the weak LLN $\frac{1}{n}\sum_{i=1}^{n}\|x^i\|^2 \xrightarrow{p} \mathbb{E}\left[\|X\|^2\right]$ is satisfied (see Chapter 3 of White (2014) for various assumptions on the data $\mathcal{D}_n$ and the distribution P_X 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 (van de Geer 2000). Assumption 4 is implied by the stronger assumption of uniform convergence of the estimate $\hat{f}_n$ to the function f^* on the support $\mathcal{X}$ of the covariates, i.e., when $\sup_{x \in \mathcal{X}}\|f^*(x) - \hat{f}_n(x)\| \xrightarrow{p} 0$. Appendix EC.3 expands on the above arguments and shows that Assumption 4 also holds for Lasso, kNN, and RF regression under certain conditions.

The following result implies that the quadratic mean of the deviation terms $\{\tilde{\varepsilon}^i_n(x)\}_{i=1}^n$ vanishes in the limit in probability for a.e. $x \in \mathcal{X}$ under Assumption 4.

LEMMA 1. *For any $x \in \mathcal{X}$, the mean-squared deviation can be bounded from above as*

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

Proof. Follows from $\|\tilde{\varepsilon}^i_n(x)\| \leq \|f^*(x) - \hat{f}_n(x)\| + \|f^*(x^i) - \hat{f}_n(x^i)\|$, $\forall i \in [n]$, squaring both sides of this inequality, and using the arithmetic mean-quadratic mean (AM-QM) inequality. □

Our next result establishes 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 $\mathcal{Z}$. We leave this proof in the main text to illustrate our proof technique and comment on alternative assumptions under which this result holds.

PROPOSITION 1. *Suppose Assumptions 3 and 4 and either Assumption 1 or Assumption 2 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 $\sup_{z \in \mathcal{Z}} |\hat{g}^{ER}_n(z;x) - g(z;x)| \xrightarrow{p} 0$ for a.e. $x \in \mathcal{X}$. Note that

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

The second term on the r.h.s. of the above inequality vanishes in the limit in probability under Assumption 3. 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. First, suppose Assumption 1 holds. We then have for a.e. $x \in \mathcal{X}$:

$$\sup_{z \in \mathcal{Z}} \frac{1}{n}\sum_{i=1}^n \left| c\left(z, \hat{f}_n(x) + \hat{\varepsilon}^i_n\right) - c\left(z, f^*(x) + \varepsilon^i\right)\right| \leq \sup_{z \in \mathcal{Z}} \frac{1}{n}\sum_{i=1}^n L(z)\|\tilde{\varepsilon}^i_n(x)\| \leq \left(\sup_{z \in \mathcal{Z}} L(z)\right)\sqrt{\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}^i_n(x)\|^2},$$

where the last step follows from the AM-QM inequality. Now, the result follows from Lemma 1, Assumption 4, and the continuous mapping theorem. Next, suppose instead that Assumption 2 holds. Note that for a.e. $x \in \mathcal{X}$, we a.s. have for n large enough:

$$\begin{aligned}\sup_{z \in \mathcal{Z}} \frac{1}{n}\sum_{i=1}^n \left| c\left(z, \hat{f}_n(x) + \hat{\varepsilon}^i_n\right) - c\left(z, f^*(x) + \varepsilon^i\right)\right| &\leq \sup_{z \in \mathcal{Z}} \frac{1}{n}\sum_{i=1}^n L_{\delta(x)}(z, f^*(x) + \varepsilon^i)\|\tilde{\varepsilon}^i_n(x)\| \\ &\leq \sup_{z \in \mathcal{Z}} \sqrt{\frac{1}{n}\sum_{i=1}^n L^2_{\delta(x)}(z, f^*(x) + \varepsilon^i)}\,\sqrt{\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}^i_n(x)\|^2},\end{aligned}$$

where the last step follows from the Cauchy-Schwarz inequality. The result then follows from Assumptions 2 and 4, Lemma 1, the continuous mapping theorem, and $O_p(1)o_p(1) = o_p(1)$. □

Remark 1. Assumption (4b) can be weakened to $\frac{1}{n}\sum_{i=1}^{n}\|f^*(x^i)-\hat{f}_n(x^i)\| \xrightarrow{p} 0$ when Assumption 1 holds. We stick with the stronger Assumption (4b) throughout for uniformity.

Proposition 1 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 1. *Suppose Assumptions 3 and 4 and either Assumption 1 or Assumption 2 hold. Then, we have* $\hat{v}_n^{ER}(x) \xrightarrow{p} v^*(x)$, $\mathbb{D}\left(\hat{S}_n^{ER}(x), S^*(x)\right) \xrightarrow{p} 0$, *and* $\sup\limits_{z \in \hat{S}_n^{ER}(x)} g(z;x) \xrightarrow{p} v^*(x)$ *for a.e.* $x \in \mathcal{X}$.

The proof of Theorem 1 follows a similar outline as the proof of Theorem 5.3 of Shapiro et al. (2009), with a key difference being 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 1 hold even if the set $\mathcal{Z}$ is unbounded (see the discussion following Theorem 5.3 of Shapiro et al. 2009). While we consider convergence in probability instead of almost sure convergence (because the statistics literature is typically concerned with conditions under which Assumption 4 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, 3, and 4.

3.2. Rates of convergence

We next 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 5. *The function c in problem* (1) *and the data $\mathcal{D}_n$ satisfy:*

(5a) the function $c(\cdot, y)$ is continuous on the set $\mathcal{Z}$ for each $y \in \mathbb{R}^{d_y}$,

(5b) the following functional central limit theorem (CLT) for the full-information SAA objective:

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

where $V(\cdot;x)$ is a random element of $L^\infty(\mathcal{Z})$, the Banach space of essentially bounded functions on $\mathcal{Z}$ equipped with the supremum norm.

Appendix EC.2 verifies that Assumption (5a) holds for Example 1 under mild conditions. It can be weakened to assume that $c(\cdot, y)$ is continuous for each y in a neighborhood of $\mathcal{Y}$ if Assumption (2a) holds. Assumption (5b) holds, for instance, when the data $\mathcal{D}_n$ 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 $\mathbb{E}\left[\left(c(\tilde{z}, f^*(x)+\varepsilon)\right)^2\right] < +\infty$ (see page 164 of Shapiro et al. (2009) for details). Theorem 1 of Doukhan et al. (1995), Theorem 2.1 of Arcones and Yu (1994), Theorem 9 of Arcones (1994), and Corollary 2.3 of Andrews and Pollard (1994) provide conditions under which the functional CLT holds under mixing assumptions on the data $\mathcal{D}_n$. Theorems 1.5.4 and 1.5.6 of van der Vaart and Wellner (1996) present a general set of conditions under which the functional CLT holds.

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

Assumption 6. *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 criteria:*

(6a) Pointwise error rate: $\|f^*(x) - \hat{f}_n(x)\|^2 = O_p\left(n^{-\alpha}\right)$ for a.e. $x \in \mathcal{X}$,

(6b) Mean-squared estimation error rate: $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \|f^*(x^i) - \hat{f}_n(x^i)\|^2 = O_p(n^{-\alpha})$.

Note that the $O_p(\cdot)$ terms in Assumption 6 hide factors proportional to the dimension d_y of the random vector Y. Along with Lemma 1, Assumption 6 implies that $\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}^i_n(x)\|^2 = O_p(n^{-\alpha})$ for a.e. $x \in \mathcal{X}$. For our running example of OLS regression, Assumption 6 holds with $\alpha = 1$, independent of the dimension d_x of X, under mild assumptions on the data $\mathcal{D}_n$ and the distribution P_X of the covariates (see Chapter 5 of White 2014). 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 this assumption 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 (Raskutti et al. 2012) can hope to break the curse of dimensionality and achieve rates with $\alpha = 1$. Appendix EC.3 verifies that Assumption 6 holds for these prediction setups with the stated constants α.

Our main result of this section extends Theorem 5.7 of Shapiro et al. (2009) 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. In the next section we discuss how these dimensions affect the rate of convergence via a non-asymptotic/finite sample analysis. Note that the convergence rate analysis in Theorem 5.7 of Shapiro et al. (2009) for the full-information SAA problem (2) is sharper in the sense that it also characterizes the asymptotic distribution of the optimal objective value, see equations (5.25) and (5.26) therein.

Theorem 2. *Suppose Assumptions 5 and 6 hold, the objective function of the true problem* (1) *is continuous on $\mathcal{Z}$ for a.e. $x \in \mathcal{X}$, and either Assumption 1 or Assumption 2 holds. Then, we have $\hat{v}^{ER}_n(x) = v^*(x) + \tilde{o}_p(n^{-\frac{\alpha}{2}})$ for a.e. $x \in \mathcal{X}$.*

Proposition 1 implies that $|g(\hat{z}^{ER}_n(x);x) - \hat{v}^{ER}_n(x)| = \tilde{o}_p(n^{-\frac{\alpha}{2}})$ under the assumptions of Theorem 2, which then implies that the estimator $\hat{z}^{ER}_n(x)$ satisfies $|g(\hat{z}^{ER}_n(x);x) - v^*(x)| = \tilde{o}_p(n^{-\frac{\alpha}{2}})$. Note that Theorem 7.43 of Shapiro et al. (2009) lists conditions under which the function $g(\cdot;x)$ is continuous on $\mathcal{Z}$ for a.e. $x \in \mathcal{X}$.

3.3. Finite sample guarantees

Finally, we establish a lower bound on the probability that the solution of the ER-SAA problem (4) is nearly optimal to the true problem (1). These assumptions are motivated by the analysis in Section 2 of Homem-de-Mello (2008) and Section 7.2.9 of Shapiro et al. (2009).

Assumption 7. *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 $\mathbb{P}\Big\{\sup_{z\in\mathcal{Z}} |g^*_n(z;x) - g(z;x)| > \kappa\Big\} \leq K(\kappa,x)\exp\left(-n\beta(\kappa,x)\right)$, $\forall n \in \mathbb{N}$.*

Lemma 2.4 of Homem-de-Mello (2008) provides conditions under which Assumption 7 holds (also see Section 7.2.9 of Shapiro et al. 2009). In particular, Homem-de-Mello (2008) shows that Assumption 7 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 bound conditions hold. When the data $\mathcal{D}_n$ is i.i.d., Section 7.2.9 of Shapiro et al. (2009) presents conditions under which these pointwise exponential bound conditions are satisfied via Cramér's large deviation theorem. Bryc and Dembo (1996) presents mixing conditions on the observations $\mathcal{D}_n$ under which these assumptions are also satisfied (also see the references therein). The Gärtner-Ellis Theorem (see Section 2.3 of Dembo and Zeitouni 2010) provides an alternative avenue for verifying Assumption 7 for non-i.i.d. data $\mathcal{D}_n$ (Dai et al. 2000). If we also assume that the random variables $c(z,f^*(x)+\varepsilon) - \mathbb{E}\left[c(z,f^*(x)+\varepsilon)\right]$ are 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. (2009).

We make the following large deviation assumption on the regression procedure (3) that is similar in spirit to Assumption 7.

Assumption 8. *The regression procedure* (3) *possesses the following large deviation properties: for any constant $\kappa > 0$, there exist positive constants $K_f(\kappa,x)$, $\bar{K}_f(\kappa)$, $\beta_f(\kappa,x)$, and $\bar{\beta}_f(\kappa)$ satisfying:*

(8a) Pointwise error bound: $\mathbb{P}\big\{\|f^*(x) - \hat{f}_n(x)\|^2 > \kappa^2\big\} \leq K_f(\kappa,x)\exp\left(-n\beta_f(\kappa,x)\right)$ for a.e. $x \in \mathcal{X}$,

(8b) Mean-squared estimation error bound: $\mathbb{P}\Big\{\frac{1}{n}\sum_{i=1}^n \|f^*(x^i) - \hat{f}_n(x^i)\|^2 > \kappa^2\Big\} \leq \bar{K}_f(\kappa)\exp\left(-n\bar{\beta}_f(\kappa)\right)$.

In Appendix EC.3, we verify that this assumption holds for OLS regression and the Lasso with constants $\beta_f(\kappa,x)$ and $\bar{\beta}_f(\kappa)$ scaling as $O(\kappa^2)$ under suitable assumptions on the errors ε. Note that

Assumption 8 strengthens Assumption 6 by imposing restrictions on the tails of the estimators. Using the union bound, Assumption 8 implies along with Lemma 1 that for a.e. $x \in \mathcal{X}$:

$$\mathbb{P}\Big\{\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}_n^i(x)\|^2 > \kappa^2\Big\} \leq K_f\left(\frac{\kappa}{2}, x\right)\exp\left(-n\beta_f\left(\frac{\kappa}{2}, x\right)\right) + \bar{K}_f\left(\frac{\kappa}{2}\right)\exp\left(-n\bar{\beta}_f\left(\frac{\kappa}{2}\right)\right).$$

We also need the following strengthening of Assumption 2 for our finite sample results (see Appendix EC.2 for conditions under which it holds).

Assumption 2′. *There exists a function $\delta: \mathcal{X} \to \mathbb{R}_+$ such that for a.e. $x \in \mathcal{X}$, the regression step and the data $\mathcal{D}_n$ satisfy $\|\tilde{\varepsilon}_n^i(x)\| \leq \delta(x)$, $\forall n \in \mathbb{N}$ and $i \in [n]$. Furthermore, Assumption (2b) holds with Lipschitz constant satisfying $\sup_{z\in\mathcal{Z},\varepsilon\in\Xi} L_{\delta(x)}(z, f^*(x)+\varepsilon) < +\infty$, for a.e. $x \in \mathcal{X}$.*

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 7.

Lemma 2. *Suppose Assumption 8 and either Assumption 1 or Assumption 2′ hold. Then for any constant $\kappa > 0$ and a.e. $x \in \mathcal{X}$, there exist positive constants $\bar{K}(\kappa, x)$ and $\bar{\beta}(\kappa, x)$ satisfying*

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

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 (2008) and Section 5.3 of Shapiro et al. (2009), to the ER-SAA setting.

Theorem 3. *Suppose Assumptions 7 and 8 and either Assumption 1 or Assumption 2′ hold. Then, for a.e. $x \in \mathcal{X}$, given $\eta > 0$, there exist constants $Q(\eta, x) > 0$ and $\gamma(\eta, x) > 0$ such that*

$$\mathbb{P}\left\{\text{dist}(\hat{z}_n^{ER}(x), S^*(x)) \geq \eta\right\} \leq Q(\eta, x)\exp(-n\gamma(\eta, x)), \quad \forall n \in \mathbb{N}.$$

We now specialize the results in this section to the setting where the ER-SAA formulation is applied to two-stage stochastic LP with OLS, Lasso, or kNN regression as the prediction setup. We make stronger than necessary assumptions on the regression setups to improve readability.

Proposition 2. *Consider Example 1. Suppose the set $\mathcal{Z}$ is compact with diameter D and for a.e. $x \in \mathcal{X}$, the random variable $c(z, f^*(x)+\varepsilon) - \mathbb{E}\left[c(z, f^*(x)+\varepsilon)\right]$ is sub-Gaussian with variance proxy $\sigma_c^2(x)$ for each $z \in \mathcal{Z}$. Let ε^i, $i \in [n]$, be i.i.d. sub-Gaussian random vectors with variance proxy σ^2, $\delta \in (0,1)$ be the desired reliability level, and $S^\kappa(x) := \{z \in \mathcal{Z} : g(z;x) \leq v^*(x) + \kappa\}$ denote the set of κ-optimal solutions to the true problem* (1).

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, for sample size n satisfying*
$$n \geq \frac{O(1)\sigma_c^2(x)}{\kappa^2}\left[d_z \log\left(\frac{O(1)D}{\kappa}\right) + \log\left(\frac{O(1)}{\delta}\right)\right] + \frac{O(1)\sigma^2 d_y}{\kappa^2}\left[\log\left(\frac{O(1)}{\delta}\right) + d_x\right],$$
we have $\mathbb{P}\left\{\hat{S}_n^{ER}(x) \subseteq S^{\kappa}(x)\right\} \geq 1 - \delta$*.*
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, $\mathbb{E}\left[|X_j|^2\right] > 0$, $\forall j \in [d_x]$, and the matrix $\mathbb{E}\left[XX^T\right] - \tau \operatorname{diag}(\mathbb{E}\left[XX^T\right])$ is positive semidefinite for some constant $\tau \in (0,1]$. Then, for sample size n satisfying*
$$n \geq \frac{O(1)\sigma_c^2(x)}{\kappa^2}\left[d_z \log\left(\frac{O(1)D}{\kappa}\right) + \log\left(\frac{O(1)}{\delta}\right)\right] + \frac{O(1)\sigma^2 s d_y}{\kappa^2}\left[\log\left(\frac{O(1)}{\delta}\right) + \log(d_x)\right],$$
we have $\mathbb{P}\left\{\hat{S}_n^{ER}(x) \subseteq S^{\kappa}(x)\right\} \geq 1 - \delta$*.*
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, for sample size $n \geq O(1)\left(\frac{O(1)}{\kappa}\right)^{\frac{d_x}{1-\gamma}}$, $\frac{n^{\gamma}}{\log(n)} \geq \frac{O(1)d_x d_y \sigma^2}{\kappa^2}$, and*
$$n \geq \frac{O(1)\sigma_c^2(x)}{\kappa^2}\left[d_z \log\left(\frac{O(1)D}{\kappa}\right) + \log\left(\frac{O(1)}{\delta}\right)\right] + \left(\frac{O(1)\sigma^2 d_y}{\kappa^2}\right)^{\frac{1}{\gamma}}\left[d_x \log\left(\frac{O(1)}{d_x}\right) + \log\left(\frac{O(1)}{\delta}\right)\right]^{\frac{1}{\gamma}} + \left(\frac{O(1)d_y}{\kappa^2}\right)^{d_x}\left[\frac{d_x}{2}\log\left(\frac{O(1)d_x d_y}{\kappa^2}\right) + \log\left(\frac{O(1)}{\delta}\right)\right],$$
we have $\mathbb{P}\left\{\hat{S}_n^{ER}(x) \subseteq S^{\kappa}(x)\right\} \geq 1 - \delta$*.*

The proof of Proposition 2 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 full-information SAA problem (2). The sample size estimates given in Proposition 2 involve the sum of these two contributions, the first of which is the classical estimate
$$n \geq \frac{O(1)\sigma_c^2(x)}{\kappa^2}\left[d_z \log\left(\frac{O(1)D}{\kappa}\right) + \log\left(\frac{O(1)}{\delta}\right)\right]$$
for solutions of the full-information SAA problem (2) to possess a similar guarantee (cf. Section 5.3 of Shapiro et al. 2009). Our learning of the regression function f^* introduces additional terms in the estimate that depend on the dimensions d_y and d_x of the random vector Y and the covariates X.

Proposition 2 also illustrates the tradeoff between using parametric and nonparametric regression approaches within the ER-SAA framework. *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 X, 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 $\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}_n^i(x)\|^2 \xrightarrow{p} 0$ at a certain rate are adapted to ensure that the mean-squared deviation terms $\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}_n^{i,J}(x)\|^2$ and $\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}_n^{i,J+}(x)\|^2$ converge to zero in probability at a certain rate, where

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

Similar to the deviation terms $\tilde{\varepsilon}_n^i(x)$ introduced at the start of Section 3, the deviation terms $\tilde{\varepsilon}_n^{i,J}(x)$ and $\tilde{\varepsilon}_n^{i,J+}(x)$ can be interpreted as the sum of a *prediction error* at the point $x \in \mathcal{X}$ and the *leave-one-out estimation error* at the training point $x^i \in \mathcal{X}$. Because (cf. Lemma 1)

$$\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}_n^{i,J}(x)\|^2 \le 2\|f^*(x) - \hat{f}_n(x)\|^2 + \frac{2}{n}\sum_{i=1}^{n}\|f^*(x^i) - \hat{f}_{-i}(x^i)\|^2, \quad \text{and}$$
$$\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}_n^{i,J+}(x)\|^2 \le \frac{2}{n}\sum_{i=1}^{n}\|f^*(x) - \hat{f}_{-i}(x)\|^2 + \frac{2}{n}\sum_{i=1}^{n}\|f^*(x^i) - \hat{f}_{-i}(x^i)\|^2,$$

it suffices for the assumptions on the quantities $\|f^*(x) - \hat{f}_n(x)\|^2$ and $\frac{1}{n}\sum_{i=1}^{n}\|f^*(x^i) - \hat{f}_n(x^i)\|^2$ in the previous sections to be replaced with assumptions on the terms $\|f^*(x) - \hat{f}_n(x)\|^2$ and $\frac{1}{n}\sum_{i=1}^{n}\|f^*(x^i) - \hat{f}_{-i}(x^i)\|^2$ for the J-SAA approach, and with assumptions on the terms $\frac{1}{n}\sum_{i=1}^{n}\|f^*(x) - \hat{f}_{-i}(x)\|^2$ and $\frac{1}{n}\sum_{i=1}^{n}\|f^*(x^i) - \hat{f}_{-i}(x^i)\|^2$ for the J+-SAA approach. 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 3.1, 3.2, and 3.3, we present these details in Appendix EC.1.

4. Computational experiments

We consider instances of the following resource allocation model adapted from Luedtke (2014):

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

where the second-stage function is defined as

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

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 (2014), 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. We assume that some of the variability in the demands can be explained with knowledge of covariates X_l, $l \in \mathcal{L}$, where $|\mathcal{L}| = d_x$. We assume that the demands Y are related to the covariates through

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

where $p \in \{0.5, 1, 2\}$ is a fixed parameter that determines the model class, $\varepsilon_j \sim \mathcal{N}\left(0, \sigma_j^2\right)$ is an additive error, φ^*, ζ^* and σ_j are 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 $j \in \mathcal{J}$). Throughout, we assume that $|\mathcal{L}^*| = 3$, i.e., the demands truly depend only on three covariates. We simulate i.i.d. data $\mathcal{D}_n$ 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 Appendix EC.4 in the electronic companion for details).

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 compare our data-driven SAA estimators with the kNN-based reweighted SAA (kNN-SAA) approach of Bertsimas and Kallus (2019) on a few test instances by varying the dimensions of the covariates d_x, the sample size n, the degree p, and the standard deviation σ of the errors ε. 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 (2019) and Bertsimas and McCord (2019) show that this approach is one of the better performing reweighted SAA approaches. The 'parameter k' in kNN-SAA is chosen by determining the value of k in $[\lfloor n^{0.1} \rfloor, \lceil n^{0.9} \rceil]$ that yields the smallest test error using 5-fold cross-validation when kNN is used to predict Y from X.

Solutions obtained 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. (1999) (see Appendix EC.4 for details). 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 the 2nd and 98th percentiles of the 99% UCBs 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 (Bezanson et al. 2017), use Gurobi 8.1.0 to solve LPs through the JuMP 0.18.5 interface (Dunning et al. 2017), and use `glmnet` 0.3.0 (Friedman et al. 2010) 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 1 compares the performance of the kNN-SAA and ER-SAA+OLS approaches by varying the model degree p, the covariate dimension among $d_x \in \{3, 10, 100\}$, and the sample size among $n \in \{1.5(d_x+1), 2(d_x+1), 5(d_x+1), 20(d_x+1), 100(d_x+1)\}$. Note that OLS regression estimates $d_x + 1$ parameters for each $j \in \mathcal{J}$. When the prediction model is correctly specified (i.e., $p = 1$), the ER-SAA+OLS approach unsurprisingly dominates the kNN-SAA approach. When $p \neq 1$, as anticipated, the ER-SAA+OLS approach does not yield a consistent estimator, whereas the kNN-SAA approach yields consistent estimators, albeit with a slow rate of convergence (cf. Proposition 2). However, the ER-SAA approach consistently outperforms the kNN-SAA approach when $p = 0.5$ even for the largest sample size of $n = 100(d_x + 1)$. When the degree $p = 2$, the kNN-SAA approach fares better than the ER-SAA approach only for a sample size of $n \geq 80$ when the covariate dimension is small ($d_x = 3$), and loses this advantage in the larger covariate dimensions. 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 10%, 5%, and 24% for the $p = 1$, $p = 0.5$, and $p = 2$ instances, respectively, for large values of n. This indicates that using covariate information can be advantageous in these instances.

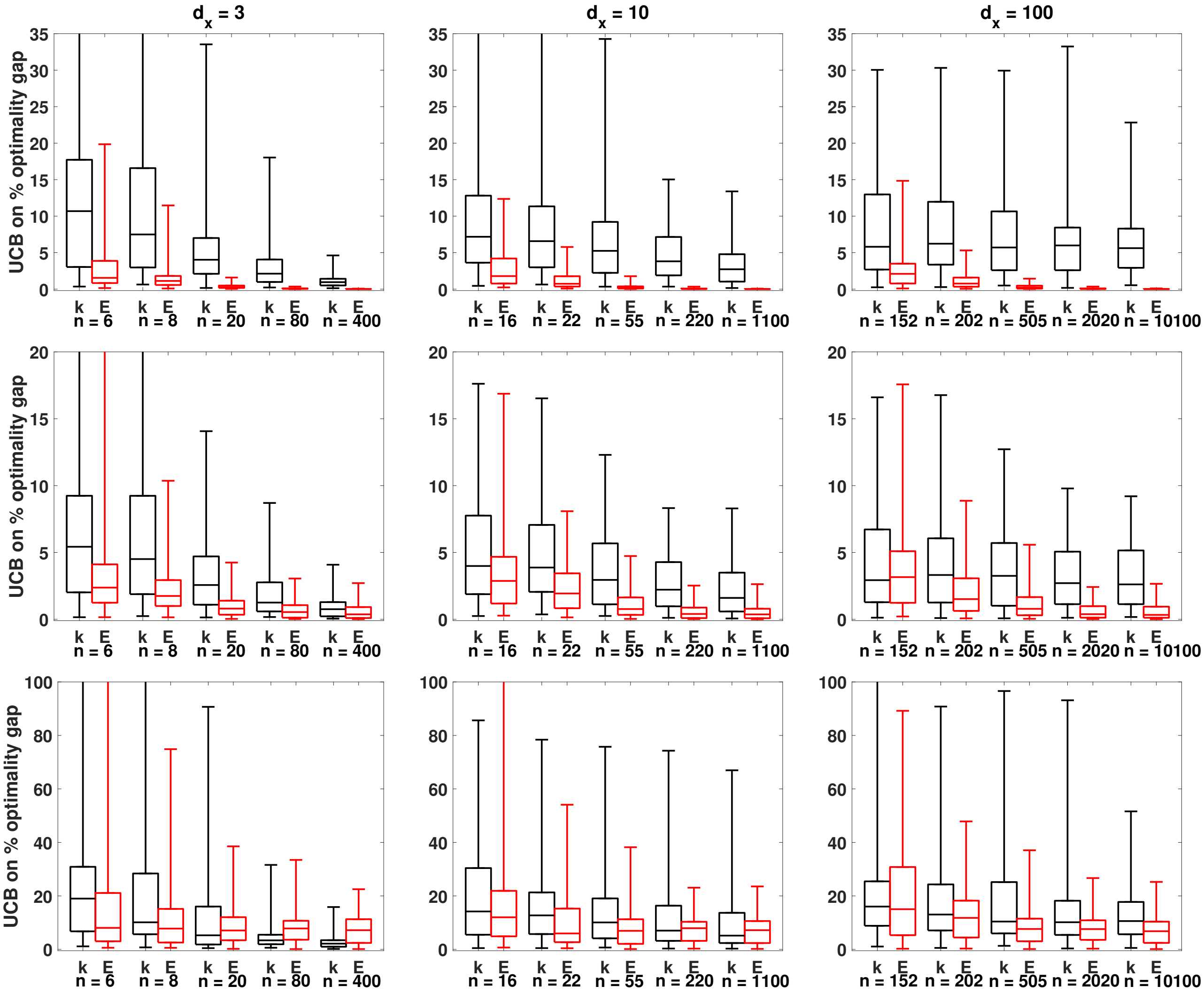

Figure 1 Comparison of the kNN-SAA (k) and ER-SAA+OLS (E) approaches. Top row: $p = 1$. Middle row: $p = 0.5$. Bottom row: $p = 2$. Left column: $d_x = 3$. Middle column: $d_x = 10$. Right column: $d_x = 100$.

Impact of the Jackknife-based formulations. Figure 2 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 $n \in \{1.2(d_x + 1), 1.3(d_x + 1), 1.5(d_x + 1), 2(d_x + 1), 3(d_x + 1)\}$. 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 98th 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.

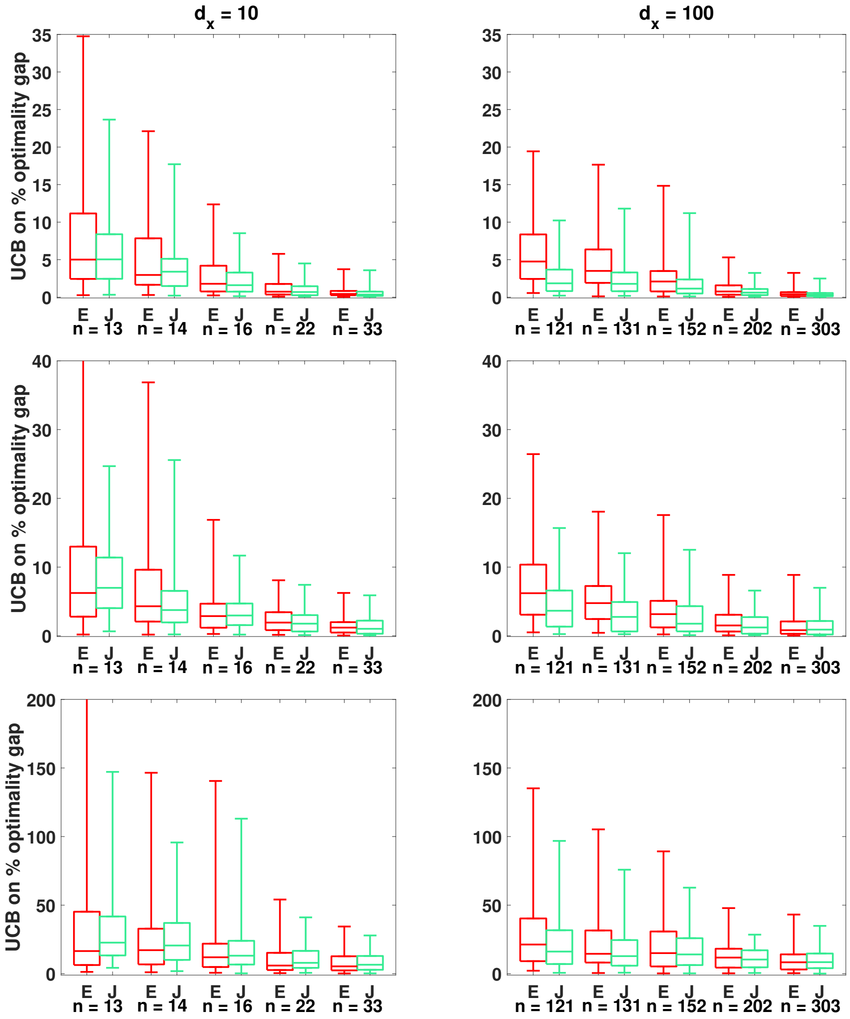

Figure 2 Comparison of the ER-SAA (E) and J-SAA (J) approaches with OLS regression. Top row: $p = 1$. Middle row: $p = 0.5$. Bottom row: $p = 2$. Left column: $d_x = 10$. Right column: $d_x = 100$.

Impact of the prediction setup. Figure 3 compares the performance of the ER-SAA+OLS and ER-SAA+Lasso approaches by varying the model degree p, the covariate dimension among $d_x \in \{10, 100\}$, and the sample size among $n \in \{1.2(d_x + 1), 1.3(d_x + 1), 1.5(d_x + 1), 2(d_x + 1), 3(d_x + 1)\}$. 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 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 also shrinks as the sample size increases.

Impact of the error variance. Figure 4 compares the performance of the kNN-SAA and ER-SAA+OLS approaches by varying the standard deviation of the errors ε among $\sigma \in \{5, 10, 20\}$ (note that the case studies thus far used $\sigma = 5$) and the sample size among $n \in \{1.5(d_x + 1), 2(d_x +$

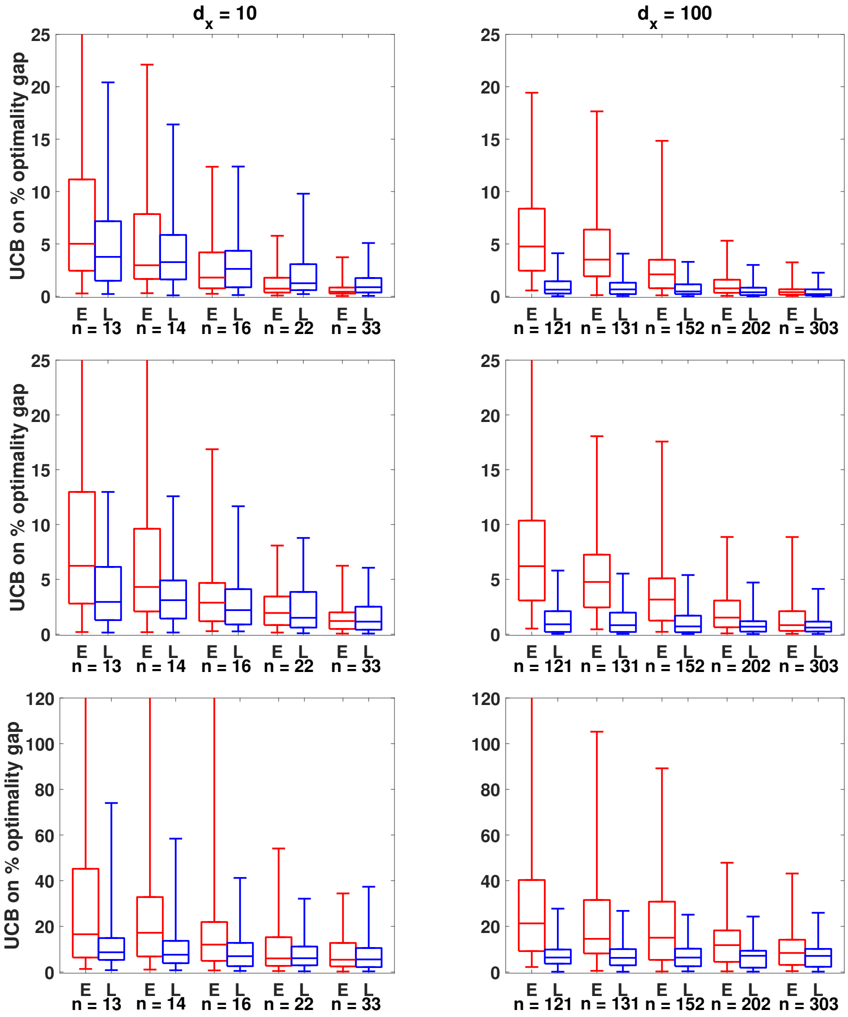

Figure 3 Comparison of the ER-SAA+OLS (E) and ER-SAA+Lasso (L) approaches. Top row: $p = 1$. Middle row: $p = 0.5$. Bottom row: $p = 2$. Left column: $d_x = 10$. Right column: $d_x = 100$.

$1), 5(d_x + 1), 20(d_x + 1), 100(d_x + 1)\}$ for $d_x = 10$ and $p = 1$. 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 kNN-SAA formulation appears to be unaffected (and even slightly improve!) with increasing error variance. A possible reason for this behavior is that the dominant term in the sample size estimate provided by Proposition 2 for kNN regression does not involve σ.

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

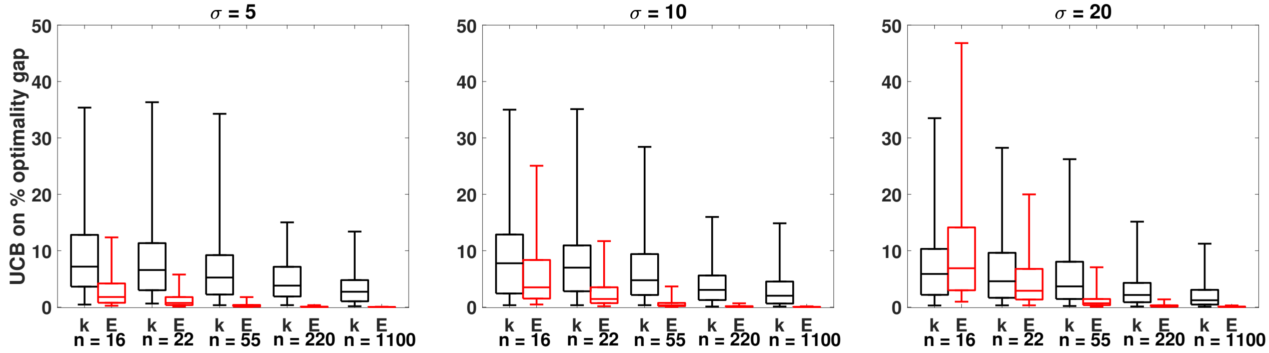

Figure 4 Effect of increasing σ on the ER-SAA+OLS (E) and kNN-SAA (k) approaches when $d_x = 10$ and $p = 1$. Left plot: $\sigma = 5$. Middle plot: $\sigma = 10$. Right plot: $\sigma = 20$.

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 that the assumptions on the prediction setup hold for other frameworks of interest is an important task to be undertaken by the DM. Ongoing work includes analysis of a distributionally robust optimization extension of the ER-SAA problem (4) that possesses practical finite sample guarantees (cf. Bertsimas et al. 2019, Dou and Anitescu 2019), and analysis of the extension of the ER-SAA approach to the multi-stage stochastic programming setting (cf. Ban et al. 2018, Bertsimas and McCord 2019). Designing asymptotically optimal estimators for problems with stochastic constraints (Homem-de-Mello and Bayraksan 2015) and for the case when decisions affect the realizations of the random variables (Bertsimas and Kallus 2019) are interesting avenues for future work.

Appendix A: Omitted proofs

A.1. Proof of Theorem 1

Before we prove Theorem 1, we present the following lemma that is needed in its proof.

Lemma 3. *Let $W \subset \mathbb{R}^{d_w}$ be a nonempty and compact set and $h : W \to \mathbb{R}$ be a lower semicontinuous function. Consider minimizing $h(w)$ over W. Let W^* denote the set of optimal solutions to this problem, i.e., $W^* := \arg\min_{w \in W} h(w)$. Suppose there exists $\delta > 0$ and $\bar{w} \in W$ such that $\operatorname{dist}(\bar{w}, W^*) \geq \delta$. Then, there exists $\kappa > 0$ such that $h(\bar{w}) \geq \min_{w \in W} h(w) + \kappa$.*

Proof of Lemma 3. Let $W_\delta := \{w \in W : \operatorname{dist}(w, W^*) \geq \delta\}$, and note that $\bar{w} \in W_\delta$. Since h is lower semicontinuous and W_δ is nonempty and compact, $\inf\limits_{w \in W_\delta} h(w)$ is attained. Furthermore, we readily have $\min\limits_{w \in W_\delta} h(w) > \min\limits_{w \in W} h(w)$. Setting $\kappa = \min\limits_{w \in W_\delta} h(w) - \min\limits_{w \in W} h(w)$ yields the desired result. □

Proof of Theorem 1. Let $z^*(x) \in S^*(x)$ and $\hat{z}^{ER}_n(x) \in \hat{S}^{ER}_n(x)$. Consider any constant $\delta > 0$. From Proposition 1, we have for a.e. $x \in \mathcal{X}$:

$$\begin{aligned}
&\mathbb{P}\left\{\left|\hat{g}^{ER}_n(z^*(x); x) - v^*(x)\right| > \delta\right\} \to 0 \implies \mathbb{P}\left\{\hat{v}^{ER}_n(x) > v^*(x) + \delta\right\} \to 0 \quad \text{and} \\
&\mathbb{P}\left\{\left|\hat{v}^{ER}_n(x) - g(\hat{z}^{ER}_n(x); x)\right| > \delta\right\} \to 0 \implies \mathbb{P}\left\{v^*(x) > \hat{v}^{ER}_n(x) + \delta\right\} \to 0.
\end{aligned}$$

These inequalities yield $\mathbb{P}\{|\hat{v}^{ER}_n(x) - v^*(x)| > \delta\} \to 0$ for a.e. $x \in \mathcal{X}$, which implies $\hat{v}^{ER}_n(x) \xrightarrow{p} v^*(x)$ for a.e. $x \in \mathcal{X}$.

Suppose for contradiction that $\mathbb{D}\left(\hat{S}^{ER}_n(x), S^*(x)\right) \overset{p}{\nrightarrow} 0$, $\forall x \in \bar{\mathcal{X}}$, 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 $\mathbb{P}\left\{\mathbb{D}\left(\hat{S}^{ER}_{n_q}(\bar{x}), S^*(\bar{x})\right) \geq \delta\right\} \geq \beta$, $\forall q \in \mathbb{N}$. Lemma 3 then implies that for a.e. $\bar{x} \in \bar{\mathcal{X}}$, there exists $\kappa(\bar{x}) > 0$ such that

$$\mathbb{P}\Big\{\sup_{z \in \hat{S}^{ER}_{n_q}(\bar{x})} g(z; \bar{x}) > v^*(\bar{x}) + \kappa(\bar{x})\Big\} \geq \beta, \quad \forall q \in \mathbb{N}. \tag{8}$$

From Proposition 1, we have for a.e. $\bar{x} \in \bar{\mathcal{X}}$:

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

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

$$\mathbb{P}\Big\{\sup_{z \in \hat{S}^{ER}_n(\bar{x})} g(z; \bar{x}) \leq v^*(\bar{x}) + \kappa(\bar{x})\Big\} \to 1,$$

which contradicts the inequality (8). The above arguments also readily imply that the ER-SAA estimators are asymptotically optimal, i.e., $\sup\limits_{z \in \hat{S}^{ER}_n(x)} g(z; x) \xrightarrow{p} v^*(x)$ for a.e. $x \in \mathcal{X}$. □

A.2. Proof of Theorem 2

Proof of Theorem 2. Let $\{\beta_n\} \downarrow 0$ be *any* positive sequence such that $\{\beta_n \sqrt{n}\} \to \infty$ and $\beta_n \sqrt{n} = o\left(n^{\frac{\alpha}{2}}\right)$. We consider the sequence $\{\beta_n\}$ to force the following two sequences of random variables to converge to zero in probability. Assumption (5a), the continuity of $g(\cdot; x)$, and the compactness of $\mathcal{Z}$ imply that the functions $\hat{g}^{ER}_n(\cdot; x)$, $g^*_n(\cdot; x)$, and $g(\cdot; x)$ are all elements of $L^\infty(\mathcal{Z})$

(in fact, they are all elements of $C(\mathcal{Z})$, the space of continuous functions on $\mathcal{Z}$ equipped with the sup-norm). Assumption (5b) along with Slutzky's lemma then yields

$$\beta_n\sqrt{n}\,(g_n^*(\cdot;x) - g(\cdot;x)) \xrightarrow{d} 0, \quad \text{for a.e. } x \in \mathcal{X}.$$

Recall that the above notation means that the sequence of functions converges in distribution to the zero element of $L^\infty(\mathcal{Z})$ with respect to the sup-norm. From the proof of Proposition 1 and using Assumption 6, we have that $\hat{g}_n^{ER}(\cdot;x) - g_n^*(\cdot;x) = O_p\left(n^{-\frac{\alpha}{2}}\right)$ when viewed as a random element of $L^\infty(\mathcal{Z})$. Because by construction $n^{-\alpha/2}\beta_n\sqrt{n} = o_p(1)$ (for deterministic sequences o_p reduces to o), $n^{\frac{\alpha}{2}}(\hat{g}_n^{ER}(\cdot;x) - g_n^*(\cdot;x)) = O_p(1)$, and by the fact that $o_p(1)O_p(1) = o_p(1)$, we have

$$\beta_n\sqrt{n}\left(\hat{g}_n^{ER}(\cdot;x) - g_n^*(\cdot;x)\right) = o_p(1), \quad \text{for a.e. } x \in \mathcal{X}.$$

Using $o_p(1) + o_p(1) = o_p(1)$ to combine the above two results, we have

$$\beta_n\sqrt{n}\left(\hat{g}_n^{ER}(\cdot;x) - g(\cdot;x)\right) = o_p(1), \quad \text{for a.e. } x \in \mathcal{X}.$$

This implies that the sequence $\beta_n\sqrt{n}\,(\hat{g}_n^{ER}(\cdot;x) - g(\cdot;x))$ of random functions in $C(\mathcal{Z})$ converges in distribution to the zero element of $C(\mathcal{Z})$. Therefore, by mirroring the arguments in the proof of Theorem 5.7 of Shapiro et al. (2009) (that uses the Delta theorem, see Theorem 7.59 of Shapiro et al. 2009), we then obtain

$$\hat{v}_n^{ER}(x) = v^*(x) + o_p(\beta_n^{-1}n^{-0.5}), \quad \text{for a.e. } x \in \mathcal{X},$$

which is almost what we set out to establish since $\beta_n\sqrt{n} = o\left(n^{\frac{\alpha}{2}}\right)$.

We now show that the stated results hold because $\{\beta_n\}$ was an *arbitrary* sequence converging to zero satisfying $\{\beta_n\sqrt{n}\} \to \infty$ and $\beta_n\sqrt{n} = o\left(n^{\frac{\alpha}{2}}\right)$. Let $\beta_n = n^{\frac{\alpha-1}{2}}(\max\{1, \log\log n\})^{-1}$, and note that such a sequence $\{\beta_n\}$ satisfies the aforementioned conditions (note that the choice of $\log\log n$ is arbitrary). This yields $\hat{v}_n^{ER}(x) = v^*(x) + o_p\left(n^{-\frac{\alpha}{2}}\log\log n\right) = v^*(x) + \tilde{o}_p\left(n^{-\frac{\alpha}{2}}\right)$ for a.e. $x \in \mathcal{X}$. □

A.3. Proofs of Lemma 2, Theorem 3, and Proposition 2

Proof of Lemma 2. Suppose Assumption 1 holds. From the proof of Proposition 1, we have:

$$\begin{aligned}\mathbb{P}\left\{\sup_{z\in\mathcal{Z}}\left|\hat{g}_n^{ER}(z;x) - g_n^*(z;x)\right| > \kappa\right\} &\le \mathbb{P}\left\{\left(\sup_{z\in\mathcal{Z}} L(z)\right)\sqrt{\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|^2} > \kappa\right\}\\ &\le \mathbb{P}\left\{\sqrt{\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|^2} > \frac{\kappa}{B_L(x)}\right\},\end{aligned}$$

where $B_L(x) := \sup\limits_{z \in \mathcal{Z}} L(z)$. Next, suppose instead that Assumption 2′ holds. From the proof of Proposition 1, we have:

$$\begin{aligned}\mathbb{P}\left\{\sup_{z\in\mathcal{Z}}\left|\hat{g}^{ER}_n(z;x) - g^*_n(z;x)\right| > \kappa\right\} &\leq \mathbb{P}\left\{\sup_{z\in\mathcal{Z}} \sqrt{\frac{1}{n}\sum_{i=1}^{n} L^2_{\delta(x)}(z, f^*(x)+\varepsilon^i)}\ \sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}^i_n(x)\|^2} > \kappa\right\} \\ &\leq \mathbb{P}\left\{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}^i_n(x)\|^2} > \frac{\kappa}{B_L(x)}\right\},\end{aligned}$$

where $B_L(x) := \sup\limits_{(z,\varepsilon)\in\mathcal{Z}\times\Xi} L_{\delta(x)}(z, f^*(x)+\varepsilon)$ is finite by virtue of Assumption 2′. In both cases, the inequality immediately following Assumption 8 yields the desired conclusion via

$$\begin{aligned}\mathbb{P}\left\{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}^i_n(x)\|^2} > \frac{\kappa}{B_L(x)}\right\} \leq K_f\left(\frac{\kappa}{2B_L(x)}, x\right)\exp\left(-n\beta_f\left(\frac{\kappa}{2B_L(x)}, x\right)\right) + \\ \bar{K}_f\left(\frac{\kappa}{2B_L(x)}\right)\exp\left(-n\bar{\beta}_f\left(\frac{\kappa}{2B_L(x)}\right)\right). \quad \square\end{aligned}$$

Proof of Theorem 3. Note that for any $\kappa > 0$:

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

Bounding the two terms on the r.h.s. of the above inequality using Assumption 7 and Lemma 2 yields for a.e. $x \in \mathcal{X}$:

$$\mathbb{P}\Big\{\sup_{z\in\mathcal{Z}}\left|\hat{g}^{ER}_n(z;x) - g(z;x)\right| > \kappa\Big\} \leq \tilde{K}(\kappa, x)\exp\left(-n\tilde{\beta}(\kappa, x)\right), \tag{9}$$

where $\tilde{K}(\kappa, x) := 2\max\left\{K(0.5\kappa, x), \bar{K}(0.5\kappa, x)\right\}$ and $\tilde{\beta}(\kappa, x) := \min\left\{\beta(0.5\kappa, x), \bar{\beta}(0.5\kappa, x)\right\}$. From inequality (9), we have for a.e. $x \in \mathcal{X}$ and any $\kappa(x) > 0$, $z^*(x) \in S^*(x)$:

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

Since $\hat{v}^{ER}_n(x) \leq \hat{g}^{ER}_n(z^*(x);x)$ by definition, this implies for a.e. $x \in \mathcal{X}$ and any $\kappa(x) > 0$:

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

Suppose $\text{dist}(\hat{z}^{ER}_n(x), S^*(x)) \geq \eta$ for some $x \in \mathcal{X}$ and some sample path. Since $g(\cdot;x)$ is lower semicontinuous on the compact set $\mathcal{Z}$ for a.e. $x \in \mathcal{X}$, Lemma 3 implies that there exists $\kappa(\eta, x) > 0$ such that $g(\hat{z}^{ER}_n(x);x) > v^*(x) + \kappa(\eta, x)$ 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. $x \in \mathcal{X}$:

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

Proof of Proposition 2. We show that for a.e. $x \in \mathcal{X}$, there exist positive constants $\tilde{K}(\kappa, x)$ and $\tilde{\beta}(\kappa, x)$ s.t.

$$\mathbb{P}\Big\{\sup_{z \in \mathcal{Z}} \big|\hat{g}_n^{ER}(z;x) - g(z;x)\big| > \kappa\Big\} \leq \tilde{K}(\kappa, x) \exp\Big(-n\tilde{\beta}(\kappa, x)\Big). \tag{10}$$

Inequality (10) then implies

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

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

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

which in turn implies that

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

We now state results that can be used to bound the constants $\tilde{K}(\kappa, x)$ and $\tilde{\beta}(\kappa, x)$ in inequality (10); we ignore their dependence on x to keep the exposition simple. Theorems 7.66 and 7.67 of Shapiro et al. (2009) imply for our setting of two-stage stochastic LP the bound

$$\mathbb{P}\left\{\sup_{z \in \mathcal{Z}} |g_n^*(z;x) - g(z;x)| > \kappa\right\} \leq O(1) \left(\frac{O(1)D}{\kappa}\right)^{d_z} \exp\left(-\frac{n\kappa^2}{O(1)\sigma_c^2(x)}\right) \tag{11}$$

for a.e. $x \in \mathcal{X}$. The following large deviation inequalities for our three different regression setups (see Appendix EC.3) can be used to specialize the bound afforded by Lemma 2:

1. OLS regression: $\mathbb{P}\Big\{\frac{1}{n}\sum_{i=1}^{n} \|\tilde{\varepsilon}_n^i(x)\|^2 > \kappa^2\Big\} \leq \exp(d_x) \exp\left(-\frac{n\kappa^2}{O(1)\sigma^2 d_y}\right)$, which follows from Remark 12 of Hsu et al. (2012), Theorem 2.2 and Remark 2.3 of Rigollet and Hütter (2017).
2. Lasso regression: $\mathbb{P}\Big\{\frac{1}{n}\sum_{i=1}^{n} \|\tilde{\varepsilon}_n^i(x)\|^2 > \kappa^2\Big\} \leq 2d_x \exp\left(-\frac{n\kappa^2}{O(1)\sigma^2 s d_y}\right)$, which follows from Theorem 2.1 and Corollary 1 of Bunea et al. (2007).
3. kNN regression: Whenever $n \geq O(1)\left(\frac{O(1)}{\kappa}\right)^{\frac{d_x}{1-\gamma}}$ and $\frac{n^{\gamma}}{\log(n)} \geq \frac{O(1) d_x d_y \sigma^2}{\kappa^2}$, we have
$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} \|\tilde{\varepsilon}_n^i(x)\|^2 > \kappa^2 d_y\right\} \leq \left(\frac{O(1)\sqrt{d_x}}{\kappa}\right)^{d_x} \exp\left(-O(1)n(O(1)\kappa)^{2d_x}\right) + O(1)n^{2d_x}\left(\frac{O(1)}{d_x}\right)^{d_x} \exp\left(-\frac{n^{\gamma}\kappa^2}{O(1)\sigma^2}\right)$$
from Lemma 10 of Bertsimas and McCord (2019).

Suppose the regression step (3) is Lasso regression. We have from Lemma 2 that

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

Along with the uniform exponential bound inequality (11), this yields for a.e. $x\in\mathcal{X}$:

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

Requiring each term in the r.h.s. of the above inequality to be $\leq\frac{\delta}{2}$ and using the union bound yields the stated conservative sample size results. Sample complexities for OLS and kNN regression can be similarly derived. □

Acknowledgments

This research is supported by the Dept. 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 Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institutes for Discovery, and the National Science Foundation. R.K. thanks Rui Chen and Prof. Garvesh Raskutti for helpful discussions.

References

Andrews DW, Pollard D (1994) An introduction to functional central limit theorems for dependent stochastic processes. *International Statistical Review* 62(1):119–132.

Arcones MA (1994) Limit theorems for nonlinear functionals of a stationary Gaussian sequence of vectors. *The Annals of Probability* 22(4):2242–2274.

Arcones MA, Yu B (1994) Central limit theorems for empirical and U-processes of stationary mixing sequences. *Journal of Theoretical Probability* 7(1):47–71.

Ban GY, Gallien J, Mersereau AJ (2018) Dynamic procurement of new products with covariate information: The residual tree method. Articles In Advance. *Manufacturing & Service Operations Management* 1–18.

Ban GY, Rudin C (2018) The big data newsvendor: Practical insights from machine learning. *Operations Research* 67(1):90–108.

Barber RF, Candes EJ, Ramdas A, Tibshirani RJ (2019) Predictive inference with the jackknife+. *arXiv preprint arXiv:1905.02928v3* 1–44.

Bertsimas D, Kallus N (2019) From predictive to prescriptive analytics. *Management Science* 1–20.

Bertsimas D, McCord C (2019) From predictions to prescriptions in multistage optimization problems. *arXiv preprint arXiv:1904.11637* 1–38.

Bertsimas D, McCord C, Sturt B (2019) Dynamic optimization with side information. *arXiv preprint arXiv:1907.07307* 1–37.

Bezanson J, Edelman A, Karpinski S, Shah V (2017) Julia: a fresh approach to numerical computing. *SIAM Review* 59(1):65–98.

Birge JR, Louveaux F (2011) *Introduction to stochastic programming* (Springer Science & Business Media).

Bryc W, Dembo A (1996) Large deviations and strong mixing. *Annales de l'IHP Probabilités et statistiques*, volume 32, 549–569.

Bunea F, Tsybakov A, Wegkamp M, et al. (2007) Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics* 1:169–194.

Dai L, Chen C, Birge J (2000) Convergence properties of two-stage stochastic programming. *Journal of Optimization Theory and Applications* 106(3):489–509.

Davarnia D, Kocuk B, Cornuéjols G (2018) Bayesian solution estimators in stochastic optimization. Optimization Online. URL: `http://www.optimization-online.org/DB_HTML/2017/11/6318.html`.

Dembo A, Zeitouni O (2010) *Large deviations techniques and applications*, volume 38 of *Stochastic Modelling and Applied Probability* (Springer), 2nd edition, URL `http://dx.doi.org/10.1007/978-3-642-03311-7`.

Diao S, Sen S (2020) Distribution-free algorithms for learning enabled predictive stochastic programming. Optimization Online. URL: `http://www.optimization-online.org/DB_HTML/2020/03/7661.html`.

Donti P, Amos B, Kolter JZ (2017) Task-based end-to-end model learning in stochastic optimization. *Advances in Neural Information Processing Systems*, 5484–5494.

Dou X, Anitescu M (2019) Distributionally robust optimization with correlated data from vector autoregressive processes. *Operations Research Letters* 47(4):294–299.

Doukhan P, Massart P, Rio E (1995) Invariance principles for absolutely regular empirical processes. *Annales de l'IHP Probabilités et statistiques*, volume 31, 393–427.

Dunning I, Huchette J, Lubin M (2017) JuMP: A modeling language for mathematical optimization. *SIAM Review* 59(2):295–320.

Elmachtoub AN, Grigas P (2017) Smart "predict, then optimize". *arXiv preprint arXiv:1710.08005* 1–38.

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1):1–22, URL `http://www.jstatsoft.org/v33/i01/`.

Györfi L, Kohler M, Krzyzak A, Walk H (2006) *A distribution-free theory of nonparametric regression* (Springer Science & Business Media).

Homem-de-Mello T (2008) On rates of convergence for stochastic optimization problems under non–independent and identically distributed sampling. *SIAM Journal on Optimization* 19(2):524–551.

Homem-de Mello T, Bayraksan G (2014) Monte Carlo sampling-based methods for stochastic optimization. *Surveys in Operations Research and Management Science* 19(1):56–85.

Homem-de-Mello T, Bayraksan G (2015) Stochastic constraints and variance reduction techniques. Fu MC, ed., *Handbook of Simulation Optimization*, 245–276 (Springer New York).

Hsu D, Kakade SM, Zhang T (2012) Random design analysis of ridge regression. *Conference on learning theory*, 9–1.

Kim K, Mehrotra S (2015) A two-stage stochastic integer programming approach to integrated staffing and scheduling with application to nurse management. *Operations Research* 63(6):1431–1451.

Lahiri SN (2013) *Resampling methods for dependent data* (Springer Science & Business Media).

Luedtke J (2014) A branch-and-cut decomposition algorithm for solving chance-constrained mathematical programs with finite support. *Mathematical Programming* 146(1-2):219–244.

Mak WK, Morton DP, Wood RK (1999) Monte Carlo bounding techniques for determining solution quality in stochastic programs. *Operations Research Letters* 24(1-2):47–56.

Raskutti G, Wainwright MJ, Yu B (2012) Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research* 13(Feb):389–427.

Rigollet P, Hütter JC (2017) High dimensional statistics. Lecture notes for MIT's 18.657 course, URL `http://www-math.mit.edu/~rigollet/PDFs/RigNotes17.pdf`.

Royset JO, Wets RJ (2014) From data to assessments and decisions: Epi-spline technology. *Bridging data and decisions*, 27–53 (INFORMS).

Schütz P, Tomasgard A, Ahmed S (2009) Supply chain design under uncertainty using sample average approximation and dual decomposition. *European Journal of Operational Research* 199(2):409–419.

Sen S, Deng Y (2018) Learning enabled optimization: Towards a fusion of statistical learning and stochastic programming. `http://www.optimization-online.org/DB_FILE/2017/03/5904.pdf`.

Shapiro A, Dentcheva D, Ruszczyński A (2009) *Lectures on stochastic programming: modeling and theory* (SIAM).

van de Geer SA (2000) *Empirical Processes in M-estimation*, volume 6 (Cambridge university press).

van der Vaart AW (1998) *Asymptotic statistics*, volume 3 (Cambridge university press).

van der Vaart AW, Wellner JA (1996) *Weak convergence and empirical processes: with applications to statistics* (Springer).

Wainwright MJ (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 (Cambridge University Press).

Walk H (2010) Strong laws of large numbers and nonparametric estimation. *Recent Developments in Applied Probability and Statistics*, 183–214 (Springer).

White H (2014) *Asymptotic theory for econometricians* (Academic press).

Electronic companion

We begin with the analysis of the Jackknife-based variants in Section EC.1. Then, in Section EC.2, we present a class of two-stage stochastic programs that satisfy our assumptions. Section EC.3 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 EC.4 by providing omitted details for the computational experiments.

Appendix EC.1: 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 (6) and (7), 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 $\{\tilde{\varepsilon}^i_n(x)\}$ considered in Section 3, we consider the following deviation sequences $\{\tilde{\varepsilon}^{i,J}_n(x)\}$ and $\{\tilde{\varepsilon}^{i,J+}_n(x)\}$:

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

We let $\hat{z}^J_n(x)$ and $\hat{z}^{J+}_n(x)$ denote an optimal solution to problem (6) and (7), respectively, and $\hat{S}^J_n(x)$ and $\hat{S}^{J+}_n(x)$ denote the corresponding sets of optimal solutions. We assume throughout that the sets $\hat{S}^J_n(x)$ and $\hat{S}^{J+}_n(x)$ are nonempty for a.e. $x \in \mathcal{X}$.

EC.1.1. Consistency and asymptotic optimality

We present conditions under which the optimal objective value of and optimal solutions to the J-SAA and J+-SAA problems (6) and (7) asymptotically converge to those of the true problem (1). We begin by adapting Assumption 2 in Section 3.1.

ASSUMPTION 2J. *Assumption 2 holds with the sequence $\{\tilde{\varepsilon}^i_n(x)\}$ substituted by $\{\tilde{\varepsilon}^{i,J}_n(x)\}$.*

ASSUMPTION 2J+. *Assumption 2 holds with the sequence $\{\tilde{\varepsilon}^i_n(x)\}$ substituted by $\{\tilde{\varepsilon}^{i,J+}_n(x)\}$.*

Part one of Assumptions 2J and 2J+ are satisfied for our running example of OLS regression if the support $\mathcal{X}$ of the covariates X is compact and all of the parameter estimates $\hat{\theta}_n$ and $\hat{\theta}_{-i}$, $i \in [n]$, a.s. lie within the same compact set for n large enough, where $\hat{\theta}_{-i}$ denotes the estimate of θ^* obtained using the training data $\mathcal{D}_n \backslash \{(y^i, x^i)\}$. Next, we make the following assumptions on the consistency of the (leave-one-out version of the) regression procedure (3) that adapts Assumption 4 for the J-SAA and J+-SAA approaches.

ASSUMPTION 4J. *The regression procedure* (3) *satisfies the following consistency properties:*

(4Ja) Pointwise error consistency: $\hat{f}_n(x) \xrightarrow{p} f^*(x)$ *for a.e.* $x \in \mathcal{X}$,

(4Jb) Mean-squared estimation error consistency: $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \|f^*(x^i) - \hat{f}_{-i}(x^i)\|^2 \xrightarrow{p} 0$.

Assumption 4J+. *The regression procedure* (3) *satisfies the following consistency properties:*

(4J+a) Pointwise error consistency: $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \|f^*(x) - \hat{f}_{-i}(x)\|^2 \xrightarrow{p} 0$ *for a.e.* $x \in \mathcal{X}$,

(4J+b) Mean-squared estimation error consistency: $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \|f^*(x^i) - \hat{f}_{-i}(x^i)\|^2 \xrightarrow{p} 0$.

Assumptions (4Jb) and (4J+b) are equivalent to Assumption (4b) when the stability assumption $\frac{1}{n}\sum_{i=1}^{n}\|\hat{f}_n(x^i) - \hat{f}_{-i}(x^i)\|^2 \xrightarrow{p} 0$ is made on the regression setup (cf. Section 5 of Barber et al. 2019). Section EC.3 demonstrates that Assumptions 4J and 4J+ hold for OLS, Lasso, kNN, and RF regression when $\mathcal{D}_n$ is i.i.d. The motivation for Assumptions 4J and 4J+ is to establish that the mean-squared deviation terms $\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}_n^{i,J}(x)\|^2$ and $\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}_n^{i,J+}(x)\|^2$ converge to zero in probability (cf. Lemma 1).

Lemma EC.1. *For any* $x \in \mathcal{X}$, *we have*

$$\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}_n^{i,J}(x)\|^2 \leq 2\|f^*(x) - \hat{f}_n(x)\|^2 + \frac{2}{n}\sum_{i=1}^{n}\|f^*(x^i) - \hat{f}_{-i}(x^i)\|^2,$$
$$\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}_n^{i,J+}(x)\|^2 \leq \frac{2}{n}\sum_{i=1}^{n}\|f^*(x) - \hat{f}_{-i}(x)\|^2 + \frac{2}{n}\sum_{i=1}^{n}\|f^*(x^i) - \hat{f}_{-i}(x^i)\|^2.$$

We now state conditions under which the sequence of objective functions of problems (6) and (7) 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 EC.1. *Suppose Assumptions 3, 4J, and 4J+, and either Assumption 1 or Assumptions 2J and 2J+ hold. Then, for a.e.* $x \in \mathcal{X}$, *the sequences of objective functions of J-SAA and J+-SAA problems* (6) *and* (7) *converge uniformly in probability to the objective function of the true problem* (1) *on the feasible region* $\mathcal{Z}$.

Proposition EC.1 helps us establish conditions under which the optimal objective values and solutions of the J-SAA and J+-SAA problems (6) and (7) converge to those of the true problem (1).

Theorem EC.1. *Suppose Assumptions 3, 4J, and 4J+, and either Assumption 1 or Assumptions 2J and 2J+ hold. Then, we have* $\hat{v}_n^J(x) \xrightarrow{p} v^*(x)$, $\hat{v}_n^{J+}(x) \xrightarrow{p} v^*(x)$, $\mathbb{D}\left(\hat{S}_n^J(x), S^*(x)\right) \xrightarrow{p} 0$, $\mathbb{D}\left(\hat{S}_n^{J+}(x), S^*(x)\right) \xrightarrow{p} 0$, $\sup_{z \in \hat{S}_n^J(x)} g(z;x) \xrightarrow{p} v^*(x)$, *and* $\sup_{z \in \hat{S}_n^{J+}(x)} g(z;x) \xrightarrow{p} v^*(x)$ *for a.e.* $x \in \mathcal{X}$.

EC.1.2. Rates of convergence

We derive rates of convergence of the optimal objective value of the sequence of J-SAA and J+-SAA problems (6) and (7) to the optimal objective value of the true problem (1) in this section. In order to enable this, we make the following assumptions on the regression procedure (3) that adapt Assumption 6 to strengthen Assumptions 4J and 4J+. Assumptions 6J and 6J+ ensure that the deviations of the J-SAA and J+-SAA problems (6) and (7) from the full-information SAA problem (2) converge at a certain rate.

ASSUMPTION 6J. *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:*

(6Ja) Pointwise error rate: $\|f^*(x)-\hat{f}_n(x)\|^2 = O_p\left(n^{-\alpha}\right)$ *for a.e.* $x\in\mathcal{X}$,

(6Jb) Mean-squared estimation error rate: $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\|f^*(x^i)-\hat{f}_{-i}(x^i)\|^2 = O_p(n^{-\alpha})$.

ASSUMPTION 6J+. *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:*

(6J+a) Pointwise error rate: $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\|f^*(x)-\hat{f}_{-i}(x)\|^2 = O_p\left(n^{-\alpha}\right)$ *for a.e.* $x\in\mathcal{X}$,

(6J+b) Mean-squared estimation error rate: $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\|f^*(x^i)-\hat{f}_{-i}(x^i)\|^2 = O_p(n^{-\alpha})$.

Appendix EC.3 demonstrates that Assumptions 6J and 6J+ hold with rates similar to those in Assumption 6 when the data $\mathcal{D}_n$ is i.i.d. Along with Lemma EC.1, Assumptions 6J and 6J+ imply that the mean-squared deviation terms for the J-SAA and J+-SAA approaches can be bounded as $\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}_n^{i,J}(x)\|^2 = O_p(n^{-\alpha})$ and $\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}_n^{i,J+}(x)\|^2 = O_p(n^{-\alpha})$ 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 EC.2. *Suppose Assumptions 5, 6J, and 6J+ hold, the objective function of the true problem* (1) *is continuous on $\mathcal{Z}$ for a.e. $x\in\mathcal{X}$, and either Assumption 1 or Assumptions 2J and 2J+ hold. Then, for a.e. $x\in\mathcal{X}$, we have $\hat{v}_n^J(x) = v^*(x)+\tilde{o}_p(n^{-\frac{\alpha}{2}})$ and $\hat{v}_n^{J+}(x) = v^*(x)+\tilde{o}_p(n^{-\frac{\alpha}{2}})$.*

Proposition EC.1 and Theorem EC.2 imply that the J-SAA and J+-SAA estimators satisfy $|g(\hat{z}_n^J(x);x)-v^*(x)| = \tilde{o}_p(n^{-\frac{\alpha}{2}})$ and $|g(\hat{z}_n^{J+}(x);x)-v^*(x)| = \tilde{o}_p(n^{-\frac{\alpha}{2}})$ for a.e. $x\in\mathcal{X}$.

EC.1.3. 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 8 to assume that the prediction error and mean-squared estimation error of the regression procedure (3) at the training points satisfy the following large deviation properties.

Assumption 8J. *The regression procedure* (3) *has the following large deviation properties: for any constant* $\kappa > 0$*, there exist positive constants* $K_f(\kappa, x)$*,* $\bar{K}_f(\kappa)$*,* $\beta_f(\kappa, x)$*, and* $\bar{\beta}_f(\kappa)$ *satisfying*

(8Ja) Pointwise error bound: $\mathbb{P}\left\{\|f^*(x) - \hat{f}_n(x)\|^2 > \kappa^2\right\} \le K_f^J(\kappa, x) \exp\left(-n\beta_f^J(\kappa, x)\right)$ *for a.e.* $x \in \mathcal{X}$*,*

(8Jb) Mean-squared estimation error bound: $\mathbb{P}\left\{\dfrac{1}{n}\sum_{i=1}^{n}\|f^*(x^i) - \hat{f}_{-i}(x^i)\|^2 > \kappa^2\right\} \le \bar{K}_f^J(\kappa) \exp\left(-n\bar{\beta}_f^J(\kappa)\right)$*.*

Assumption 8J+. *The regression procedure* (3) *has the following large deviation properties: for any constant* $\kappa > 0$*, there exist positive constants* $K_f(\kappa, x)$*,* $\bar{K}_f(\kappa)$*,* $\beta_f(\kappa, x)$*, and* $\bar{\beta}_f(\kappa)$ *satisfying*

(8J+a) Pointwise error bound: $\mathbb{P}\left\{\dfrac{1}{n}\sum_{i=1}^{n}\|f^*(x) - \hat{f}_{-i}(x)\|^2 > \kappa^2\right\} \le K_f^{J+}(\kappa, x) \exp\left(-n\beta_f^{J+}(\kappa, x)\right)$ *for a.e.* $x \in \mathcal{X}$*,*

(8J+b) Mean-squared estimation error bound: $\mathbb{P}\left\{\dfrac{1}{n}\sum_{i=1}^{n}\|f^*(x^i) - \hat{f}_{-i}(x^i)\|^2 > \kappa^2\right\} \le \bar{K}_f^{J+}(\kappa) \exp\left(-n\bar{\beta}_f^{J+}(\kappa)\right)$*.*

Assumptions 8J and 8J+ strengthen Assumptions 6J and 6J+ by imposing restrictions on the tails of the regression estimators. Please see the discussion in Appendix EC.3 for when these strengthened assumptions are satisfied with i.i.d. data $\mathcal{D}_n$. We require the following strengthening of Assumptions 2J and 2J+ for our finite sample results.

Assumption 2′J. *Assumption 2′ holds with the sequence* $\{\tilde{\varepsilon}^i_n(x)\}$ *substituted by* $\{\tilde{\varepsilon}^{i,J}_n(x)\}$*.*

Assumption 2′J+. *Assumption 2′ holds with the sequence* $\{\tilde{\varepsilon}^i_n(x)\}$ *substituted by* $\{\tilde{\varepsilon}^{i,J+}_n(x)\}$*.*

The next result presents conditions under which the maximum deviations of the J-SAA and J+-SAA objectives from the full-information SAA objective satisfy qualitatively similar large deviations bounds as that in Assumption 7.

Lemma EC.2. *Suppose Assumptions 8J and 8J+ and either Assumption 1 or Assumptions 2′J and 2′J+ hold. Then for any constant* $\kappa > 0$ *and a.e.* $x \in \mathcal{X}$*, there exist positive constants* $\bar{K}^J(\kappa, x)$*,* $\bar{K}^{J+}(\kappa, x)$*,* $\bar{\beta}^J(\kappa, x)$*, and* $\bar{\beta}^{J+}(\kappa, x)$ *satisfying*

$$\mathbb{P}\left\{\sup_{z \in \mathcal{Z}} \left|\hat{g}^J_n(z; x) - g^*_n(z; x)\right| > \kappa\right\} \le \bar{K}^J(\kappa, x) \exp\left(-n\bar{\beta}^J(\kappa, x)\right), \text{ and}$$

$$\mathbb{P}\left\{\sup_{z \in \mathcal{Z}} \left|\hat{g}^{J+}_n(z; x) - g^*_n(z; x)\right| > \kappa\right\} \le \bar{K}^{J+}(\kappa, x) \exp\left(-n\bar{\beta}^{J+}(\kappa, x)\right).$$

We now establish exponential rates of convergence in the number of samples n of the distances between solutions to the J-SAA and J+-SAA problems (6) and (7) and the set of optimal solutions to the true problem (1).

Theorem EC.3. *Suppose Assumptions 7, 8J, and 8J+ and either Assumption 1 or Assumptions 2′J and 2′J+ hold. Then, for a.e. $x \in \mathcal{X}$, given $\eta > 0$, there exist positive constants $Q^J(\eta, x)$, $Q^{J+}(\eta, x)$, $\gamma^J(\eta, x)$, and $\gamma^{J+}(\eta, x)$ such that*

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

Appendix EC.2: Application to two-stage stochastic programming problems

We present a class of stochastic programs that satisfy Assumptions 1, 2, 2′, and 5. 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. Throughout this section, we let $\bar{\mathcal{Y}}$ denote the union of the set $\mathcal{Y}$ and the set of all possible scenarios $\{\hat{f}_n(x) + \hat{\varepsilon}_n^i\}$ for the ER-SAA problem (4) for each $x \in \mathcal{X}$.

Consider first the two-stage stochastic program

$$\min_{z \in \mathcal{Z}} \mathbb{E}\left[c(z, Y)\right] := p(z) + \mathbb{E}\left[Q(z, Y)\right], \tag{EC.1}$$

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

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

We make the following assumptions on problem (EC.1).

Assumption EC.1. *The set $\mathcal{Z}$ is nonempty and compact, the matrix W has full row rank, the set $\Lambda := \{\lambda : \lambda^T W \leq q^T\}$ is nonempty, and the value function $Q(z, y) < +\infty$ for each $(z, y) \in \mathcal{Z} \times \bar{\mathcal{Y}}$.*

Assumption EC.2. *The functions h and $T(\cdot, z)$ are Lipschitz continuous on $\bar{\mathcal{Y}}$ 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 $y \in \bar{\mathcal{Y}}$ with Lipschitz constants L_p and $L_{T,z}(y)$. Additionally, the Lipschitz constants for the function T satisfy $\sup\limits_{z \in \mathcal{Z}} L_{T,y}(z) < +\infty$ and $\sup\limits_{y \in \mathcal{Y}} L_{T,z}(y) < +\infty$.*

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

Assumption EC.3. *We have $\mathbb{E}\left[\sup\limits_{z \in \mathcal{Z}} \|h(Y) - T(Y, z)\|^2\right] < +\infty$. Additionally, the random variable $V(f^*(x) + \varepsilon, \lambda, z) - \mathbb{E}_{\bar{\varepsilon} \sim P_\varepsilon}\left[V(f^*(x) + \bar{\varepsilon}, \lambda, z)\right]$ is sub-Gaussian with variance proxy $\sigma_c^2(x)$ for each $\lambda \in \operatorname{vert}(\Lambda)$ and $z \in \mathcal{Z}$ 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 fits within the above setup and readily satisfies Assumptions EC.1 and EC.2. It also satisfies Assumption EC.3 when the error ε is sub-Gaussian. Additionally, under Assumption EC.1, we have by LP duality that for each $y \in \bar{\mathcal{Y}}$:

$$Q(z,y) = \max_{\lambda \in \Lambda} \lambda^{\mathrm{T}} \left(h(y) - T(y,z)\right) = \max_{\lambda \in \mathrm{vert}(\Lambda)} \lambda^{\mathrm{T}} \left(h(y) - T(y,z)\right). \tag{EC.2}$$

Proposition EC.2. *Suppose Assumptions EC.1, EC.2, and EC.3 hold and the data $\mathcal{D}_n$ is i.i.d. Then, problem* (EC.1) *satisfies Assumptions 1, 3, 5, and 7. Furthermore, the objective function of the true problem* (1) *is continuous on $\mathcal{Z}$.*

Proof. We have by Assumptions EC.1 and EC.2 that for any $y, \bar{y} \in \bar{\mathcal{Y}}$ and $z \in \mathcal{Z}$:

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

Therefore, Assumption 1 holds since $\sup_{z \in \mathcal{Z}} L_{T,y}(z) < +\infty$ and $\max_{\lambda \in \mathrm{vert}(\Lambda)} \|\lambda\| < +\infty$. Note that Assumption (5a) readily holds since p is Lipschitz continuous on $\mathcal{Z}$ and the dual representation (EC.2) implies that $Q(\cdot, y)$ is a finite maximum of continuous functions for each $y \in \bar{\mathcal{Y}}$ by virtue of Assumption EC.2. Assumption 3 then holds by virtue of Assumptions EC.1 and EC.3 and Theorem 7.48 of Shapiro et al. (2009), which also implies that the objective function of problem (1) is continuous on $\mathcal{Z}$. Next, note that for any $z, \bar{z} \in \mathcal{Z}$ and a.e. $y \in \mathcal{Y}$:

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

Consequently, Assumption (5b) holds by virtue of Assumptions EC.1, EC.2, and EC.3, see page 164 of Shapiro et al. (2009) for details. Finally, note that for any $z \in \mathcal{Z}$ and a.e. $x \in \mathcal{X}$:

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

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

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

where the last inequality follows from Assumption EC.3. Therefore, Assumption 7 follows from Assumptions EC.1, EC.2, and EC.3 and Theorem 7.65 of Shapiro et al. (2009). □

The assumption $\sup_{y\in\mathcal{Y}} L_{T,z}(y) < +\infty$ can be relaxed to assume that the moment generating function of $L_{T,z}$ is finite valued in a neighborhood of zero, see Assumption (C3) and Theorem 7.65 in Section 7.2.9 of Shapiro et al. (2009). The discussion in Section 3 following Assumptions 3, 5, and 7 provides avenues for relaxing the i.i.d. assumption on the data $\mathcal{D}_n$. The conclusions of Proposition EC.2 can also be established for the case of objective uncertainty (i.e., only the objective coefficients q depend on Y) if Assumptions EC.1, EC.2, and EC.3 are suitably modified.

Generalization to a broader class of stochastic programs. Consider the setting where the feasible region $\mathcal{Z}$ is nonempty and compact and the function $c(z,\cdot)$ in the objective of the stochastic program (1) is multivariate polynomial with the coefficients of the polynomial being Lipschitz continuous functions of the decision variables z on the set $\mathcal{Z}$. Furthermore, suppose the random variable ε has a sub-exponential distribution. We argue that this class of stochastic programs satisfies Assumptions 2, 2′, 3, and 5. The arguments below can be generalized to the setting where $c(z,\cdot)$ is only *piecewise-polynomial* and the distribution of ε is *sufficiently light-tail.*

We first note that Assumption (5a) immediately holds for the above setup. Assumptions (A1) and (A2) in page 164 of Shapiro et al. (2009) also readily hold since moments of all orders are finite for sub-exponential random variables. Consequently, Assumption (5b) also holds. Assumption 3 holds by similar arguments through Theorem 7.48 of Shapiro et al. (2009). For the remainder of this section, we focus on establishing that Assumptions 2 and 2′ hold.

We begin by noting that $L^2_{\delta(x)}(z,f^*(x)+\varepsilon) \leq \max_{\bar y\in\mathcal{B}_{\delta(x)}(f^*(x)+\varepsilon)} \|\nabla_y c(z,\bar y)\|^2$, which implies

$$\sup_{z\in\mathcal{Z}} L^2_{\delta(x)}(z,f^*(x)+\varepsilon) \leq \sup_{z\in\mathcal{Z}}\max_{\bar y\in\mathcal{B}_{\delta(x)}(0)} \|\nabla_y c(z,f^*(x)+\varepsilon+\bar y)\|^2.$$

Let $H(x,\varepsilon) := \sup_{z\in\mathcal{Z}}\max_{\bar y\in\mathcal{B}_{\delta(x)}(0)} \|\nabla_y c(z,f^*(x)+\varepsilon+\bar y)\|^2$. Note that $H(x,\cdot)$ is nonnegative and well-defined for a.e. $x\in\mathcal{X}$, and it can be upper bounded by a polynomial function of the random vector ε for a.e. $x\in\mathcal{X}$. Therefore, Assumption (2b) follows by the weak LLN since absolute moments of all orders are finite for sub-exponential random variables. The second part of Assumption 2′ also holds if the support Ξ of ε is compact as we then have $\sup_{(z,\varepsilon)\in\mathcal{Z}\times\Xi}\max_{\bar y\in\mathcal{B}_\delta(0)} \|\nabla_y c(z,f^*(x)+\varepsilon+\bar y)\| < +\infty$ for a.e. $x\in\mathcal{X}$.

Appendix EC.3: Some prediction setups that satisfy our assumptions

We verify that Assumptions 4, 6, and 8 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 (van der Vaart 1998, van de Geer 2000), which encapsulate a rich class of prediction techniques. We often also consider the special case where the true model can be written as $Y = f(X;\theta^*) + \varepsilon$ and the goal of the regression procedure (3) is to estimate the finite-dimensional parameter $\theta^* \in \Theta$ using the data $\mathcal{D}_n$. To summarize, we largely consider the regression setup (possibly with a regularization term)

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

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

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

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 $\{\theta^*\}$. 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* the errors ε are independent of X and no restriction is made on the design points $\{x^i\}_{i=1}^n$.

EC.3.1. Parametric regression techniques

We verify that Assumptions 4, 6 and 8 and their counterparts for the J-SAA and J+-SAA problems hold for OLS regression, the Lasso and generalized linear regression models under suitable assumptions. Theorem 2.6 and Corollary 2.8 of Rigollet and Hütter (2017) present conditions under which these assumptions hold for best subset selection regression, and Theorem 2.14 therein presents similar guarantees for the Bayes Information Criterion estimator. Koltchinskii (2009) verifies these assumptions for the Dantzig selector under certain conditions. Hsu et al. (2012) verifies these conditions for ridge regression. Negahban et al. (2012) provides results for regularized M-estimators in the high-dimensional setting.

EC.3.1.1. Ordinary least squares regression We present sufficient conditions from White (2014), Hsu et al. (2012), and Rigollet and Hütter (2017) under which Assumptions 4, 6, and 8 hold. Note that Theorems 2.31 and 4.25 of White (2014) 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 (2014) also present analyses that can handle *instrumental variables*, which can be used to verify Assumptions 4, 6, and 8 when the errors ε are correlated with the features X. We have the following result:

Proposition EC.3. *Suppose $f^*(X) = \theta^* X$ and we use OLS regression to estimate θ^*.*

1. *Suppose $\mathbb{E}[|X_i \varepsilon_j|] < +\infty$, $\forall i \in [d_x]$ and $j \in [d_y]$, $\mathbb{E}[\|X\|^2] < +\infty$, and $\mathbb{E}[XX^T]$ is positive definite. If $\{(x^i, \varepsilon^i)\}_{i=1}^n$ is either i.i.d., or a stationary ergodic sequence, then $\hat{\theta}_n$ a.s. exists for n large enough and $\hat{\theta}_n \xrightarrow{a.s.} \theta^*$. Consequently, Assumption 4 holds.*
2. *Suppose $\mathbb{E}[|X_i \varepsilon_j|^2] < +\infty$, $\forall i \in [d_x]$ and $j \in [d_y]$, $\mathbb{E}[\|X\|^2] < +\infty$, $\mathbb{E}[XX^T]$ is positive definite, the covariance matrix of the random variable $\sum_{j=1}^{d_y} X\varepsilon_j$ is positive definite, and $\{(x^i, \varepsilon^i)\}_{i=1}^n$ is i.i.d. Then Assumption 6 holds with $\alpha = 1$.*
3. *Suppose $\{(x^i, \varepsilon^i)\}_{i=1}^n$ is i.i.d., the error ε is sub-Gaussian with variance proxy σ^2, 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 Assumption 8 holds 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. The first part follows from Theorems 3.5 and 3.37 of White (2014). The second part follows from Theorem 5.3 of White (2014). The third part follows from Remark 12 of Hsu et al. (2012). If we assume that ε and X are independent, then the third part also follows from Theorem 2.2 and Remark 2.3 of Rigollet and Hütter (2017). Although Rigollet and Hütter (2017) consider the fixed design case, their proof readily extends to the above setting since no restrictions were placed on the design. □

Theorems 3.49 and 3.78 of White (2014) present sufficient conditions under which Assumption 4 holds under mixing and martingale conditions on the data $\mathcal{D}_n$. Theorem 5.17 and Exercise 5.21 of White (2014) present sufficient conditions under which Assumption 6 holds with $\alpha = 1$ for ergodic and mixing data $\mathcal{D}_n$, respectively. Note that results in Bryc and Dembo (1996) and Dembo and Zeitouni (2010) can be used to establish Assumption 8 for the non-i.i.d. setting.

The above results can be used in conjunction with the techniques in Section EC.3.2 to verify Assumptions 4J, 4J+, 6J, 6J+, 8J, and 8J+ for i.i.d. data $\mathcal{D}_n$. In the remainder of this section, we specialize the verification of these assumptions for OLS regression when problem (1) is a two-stage stochastic LP (see Example 1). We assume that $d_y = 1$ for ease of exposition.

Following Remark 1 and the discussion in Section EC.1, 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 $\bar{X}_{[i]} = (x^i)^{\mathrm{T}}$, $h^i := (\bar{X}(\bar{X}^{\mathrm{T}}\bar{X})^{-1}\bar{X}^{\mathrm{T}})_{ii}$ denote the ith *leverage score*, and $e^i := y^i - \hat{\theta}_n^{\mathrm{T}} x^i$ denote the residual of the model $\hat{f}_n$ at the ith data point. From Section 10.6.3 of Seber and Lee (2003), we have

$$\frac{1}{n}\sum_{i=1}^n \|\hat{f}_n(x^i) - \hat{f}_{-i}(x^i)\| = \frac{1}{n}\sum_{i=1}^n \frac{h^i|e^i|}{1-h^i} \leq \sqrt{\frac{1}{n}\sum_{i=1}^n (h^i)^2}\sqrt{\frac{1}{n}\sum_{i=1}^n \left(\frac{e^i}{1-h^i}\right)^2} \leq \frac{d_x}{\sqrt{n}}\sqrt{\frac{1}{n}\sum_{i=1}^n \left(\frac{e^i}{1-h^i}\right)^2},$$

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

where Tr denotes the trace operator. The quantity $\sqrt{\frac{1}{n}\sum_{i=1}^n \left(\frac{e^i}{1-h^i}\right)^2}$ 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.

EC.3.1.2. The Lasso and high-dimensional generalized linear models Following van de Geer (2008) and Bunea et al. (2007), we consider generalized linear models with an ℓ_1-penalty. We assume that $d_y = 1$ for ease of exposition. The setup is as follows: the model class $\mathcal{F} := \left\{ f : f(\cdot;\theta) := \sum_{k=1}^m \theta_k \psi_k(\cdot), \theta \in \Theta \right\}$, where $\{\psi_k(\cdot)\}_{k=1}^m$ is a sequence of real-valued basis functions with domain $\mathcal{X}$, the data $\mathcal{D}_n$ 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 van de Geer 2008), and the estimate $\hat{\theta}_n$ of θ^* is obtained as

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

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 (2008) and Theorems 2.1, 2.2, and 2.3 of Bunea et al. (2007) present conditions under which Assumptions 4, 6 and 8 hold for the above setting.

In the remainder of this section, we specialize the results of Bunea et al. (2007) to the traditional Lasso setup (Tibshirani 1996). In this setup, $m = d_x$, $\psi_k(x) = x_k$, $\Theta = \mathbb{R}^{d_x}$, and $\ell(y, \hat{y}) = \|y - \hat{y}\|^2$.

Proposition EC.4. *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 ε is sub-Gaussian with variance proxy σ^2. 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 Assumption 4 holds, Assumption 6 holds with $\alpha = 1$, and Assumption 8 holds 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. (2007). □

Chatterjee (2013) establishes consistency of the Lasso under the following weaker assumptions: the data $\mathcal{D}_n$ 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 6 holds at a slower rate with $\alpha = 0.5$. Theorem 2.15 of Rigollet and Hütter (2017) can then be used to show that Assumption 8 holds 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 (2015) present conditions under which Assumptions 4, 6, and 8 can be verified for time series data $\mathcal{D}_n$. The above results can be used in conjunction with the discussion in Section EC.3.2 to derive rates of convergence and finite sample guarantees for the Jackknife-based estimators for i.i.d. data $\mathcal{D}_n$.

EC.3.2. Theory for general M-estimators

We use results from Chapter 5 of van der Vaart (1998), Chapter 3 of van der Vaart and Wellner (1996), and Shapiro et al. (2009) to verify Assumptions 4, 6 and 8 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 4 and 6, 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 $\|\theta^* - \hat{\theta}_n\| \xrightarrow{p} 0$.

Verifying Assumption 4. Theorem 5.7 of van der Vaart (1998) presents conditions under which $\hat{\theta}_n \xrightarrow{p} \theta^*$ for i.i.d. data $\mathcal{D}_n$ (cf. Theorems 5.3 and 5.4 of Shapiro et al. 2009). Similar to the discussion following Assumption 3, this result also holds when $\mathcal{D}_n$ satisfies certain mixing/stationarity assumptions. Section 5.2 of van der Vaart (1998) also presents alternative conditions for $\hat{\theta}_n \xrightarrow{p} \theta^*$.

Verifying Assumption 6. We discuss conditions under which $\|\hat{\theta}_n - \theta^*\| \xrightarrow{p} 0$ at certain rates. Theorem 5.23 of van der Vaart (1998) presents regularity conditions under which this convergence holds at the conventional $n^{-0.5}$ rate, in which case Assumption 6 holds with $\alpha = 1$ (cf. Theorem 5.8 of Shapiro et al. 2009). Once again, the above conclusion holds when the observations $\mathcal{D}_n$ satisfy certain mixing/stationarity assumptions. Chapter 5 of van der Vaart (1998) and Chapter 3.2 of van der Vaart and Wellner (1996) provide some examples of M-estimators that possess this rate of convergence. Theorem 5.52 and Chapter 25 of van der Vaart (1998) present conditions under which Assumption 6 holds with constant $\alpha < 1$ (including the setting of semiparametric regression).

Verifying Assumption 8. 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 7. Specifically, suppose for any constant $\kappa > 0$, there exist positive constants $\hat{K}(\kappa)$ and $\hat{\beta}(\kappa)$ such that

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

see the discussion surrounding Assumption 7 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 Shapiro et al. 2009). Theorem 2.3 of Homem-de-Mello (2008) then implies that Assumption 8 holds whenever the sample average term $\frac{1}{n}\sum_{i=1}^{n} L_f^2(x^i)$ is bounded. We note that results in Bryc and Dembo (1996), Dembo and Zeitouni (2010) can be used to establish such uniform exponential bounds for mixing data $\mathcal{D}_n$ by adapting Lemma 2.4 of Homem-de-Mello (2008).

Verifying the assumptions for the Jackknife-based methods. We now present techniques for verifying Assumptions 4J, 4J+, 6J, 6J+, 8J, and 8J+ when the data $\mathcal{D}_n$ is i.i.d. Noting from Markov's inequality that

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}\|f^*(x^i) - \hat{f}_{-i}(x^i)\|^2 > \kappa^2\right\} \leq \frac{1}{n\kappa^2}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{D}_N}\left[\|f^*(x^i) - \hat{f}_{-i}(x^i)\|^2\right] = \frac{1}{\kappa^2}\mathbb{E}_{\mathcal{D}_{n-1},X}\left[\|f^*(X) - \hat{f}_{n-1}(X)\|^2\right],$$

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}\|f^*(x) - \hat{f}_{-i}(x)\|^2 > \kappa^2\right\} \leq \frac{1}{n\kappa^2}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{D}_n}\left[\|f^*(x) - \hat{f}_{-i}(x)\|^2\right] = \frac{1}{\kappa^2}\mathbb{E}_{\mathcal{D}_{n-1}}\left[\|f^*(x) - \hat{f}_{n-1}(x)\|^2\right],$$

we have that Assumptions 6J and 6J+ on the Jackknife-based methods hold if, for a.e. $x \in \mathcal{X}$, the expectations $\mathbb{E}_{\mathcal{D}_{n-1},X}\left[\|f^*(X) - \hat{f}_{n-1}(X)\|^2\right]$ and $\mathbb{E}_{\mathcal{D}_{n-1}}\left[\|f^*(x) - \hat{f}_{n-1}(x)\|^2\right]$ converge to zero at suitable rates. Under the aforementioned Lipshitz continuity assumption on the function $f(x;\cdot)$ at θ^* and the assumption that $\mathbb{E}\left[L_f^2(X)\right] < +\infty$, it suffices to establish rates of convergence for the expectation term $\mathbb{E}_{\mathcal{D}_{n-1}}\left[\|\theta^* - \hat{\theta}_{n-1}\|^2\right]$. 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. (2012) for instance. Chapter 14 of Biau and

Devroye (2015) provides similar rate results for kNN regression. Alternatively, we can also bound the terms appearing in the assumptions for the Jackknife-based formulations as

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

$$\frac{1}{n}\sum_{i=1}^{n}\|f(x^i;\theta^*)-f(x^i;\hat{\theta}_{-i})\|^2 \leq \frac{1}{n}\sum_{i=1}^{n}L_f^2(x^i)\|\theta^*-\hat{\theta}_{-i}\|^2 \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}L_f^4(x^i)}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\theta^*-\hat{\theta}_{-i}\|^4},$$

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 terms $\frac{1}{n}\sum_{i=1}^{n}\|\theta^*-\hat{\theta}_{-i}\|^4$ and $\frac{1}{n}\sum_{i=1}^{n}L_f^4(x^i)$. A third direct approach is to use the weaker bounds

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}\|f^*(x^i)-\hat{f}_{-i}(x^i)\|^2>\kappa^2\right\} \leq \sum_{i=1}^{n}\mathbb{P}\left\{\|f^*(x^i)-\hat{f}_{-i}(x^i)\|^2>\kappa^2\right\} = n\mathbb{P}\left\{\|f^*(X)-\hat{f}_{n-1}(X)\|^2>\kappa^2\right\},$$

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n}\|f^*(x)-\hat{f}_{-i}(x)\|^2>\kappa^2\right\} \leq \sum_{i=1}^{n}\mathbb{P}\left\{\|f^*(x)-\hat{f}_{-i}(x)\|^2>\kappa^2\right\} = n\mathbb{P}\left\{\|f^*(x)-\hat{f}_{n-1}(x)\|^2>\kappa^2\right\}.$$

Finally, note that it is sufficient to establish rates and finite sample guarantees for the terms $\frac{1}{n}\sum_{i=1}^{n}\|\hat{f}_n(x^i)-\hat{f}_{-i}(x^i)\|^2$ and $\frac{1}{n}\sum_{i=1}^{n}\|\hat{f}_n(x)-\hat{f}_{-i}(x)\|^2$ when Assumptions 4, 6, and 8 are enforced.

EC.3.3. Nonparametric regression techniques

We verify that Assumptions 4 and 6 hold for kNN regression, CART, and RF regression, and state a large deviation result similar to Assumption 8 for kNN regression. The discussion in Section EC.3.2 then provides an avenue for verifying the corresponding assumptions for the J-SAA and J+-SAA problems for i.i.d. data $\mathcal{D}_n$. Note that Walk (2010), Györfi et al. (2006), and Chen and Shah (2018) can be used to verify some of these assumptions for kernel regression and semi-recursive Devroye-Wagner estimates for mixing data $\mathcal{D}_n$, Raskutti et al. (2012) can be used to verify these assumptions for sparse additive nonparametric regression, Chapter 13 of Wainwright (2019) can be used to verify these assumptions for (regularized) nonparametric least squares regression, and Seijo and Sen (2011) and Mazumder et al. (2019) can be used to verify these assumptions for convex regression. In what follows, we only consider the setting where the data $\mathcal{D}_n$ 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 $\hat{f}_n(x) := \dfrac{1}{k}\displaystyle\sum_{i=1}^{k} y^{(i)}(x)$, where $\{(y^{(i)}(x), x^{(i)}(x))\}_{i=1}^{n}$ 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 EC.5. *Suppose the data $\mathcal{D}_n$ is i.i.d. and the support $\mathcal{X}$ of the covariates is compact. Consider the setting where we use kNN regression to estimate the regression function f^* with the parameter 'k' satisfying $\lim\limits_{n\to\infty}\frac{k}{\log(n)} = \infty$ and $k = o(n)$.*

1. *Suppose the function* f^* *is continuous on* $\mathcal{X}$*. If the distribution of the errors* ε *satisfies* $\sup_{x \in \mathcal{X}} \mathbb{E}\left[\exp(\lambda|\varepsilon_j|) \mid X = x\right] < +\infty$ *for each* $j \in [d_y]$ *and some* $\lambda > 0$*, then Assumption 4 holds.*
2. *Suppose the function* f^* *is twice continuously differentiable on* $\mathcal{X}$*, the random vector* X *has a density that is twice continuously differentiable, and the error* ε *is sub-Gaussian. Then, there exists a choice of the parameter 'k' such that Assumption 6 holds with* $\alpha = \frac{O(1)}{d_x}$*.*
3. *Suppose the function* f^* *is Lipschitz continuous on* $\mathcal{X}$*, the error* ε *is sub-Gaussian with variance proxy* σ^2*, 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 (2015). The second part follows from Theorems 14.3 and 14.5 of Biau and Devroye (2015) and Markov's inequality. The last part follows from Lemma 10 of Bertsimas and McCord (2019). □

Lemma 7 of Bertsimas and McCord (2019) presents conditions under which CART regression satisfies Assumption 4. Along with Theorem 8 of Wager and Athey (2018), the above result can be used to show that Assumption 6 holds for CART regression with $\alpha = \frac{O(1)}{d_x}$. Lemma 9 of Bertsimas and McCord (2019) presents conditions under which RF regression satisfies Assumption 4. Once again, we can use this result along with Theorem 8 of Wager and Athey (2018) to show that Assumption 6 holds for RF regression with $\alpha = \frac{O(1)}{d_x}$.

Appendix EC.4: 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. (2009) (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 (Mak et al. 1999).

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

```
**Input**: Covariate realization $X = x$ and data-driven solution $\hat{z}_n(x)$ for a particular realization of the data $\mathcal{D}_n$.
**Output**: $\hat{B}_{99}(x)$, which is a normalized estimate of the 99% UCB on the optimality gap of $\hat{z}_n(x)$.
**for** $k = 1, \cdots, 30$ **do**
    Draw 1000 i.i.d. samples $\bar{\mathcal{D}}^k := \{\bar{\varepsilon}^{k,i}\}_{i=1}^{1000}$ of $\varepsilon$ according to the distribution $P_\varepsilon$.
    Estimate the optimal value $v^*(x)$ by solving the full-information SAA problem (2) using the data $\bar{\mathcal{D}}^k$:
        $$\bar{v}^k(x) := \min_{z \in \mathcal{Z}} \frac{1}{1000} \sum_{i=1}^{1000} c(z, f^*(x) + \bar{\varepsilon}^{k,i}).$$
    Estimate the out-of-sample cost of the solution $\hat{z}_n(x)$ using the data $\bar{\mathcal{D}}^k$:
        $$\hat{v}^k(x) := \frac{1}{1000} \sum_{i=1}^{1000} c(\hat{z}_n(x), f^*(x) + \bar{\varepsilon}^{k,i}).$$
    Estimate the optimality gap of the solution $\hat{z}_n(x)$ as $\hat{G}^k(x) = \hat{v}^k(x) - \bar{v}^k(x)$.
**end for**
Construct the normalized estimate of the 99% UCB on the optimality gap of $\hat{z}_n(x)$ as
    $$\hat{B}_{99}(x) := \frac{100}{|\bar{v}(x)|} \left( \frac{1}{30} \sum_{k=1}^{30} \hat{G}^k(x) + 2.462 \sqrt{\frac{\mathrm{var}(\{\hat{G}^k(x)\})}{30}} \right),$$
where $\bar{v}(x) := \frac{1}{30} \sum_{k=1}^{30} \bar{v}^k(x)$ and $\mathrm{var}(\{\hat{G}^k(x)\})$ denotes the variance of the gaps $\{\hat{G}^k(x)\}_{k=1}^{30}$.
```

References

Barber RF, Candes EJ, Ramdas A, Tibshirani RJ (2019) Predictive inference with the jackknife+. *arXiv preprint arXiv:1905.02928v3* 1–44.

Basu S, Michailidis G (2015) Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics* 43(4):1535–1567.

Bertsimas D, McCord C (2019) From predictions to prescriptions in multistage optimization problems. *arXiv preprint arXiv:1904.11637* 1–38.

Biau G, Devroye L (2015) *Lectures on the nearest neighbor method* (Springer).

Bryc W, Dembo A (1996) Large deviations and strong mixing. *Annales de l'IHP Probabilités et statistiques*, volume 32, 549–569.

Bunea F, Tsybakov A, Wegkamp M, et al. (2007) Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics* 1:169–194.

Chatterjee S (2013) Assumptionless consistency of the lasso. *arXiv preprint arXiv:1303.5817* .

Chen GH, Shah D (2018) Explaining the success of nearest neighbor methods in prediction. *Foundations and Trends in Machine Learning* 10(5-6):337–588.

Dembo A, Zeitouni O (2010) *Large deviations techniques and applications*, volume 38 of *Stochastic Modelling and Applied Probability* (Springer), 2nd edition, URL `http://dx.doi.org/10.1007/978-3-642-03311-7`.

Györfi L, Kohler M, Krzyzak A, Walk H (2006) *A distribution-free theory of nonparametric regression* (Springer Science & Business Media).

Homem-de-Mello T (2008) On rates of convergence for stochastic optimization problems under non–independent and identically distributed sampling. *SIAM Journal on Optimization* 19(2):524–551.

Hsu D, Kakade SM, Zhang T (2012) Random design analysis of ridge regression. *Conference on learning theory*, 9–1.

Koltchinskii V (2009) The Dantzig selector and sparsity oracle inequalities. *Bernoulli* 15(3):799–828.

Lewandowski D, Kurowicka D, Joe H (2009) Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis* 100(9):1989–2001.

Mak WK, Morton DP, Wood RK (1999) Monte Carlo bounding techniques for determining solution quality in stochastic programs. *Operations Research Letters* 24(1-2):47–56.

Mazumder R, Choudhury A, Iyengar G, Sen B (2019) A computational framework for multivariate convex regression and its variants. *Journal of the American Statistical Association* 114(525):318–331.

Negahban SN, Ravikumar P, Wainwright MJ, Yu B, et al. (2012) A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science* 27(4):538–557.

Raskutti G, Wainwright MJ, Yu B (2012) Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research* 13(Feb):389–427.

Rigollet P, Hütter JC (2017) High dimensional statistics. Lecture notes for MIT's 18.657 course, URL `http://www-math.mit.edu/~rigollet/PDFs/RigNotes17.pdf`.

Seber GA, Lee AJ (2003) *Linear regression analysis* (John Wiley & Sons).

Seijo E, Sen B (2011) Nonparametric least squares estimation of a multivariate convex regression function. *The Annals of Statistics* 39(3):1633–1657.

Shapiro A, Dentcheva D, Ruszczyński A (2009) *Lectures on stochastic programming: modeling and theory* (SIAM).

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

van de Geer SA (2000) *Empirical Processes in M-estimation*, volume 6 (Cambridge university press).

van de Geer SA (2008) High-dimensional generalized linear models and the lasso. *The Annals of Statistics* 36(2):614–645.

van der Vaart AW (1998) *Asymptotic statistics*, volume 3 (Cambridge university press).

van der Vaart AW, Wellner JA (1996) *Weak convergence and empirical processes: with applications to statistics* (Springer).

Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113(523):1228–1242.

Wainwright MJ (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48 (Cambridge University Press).

Walk H (2010) Strong laws of large numbers and nonparametric estimation. *Recent Developments in Applied Probability and Statistics*, 183–214 (Springer).

White H (2014) *Asymptotic theory for econometricians* (Academic press).