Heteroscedasticity-aware residuals-based contextual stochastic optimization

Rohit Kannan[1], Güzin Bayraksan[2], and James R. Luedtke[3]

[1]Wisconsin Institute for Discovery, University of Wisconsin-Madison, Madison, WI, USA.
E-mail: rohitk@alum.mit.edu
[2]Department of Integrated Systems Engineering, The Ohio State University, Columbus, OH, USA.
E-mail: bayraksan.1@osu.edu
[3]Department of Industrial & Systems Engineering and Wisconsin Institute for Discovery,
University of Wisconsin-Madison, Madison, WI, USA. E-mail: jim.luedtke@wisc.edu

January 8, 2021

Abstract

We explore generalizations of some integrated learning and optimization frameworks for data-driven contextual stochastic optimization that can adapt to heteroscedasticity. We identify conditions on the stochastic program, data generation process, and the prediction setup under which these generalizations possess asymptotic and finite sample guarantees for a class of stochastic programs, including two-stage stochastic mixed-integer programs with continuous recourse. We verify that our assumptions hold for popular parametric and nonparametric regression methods.

Key words: Data-driven stochastic programming, distributionally robust optimization, covariates, regression, heteroscedasticity, convergence rate, large deviations

1 Introduction

We study data-driven stochastic programming in the presence of covariate/contextual information and examine heteroscedastic cases. Specifically, we consider the setting where we have a finite number of observations of the uncertain parameters Y within an optimization model along with simultaneous observations of random covariates X. Given a new random observation $X = x$, our goal is to solve the *conditional stochastic program*

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

Here, z denotes the decision vector, $\mathcal{Z} \subseteq \mathbb{R}^{d_z}$ is the feasible region, and $c : \mathbb{R}^{d_z} \times \mathbb{R}^{d_y} \to \overline{\mathbb{R}}$ is an extended real-valued function. An example application of this framework is production planning under demand uncertainty [5], where products' demands (Y) can be predicted using covariates (X) such as historical demands, location, and web chatter before making decisions (z) on production and inventory levels. Another application is grid scheduling under wind uncertainty [10], where covariates (X) such as weather observations, seasonality, and location can be used to predict available wind power (Y) before creating generator schedules (z). Heteroscedasticity arises, for instance, when the variability of product demands or wind power availability depends significantly on the location, seasonality, or other covariates.

Kannan et al. [17, 18] consider data-driven approaches that integrate a machine learning prediction model within a sample average approximation (SAA) or distributionally robust optimization (DRO) setup to approximate the solution to the conditional stochastic program (SP); see also [1, 26]. They first fit a statistical/machine learning model to predict Y given X and use this model and its residuals to construct scenarios for Y given $X = x$. Then, they use these scenarios within an SAA or DRO framework to approximate the solution to (SP). We refer the readers to [e.g., 1, 5, 17, 18, 26] for a review of other data-driven approximations to (SP).

The data-driven formulations in Kannan et al. [17, 18] assume that the dependence of the random vector Y on the random covariates X can be modeled as $Y = f^*(X) + \varepsilon$, where $f^*(x) := \mathbb{E}[Y \mid X = x]$ is the regression function and ε are zero-mean errors. These approaches crucially require the errors ε to be *independent* of the covariates X. Motivated by applications where such an assumption may fail to hold, we explore generalizations of these approaches that do not require this independence assumption.

Notation. Let $[n] := \{1, \dots, n\}$, $\|\cdot\|$ denote the Euclidean or operator ℓ_2-norm, $\text{proj}_{\mathcal{S}}(v)$ denote the orthogonal projection of v onto a nonempty closed convex set $\mathcal{S}$, I denote an identity matrix of appropriate dimension, v^{T} denote the transpose of a vector v, and $A \succ 0$ denote that the matrix A is positive definite. Let δ denote the Dirac measure. For scalars $c_1, \dots, c_l$, we write $\text{diag}(c_1, \dots, c_l)$ to denote the $l \times l$ diagonal matrix with ith diagonal entry equal to c_i. For sets $\mathcal{A}, \mathcal{B} \subseteq \mathbb{R}^{d_z}$, let $\mathbb{D}(\mathcal{A}, \mathcal{B}) := \sup_{v \in \mathcal{A}} \text{dist}(v, \mathcal{B})$ denote the deviation of $\mathcal{A}$ from $\mathcal{B}$, where $\text{dist}(v, \mathcal{B}) := \inf_{w \in \mathcal{B}} \|v - w\|$. The abbreviations 'a.e.', 'a.s.', 'LLN', 'i.i.d.', and 'r.h.s.' are shorthand for 'almost everywhere', 'almost surely', 'law of large numbers', 'independent and identically distributed', and 'right-hand side'. For a random vector V with probability measure P_V, we write a.e. $v \in V$ to denote P_V-a.e. $v \in V$. The symbols $\xrightarrow{p}$ and $\xrightarrow{a.s.}$ denote convergence in probability and almost surely with respect to the probability measure generating the joint data on (Y, X). For random sequences $\{V_n\}$ and $\{W_n\}$, we write $V_n = o_p(W_n)$ and $V_n = O_p(W_n)$ to convey that $V_n = R_n W_n$ with $\{R_n\}$ converging in probability to zero, or being bounded in probability, respectively. We write $O(1)$ for generic constants.

2 Heteroscedasticity-aware residuals-based approximations

2.1 Framework and approximations

To handle heteroscedasticity, we assume that the random vector Y is related to the random covariates X as[1] $Y = f^*(X) + Q^*(X)\varepsilon$, where f^* denotes the regression function, $Q^*(X)$ is the square root of the conditional covariance matrix of the error term, and the zero-mean random errors ε are independent of the covariates X. This type of model is common in statistics; see, e.g., [2, 6, 8, 30]. The functions f^* and Q^* are assumed to belong to known classes of functions $\mathcal{F}$ and $\mathcal{Q}$, respectively (which may be infinite-dimensional and depend on the sample size n). Let $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$, $\mathcal{X} \subseteq \mathbb{R}^{d_x}$, and $\Xi \subseteq \mathbb{R}^{d_y}$ denote the supports of Y, X, and ε, respectively. Additionally, let $P_{Y|X=x}$ denote the conditional distribution of Y given $X = x$ and P_X and P_ε denote the distributions of X and ε, respectively. We assume that $\mathcal{Y}$ is nonempty and convex and $Q^*(x) \succ 0$ for a.e. $x \in \mathcal{X}$.

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

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

where the expectation above is computed with respect to the distribution P_ε of ε. We assume that the feasible set $\mathcal{Z} \subset \mathbb{R}^{d_z}$ is nonempty and compact, $\mathbb{E}[|c(z, f^*(x) + \varepsilon)|] < +\infty$ for each $z \in \mathcal{Z}$ and a.e. $x \in \mathcal{X}$, and the function $g(\cdot; x)$ is lower semicontinuous on $\mathcal{Z}$ for a.e. $x \in \mathcal{X}$. These assumptions ensure that problem (1) is well-defined and its set of optimal solutions $S^*(x)$ 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) and $\{\varepsilon^i\}_{i=1}^n$ denote the corresponding realizations of the errors ε. Note that these realizations of ε satisfy

$$\varepsilon^i = \big[Q^*(x^i)\big]^{-1}(y^i - f^*(x^i)), \quad \forall i \in [n].$$

If we know the functions f^* and Q^*, then we can construct the following *full-information SAA* (FI-SAA) to problem (1) using the data $\mathcal{D}_n$:

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

[1] We focus our attention on this popular model of heteroscedasticity even though our framework applies more generally, e.g., to relationships of the form $Y = m^*(X, \varepsilon)$ with the mapping $m^*(x, \cdot)$ being *invertible* for a.e. $x \in \mathcal{X}$ and satisfying some regularity conditions.

Because the functions f^* and Q^* are unknown, we first estimate them by $\hat{f}_n$ and $\hat{Q}_n$, respectively, using a regression method on the data $\mathcal{D}_n$ (see Section 4 for details). Assuming that the estimate $\hat{Q}_n$ is a.s. positive definite on $\mathcal{X}$ (i.e., it a.s. satisfies $\hat{Q}_n(x) \succ 0$ for a.e. $x \in \mathcal{X}$), we then use the empirical estimates

$$\hat{\varepsilon}_n^i := \left[\hat{Q}_n(x^i)\right]^{-1}(y^i - \hat{f}_n(x^i)), \quad \forall i \in [n],$$

of $\{\varepsilon^i\}_{i=1}^n$ to construct the following *empirical residuals-based SAA* (ER-SAA) to problem (1) in the heteroscedastic setting (cf. [17, 18])[2]:

$$\hat{v}_n^{ER}(x) := \min_{z \in \mathcal{Z}} \left\{ \hat{g}_n^{ER}(z; x) := \frac{1}{n} \sum_{i=1}^n c\big(z, \text{proj}_{\mathcal{Y}}(\hat{f}_n(x) + \hat{Q}_n(x)\hat{\varepsilon}_n^i)\big) \right\}. \tag{3}$$

Let $\hat{z}_n^{ER}(x)$ denote an optimal solution to problem (3) and $\hat{S}_n^{ER}(x)$ denote its optimal solution set. Additionally, let $P_n^*(x)$ and $\hat{P}_n^{ER}(x)$ denote the estimates of the conditional distribution $P_{Y|X=x}$ of Y given $X = x$ corresponding to the FI-SAA problem (2) and ER-SAA problem (3), respectively, i.e.,

$$P_n^*(x) := \frac{1}{n} \sum_{i=1}^n \delta_{f^*(x) + Q^*(x)\varepsilon^i} \quad \text{and} \quad \hat{P}_n^{ER}(x) := \frac{1}{n} \sum_{i=1}^n \delta_{\text{proj}_{\mathcal{Y}}(\hat{f}_n(x) + \hat{Q}_n(x)\hat{\varepsilon}_n^i)}.$$

When we only have a limited number of observations n, the following residuals-based DRO formulation provides an alternative to the ER-SAA problem (3) that can yield solutions with better out-of-sample performance (cf. [18]):

$$\min_{z \in \mathcal{Z}} \sup_{Q \in \hat{\mathcal{P}}_n(x)} \mathbb{E}_{Y \sim Q}\left[c(z, Y)\right], \tag{4}$$

where $\hat{\mathcal{P}}_n(x)$ is an ambiguity set for $P_{Y|X=x}$. Following [18], we call problem (4) with $\hat{\mathcal{P}}_n(x)$ centered at $\hat{P}_n^{ER}(x)$ the *empirical residuals-based DRO* (ER-DRO) problem.

2.2 Theoretical results for the heteroscedastic setting

For the homoscedastic case, i.e., when $Q^* \equiv \hat{Q}_n \equiv I$ and so the model class $\mathcal{Q}$ comprises only the constant function $Q : x \mapsto I$, $\forall x \in \mathcal{X}$, Kannan et al. [17, 18] investigate conditions under which the optimal value of problems (3) and (4) asymptotically converge in probability to those of the true problem (1). They also identify conditions under which every accumulation point of a sequence of optimal solutions to problems (3) and (4) is in probability an optimal solution to problem (1) and outline conditions under which solutions to problems (3) and (4) possess finite sample guarantees. An integral part of this analysis is bounding a distance between the empirical distributions $\hat{P}_n^{ER}(x)$ and $P_n^*(x)$.

By the Lipschitz continuity of orthogonal projections, we have for each $x \in \mathcal{X}$

$$\|\text{proj}_{\mathcal{Y}}(\hat{f}_n(x) + \hat{Q}_n(x)\hat{\varepsilon}_n^i) - (f^*(x) + Q^*(x)\varepsilon^i)\| \leq \|\tilde{\varepsilon}_n^i(x)\|, \qquad \forall i \in [n],$$

where the ith deviation term $\tilde{\varepsilon}_n^i(x)$ is given by

$$\tilde{\varepsilon}_n^i(x) := (\hat{f}_n(x) + \hat{Q}_n(x)\hat{\varepsilon}_n^i) - (f^*(x) + Q^*(x)\varepsilon^i).$$

The analysis in [17, 18] implies that under certain assumptions on the stochastic program (1), asymptotic and finite sample guarantees on the power mean deviation term $(\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|^p)^{1/p}$ for a suitable value of $p \geq 1$ translate to asymptotic and finite sample guarantees on the optimal value and optimal solutions to the ER-SAA problem (3) and the ER-DRO problem (4). Specifically, the analyses in [17, 18] imply that theoretical guarantees on the term $(\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|^p)^{1/p}$ for $p = 1$ and $p = 2$ translate to theoretical guarantees on solutions to problems (3) and (4) for a class of two-stage stochastic mixed-integer programs (MIPs) with continuous recourse and, in the ER-DRO setting, to broad families of ambiguity sets.

We now provide concrete examples of how guarantees on the *mean deviation term* $\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|$ (i.e., when $p = 1$) translate to guarantees on the ER-SAA problem (3). In addition to focusing on the ER-SAA problem (3) for brevity, we narrow our attention to stochastic programs (1) whose objective function satisfies the following Lipschitz condition.

[2] We can also construct similar generalizations of the Jackknife-based SAAs in [17].

Assumption 1. For each $z \in \mathcal{Z}$, the function $c(z, \cdot)$ is Lipschitz continuous on $\mathcal{Y}$ with Lipschitz constant $L(z)$ satisfying $\sup_{z \in \mathcal{Z}} L(z) < +\infty$.

As an example, Appendix EC.2 of [17] verifies that Assumption 1 holds for two-stage stochastic MIPs with continuous recourse under mild conditions. For extensions of the below results to a broader class of stochastic programs (1) and to the ER-DRO problem (4), we refer the readers to [17, 18].

We now list conditions on the FI-SAA problem (2) under which consistency and asymptotic optimality, rates of convergence, and finite sample guarantees—to be defined precisely in respective theorems below—can be achieved for the ER-SAA approximation (3) of the true problem (1) in the heteroscedastic setting. As mentioned, a key component of this analysis requires respective conditions to be satisfied by the mean deviation term; these are investigated in Section 3. Section 4 presents examples of regression/learning setups that satisfy the assumptions set forth for the heteroscedastic setting.

We begin with a uniform weak LLN assumption on the FI-SAA objective (see Assumption 3 of [17] and the surrounding discussion for conditions under which it holds). Along with suitable convergence of the mean deviation term, this assumption helps us establish *uniform* convergence in probability of the sequence of objective functions of the ER-SAA problem (3) to the objective function of the true problem (1) on the feasible set $\mathcal{Z}$ (see Proposition 1 of [17]). This in turn provides the building block for consistency and asymptotic optimality.

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

Our first result implies that consistency of the mean deviation term $\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|$ translates to consistency and asymptotic optimality of solutions to the ER-SAA problem (3).

Theorem 1. [**Consistency and asymptotic optimality**] Suppose Assumptions 1 and 2 hold and the mean deviation term converges to zero in probability, i.e., $\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\| \xrightarrow{p} 0$ for a.e. $x \in \mathcal{X}$. Then for a.e. $x \in \mathcal{X}$

$$\hat{v}_n^{ER}(x) \xrightarrow{p} v^*(x), \quad \mathbb{D}\big(\hat{S}_n^{ER}(x), S^*(x)\big) \xrightarrow{p} 0, \quad \text{and} \quad \sup_{z \in \hat{S}_n^{ER}(x)} g(z; x) \xrightarrow{p} v^*(x).$$

Proof. See the proofs of Proposition 1 and Theorem 1 of Kannan et al. [17]. □

Next, we refine Assumption 2 to assume that the sequence of objective functions of the FI-SAA problem (2) converges to the objective function of the true problem (1) at a suitable rate (see Assumption 5 of [17] and the surrounding discussion for conditions under which it holds).

Assumption 3. The function c in problem (1) and the data $\mathcal{D}_n$ satisfy the following functional central limit theorem for the FI-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 $g_n^*(\cdot; x)$, $g(\cdot; x)$, and $V(\cdot; x)$ are (random) elements of $L^\infty(\mathcal{Z})$, the Banach space of essentially bounded functions on $\mathcal{Z}$ equipped with the supremum norm.

Our second result implies that rates of convergence of the mean deviation term $\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|$ to zero directly translate to rates of convergence of the suboptimality of ER-SAA solutions to zero.

Theorem 2. [**Rate of convergence**] Suppose Assumptions 1 and 3 hold and there exists a constant $r \in (0, 1]$ such that $\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\| = O_p(n^{-r/2})$ for a.e. $x \in \mathcal{X}$. Then, for a.e. $x \in \mathcal{X}$

$$\big|\hat{v}_n^{ER}(x) - v^*(x)\big| = O_p(n^{-r/2}) \quad \text{and} \quad \big|g(\hat{z}_n^{ER}(x); x) - v^*(x)\big| = O_p(n^{-r/2}).$$

Proof. Follows from the proof of Theorem 11 of Kannan et al. [18] (cf. Theorem 2 of [17]). □

Finally, we refine Assumption 3 to assume that the sequence of objectives of the FI-SAA problem (2) possess a finite sample guarantee (see [17, Assumption 7] and the discussion after it for conditions under which it holds).

Assumption 4. The FI-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

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

Our final result of this section implies that finite sample guarantees on the mean deviation term $\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|$ translate to finite sample guarantees on solutions to the ER-SAA problem (3).

Theorem 3. [**Finite sample guarantee**] Suppose Assumptions 1 and 4 hold and for any constant $\kappa > 0$ and a.e. $x \in \mathcal{X}$, there exist positive constants $\tilde{K}(\kappa, x)$ and $\tilde{\beta}(\kappa, x)$ such that

$$\mathbb{P}\bigg\{\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\| > \kappa\bigg\} \leq \tilde{K}(\kappa, x)\exp\big(-n\tilde{\beta}(\kappa, x)\big), \quad \forall n \in \mathbb{N}.$$

Then, for a.e. $x \in \mathcal{X}$, given constant $\eta > 0$, there exist positive constants $Q(\eta, x)$ and $\gamma(\eta, x)$ (depending on K, $\tilde{K}$, β, and $\tilde{\beta}$) such that

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

Proof. See Theorem 3 of Kannan et al. [17]. □

In the remainder of this note, we identify conditions under which the asymptotic and finite sample guarantees required by Theorems 1, 2, and 3 hold for the mean deviation term $\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|$. A similar analysis can be carried out for the root-mean-square deviation term $(\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|^2)^{1/2}$, which is required by [18] for the analysis of phi-divergence-based ER-DRO problems (4) for stochastic programs satisfying Assumption 1. We omit these details for brevity.

3 Guarantees for the mean deviation term

In this section, we investigate conditions under which the mean deviation term $\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\|$ converges to zero in probability at a certain rate and possesses finite sample guarantees. We begin by bounding the mean deviation in terms of the functions f^* and Q^*, their regression estimates $\hat{f}_n$ and $\hat{Q}_n$, and the data $\mathcal{D}_n$. Throughout, we implicitly assume that the estimate $\hat{Q}_n$ a.s. satisfies $\hat{Q}_n(x) \succ 0$ for a.e. $x \in \mathcal{X}$, which can be guaranteed by an appropriate choice of the model class $\mathcal{Q}$.

3.1 Bounding the mean deviation term

We begin by noting that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \|\tilde{\varepsilon}_n^i(x)\| &= \frac{1}{n}\sum_{i=1}^n \|(\hat{f}_n(x) + \hat{Q}_n(x)\hat{\varepsilon}_n^i) - (f^*(x) + Q^*(x)\varepsilon^i)\| \\ &\leq \|\hat{f}_n(x) - f^*(x)\| + \frac{1}{n}\sum_{i=1}^n \|\hat{Q}_n(x)\hat{\varepsilon}_n^i - Q^*(x)\varepsilon^i\|. \end{aligned} \tag{5}$$

We now bound the second term on the r.h.s. of inequality (5). We have

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

$$=\frac{1}{n}\sum_{i=1}^{n}\big\|\big(\hat{Q}_n(x)\big[\hat{Q}_n(x^i)\big]^{-1}-Q^*(x)\big[Q^*(x^i)\big]^{-1}\big)Q^*(x^i)\varepsilon^i\big\|+\frac{1}{n}\sum_{i=1}^{n}\big\|\hat{Q}_n(x)\big[\hat{Q}_n(x^i)\big]^{-1}(f^*(x^i)-\hat{f}_n(x^i))\big\|, \tag{6}$$

where the final step follows from the definition of $\{\varepsilon^i\}_{i=1}^n$. We have for each $i \in [n]$

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

Plugging the above equality into inequality (6), we get

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\|\hat{Q}_n(x)\hat{\varepsilon}_n^i-Q^*(x)\varepsilon^i\|\\
\leq&\frac{1}{n}\sum_{i=1}^{n}\Big(\|\hat{Q}_n(x)\|\big\|\big[\hat{Q}_n(x^i)\big]^{-1}-\big[Q^*(x^i)\big]^{-1}\big\|\|Q^*(x^i)\|+\|\hat{Q}_n(x)-Q^*(x)\|\Big)\|\varepsilon^i\|+\\
&\frac{1}{n}\sum_{i=1}^{n}\|\hat{Q}_n(x)\|\big\|\big[\hat{Q}_n(x^i)\big]^{-1}\big\|\|f^*(x^i)-\hat{f}_n(x^i)\|\\
\leq&\|\hat{Q}_n(x)\|\left(\frac{1}{n}\sum_{i=1}^{n}\big\|\big[\hat{Q}_n(x^i)\big]^{-1}-\big[Q^*(x^i)\big]^{-1}\big\|^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\|Q^*(x^i)\|^4\right)^{1/4}\left(\frac{1}{n}\sum_{i=1}^{n}\|\varepsilon^i\|^4\right)^{1/4}+ \qquad (7)\\
&\|\hat{Q}_n(x)-Q^*(x)\|\left(\frac{1}{n}\sum_{i=1}^{n}\|\varepsilon^i\|\right)+\|\hat{Q}_n(x)\|\left(\frac{1}{n}\sum_{i=1}^{n}\big\|\big[\hat{Q}_n(x^i)\big]^{-1}\big\|^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\|f^*(x^i)-\hat{f}_n(x^i)\|^2\right)^{1/2},
\end{aligned}$$

where the last step above follows by repeated application of the Cauchy-Schwarz inequality. Finally, using inequality (7) in inequality (5), we get

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

In the remainder of this section, we rely on inequality (8) to identify conditions under which the mean deviation term $\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}_n^i(x)\|$ possesses asymptotic and finite sample guarantees. We postpone the verification of these assumptions to Section 4.

Before we proceed, we mention alternative ways to bound the mean deviation term $\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}_n^i(x)\|$ that may be easier to verify in some contexts. By slightly changing some of the steps leading to inequality (7), the third term on the r.h.s. of inequality (8) can be replaced with the term

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

When the second term in the expression above possesses the requisite asymptotic and finite sample guarantees (see, e.g., [30, Section 3]), this yields an alternative form of inequality (8) that requires milder assumptions on the distribution of the errors ε. For another alternative, notice that the first term on the r.h.s. of inequality (6) can also be bounded from above as

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\big\|\big(\hat{Q}_n(x)\big[\hat{Q}_n(x^i)\big]^{-1}-Q^*(x)\big[Q^*(x^i)\big]^{-1}\big)Q^*(x^i)\varepsilon^i\big\|\\
=&\frac{1}{n}\sum_{i=1}^{n}\big\|\big(\hat{Q}_n(x)\big[\hat{Q}_n(x^i)\big]^{-1}Q^*(x^i)-Q^*(x)\big)\varepsilon^i\big\|
\end{aligned}$$

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

where the final step follows by the Cauchy-Schwarz inequality. Additionally, the second term on the r.h.s. of inequality (6) can also be bounded from above as

$$
\frac{1}{n}\sum_{i=1}^{n}\|\hat{Q}_n(x)\big[\hat{Q}_n(x^i)\big]^{-1}(f^*(x^i)-\hat{f}_n(x^i))\|\leq\|\hat{Q}_n(x)\|\Big(\sup_{\bar{x}\in\mathcal{X}}\|\big[\hat{Q}_n(\bar{x})\big]^{-1}\|\Big)\left(\frac{1}{n}\sum_{i=1}^{n}\|f^*(x^i)-\hat{f}_n(x^i)\|\right).
$$

Using these bounds in inequality (6), we conclude that the mean deviation $\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}_n^i(x)\|$ can also be bounded from above as

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}_n^i(x)\|\leq\|\hat{f}_n(x)-f^*(x)\|+\|\hat{Q}_n(x)-Q^*(x)\|\left(\frac{1}{n}\sum_{i=1}^{n}\|\varepsilon^i\|\right)+&\\
\|\hat{Q}_n(x)\|\Big(\sup_{\bar{x}\in\mathcal{X}}\|\big[\hat{Q}_n(\bar{x})\big]^{-1}\|\Big)\left(\frac{1}{n}\sum_{i=1}^{n}\|Q^*(x^i)-\hat{Q}_n(x^i)\|^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\|\varepsilon^i\|^2\right)^{1/2}+&\\
\|\hat{Q}_n(x)\|\Big(\sup_{\bar{x}\in\mathcal{X}}\|\big[\hat{Q}_n(\bar{x})\big]^{-1}\|\Big)\left(\frac{1}{n}\sum_{i=1}^{n}\|f^*(x^i)-\hat{f}_n(x^i)\|\right).& \qquad (9)
\end{aligned}
$$

Inequality (9) can be used to derive alternative conditions under which our asymptotic and finite sample guarantees hold. For instance, asymptotic and finite sample guarantees on the uniform convergence of the estimate $\hat{Q}_n$ to Q^* on $\mathcal{X}$ directly translate to the requisite asymptotic and finite sample guarantees on the quantities involving the estimate $\hat{Q}_n$ in (9). These conditions again necessitate milder assumptions on the distribution of the errors ε relative to (8); however, they require the function Q^* and its regression estimate $\hat{Q}_n$ to be (asymptotically) a.s. uniformly invertible (cf. [22]), i.e., $\sup_{\bar{x}\in\mathcal{X}}\|[Q^*(\bar{x})]^{-1}\|<+\infty$ and a.s. (for n large enough) $\sup_{\bar{x}\in\mathcal{X}}\|\big[\hat{Q}_n(\bar{x})\big]^{-1}\|<+\infty$. We omit these details for brevity and continue with inequality (8) for the rest of our analysis.

3.2 Consistency

We begin with assumptions that guarantee that the mean deviation term $\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}_n^i(x)\|$ converges to zero in probability.

Assumption 5. The function Q^* and the data $\mathcal{D}_n$ satisfy the weak LLNs

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

Assumption 6. The samples $\{\varepsilon^i\}_{i=1}^{n}$ satisfy the weak LLN $\frac{1}{n}\sum_{i=1}^{n}\|\varepsilon^i\|^4\xrightarrow{p}\mathbb{E}[\|\varepsilon\|^4]$.

Assumptions 5 and 6 are mild weak LLN assumptions that hold, for instance, when the samples $\{(x^i,\varepsilon^i)\}$ are i.i.d. and the quantities $\mathbb{E}[\|Q^*(X)\|^4]$, $\mathbb{E}\big[\|\big[Q^*(X)\big]^{-1}\|^2\big]$, and $\mathbb{E}[\|\varepsilon\|^4]$ are finite. They also hold for non-i.i.d. data arising from mixing/stationary processes that satisfy suitable assumptions (see the discussion following Assumption 3 of [17]). We also require the following consistency assumption on the regression estimates $\hat{f}_n$ and $\hat{Q}_n$ (cf. Assumption 4 of [17]).

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

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

The following result will prove useful in our analysis.

Lemma 4. We have

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

Proof. The triangle inequality for the operator norm implies

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

Therefore, the following component-wise inequality holds:

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

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

Applying Assumptions 5, 6, and 7 to inequality (8) immediately yields the following result.

Theorem 5. Suppose Assumptions 5, 6, and 7 hold. Then $\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}_n^i(x)\| \xrightarrow{p} 0$ for a.e. $x \in \mathcal{X}$.

Proof. Follows from inequality (8), Assumptions 5, 6, and 7, Lemma 4, the continuous mapping theorem, and the fact that $O_p(1)O_p(1) = O_p(1)$, $O_p(1)o_p(1) = o_p(1)$, and $o_p(1) + o_p(1) = o_p(1)$. □

3.3 Rates of convergence

We refine Assumption 7 to obtain rates of convergence (cf. Assumption 6 of [17]).

Assumption 8. There is a constant[3] $0 < r \leq 1$ such that the regression estimates $\hat{f}_n$ and $\hat{Q}_n$ satisfy the following convergence rate criteria:

$$\begin{aligned} &\|\hat{f}_n(x) - f^*(x)\| = O_p(n^{-r/2}) \quad \text{and} \quad \|\hat{Q}_n(x) - Q^*(x)\| = O_p(n^{-r/2}), \quad \text{for a.e. } x \in \mathcal{X}, \\ &\frac{1}{n}\sum_{i=1}^{n}\|\hat{f}_n(x^i) - f^*(x^i)\|^2 = O_p(n^{-r}), \quad \frac{1}{n}\sum_{i=1}^{n}\|[\hat{Q}_n(x^i)]^{-1} - [Q^*(x^i)]^{-1}\|^2 = O_p(n^{-r}). \end{aligned}$$

Inequality (8) along with Assumptions 5, 6, and 8 readily yields the following result.

Theorem 6. Suppose Assumptions 5, 6, and 8 hold. Then $\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}_n^i(x)\| = O_p(n^{-r/2})$ for a.e. $x \in \mathcal{X}$.

Proof. Follows by applying Assumptions 5, 6, and 8, Lemma 4, the continuous mapping theorem, and the fact that $O_p(1) + O_p(n^{-r/2}) = O_p(1)$, $O_p(1)O_p(n^{-r/2}) = O_p(n^{-r/2})$, and $O_p(n^{-r/2}) + O_p(n^{-r/2}) = O_p(n^{-r/2})$ to inequality (8). □

3.4 Finite sample guarantees

We make the following additional assumptions to establish a finite sample guarantee (cf. Assumption 8 of [17]).

Assumption 9. The regression estimates $\hat{f}_n$ and $\hat{Q}_n$ possess the following finite sample properties: for any constant $\kappa > 0$, there exist positive constants $K_f(\kappa, x)$, $\bar{K}_f(\kappa)$, $\beta_f(\kappa, x)$, $\bar{\beta}_f(\kappa)$, $K_Q(\kappa, x)$, $\bar{K}_Q(\kappa)$, $\beta_Q(\kappa, x)$, and $\bar{\beta}_Q(\kappa)$ such that for each $n \in \mathbb{N}$

$$\begin{aligned} \mathbb{P}\{\|f^*(x) - \hat{f}_n(x)\| > \kappa\} &\leq K_f(\kappa, x)\exp\left(-n\beta_f(\kappa, x)\right), \quad \text{for a.e. } x \in \mathcal{X}, \\ \mathbb{P}\{\|Q^*(x) - \hat{Q}_n(x)\| > \kappa\} &\leq K_Q(\kappa, x)\exp\left(-n\beta_Q(\kappa, x)\right), \quad \text{for a.e. } x \in \mathcal{X}, \end{aligned}$$

[3] The constant r is independent of n, but could depend on the covariate dimension d_x.

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

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

The next two assumptions strengthen Assumptions 5 and 6 to assume finite sample properties for the quantities involved.

Assumption 10. For any constant $\kappa > 0$, there exist positive constants $\gamma_Q(\kappa)$ and $\bar{\gamma}_Q(\kappa)$ such that for each $n \in \mathbb{N}$

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

$$\mathbb{P}\bigg\{\bigg(\frac{1}{n}\sum_{i=1}^{n}\|Q^*(x^i)\|^4\bigg)^{1/4} > \left(\mathbb{E}\left[\|Q^*(X)\|^4\right]\right)^{1/4} + \kappa\bigg\} \leq \exp(-n\bar{\gamma}_Q(\kappa)).$$

Assumption 11. For any constant $\kappa > 0$, there exist positive constants $\gamma_\varepsilon(\kappa)$ and $\bar{\gamma}_\varepsilon(\kappa)$ such that for each $n \in \mathbb{N}$

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

The first part of Assumption 10 holds, e.g., if for each $\kappa > 0$, there is a constant $\gamma_Q(\kappa) > 0$ such that

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

The function $\gamma_Q(\cdot)$ in the inequality above is related to the so-called rate function in large deviations theory (see Section 7.2.8 of [27]). Similar conclusions hold for the probability inequalities involving the terms $\frac{1}{n}\sum_{i=1}^{n}\|Q^*(x^i)\|^4$ and $\frac{1}{n}\sum_{i=1}^{n}\|\varepsilon^i\|^4$ in Assumptions 10 and 11. From large deviations theory, we can also conclude that the constants $\gamma_Q(\kappa)$, $\bar{\gamma}_Q(\kappa)$, $\gamma_\varepsilon(\kappa)$, and $\bar{\gamma}_\varepsilon(\kappa)$ in Assumptions 10 and 11 are guaranteed to exist for i.i.d. data $\mathcal{D}_n$ and for each constant $\kappa > 0$ whenever the following light-tail conditions hold: $\mathbb{E}\big[\exp(\big\|\big[Q^*(X)\big]^{-1}\big\|^p)\big] < +\infty$ for some $p > 2$, $\mathbb{E}[\exp(\|Q^*(X)\|^p)] < +\infty$ for some $p > 4$, and $\mathbb{E}[\exp(\|\varepsilon\|^p)] < +\infty$ for some $p > 4$. The discussion following Assumption 7 of [17] provides avenues for verifying Assumptions 10 and 11 for non-i.i.d. data $\mathcal{D}_n$.

We are now ready to state our finite sample guarantee.

Theorem 7. Suppose Assumptions 9, 10, and 11 hold. Then, for any constant $\kappa > 0$ and a.e. $x \in \mathcal{X}$, there exist positive constants $\tilde{K}(\kappa, x)$ and $\tilde{\beta}(\kappa, x)$ such that

$$\mathbb{P}\bigg\{\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}_n^i(x)\| > \kappa\bigg\} \leq \tilde{K}(\kappa, x)\exp(-n\tilde{\beta}(\kappa, x)).$$

Proof. Using (8) and the inequality $\mathbb{P}\{V + W > c_1 + c_2\} \leq \mathbb{P}\{V > c_1\} + \mathbb{P}\{W > c_2\}$ for any random variables V, W and constants c_1, c_2, we get

$$\mathbb{P}\bigg\{\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}_n^i(x)\| > \kappa\bigg\}$$

$$\leq \mathbb{P}\Big\{\|\hat{f}_n(x) - f^*(x)\| > \frac{\kappa}{4}\Big\} + \mathbb{P}\bigg\{\bigg(\frac{1}{n}\sum_{i=1}^{n}\|\varepsilon^i\|\bigg)\|\hat{Q}_n(x) - Q^*(x)\| > \frac{\kappa}{4}\bigg\} +$$

$$\mathbb{P}\bigg\{\|\hat{Q}_n(x)\|\bigg(\frac{1}{n}\sum_{i=1}^{n}\big\|\big[\hat{Q}_n(x^i)\big]^{-1} - \big[Q^*(x^i)\big]^{-1}\big\|^2\bigg)^{1/2}\bigg(\frac{1}{n}\sum_{i=1}^{n}\|Q^*(x^i)\|^4\bigg)^{1/4}\bigg(\frac{1}{n}\sum_{i=1}^{n}\|\varepsilon^i\|^4\bigg)^{1/4} > \frac{\kappa}{4}\bigg\} +$$

$$\mathbb{P}\Big\{\|\hat{Q}_n(x)\|\Big(\frac{1}{n}\sum_{i=1}^{n}\big\|\big[\hat{Q}_n(x^i)\big]^{-1}\big\|^2\Big)^{1/2}\Big(\frac{1}{n}\sum_{i=1}^{n}\|f^*(x^i)-\hat{f}_n(x^i)\|^2\Big)^{1/2} > \frac{\kappa}{4}\Big\}. \tag{10}$$

For a.e. $x \in \mathcal{X}$, the first term on the r.h.s. of inequality (10) can be bounded using Assumption 9 as

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

Next, consider the second term on the r.h.s. of inequality (10). We have for a.e. $x \in \mathcal{X}$

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

where the second inequality follows from Assumption 11 and the final step follows from Assumption 9.

The third term on the r.h.s. of inequality (10) can be bounded for a.e. $x \in \mathcal{X}$ as

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

where the second inequality follows from Assumptions 9, 10, and 11, the final inequality follows from Assumption 9, and

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

Finally, the fourth term on the r.h.s. of inequality (10) can be bounded for a.e. $x \in \mathcal{X}$ as

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

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

where the second inequality follows from Assumption 9 and Lemma 4, the final inequality follows from Assumptions 9 and 10 and the probability inequality stated at the beginning of this proof, and

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

Putting the above bounds together in inequality (10), we have for a.e. $x\in\mathcal{X}$

$$\begin{aligned}\mathbb{P}\bigg\{\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}_n^i(x)\|>\kappa\bigg\}\leq{}&\exp(-n\gamma_\varepsilon(\kappa))+\exp(-n\bar{\gamma}_\varepsilon(\kappa))+\exp(-n\gamma_Q(\kappa))+\exp(-n\bar{\gamma}_Q(\kappa))+\\&K_f(\tfrac{\kappa}{4},x)\exp(-n\beta_f(\tfrac{\kappa}{4},x))+\bar{K}_f(h_2(\kappa,x))\exp(-n\bar{\beta}_f(h_2(\kappa,x)))+\\&K_Q\big(\tfrac{\kappa}{4(\mathbb{E}[\|\varepsilon\|]+\kappa)},x\big)\exp\big(-n\beta_Q(\tfrac{\kappa}{4(\mathbb{E}[\|\varepsilon\|]+\kappa)},x)\big)+2K_Q(\kappa,x)\exp\left(-n\beta_Q(\kappa,x)\right)+\\&\bar{K}_Q(\kappa)\exp(-n\bar{\beta}_Q(\kappa))+\bar{K}_Q(h_1(\kappa,x))\exp(-n\bar{\beta}_Q(h_1(\kappa,x))),\end{aligned}\tag{11}$$

which then implies the desired result. □

Suppose we make the mild assumptions that the functions $\bar{K}_f(\cdot)$, $K_Q(\cdot,x)$, and $\bar{K}_Q(\cdot)$ in Assumption 9 are monotonically nonincreasing on $\mathbb{R}_+$ and the functions $\bar{\beta}_f(\cdot)$, $\beta_Q(\cdot,x)$, and $\bar{\beta}_Q(\cdot)$ therein are monotonically nondecreasing on $\mathbb{R}_+$ (cf. Appendix EC.3 of [17]). For a.e. $x\in\mathcal{X}$ and tolerance κ satisfying

$$\kappa<\min\Big\{\mathbb{E}[\|\varepsilon\|],\|Q^*(x)\|,\big(\mathbb{E}[\|Q^*(X)\|^4]\big)^{1/4},\big(\mathbb{E}\big[\big\|\big[Q^*(X)\big]^{-1}\big\|^2\big]\big)^{1/2}\Big\},$$

we can use inequality (11) to derive the bound

$$\begin{aligned}\mathbb{P}\bigg\{\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}_n^i(x)\|>\kappa\bigg\}\leq{}&\exp(-n\gamma_\varepsilon(\kappa))+\exp(-n\bar{\gamma}_\varepsilon(\kappa))+\exp(-n\gamma_Q(\kappa))+\exp(-n\bar{\gamma}_Q(\kappa))+\\&K_f(\tfrac{\kappa}{4},x)\exp(-n\beta_f(\tfrac{\kappa}{4},x))+\bar{K}_f(\bar{h}_2(\kappa,x))\exp(-n\bar{\beta}_f(\bar{h}_2(\kappa,x)))+\\&K_Q\big(\tfrac{\kappa}{8\mathbb{E}[\|\varepsilon\|]},x\big)\exp\big(-n\beta_Q(\tfrac{\kappa}{8\mathbb{E}[\|\varepsilon\|]},x)\big)+2K_Q(\kappa,x)\exp\left(-n\beta_Q(\kappa,x)\right)+\\&\bar{K}_Q(\kappa)\exp(-n\bar{\beta}_Q(\kappa))+\bar{K}_Q(\bar{h}_1(\kappa,x))\exp(-n\bar{\beta}_Q(\bar{h}_1(\kappa,x))),\end{aligned}$$

where

$$\bar{h}_1(\kappa,x):=\frac{\kappa}{32\|Q^*(x)\|\big(\mathbb{E}[\|Q^*(X)\|^4]\big)^{1/4}\big(\mathbb{E}[\|\varepsilon\|^4]\big)^{1/4}},\quad \bar{h}_2(\kappa,x):=\frac{\kappa}{24\|Q^*(x)\|\Big(\mathbb{E}\big[\big\|\big[Q^*(X)\big]^{-1}\big\|^2\big]\Big)^{1/2}}.$$

Therefore, for a.e. $x\in\mathcal{X}$ and $\kappa<\min\{\mathbb{E}[\|\varepsilon\|],\|Q^*(x)\|,\big(\mathbb{E}[\|Q^*(X)\|^4]\big)^{1/4},\big(\mathbb{E}\big[\big\|\big[Q^*(X)\big]^{-1}\big\|^2\big]\big)^{1/2}\}$

$$\mathbb{P}\bigg\{\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}_n^i(x)\|>\kappa\bigg\}\leq\tilde{K}(\kappa,x)\exp(-n\tilde{\beta}(\kappa,x)),$$

with $\tilde{K}(\kappa,x):=4+K_f(\frac{\kappa}{4},x)+\bar{K}_f(\bar{h}_2(\kappa,x))+K_Q\big(\frac{\kappa}{8\mathbb{E}[\|\varepsilon\|]},x\big)+2K_Q(\kappa,x)+\bar{K}_Q(\kappa)+\bar{K}_Q(\bar{h}_1(\kappa,x))$ and

$$\tilde{\beta}(\kappa,x):=\min\Big\{\gamma_\varepsilon(\kappa),\bar{\gamma}_\varepsilon(\kappa),\gamma_Q(\kappa),\bar{\gamma}_Q(\kappa),\beta_f(\tfrac{\kappa}{4},x),\bar{\beta}_f(\bar{h}_2(\kappa,x)),\beta_Q\big(\tfrac{\kappa}{8\mathbb{E}[\|\varepsilon\|]},x\big),\beta_Q(\kappa,x),\bar{\beta}_Q(\kappa),\bar{\beta}_Q(\bar{h}_1(\kappa,x))\Big\}.$$

Unlike the functions $h_1(\cdot,x)$ and $h_2(\cdot,x)$, the functions $\bar{h}_1(\cdot,x)$ and $\bar{h}_2(\cdot,x)$ are linear. Consequently, for small-enough tolerances $\kappa>0$ and a given risk level $\alpha\in(0,1)$, the above simplification enables an easier interpretation of the sample size n required for $\mathbb{P}\{\frac{1}{n}\sum_{i=1}^{n}\|\tilde{\varepsilon}_n^i(x)\|>\kappa\}\leq\alpha$. This can in turn enable a more interpretable estimate of the sample size n required for solutions of the ER-SAA problem (3) and the ER-DRO problem (4) to be approximately optimal to the true problem (1) with probability $1-\alpha$ (cf. Proposition 2 of [17]).

4 Some regression setups that satisfy our assumptions

In this section, we verify that Assumptions 7, 8, and 9 hold for some regression setups. We do not attempt to be exhaustive. We first discuss methods for estimating the regression function f^* and note their asymptotic and finite sample guarantees. We then list some popular models for the class of functions $\mathcal{Q}$, discuss approaches for estimating the matrix-valued function Q^*, and note their theoretical guarantees.

4.1 Estimating the regression function

We identify conditions under which the parts of Assumptions 7, 8, and 9 involving the regression estimate $\hat{f}_n$ hold for some prediction setups. Although these assumptions on $\hat{f}_n$ are the same as those in Assumptions 4, 6, and 8 of [17], we focus on regression setups that work in the heteroscedastic setting.

Ordinary least squares (OLS) regression. When the regression function f^* is linear, its OLS estimate $\hat{f}_n$ satisfies Assumptions 7 and 8 with constant $r = 1$ (see Proposition EC.3. of [17] for details). Furthermore, Theorem 11 and Remark 12 of [14] can be used to readily identify conditions under which the estimates $\hat{f}_n$ possess a finite sample guarantee like in Assumption 9. However, OLS regression does not yield an efficient estimator[4] of f^* in the heteroscedastic case [23]. An alternative to OLS regression is feasible weighted least squares (FWLS) regression [22, 23], which results in asymptotically efficient estimates when the estimate $\hat{Q}_n$ of Q^* is consistent. These asymptotic results of FWLS regression continue to hold at the expense of asymptotic efficiency even if the estimate $\hat{Q}_n$ of Q^* may be inconsistent (see, e.g., Section 3.3 of [23]).

Sparse regression methods. Proposition EC.4. of [17] lists conditions under which the ordinary Lasso regression estimate $\hat{f}_n$ satisfies Assumptions 7 and 8 with constant $r = 1$ and a finite sample guarantee like in Assumption 9. Theorem 1 of Belloni et al. [3] outlines conditions under which similar asymptotic results hold for the heteroscedasticity-adapted Lasso. Medeiros and Mendes [19] and Ziel [31] present asymptotic analyses of the adaptive Lasso for time series data. Their analyses applies to GARCH-type processes. Theorems 2 and 3 of [19] and Theorem 1 of [31] present conditions under which the estimate $\hat{f}_n$ satisfies Assumptions 7 and 8 with $r = 1$. Belloni et al. [4] present asymptotic and finite sample guarantees for the heteroscedasticity-adapted square-root Lasso. Finally, Dalalyan et al. [8] introduce a scaled heteroscedastic Dantzig selector. Theorem 5.2 therein presents large deviation bounds for both regression estimates $\hat{f}_n$ and $\hat{Q}_n$ under certain sparsity assumptions[5].

Other M-estimators. The conclusions for OLS regression carry over to more general M-estimators. In particular, Appendix EC.2 of [17] presents conditions under which Assumptions 7 and 8 continue to hold with $r = 1$. Similar to the special case of OLS regression, vanilla M-estimators may no longer be efficient—feasible weighted M-estimation is an asymptotically efficient alternative. Theorems 1, 3, and 5 of Sun et al. [28] and Theorem 2.1 of Zhou et al. [30] present large deviation results of the form Assumption 9 for adaptive Huber regression when the function f^* is linear. Remarkably, their results hold even for heavy-tailed error distributions. Finally, Schick [25] considers a semiparametric regression setup for f^* and establishes rates of convergence of weighted least squares estimates.

kNN regression. Proposition EC.5. of [17] summarizes conditions under which the kNN regression estimate $\hat{f}_n$ of f^* satisfies Assumptions 7 and 8 with constant $r = O(1)/d_x$. It also notes conditions under which $\hat{f}_n$ possesses a finite sample guarantee like in Assumption 9 (cf. Corollary 1 of [15]).

Kernel regression. Hansen [13] studies conditions under which kernel regression estimates are uniformly consistent given dependent data $\mathcal{D}_n$ satisfying mixing conditions. Theorems 1, 2, and 4 therein can be used to show that the kernel regression estimate $\hat{f}_n$ satisfies Assumptions 7 and 8 with constant $r = O(1)/d_x$. Mokkadem et al. [20] study large deviations results for some kernel regression estimates.

[4] Here, by the term *efficient estimator*, we mean a minimum variance unbiased estimator.

[5] Although [8] consider the fixed design setting, their analysis can be modified to accommodate random designs under suitable assumptions on the distribution P_X of the covariates X (see Section 4 of [8]).

4.2 Estimating the conditional covariance matrix of the errors

In this section, we identify conditions under which the parts of Assumptions 7, 8, and 9 involving the regression estimate $\hat{Q}_n$ hold for some prediction setups. These assumptions for $\hat{Q}_n$—in particular, Assumption 9—are not as well-studied in the literature as those for $\hat{f}_n$. Therefore, they are typically harder to verify than their counterparts in Section 4.1. Because deriving theoretical properties of estimators for the heteroscedastic setting and deriving finite sample properties of estimators in general are areas of topical interest, we envision that future research will enable easier verification of these assumptions.

For simplicity, we only consider function classes $\mathcal{Q}$ that comprise diagonal covariance matrices (cf. [30]), although the theoretical developments in Section 3 apply more generally. Bauwens et al. [2] review some model classes $\mathcal{Q}$ with non-diagonal covariance matrices that are popular in time series modeling.

Example 1. [Parametric Models] The model class is

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

where $q_j : \mathbb{R}^{d_x} \to \mathbb{R}_+$ for each $j \in [d_y]$. Forms of the functions q_j of interest include [21, 23]:

i. $(q_j(X))^2 = \sigma_j^2(1 + \theta_j^{\mathrm{T}} X)^2$ for parameters (σ_j, θ_j),

ii. $(q_j(X))^2 = \exp(\sigma_j + \theta_j^{\mathrm{T}} X)$ for parameters (σ_j, θ_j),

iii. $(q_j(X))^2 = \exp\bigl(\sigma_j + \theta_j^{\mathrm{T}} \log(X)\bigr)$ for parameters (σ_j, θ_j).

For the rest of the note, we absorb the parameter σ_j into θ_j for simplicity of presentation. With this notation, the above examples can be cast in the general form $(q_j(X))^2 = h_j(\theta_j^{\mathrm{T}} g_j(X))$ for known functions h_j and g_j and a parameter θ_j that is to be estimated. The above setup can also accommodate cases where the parameters of the function Q^* include some of the parameters of the function f^*.

Example 2. [Nonparametric Model] The model class is

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

where each $q_j : \mathbb{R}^{d_x} \to \mathbb{R}_+$ is assumed to be 'sufficiently smooth'. Chapter 8 of Fan and Yao [11] presents some popular models for the functions q_j in a time series context.

Suppose for ease of exposition that the covariance matrix of the errors ε is the identity matrix. Then, for each $j \in [d_y]$ and any $\bar{x} \in \mathcal{X}$, we have $\mathbb{E}\left[(Y_j - f_j^*(X))^2 \mid X = \bar{x}\right] = (q_j^*(\bar{x}))^2$ for the components $q_j^*(\bar{x})$ of $Q^*(\bar{x})$ in Examples 1 and 2. This motivates the estimation of each function q_j^* by regressing the squared residuals $(y_j^i - \hat{f}_{j,n}(x^i))^2$ on the covariate observation x^i. For the parametric setup in Example 1, this nonlinear regression problem can often be transformed into a linear regression problem. An alternative for the parametric regression setup is to estimate the parameters θ_j in $\hat{Q}_n$ concurrently with the parameters of the estimate $\hat{f}_n$ using an M-estimation procedure. Section 3 of Davidian and Carroll [9] outlines several approaches for estimating the parameters in Example 1, including the methods mentioned above. Chapter 8 of Fan and Yao [11] discusses nonparametric regression methods for estimating each function q_j.

We now outline approaches for verifying that the estimate $\hat{Q}_n$ satisfies Assumptions 7, 8, and 9. Consider first the parametric setup in Example 1. Suppose the function $Q^*(\cdot) \equiv Q(\cdot; \theta^*)$ for some function Q and the goal is to estimate the parameter θ^*. Let $\hat{\theta}_n$ denote the estimate of θ^* corresponding to the regression estimate $\hat{Q}_n$, i.e., $\hat{Q}_n(\cdot) \equiv Q(\cdot; \hat{\theta}_n)$. Suppose for a.e. realization $x \in \mathcal{X}$, the function $Q(x; \cdot)$ is Lipschitz continuous with Lipschitz constant $L_Q(x)$ and its inverse $[Q(x; \cdot)]^{-1}$ is also Lipschitz continuous with Lipschitz constant $\bar{L}_Q(x)$. These assumptions hold for the model classes in Example 1 if the parameters θ therein are restricted to lie in suitable compact sets[6]. Because

$$\|\hat{Q}_n(x) - Q^*(x)\| \leq L_Q(x)\|\hat{\theta}_n - \theta^*\| \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} \bigl\| \bigl[\hat{Q}_n(x^i)\bigr]^{-1} - \bigl[Q^*(x^i)\bigr]^{-1} \bigr\|^2 \leq \Bigl(\frac{1}{n}\sum_{i=1}^{n} \bar{L}_Q^2(x^i)\Bigr)\|\hat{\theta}_n - \theta^*\|^2,$$

[6] As noted in [17, Appendix EC.3.2.], it suffices to assume that the above Lipschitz continuity assumptions hold locally for the asymptotic results.

asymptotic and finite sample guarantees on the estimator $\hat{\theta}_n$ of θ^* directly translate to the asymptotic and finite sample guarantees on the estimate $\hat{Q}_n$ in Assumptions 7, 8, and 9. When the functions f^* and Q^* are jointly estimated using M-estimators, the results listed in Appendix EC.3.2. of [17] provide conditions under which the estimator $\hat{\theta}_n$ of θ^* is consistent and Assumptions 7 and 8 hold with $r = 1$. They also present a hard-to-verify uniform exponential bound condition under which $\hat{\theta}_n$ possesses a finite sample guarantee. Carroll and Ruppert [6] consider robust M-estimators for θ^* that possess a similar rate of convergence when f^* is linear. Dalalyan et al. [8] present asymptotic and finite sample guarantees for a scaled Dantzig estimator of θ^* under some sparsity assumptions. Finally, Fan et al. [12] present a quasi-maximum likelihood approach for estimating the parameters of GARCH models and investigate their asymptotic properties.

Next, consider the nonparametric setup in Example 2, and suppose the function Q^* and its regression estimate $\hat{Q}_n$ are (asymptotically) a.s. uniformly invertible[7]. We have

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

Therefore, asymptotic and finite sample guarantees for $\|\hat{Q}_n(x) - Q^*(x)\|$ and $\frac{1}{n}\sum_{i=1}^{n}\|\hat{Q}_n(x^i) - Q^*(x^i)\|^2$ are sufficient for verifying Assumptions 7, 8, and 9. Theorem 8.5 of Fan and Yao [11] can be used to identify conditions under which these asymptotic guarantees hold for local linear estimators on time series data when the dimension of the covariates $d_x = 1$. They also note approaches for estimating Q^* when $d_x > 1$. Theorem 2 of Ruppert et al. [24] can be used to verify Assumptions 7 and 8 for local polynomial smoothers. Proposition 2.1 and Theorem 3.1 of Jin et al. [16] identify conditions under which Assumptions 7 and 8 hold for a local likelihood estimator. Van Keilegom and Wang [29] consider semi-parametric models for both f^* and Q^*. Theorems 3.1 and 3.2 therein can be used to verify Assumptions 7 and 8 for the estimates $\hat{Q}_n$. Section 3 of Zhou et al. [30] presents robust estimators of Q^* when f^* is linear and notes that these estimators $\hat{Q}_n$ possess asymptotic and finite sample guarantees in the form of Assumptions 7, 8, and 9. Finally, Theorem 3.1 of Chesneau et al. [7] can be used to derive asymptotic guarantees for wavelet estimators of Q^*.

5 Conclusion

In this note, we propose generalizations of the ER-SAA and ER-DRO frameworks in [17, 18] that can handle heteroscedastic errors, focusing mainly on ER-SAA for brevity. We identify sufficient conditions under which solutions to these approximations possess asymptotic and finite sample guarantees for a class of two-stage stochastic MIPs with continuous recourse. Furthermore, we outline conditions under which these assumptions hold for some regression setups, including OLS, Lasso, and kNN regression.

Future work includes verification of the large deviation Assumption 9 for the regression estimate $\hat{Q}_n$ for additional prediction setups, consideration of more general relationships between the random vector Y and the random covariates X, and investigation of the computational performance of the generalizations of the ER-SAA and ER-DRO problems on a practical application involving heteroscedasticity.

Acknowledgments

We thank Prof. Erick Delage for encouraging us to investigate extensions of the ER-SAA formulation to the heteroscedastic setting. This research is supported by the Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program under Contract Number DE-AC02-06CH11357.

[7] Although inequality (9) in Section 3 can yield similar guarantees under such uniform invertibility assumptions, we stick with Assumptions 7, 8, and 9 dictated by inequality (8) for simplicity.

References

[1] G.-Y. Ban, J. Gallien, and A. J. Mersereau. Dynamic procurement of new products with covariate information: The residual tree method. *Manufacturing & Service Operations Management*, 21(4):798–815, 2019.

[2] L. Bauwens, S. Laurent, and J. V. Rombouts. Multivariate GARCH models: a survey. *Journal of Applied Econometrics*, 21(1):79–109, 2006.

[3] A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012.

[4] A. Belloni, V. Chernozhukov, and L. Wang. Pivotal estimation via square-root lasso in nonparametric regression. *The Annals of Statistics*, 42(2):757–788, 2014.

[5] D. Bertsimas and N. Kallus. From predictive to prescriptive analytics. *Management Science*, 66(3):1025–1044, 2020.

[6] R. J. Carroll and D. Ruppert. Robust estimation in heteroscedastic linear models. *The Annals of Statistics*, pages 429–441, 1982.

[7] C. Chesneau, S. El Kolei, J. Kou, and F. Navarro. Nonparametric estimation in a regression model with additive and multiplicative noise. *Journal of Computational and Applied Mathematics*, page 112971, 2020.

[8] A. Dalalyan, M. Hebiri, K. Meziani, and J. Salmon. Learning heteroscedastic models by convex programming under group sparsity. In *Proceedings of the 30th international conference on machine learning*, pages 379–387, 2013.

[9] M. Davidian and R. J. Carroll. Variance function estimation. *Journal of the American Statistical Association*, 82 (400):1079–1091, 1987.

[10] P. Donti, B. Amos, and J. Z. Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, pages 5484–5494, 2017.

[11] J. Fan and Q. Yao. *Nonlinear time series: nonparametric and parametric methods*. Springer Science & Business Media, 2008.

[12] J. Fan, L. Qi, and D. Xiu. Quasi-maximum likelihood estimation of GARCH models with heavy-tailed likelihoods. *Journal of Business & Economic Statistics*, 32(2):178–191, 2014.

[13] B. E. Hansen. Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, pages 726–748, 2008.

[14] D. Hsu, S. M. Kakade, and T. Zhang. Random design analysis of ridge regression. In *Proceedings of the 25th annual conference on learning theory*, volume 23, pages 9.1–9.24, 2012.

[15] H. Jiang. Non-asymptotic uniform rates of consistency for k-nn regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3999–4006, 2019.

[16] S. Jin, L. Su, and Z. Xiao. Adaptive nonparametric regression with conditional heteroskedasticity. *Econometric Theory*, 31(6):1153, 2015.

[17] R. Kannan, G. Bayraksan, and J. R. Luedtke. Data-driven sample average approximation with covariate information. Optimization Online. URL: `http://www.optimization-online.org/DB_HTML/2020/07/7932.html`, 2020.

[18] R. Kannan, G. Bayraksan, and J. R. Luedtke. Residuals-based distributionally robust optimization with covariate information. Optimization Online. URL: `http://www.optimization-online.org/DB_HTML/2020/11/8136.html`, 2020.

[19] M. C. Medeiros and E. F. Mendes. ℓ_1-regularization of high-dimensional time-series models with non-Gaussian and heteroskedastic errors. *Journal of Econometrics*, 191(1):255–271, 2016.

[20] A. Mokkadem, M. Pelletier, and B. Thiam. Large and moderate deviations principles for kernel estimators of the multivariate regression. *Mathematical Methods of Statistics*, 17(2):146–172, 2008.

[21] J. L. Powell. Models, testing, and correction of heteroskedasticity. Lecture notes, Department of Economics, University of California, Berkeley. URL: `https://eml.berkeley.edu/~powell/e240b_sp10/hetnotes.pdf`, 2010.

[22] P. M. Robinson. Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica: Journal of the Econometric Society*, pages 875–891, 1987.

[23] J. P. Romano and M. Wolf. Resurrecting weighted least squares. *Journal of Econometrics*, 197(1):1–19, 2017.

[24] D. Ruppert, M. P. Wand, U. Holst, and O. Hössjer. Local polynomial variance-function estimation. *Technometrics*, 39(3):262–273, 1997.

[25] A. Schick. Weighted least squares estimates in partly linear regression models. *Statistics & Probability Letters*, 27(3): 281–287, 1996.

[26] S. Sen and Y. Deng. Learning enabled optimization: Towards a fusion of statistical learning and stochastic programming. Optimization Online. URL: `http://www.optimization-online.org/DB_FILE/2017/03/5904.pdf`, 2017.

[27] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.

[28] Q. Sun, W.-X. Zhou, and J. Fan. Adaptive Huber regression. *Journal of the American Statistical Association*, 115 (529):254–265, 2020.

[29] I. Van Keilegom and L. Wang. Semiparametric modeling and estimation of heteroscedasticity in regression analysis of cross-sectional data. *Electronic Journal of Statistics*, 4:133–160, 2010.

[30] W.-X. Zhou, K. Bose, J. Fan, and H. Liu. A new perspective on robust M-estimation: Finite sample theory and applications to dependence-adjusted multiple testing. *Annals of Statistics*, 46(5):1904, 2018.

[31] F. Ziel. Iteratively reweighted adaptive Lasso for conditional heteroscedastic time series with applications to AR-ARCH type processes. *Computational Statistics & Data Analysis*, 100:773–793, 2016.