Smooth Nested Simulation: Bridging Cubic and Square Root Convergence Rates in High Dimensions

Wenjia Wang
The Hong Kong University of Science and Technology, wenjiawang@ust.hk

Yanyuan Wang, Xiaowei Zhang
Faculty of Business and Economics, The University of Hong Kong, yanyuan@connect.hku.hk, xiaoweiz@hku.hk

Nested simulation concerns estimating functionals of a conditional expectation via simulation. In this paper, we propose a new method based on kernel ridge regression to exploit the smoothness of the conditional expectation as a function of the multidimensional conditioning variable. Asymptotic analysis shows that the proposed method can effectively alleviate the curse of dimensionality on the convergence rate as the simulation budget increases, provided that the conditional expectation is sufficiently smooth. The smoothness bridges the gap between the cubic root convergence rate (that is, the optimal rate for the standard nested simulation) and the square root convergence rate (that is, the canonical rate for the standard Monte Carlo simulation). We demonstrate the performance of the proposed method via numerical examples from portfolio risk management and input uncertainty quantification.

Key words: nested simulation; smoothness; kernel ridge regression; convergence rate

1. Introduction

Many simulation applications involve estimating a functional of a conditional expectation. The functional can be in the form of an expectation or a quantile. Because in general, no analytical expression is available for either the conditional expectation or the functional, the estimation requires two levels of simulation. That is, one first simulates in the *outer* level the random variable being conditioned on and then simulates in the *inner* level some other random object of interest conditioning on each simulated sample—also known as *scenario*—of the former. This class of problems is referred to as *nested simulation*. Specifically, in the present paper we examine the problem of using simulation to estimate quantities of the form

$$\theta = \mathcal{T}(\mathbb{E}[Y|X]), \tag{1}$$

where X is a $\mathbb{R}^d$-valued random variable with $d \geq 1$, Y is a $\mathbb{R}$-valued random variable, and $\mathcal{T}$ is a functional that maps a probability distribution to a real number.

Two representative examples of quantities of the form (1) stem from financial risk management (Lan et al. 2010, Hong et al. 2017, Dang et al. 2020) and input uncertainty quantification for stochastic simulation (Barton 2012, Barton et al. 2014, Xie et al. 2014).

Example 1 (Portfolio Risk Management). A risk manager is interested in assessing the risk of a portfolio of securities at some future time T_0 known as the risk horizon. The current value of the portfolio, V_0, is known to the risk manager. However, its value at the risk horizon, V_{T_0}, is a random variable that depends on X, a collection of risk factors such as interest rates, equity prices, and commodity prices the values of which are realized between time 0 and time T_0. Moreover, it can usually be expressed as a conditional expectation under a "risk-neutral measure": $V_{T_0}(X) = \mathbb{E}[W|X]$, where W is the discounted cash flow between time T_0 and some final horizon T (e.g., the expiration date of the derivatives). When the portfolio includes complex financial derivatives, as is often the case, $V_{T_0}(X)$ does not possess an analytical form, and its evaluation relies on Monte Carlo simulation.

Suppose that the portfolio does not generate interim cash flows prior to time T_0 and that the risk-free rate is r. The loss of the portfolio at the risk horizon in scenario X is then $Z = V_0 - V_{T_0}(X)$, which can be written as $Z = \mathbb{E}[Y|X]$ if we let $Y = V_0 - W$. The risk can be assessed in various ways, such as probability of a large loss $\mathbb{P}(Z \geq z_0)$, expected excess loss $\mathbb{E}[\max(Z - z_0, 0)]$, squared tracking error $\mathbb{E}[(Z - z_0)^2]$ for some threshold or target z_0, value-at-risk (VaR), or conditional value-at-risk (CVaR) of Z at some risk level τ. In the first three cases, the functional $\mathcal{T}$ in (1) is in the form of $\mathcal{T}(Z) = \mathbb{E}[\eta(Z)]$ for some function η, whereas in the last two cases, $\mathcal{T}$ represents VaR or CVaR. Nested simulation is needed to compute these quantities; one first simulates realizations of X and then—conditional on each realization—evaluates $V_{T_0}(X)$ via simulation. □

Example 2 (Input Uncertainty Quantification). A decision-maker uses simulation to estimate the performance of a complex service system (e.g., health care facilities or ride-sharing platforms) that is driven by a random input process (e.g., the arrival of patients/customers/drivers). Suppose that the distribution of the input process is parameterized by some parameter X (e.g., the arrival rates for different times of day of a non-homogeneous Poisson process). Suppose also that the performance measure of interest can be expressed as $\mathbb{E}[Y|X]$ (e.g., the mean waiting time or the order fulfillment rate), where the expectation is taken with respect to the input distribution given X. However, in general, X is not known and must be estimated from a sample of the input distribution. This results in the issue of input uncertainty—the uncertainty about X—and it often has a substantial impact on the accuracy of the estimated performance of the system.

To quantify the impact of input uncertainty on simulation outputs, one may adopt the method of Bayesian model averaging (Chick 2001, 2006) and compute $\mathbb{E}_{X\sim\mathsf{P}}[\mathbb{E}[Y|X]]$, where P is the posterior distribution of X given the sample of the input distribution. In this case, the functional $\mathcal{T}$ in (1) is in the form of $\mathcal{T}(Z) = \mathbb{E}[Z]$. One may also construct a 90% credible interval (l, u) for $\mathbb{E}_{X\sim\mathsf{P}}[\mathbb{E}[Y|X]]$, where l and u are, respectively, the 5% and 95% quantiles of $\mathbb{E}[Y|X]$ under $X \sim \mathsf{P}$. In this case, $\mathcal{T}$ is in the form of a quantile (or equivalently, VaR). One may use nested simulation to compute these quantities (Xie et al. 2014, Andradóttir and Glynn 2016). □

In addition to the preceding examples, there is a connection between nested simulation and conditional Monte Carlo, a general technique for variance reduction (Asmussen and Glynn 2007, Chapter 5). Any expectation $\mathbb{E}[Y]$ can be written as $\mathbb{E}[\mathbb{E}[Y|X]]$, and $\mathbb{V}\mathrm{ar}[\mathbb{E}[Y|X]] \leq \mathbb{V}\mathrm{ar}[Y]$ due to the law of total variance. Therefore, $\mathbb{E}[Y|X]$ is an unbiased estimator having a lower variance than Y. It is usually used when X is strongly correlated with Y and the conditional expectation can be computed exactly or estimated efficiently. Conditional Monte Carlo can also be used as a smoothing technique for gradient estimation (Fu and Hu 1997, Fu et al. 2009).

A central question to address in nested simulation concerns allocation of the simulation budget in terms of how many outer-level scenarios to simulate and how many inner-level samples to simulate for each outer-level scenario. In the present paper, we focus on *uniform sampling*, a standard treatment in the literature (Gordy and Juneja 2010, Broadie et al. 2015, Andradóttir and Glynn 2016, Zhu et al. 2020). That is, an equal number of inner-level samples are used for each outer-level scenario. The structural simplicity makes it easily parallelizable to leverage modern computing platforms (Lan 2010).

Given a simulation budget Γ, it can be shown under general conditions that to minimize the root mean squared error (RMSE) the asymptotically optimal outer-level sample size n should grow at a rate of $\Gamma^{2/3}$ (and therefore the inner-level sample size m for each outer-level scenario should grow at a rate of $\Gamma^{1/3}$). Under this rule of allocating the simulation budget, the convergence rate of the standard nested simulation is $\Gamma^{-1/3}$ (Gordy and Juneja 2010, Zhang et al. 2021). This cubic root convergence is markedly slower than the square root convergence of a typical Monte Carlo simulation for estimating an expectation without the conditioning. The deterioration in convergence rate is caused by the outer-level estimation bias, which is introduced by the nonlinear transformation $\mathcal{T}$ taking effect on the error associated with using the inner-level simulation to estimate the conditional expectation.

To reduce the outer-level bias—without increasing the simulation budget—one may seek to utilize the inner-level samples in a more efficient manner based on the following simple insight. In the standard nested simulation, the inner-level samples that are simulated for an outer-level scenario $\mathbf{x}_i$ are used *only* for estimating $\mathbb{E}[Y|X = \mathbf{x}_i]$, the conditional expectation for that scenario. These inner-level samples, however, may carry information about the conditional expectation associated with another outer-level scenario $\mathbf{x}_j$ if we anticipate $f(\mathbf{x}) := \mathbb{E}[Y|X = \mathbf{x}]$ to be *smooth* with respect to $\mathbf{x}$. The standard nested simulation precludes an exchange of information between different outer-level scenarios. Instead, we may treat the estimation of the conditional expectation via the inner-level simulation as a machine-learning task. When predicting $f(\mathbf{x})$ for any $\mathbf{x}$, this perspective allows us to benefit from the inner-level samples associated with *all* the simulated outer-level scenarios, even if $\mathbf{x}$ itself is not one of them.

1.1. Main Contributions

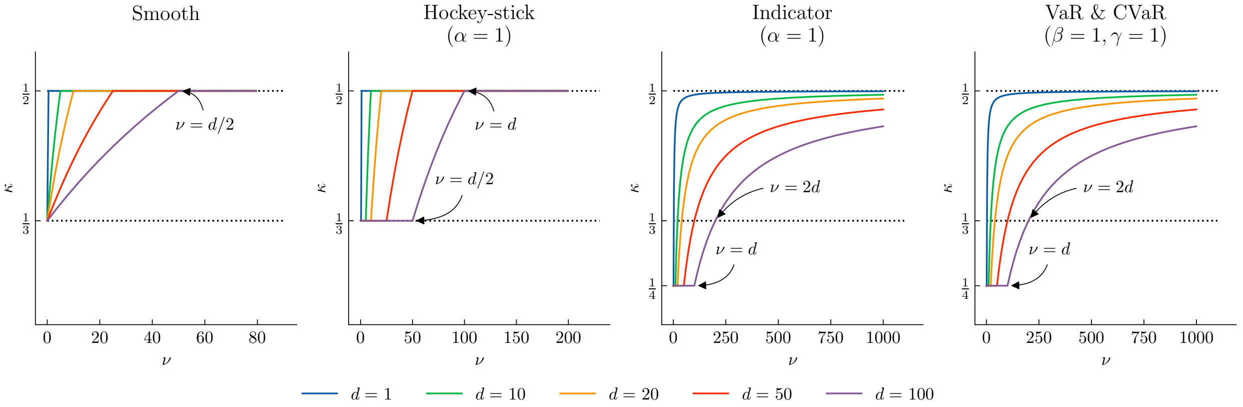
Our first contribution is to propose a new method for nested simulation. The new method employs kernel ridge regression (KRR, Kanagawa et al. 2018), a popular machine-learning method, to estimate f. Given inner-level samples, KRR seeks the best function in the reproducing kernel Hilbert space (RKHS) of one's choosing via regularized least squares in a way similar to ridge regression (Hastie 2020); hence the name. A notable feature of KRR is that it allows one to easily leverage the smoothness (i.e., the degree of differentiability) of f, which is essential for improving the convergence rate. For implementation of the KRR-driven method, we also develop a new $\mathcal{T}$-dependent cross-validation technique for hyperparameter selection. The new technique significantly outperforms the standard cross-validation, and it may be interesting in its own right.

Our second contribution is to analyze the asymptotic properties of the KRR-driven estimator $\hat{\theta}_{n,m}$ for nested simulation, including its convergence rate and the corresponding budget allocation rule (n,m). These properties demonstrate that the use of KRR in nested simulation bridges the gap between the cubic root and square root convergence rates. Specifically, we establish upper bounds on $|\hat{\theta}_{n,m} - \theta|$ and identify the growth rates of m and n as the budget Γ increases. Because the conditioning variable X is high-dimensional (i.e., d is large) in many applications, we examine the curse of dimensionality on the performance of the proposed method. These upper bounds on the convergence rate reveal a mitigating effect of the smoothness on the curse of dimensionality. For any fixed d, the gap between the cubic root and square root convergence rates diminishes gradually as the smoothness increases, and the convergence rate of the KRR-driven nested simulation recovers (or at least approaches) $\Gamma^{-1/2}$. However, if the smoothness of f is relatively low, the use of KRR may have a detrimental effect; therefore, using the standard nested simulation might be a better option. See Figure 1 for an illustration.

The theoretical framework that we develop in this paper is general. Built upon empirical process theory (van de Geer 2000), the framework permits analysis of a variety of forms of the functional $\mathcal{T}$, including not only the expectation of a function but risk measures such as VaR and CVaR. In particular, the analysis for estimating VaR/CVaR using machine learning-driven nested simulation has not been available in the literature. In addition, the framework can potentially be adopted to study the use of other machine-learning methods for estimating f in nested simulation.

Our third contribution is that we conduct extensive numerical experiments to assess the performance of the KRR-driven method, using examples from both portfolio risk management and

Figure 1 **The Parameter κ in $|\hat{\theta}_{n,m} - \theta| = O_{\mathbb{P}}(\Gamma^{-\kappa}(\log\Gamma)^{\tilde{\kappa}})$.**



Note. The four charts correspond to the nested simulation of different forms (see Section 2 for details). The first three charts correspond to the cases that $\mathcal{T}(\cdot) = \mathbb{E}[\eta(\cdot)]$ with η being a smooth (twice-differentiable) function, a hockey-stick function, and an indicator function, respectively. The last chart corresponds to the case that $\mathcal{T}(\cdot) = \mathsf{VaR}_\tau(\cdot)$ or $\mathcal{T}(\cdot) = \mathsf{CVaR}_\tau(\cdot)$ for some risk level $\tau \in (0,1)$. The parameter ν (see Section 3 for its definition) determines the smoothness of f, and (α, β, γ) are the parameters involved in technical conditions, and their typical values are 1. The results hold when the budget allocation rule (n, m) and the regularization parameter of KRR are properly specified (see Theorems 1–4 in Section 4.)

input uncertainty quantification. The dimensionality of the conditioning variables involved in these examples is as high as 100. These experiments complement our theoretical analysis and demonstrate that the proposed method is indeed a viable option for nested simulation.

1.2. Related Works

The literature on nested simulation has been growing quickly in recent years (Lee and Glynn 2003, Lesnevski et al. 2007, Gordy and Juneja 2010, Sun et al. 2011, Broadie et al. 2011, 2015, Andradóttir and Glynn 2016, Zhu et al. 2020, Zhang et al. 2021, Feng and Song 2021). The study most relevant to ours is that of Hong et al. (2017). They examine the use of kernel smoothing[1] in nested simulation and particularly the curse of dimensionality on the convergence rate, which is shown to be $\Gamma^{-\min\left(\frac{1}{2}, \frac{2}{2+d}\right)}$. This suggests that the use of kernel smoothing is beneficial to nested simulation only for low-dimensional problems, whereas in high dimensions ($d \geq 5$) it becomes detrimental, yielding a convergence rate even slower than $\Gamma^{-1/3}$ (i.e., the rate of the standard nested simulation). A root cause for the severe curse of dimensionality is that kernel smoothing is unable to track derivatives of f of an order higher than two, even if they exist. In contrast, KRR does not suffer from this issue[2]. We show that the curse of dimensionality can be greatly mitigated by

[1] The notion of "kernel" in kernel smoothing should not be confused with that in kernel ridge regression. We refer to Berlinet and Thomas-Agnan (2004, Chapter 3) for a discussion on their differences.

[2] In addition to KRR, another machine-learning method that can exploit the smoothness property is local polynomial regression. It generalizes kernel smoothing and approximates $f(\mathbf{x})$ locally with a polynomial in $\mathbf{x}$, rather than a constant

the smoothness of f. Regardless of the value of d, the convergence rate of the KRR-driven nested simulation may recover or get arbitrarily close to $\Gamma^{-1/2}$, provided that f is sufficiently smooth, which is reasonably the case for typical applications of nested simulation, as demonstrated in our numerical experiments.

The convergence rate of KRR has been extensively studied under various assumptions in machine-learning literature (van de Geer 2000, Caponnetto and De Vito 2007, Steinwart et al. 2009, Zhang et al. 2015, Tuo et al. 2020). However, nested simulation presents a unique setting that differs substantially from typical machine-learning tasks in two aspects. First, the objective is different. Whereas KRR is generally used to estimate the unknown function f, we aim to estimate $\theta = \mathcal{T}(f(X))$. The presence of the nonlinear functional $\mathcal{T}$ changes the relative importance of bias and variance in the estimation of f. The estimation error is measured differently when taking $\mathcal{T}$ into account. Simply plugging the known results of KRR into the nested simulation setting does not yield adequate performance. Second, in nested simulation we study budget allocation rules, and the number of observations of $f(\mathbf{x}_i)$ (i.e., inner-level samples) for each outer-level scenario $\mathbf{x}_i$ is a key decision variable. In contrast, in typical KRR settings no repeated observations f are allowed at the same location. Hence, in order to improve the convergence rate of the KRR-driven nested simulation, we need to *jointly* select both the regularization parameter of KRR—which itself is critical to the convergence rate of KRR—and the inner-level sample size per outer-level scenario. The existence of these two differences significantly complicates our theoretical analysis, and we develop new technical results to cope with the complication.

1.3. Notation and Organization

Throughout the paper, we use the following notation. For two positive sequences a_n and b_n, we write $a_n = O(b_n)$ and $a_n \asymp b_n$ if there exist some constants $C, C' > 0$ such that $a_n \le C b_n$ and $C' \le a_n/b_n \le C$, respectively, for all n large enough. Moreover, $a_n = O_{\mathbb{P}}(b_n)$ means that for any $\varepsilon > 0$, there exists $C > 0$ such that $\mathbb{P}(a_n > C b_n) < \varepsilon$ for all n large enough. We also write $a \wedge b$ to denote $\min(a, b)$. For a vector v, which is treated as a column vector by default, we use $v^\intercal$ and $\|v\| := \sqrt{v^\intercal v}$ to denote its transpose and its Euclidean norm, respectively.

The remainder of the paper is organized as follows. In Section 2, we introduce the background of nested simulation and formulate the research question. In Section 3, we propose the KRR-driven method. In Section 4, we analyze its asymptotic properties. In Section 5, we propose the $\mathcal{T}$-cross-validation technique for hyperparameter selection. In Section 6, we conduct numerical experiments to assess the performance of the proposed method. In Section 7, we conclude the paper. Technical results and proofs are presented in the Appendix and the e-companion to this paper.

as kernel smoothing does. The order of the polynomial should be set in accordance with the degree of differentiability of f (see Györfi et al. 2002, Chapter 5). However, unlike KRR, which is nonparametric and parsimonious, it can be challenging to fit a local polynomial regression model in multiple dimensions because it takes an enormous number of parameters to represent a high-order multivariate polynomial.

2. Problem Formulation

Motivated by typical applications of nested simulation (see Examples 1 and 2), in this paper we consider the following forms of the functional $\mathcal{T}$ in equation (1):

(i) $\mathcal{T}(Z) = \mathbb{E}[\eta(Z)]$ for some function $\eta : \mathbb{R}^d \mapsto \mathbb{R}$.

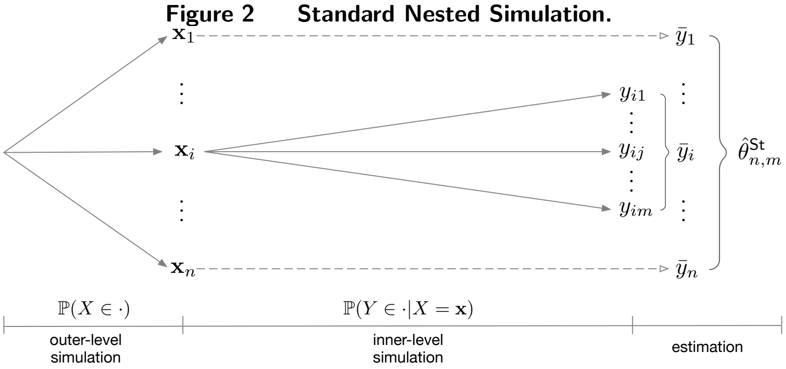
(ii) $\mathcal{T}(Z) = \mathsf{VaR}_\tau(Z) \coloneqq \inf\{z \in \mathbb{R} : \mathbb{P}(Z \le z) \ge \tau\}$, that is, the VaR of Z at level $\tau \in (0,1)$.

(iii) $\mathcal{T}(Z) = \mathsf{CVaR}_\tau(Z) \coloneqq \mathbb{E}[Z|Z > \mathsf{VaR}_\tau(Z)]$, that is, the CVaR of Z at level $\tau \in (0,1)$.

2.1. Standard Nested Simulation

The standard nested simulation involves the following two steps to generate data. (i) Generate n independent and identically distributed (i.i.d.) outer-level scenarios $\{\mathbf{x}_i : i = 1, \ldots, n\}$. (ii) For each $\mathbf{x}_i$, generate m i.i.d. inner-level samples from the conditional distribution of Y given $X = \mathbf{x}_i$, denoted by $\{y_{ij} : j = 1, \ldots, m\}$. Then, θ can be estimated based on the averages $\bar{y}_i$, where $\bar{y}_i \coloneqq \frac{1}{m}\sum_{j=1}^m y_{ij}$ (see Figure 2).

Figure 2 Standard Nested Simulation.



Specifically, let $\bar{y}_{(1)} \le \cdots \le \bar{y}_{(n)}$ denote the order statistics of $\{\bar{y}_1, \ldots, \bar{y}_n\}$. Then, $\bar{y}_{(\lceil \tau n \rceil)}$ is the sample quantile[3] at level τ, where $\lceil z \rceil$ denotes the least integer greater than or equal to z.

The standard nested simulation estimates θ via

$$\hat{\theta}^{\mathsf{St}}_{n,m} \coloneqq \begin{cases} n^{-1}\sum_{i=1}^n \eta(\bar{y}_i), & \text{if } \mathcal{T}(\cdot) = \mathbb{E}[\eta(\cdot)], \\ \bar{y}_{(\lceil \tau n \rceil)}, & \text{if } \mathcal{T}(\cdot) = \mathsf{VaR}_\tau(\cdot), \\ \bar{y}_{(\lceil \tau n \rceil)} + (1-\tau)^{-1} n^{-1} \sum_{i=1}^n (\bar{y}_i - \bar{y}_{(\lceil \tau n \rceil)})^+, & \text{if } \mathcal{T}(\cdot) = \mathsf{CVaR}_\tau(\cdot), \end{cases} \tag{2}$$

where $(z)^+ = \max(z, 0)$. The estimator for the case of CVaR is valid because it can be shown (Hong et al. 2014) that

$$\mathsf{CVaR}_\tau(Z) = \mathsf{VaR}_\tau(Z) + (1-\tau)^{-1}\,\mathbb{E}[(Z - \mathsf{VaR}_\tau(Z))^+].$$

A central problem for nested simulation is budget allocation. Given a simulation budget Γ, what are the optimal values for n and m in order to minimize the error in estimating θ? For

[3] One may use more sophisticated quantile estimators via variance reduction techniques, such as importance sampling (Glasserman et al. 2000, Jin et al. 2003). However, the analysis of these extensions in the context of nested simulation is beyond the scope of this paper.

typical applications, the simulation time required to generate one inner-level sample is substantially greater—often by orders of magnitude—than that required to generate one outer-level scenario. Therefore, it is usually assumed in the literature that the latter is negligible relative to the former. We also adopt this setup and suppose, without loss of generality, that $\Gamma = nm$.

It is shown in Gordy and Juneja (2010) and Zhang et al. (2021) that to minimize the asymptotic RMSE of $\hat{\theta}^{\mathsf{St}}_{n,m}$ in the standard nested simulation, the simulation budget should be allocated in such way that $n \asymp \Gamma^{2/3}$ and $m \asymp \Gamma^{1/3}$ as $\Gamma \to \infty$, in which case

$$\mathrm{RMSE}[\hat{\theta}^{\mathsf{St}}_{n,m}] = \Big(\mathbb{E}\big[(\hat{\theta}^{\mathsf{St}}_{n,m} - \theta)^2\big]\Big)^{1/2} \asymp \Gamma^{-1/3}.$$

Therefore, to achieve an RMSE of size ϵ, the simulation budget needs to grow like $O(\epsilon^{-3})$. This stands in clear contrast to the standard Monte Carlo simulation for estimating unconditional expectations, for which the RMSE diminishes at a rate of $\Gamma^{-1/2}$ and thus the corresponding sample complexity is $O(\epsilon^{-2})$.

The deterioration from the square root rate of convergence to the cubic root rate is caused by the presence of the additional outer-level simulation. In the standard nested simulation, despite the absence of the inner-level estimation bias ($\mathbb{E}[Y|X = \mathbf{x}_i]$ is estimated via $\bar{y}_i$ for each $\mathbf{x}_i$), the nonlinear transform $\mathcal{T}$ that takes effect in the outer level indeed introduces bias in estimating $\mathcal{T}(\mathbb{E}[Y|X])$. It may also exacerbate the impact of the inner-level estimation variance. Both these complications demand more simulation samples to be overcome and thus, worsen the convergence rate.

2.2. Strategies for Enhancement

Our goal in the present paper is to accelerate nested simulation to achieve the square root convergence rate—the canonical rate of Monte Carlo simulation. To fulfill the goal, we rely on two strategies. The first strategy is to view the purpose of the inner-level simulation to be estimating $f(\cdot) = \mathbb{E}[Y|X = \cdot]$ as a whole instead of estimating $f(\mathbf{x}_i)$ separately for each $\mathbf{x}_i$. This subtle change in viewpoint reveals that the inner-level estimation is essentially a nonparametric regression problem, open for various machine-learning methods to take on the task. It further implies that we can and should estimate $f(\mathbf{x}_i)$ using not only the inner-level samples specific to the outer-level scenario $\mathbf{x}_i$ but those from *all* scenarios. The use of machine learning in the inner level also presents us with an opportunity to improve the bias–variance trade-off—specific to the functional $\mathcal{T}$—in the outer level.

The second strategy is to leverage structural information about f to alleviate the curse of dimensionality on the convergence rate, which is common in nonparametric regression and arises when the conditioning variable X is high-dimensional. Specifically, we assume that the *smoothness*—that is, the degree of differentiability—of f is known. Knowing the smoothness allows us to properly postulate a function space within which we search for the target f. In particular, the function space induced by the smoothness property is a subspace of the $\mathcal{L}_2$ space (i.e., the set of all square-integrable functions); and the higher the smoothness, the smaller the induced function space. Therefore, knowing the smoothness may prevent us from searching an unnecessarily vast function space, which would incur higher sample complexity. Suppose, for example, f is twice-differentiable. We may then devise a machine-learning method to find the best sample-based approximation to f within—instead of $\mathcal{L}_2$—the set of twice-differentiable functions. Nevertheless, if we are unaware of the smoothness information or use a method unable to take advantage of it, we are essentially seeking "a needle in a bigger haystack."

Our choosing to exploit the smoothness of f to accelerate nested simulation is motivated by both practical and theoretical considerations. Indeed, in typical applications of nested simulation, f usually has high-order derivatives. Consider, for example, a financial portfolio that consists of a number of options that are written on some assets. Let $f(\mathbf{x})$ be the sum of the values of these options if the prices of the underlying assets are $\mathbf{x}$ at the risk horizon. Then, f is at least twice-differentiable under typical asset pricing models, such as the Black–Scholes model or the Heston model (see Glasserman 2003, Chapter 7 for details). Meanwhile, from a theoretical viewpoint, conditions regarding smoothness may be more general and easier to impose than other structural properties, such as convexity and additivity. The above considerations lead to our use of KRR. Its strong theoretical underpinnings that are built upon the smoothness property will facilitate our asymptotic analysis when it is used in nested simulation.

3. A Kernel Ridge Regression Approach

KRR seeks a sample-based approximation to f in an RKHS, which is constructed as follows. We first specify a positive definite kernel $k:\Omega\times\Omega\mapsto\mathbb{R}$, where $\Omega\subset\mathbb{R}^d$ is the domain of the conditioning variable X. We then define the space $\mathcal{N}_k^0(\Omega)$ of all functions of the form $x\mapsto\sum_{i=1}^n\beta_i k(\cdot,\mathbf{x}_i)$ for some $n\geq 1$, $\beta_1,\ldots,\beta_n\in\mathbb{R}$, and $\mathbf{x}_1,\ldots,\mathbf{x}_n\in\Omega$ and equip $\mathcal{N}_k^0(\Omega)$ with the inner product defined as

$$\left\langle\sum_{i=1}^n\beta_i k(\cdot,\mathbf{x}_i),\sum_{j=1}^{\tilde{n}}\tilde{\beta}_j k(\cdot,\tilde{\mathbf{x}}_i)\right\rangle_{\mathcal{N}_k^0(\Omega)}:=\sum_{i=1}^n\sum_{j=1}^{\tilde{n}}\beta_i\tilde{\beta}_j k(\mathbf{x}_i,\tilde{\mathbf{x}}_j).$$

The norm of $\mathcal{N}_k^0(\Omega)$ is induced by the inner product, that is, $\|g\|^2_{\mathcal{N}_k^0(\Omega)}:=\langle g,g\rangle_{\mathcal{N}_k^0(\Omega)}$ for all $g\in\mathcal{N}_k^0(\Omega)$. Finally, the RKHS induced by k, denoted by $\mathcal{N}_k(\Omega)$, is defined as the closure of $\mathcal{N}_k^0(\Omega)$ with respect to the norm $\|\cdot\|_{\mathcal{N}_k^0(\Omega)}$. See Berlinet and Thomas-Agnan (2004) for a thorough exposition on RKHSs.

3.1. RKHSs as Spaces of Smooth Functions

The choice of kernel k represents one's knowledge about the unknown function f. For example, if k is the linear kernel, then $\mathcal{N}_k(\Omega)$ is the space of all linear functions (see Berlinet and Thomas-Agnan 2004, Chapter 7 for more details). In the present paper, we consider the Matérn class of kernels because the RKHSs that they induce consist of functions that possess a prescribed smoothness property, determined by a parameter $\nu > 0$.

The Matérn kernel of smoothness ν is defined as $k(\mathbf{x}, \tilde{\mathbf{x}}) = \Psi(\mathbf{x} - \tilde{\mathbf{x}})$, where

$$\Psi(\mathbf{x}) := \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(\frac{\sqrt{2\nu}\|\mathbf{x}\|}{\ell} \right)^{\nu} \mathsf{K}_\nu\left(\frac{\sqrt{2\nu}\|\mathbf{x}\|}{\ell} \right), \quad \forall \mathbf{x} \in \mathbb{R}^d, \tag{3}$$

where $\ell > 0$, $\Gamma(\cdot)$ is the gamma function, and $\mathsf{K}_\nu(\cdot)$ is the modified Bessel function of the second kind of order ν. In practice, ν is often taken as a half-integer, that is, $\nu = p + 1/2$ for some nonnegative integer p. In this case, $\Psi(\mathbf{x})$ can be expressed in terms of elementary functions:

$$\Psi(\mathbf{x}) = \exp\left(\frac{-\sqrt{2p+1}\|\mathbf{x}\|}{\ell} \right) \frac{p!}{(2p)!} \sum_{i=0}^{p} \frac{(p+i)!}{i!(p-i)!} \left(\frac{2\sqrt{2p+1}\|\mathbf{x}\|}{\ell} \right)^{p-i}, \quad p = 0, 1, 2, \ldots$$

See Rasmussen and Williams (2006, page 85). For instance,

$$\Psi(\mathbf{x}) = \begin{cases} \exp\left(\frac{-\|\mathbf{x}\|}{\ell} \right), & \text{if } \nu = 1/2, \\ \left(1 + \frac{\sqrt{3}\|\mathbf{x}\|}{\ell} \right) \exp\left(-\frac{\sqrt{3}\|\mathbf{x}\|}{\ell} \right), & \text{if } \nu = 3/2. \end{cases}$$

We denote the corresponding RKHS by $\mathcal{N}_\Psi(\Omega)$. Its smoothness property is reflected by the fact that $\mathcal{N}_\Psi(\Omega)$ is *norm-equivalent*[4] to the *Sobolev space* of order $s = \nu + d/2$, denoted by $\mathcal{H}^s(\Omega)$. If s is an integer[5], then $\mathcal{H}^s(\Omega)$ is defined as

$$\mathcal{H}^s(\Omega) := \left\{ g \in \mathcal{L}_2(\Omega) : \|g\|^2_{\mathcal{H}^s(\Omega)} := \sum_{|\boldsymbol{\alpha}| \le s} \|\partial^{\boldsymbol{\alpha}} g\|^2_{\mathcal{L}_2(\Omega)} < \infty \right\},$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$ is a multi-index, $|\boldsymbol{\alpha}| = \sum_{j=1}^d \alpha_j$, and $\partial^{\boldsymbol{\alpha}} g$ denotes the *weak*[6] partial derivative $\partial^{\boldsymbol{\alpha}} g = \frac{\partial^{|\boldsymbol{\alpha}|} g}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}$. Hence, if we assume $f \in \mathcal{N}_\Psi(\Omega)$, we effectively assume that f is square-integrable and is weakly differentiable up to order $\nu + d/2$.

3.2. KRR-driven Nested Simulation

Suppose that the unknown function $f \in \mathcal{N}_\Psi(\Omega)$ and that the inner-level samples satisfy

$$y_{ij} = f(\mathbf{x}_i) + \epsilon_{ij}, \quad i = 1, \ldots, n, \ j = 1, \ldots, m, \tag{4}$$

where ϵ_{ij}'s are independent zero-mean random variables that may not be identically distributed.

[4] The norm equivalence means that $\mathcal{N}_\Psi(\Omega) = \mathcal{H}^s(\Omega)$ as a set of functions, and there exist some positive constants C_1, C_2 such that $C_1\|g\|_{\mathcal{H}^s(\Omega)} \le \|g\|_{\mathcal{N}_\Psi(\Omega)} \le C_2\|g\|_{\mathcal{H}^s(\Omega)}$ for all $g \in \mathcal{N}_\Psi(\Omega)$. See Kanagawa et al. (2018) for details.

[5] See Adams and Fournier (2003, Chapter 7) for the definition of Sobolev spaces of a fractional order.

[6] The weak differentiability should not be confused with the classic notion of differentiability (see Adams and Fournier 2003, page 19).

KRR estimates f via regularized least squares of the following form:

$$\min_{g\in\mathcal{N}_\Psi(\Omega)} \frac{1}{n}\sum_{i=1}^{n}\left(\bar{y}_i - g(\mathbf{x}_i)\right)^2 + \lambda\|g\|^2_{\mathcal{N}_\Psi(\Omega)}, \tag{5}$$

where $\lambda > 0$ is the regularization parameter, which controls the penalty imposed to the "model complexity" of a candidate solution to avoid overfitting. The optimal solution to (5) is

$$\begin{aligned}\hat{f} &:= \operatorname*{arg\,min}_{g\in\mathcal{N}_\Psi(\Omega)} \left(\frac{1}{n}\sum_{i=1}^{n}(\bar{y}_i - g(\mathbf{x}_i))^2 + \lambda\|g\|^2_{\mathcal{N}_\Psi(\Omega)}\right)\\ &= \mathbf{r}(\mathbf{x})^\intercal(\mathbf{R} + n\lambda\mathbf{I})^{-1}\bar{\mathbf{y}},\end{aligned} \tag{6}$$

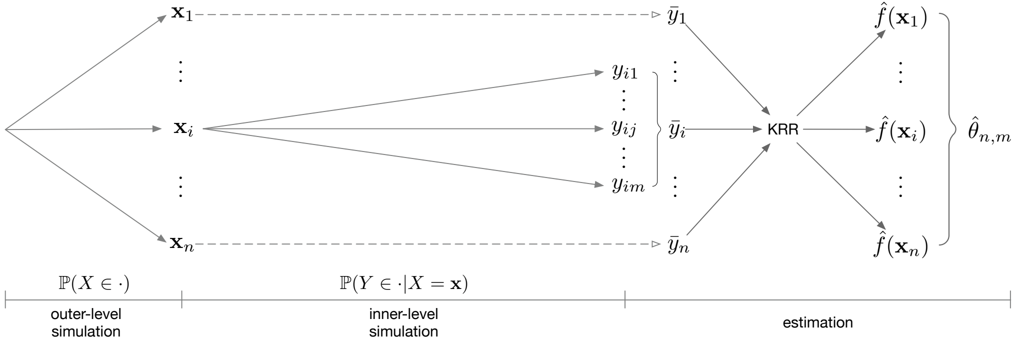
where $\mathbf{r}(\mathbf{x}) = (\Psi(\mathbf{x}-\mathbf{x}_1),\ldots,\Psi(\mathbf{x}-\mathbf{x}_n))^\intercal$, $\mathbf{R} = (\Psi(\mathbf{x}_i - \mathbf{x}_j))_{i,j=1}^n \in \mathbb{R}^{n\times n}$, and the second identity follows from the representer theorem[7] (Schölkopf and Smola 2002, Section 4.2).

After computing the KRR estimator $\hat{f}$, we let $\hat{f}_{(\lceil\tau n\rceil)}$ denote the $\lceil\tau n\rceil$-th order statistic of $\{\hat{f}(\mathbf{x}_1),\ldots,\hat{f}(\mathbf{x}_n)\}$, where $\lceil z\rceil$ denotes the least integer greater than or equal to z. Formally, we propose the KRR-driven nested simulation estimator for θ as follows:

$$\hat{\theta}_{n,m} := \begin{cases} n^{-1}\sum_{i=1}^n \eta(\hat{f}(\mathbf{x}_i)), & \text{if } \mathcal{T}(\cdot) = \mathbb{E}[\eta(\cdot)],\\ \hat{f}_{(\lceil\tau n\rceil)}, & \text{if } \mathcal{T}(\cdot) = \mathsf{VaR}_\tau(\cdot),\\ \hat{f}_{(\lceil\tau n\rceil)} + (1-\tau)^{-1}n^{-1}\sum_{i=1}^n(\hat{f}(\mathbf{x}_i) - \hat{f}_{(\lceil\tau n\rceil)})^+, & \text{if } \mathcal{T}(\cdot) = \mathsf{CVaR}_\tau(\cdot).\end{cases} \tag{7}$$

Note that in the standard nested simulation, $f(\mathbf{x}_i)$ is estimated via the average of the inner-level samples for the specific scenario $\mathbf{x}_i$; that is, $\hat{f}(\mathbf{x}_i) = \bar{y}_i$. In the KRR-driven nested simulation, the estimation of $f(\mathbf{x}_i)$ is based on all the data points $\{(\mathbf{x}_i, \bar{y}_i) : i = 1,\ldots,n\}$, thereby leveraging spatial information globally. See Figure 3 for an illustration.

Figure 3 KRR-driven Nested Simulation.



[7] The representer theorem stipulates that given a finite set of observations of a function in an RKHS, the task of finding the best approximation of the function in the RKHS, which is an infinite-dimensional optimization problem in general, can be reduced to a finite-dimensional problem that possesses an explicit optimal solution.

Remark 1. The implementation of the KRR-driven nested simulation requires—in addition to the budget allocation rule—the specification of three *hyperparameters* (λ,ν,ℓ). Theoretically, the convergence rate of $\hat{\theta}_{n,m}$ depends critically on ν and λ, but it is independent of ℓ. In the asymptotic analysis in Section 4, we assume the smoothness parameter ν of f is *known*. Given ν, our analysis reveals the proper order of magnitude of λ (but not the exact value) as Γ increases. The independence of the convergence rate on ℓ arises from the fact that the RKHSs induced by the Matérn kernels are identical for different values of ℓ but the same value of ν. However, from a practical point of view, the three hyperparameters all have an impact on the finite-sample performance of $\hat{\theta}_{n,m}$. They can be specified via cross-validation, with one complication. The goal of nested simulation is to estimate $\theta=\mathcal{T}(f(X))$ instead of f itself. Typical cross-validation focuses on the estimation accuracy of the latter, which may not be the right metric for assessing the estimator of f in the setting of nested simulation. Instead, in Section 5 we propose a new technique called $\mathcal{T}$-dependent cross-validation, which significantly outperforms the standard cross-validation.

Remark 2. KRR is closely related to *stochastic kriging* (Ankenman et al. 2010), which has been used in nested simulation with empirical success (Liu and Staum 2010, Barton et al. 2014, Xie et al. 2014). Stochastic kriging is a Bayesian method; assuming the prior distribution of f is a Gaussian process with mean 0, it estimates f using the posterior mean $\hat{f}_{\mathsf{SK}}(\mathbf{x}):=\mathbf{r}(\mathbf{x})^{\intercal}(\mathbf{R}+\mathbf{\Sigma})^{-1}\bar{\mathbf{y}}$, where $\mathbf{\Sigma}$ is the $n\times n$ diagonal matrix where the j-th diagonal element is $\mathbb{V}\mathrm{ar}[\bar{\epsilon}_i]$ with $\bar{\epsilon}_i=m^{-1}\sum_{j=1}^{m}\epsilon_{ij}$. Compared to $\hat{f}_{\mathsf{SK}}(\mathbf{x})$, a particular advantage of the KRR estimator (6) is the presence of the regularization parameter λ, which can be treated as a tuning parameter, providing significant flexibility to improve the estimation accuracy of f and eventually the performance of the nested simulation. Another advantage of KRR relative to stochastic kriging is that the former requires weaker conditions on the simulation noise. Whereas the latter requires ϵ_{ij} to be Gaussian with a *known* variance in order that the posterior of f should remain a Gaussian process, KRR allows ϵ_{ij} to be sub-Gaussian (see Definition 1 in Section 4) and does not need to assume a known variance.

4. Asymptotic Analysis

In this section, we analyze the performance of the KRR-driven estimator in a large computational budget asymptotic regime. We establish upper bounds on the convergence rate of $|\hat{\theta}_{n,m}-\theta|$ for the three forms of θ that are listed in Section 2. Before presenting the results, we highlight two main differences between the asymptotic analysis of the KRR-driven nested simulation and that of KRR in typical machine-learning contexts.

First, the objective of nested simulation is to estimate θ, whereas that of KRR is to estimate f. The difference in objective means a different, more careful bias–variance trade-off. The analysis of KRR in machine-learning literature can be used to handle the bias–variance trade-off in the inner-level estimation of $f(\mathbf{x})=\mathbb{E}[Y|X=\mathbf{x}]$. However, the presence of the nonlinear functional $\mathcal{T}$—which transforms the distribution of $f(X)$ to θ—essentially redefines bias and variance in the outer-level estimation of $\theta = \mathcal{T}(f(X))$ and thus breaks the inner-level bias–variance trade-off. Hence, a simple plug-in of existing analysis of KRR from machine-learning literature would not suffice in the nested simulation setting. The multiplicity of the forms of $\mathcal{T}$ also complicates our analysis significantly. As shown in Theorems 1–4, the convergence rate of the KRR-driven nested simulation varies for different forms of $\mathcal{T}$, and each is different from the convergence rate of the KRR estimator of f.

A second main difference is that multiple ($m > 1$) observations of $f(\mathbf{x}_i)$ are allowed in nested simulation, whereas only one single observation is allowed in typical KRR settings. Therefore, not only do we need to judiciously choose the regularization parameter λ—which is the main factor that determines the convergence rate of KRR in typical machine-learning settings—we also need to take into account the impact of m on the convergence rate. Hence, *joint* selection of λ and m is needed to improve the convergence rate of the KRR-driven nested simulation. This further complicates our theoretical analysis.

In the following, we state in Section 4.1 two basic assumptions that will be imposed throughout the paper. (Additional assumptions will be introduced later when needed.) We analyze the form of nested expectation (i.e., $\mathcal{T}(\cdot) = \mathbb{E}[\eta(\cdot)]$) in Section 4.2, analyze the two risk measures—VaR and CVaR—in Section 4.3, and summarize the results in Section 4.4. Due to space limitations, we give only proof sketches of the theoretical results and relegate the complete proofs to the e-companion.

4.1. Basic Assumptions

Definition 1 (Sub-Gaussian Distribution). A random variable Z is said to be *sub-Gaussian* with *variance proxy* σ^2, denoted by $Z \sim \mathsf{subG}(\sigma^2)$, if

$$\mathbb{E}\left[e^{t(Z-\mathbb{E}[Z])}\right] \le e^{\frac{\sigma^2 t^2}{2}}, \quad \forall t \in \mathbb{R}.$$

Typical examples of a sub-Gaussian distribution include Gaussian distribution or distribution with a bounded support. Note that σ^2 is not necessarily equal to—but an upper bound on—the variance of Z; that is, $\mathbb{V}\mathrm{ar}[Z] \le \sigma^2$ (Wainwright 2019, page 51).

Throughout this paper, we impose the following two assumptions.

Assumption 1. *The noise terms* $\{\epsilon_{ij} : 1 \le i \le n, 1 \le j \le m\}$ *are independent, zero-mean* $\mathsf{subG}(\sigma^2)$. *Moreover,* $\epsilon_{i1}, \ldots, \epsilon_{im}$ *are identically distributed, for each* $i = 1, \ldots, n$.

Assumption 2. *The domain of* X *is a bounded convex set* $\Omega \subset \mathbb{R}^d$. *Moreover,* X *has a probability density function that is bounded above and below away from zero.*

Assumption 3. $f \in \mathcal{N}_\Psi(\Omega)$, *the RKHS associated with the Matérn kernel of smoothness* ν.

Assumption 1 allows the simulation noise to be heteroscedastic, that is, the variances at different $\mathbf{x}_i$'s are potentially unequal (i.e., $\mathbb{V}\mathrm{ar}(\epsilon_{i1})$ may not be constant with respect to i). If we relax the sub-Gaussian assumption to the sub-exponential (i.e., light-tailed) assumption on the noise terms, then we may follow an analysis framework similar to that presented subsequently in this section to derive the convergence rate of $\hat{\theta}_{n,m}$, although the results would be technically more complicated. See van de Geer (2000, page 168) for a discussion on this kind of relaxation. The heavy-tailed case, which often appears in financial applications (Fuh et al. 2011), is significantly more challenging. Our analysis framework might still be valid, by virtue of a sharper characterization of the corresponding empirical process, such as that in Han and Wellner (2019).

In Assumption 2, the boundedness of Ω can be relaxed if we assume X is sub-Gaussian. The convexity of Ω is imposed to ensure that Sobelev spaces over Ω are well defined. This condition can also be relaxed; we may instead impose conditions on the boundary of Ω (e.g., Lipschitz-type boundary conditions) to serve the same purpose (Adams and Fournier 2003, Chapter 4). However, such relaxation would make the proofs of the theoretical results technically more involved without adding much value to the main ideas of the present paper. As discussed in Section 3.1, Assumption 3 is equivalent to the assumption that f lies in the Sobolev space $\mathcal{H}^{\nu+d/2}(\Omega)$, basically meaning that f is square-integrable and is weakly differentiable up to order $\nu+d/2$.

4.2. Nested Expectation

We now consider problems in the form of a nested expectation, $\theta = \mathbb{E}[\eta(\mathbb{E}[Y|X])]$, for some function η. The KRR-driven estimator in (7) is then $\hat{\theta}_{n,m} = n^{-1}\sum_{i=1}^n \eta(\hat{f}(\mathbf{x}_i))$. By the triangle inequality,

$$|\hat{\theta}_{n,m} - \theta| \le \underbrace{\left|\mathbb{E}[\eta(f(X))] - \frac{1}{n}\sum_{i=1}^n \eta(f(\mathbf{x}_i))\right|}_{I_1} + \underbrace{\left|\frac{1}{n}\sum_{i=1}^n \big[\eta(f(\mathbf{x}_i)) - \eta(\hat{f}(\mathbf{x}_i))\big]\right|}_{I_2}. \qquad (8)$$

While we may apply the central limit theorem to derive $I_1 = O_{\mathbb{P}}(n^{-1/2})$, asymptotic analysis of I_2 is highly nontrivial and categorically depends on the property of η. It might be intuitive to anticipate the general tendency that the smoother f is, the faster $\hat{f}$ converges to f. The presence of η, however, complicates characterization of the specific dependence of the convergence rate on the smoothness of f and the dimensionality of X. In particular, without knowledge about η, it is unclear *a priori* whether I_2 renders a square root convergence rate even if f is sufficiently smooth.

In light of Example 1 in Section 2, we study the following three cases of η. They are standard cases in the literature (Broadie et al. 2015, Hong et al. 2017, Zhang et al. 2021).

(i) η is twice-differentiable with bounded first- and second-order derivatives; that is,

$$\sup_{z\in\{f(\mathbf{x}):\mathbf{x}\in\Omega\}} |\eta'(z)| < \infty \quad \text{and} \quad \sup_{z\in\{f(\mathbf{x}):\mathbf{x}\in\Omega\}} |\eta''(z)| < \infty.$$

(ii) η is a hockey-stick function, $\eta(z) = (z - z_0)^+$, for some constant $z_0 \in \{f(\mathbf{x}) : \mathbf{x}\in\Omega\}$.

(iii) η is an indicator function, $\eta(z) = \mathbb{I}\{z \ge z_0\}$, for some constant $z_0 \in \{f(\mathbf{x}) : \mathbf{x}\in\Omega\}$.

4.2.1. Smooth Functions Assume η is twice-differentiable with bounded first- and second-order derivatives. It follows from Taylor's expansion and the triangle inequality that

$$I_2 \le \underbrace{\left|\frac{1}{n}\sum_{i=1}^n \eta'(f(\mathbf{x}_i))(f(\mathbf{x}_i)-\hat{f}(\mathbf{x}_i))\right|}_{I_{21}} + \underbrace{\left|\frac{1}{2n}\sum_{i=1}^n \eta''(\tilde{z}_i)(f(\mathbf{x}_i)-\hat{f}(\mathbf{x}_i))^2\right|}_{I_{22}}, \tag{9}$$

where $\tilde{z}_i$ is a value between $f(\mathbf{x}_i)$ and $\hat{f}(\mathbf{x}_i)$. We present two technical results below to bound the convergence rates of I_{21} and I_{22}, respectively.

Proposition 1. *Suppose $\varphi: \{f(\mathbf{x}): \mathbf{x}\in\Omega\}\mapsto\mathbb{R}$ is bounded, and Assumptions 1–3 hold. Then,*

$$\left|\frac{1}{n}\sum_{i=1}^n \varphi(f(\mathbf{x}_i))(f(\mathbf{x}_i)-\hat{f}(\mathbf{x}_i))\right| = O_{\mathbb{P}}(\lambda^{1/2}+(mn)^{-1/2}).$$

Proof Sketch of Proposition 1. The expression of $\hat{f}$ in (6) implies that

$$f(\mathbf{x}_i)-\hat{f}(\mathbf{x}_i) = \underbrace{f(\mathbf{x}_i)-\mathbf{r}(\mathbf{x}_i)^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{f}}_{M_1(\mathbf{x}_i)} - \underbrace{\mathbf{r}(\mathbf{x}_i)^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\bar{\boldsymbol{\epsilon}}}_{M_2(\mathbf{x}_i)}, \tag{10}$$

where $\mathbf{f}=(f(\mathbf{x}_1),\ldots,f(\mathbf{x}_n))^\intercal$ and $\bar{\boldsymbol{\epsilon}}=(\bar{\epsilon}_1,\ldots,\bar{\epsilon}_n)^\intercal$. For $M_1(\mathbf{x}_i)$, the representer theorem asserts that $f_\dagger(\mathbf{x}_i) := \mathbf{r}(\mathbf{x}_i)^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{f}$ minimizes $n^{-1}\sum_{i=1}^n(g(\mathbf{x}_i)-f(\mathbf{x}_i))^2+\lambda\|g\|^2_{\mathcal{N}_\Psi(\Omega)}$ over $g\in\mathcal{N}_\Psi(\Omega)$. It is then straightforward to show that $n^{-1}\sum_{i=1}^n(f_\dagger(\mathbf{x}_i)-f(\mathbf{x}_i))^2\le\lambda\|f\|^2_{\mathcal{N}_\Psi(\Omega)}$, which leads to the $\lambda^{1/2}$ term in Proposition 1. Moreover, $M_2(\mathbf{x}_i)$ can be rewritten as a weighted sum of $\bar{\epsilon}_j$'s, and thus it is a zero-mean random variable with a finite variance. The same observation holds for $n^{-1}\sum_{i=1}^n\varphi(f(\mathbf{x}_i))M_2(\mathbf{x}_i)$, and thus it can be upper bounded in probability by its standard deviation, which, via direct calculations, yields the $(mn)^{-1/2}$ term in Proposition 1. □

Proposition 2. *Suppose Assumptions 1–3 hold. Then,*

$$\frac{1}{n}\sum_{i=1}^n(f(\mathbf{x}_i)-\hat{f}(\mathbf{x}_i))^2 = O_{\mathbb{P}}\left(\lambda+(mn)^{-1}\lambda^{-\frac{d}{2\nu+d}}+(mn)^{-\frac{2\nu+d}{2\nu+2d}}\right).$$

Proof Sketch of Proposition 2. The proof is based on empirical process theory and is similar to that of Theorem 10.2 in van de Geer (2000). □

Remark 3. Combining the Cauchy–Schwarz inequality and Proposition 2 may yield an upper bound on I_{21}. However, this bound would not be as tight as that in Proposition 1, which is a result of a refined analysis.

Because η has bounded first- and second-order derivatives, we can apply Propositions 1 and 2 to conclude, with elementary algebraic calculations, that $|\hat{\theta}_{m,n}-\theta| = O_{\mathbb{P}}\big(n^{-1/2}+\lambda^{1/2}+(mn)^{-1}\lambda^{-\frac{d}{2\nu+d}}\big)$. We then reduce this upper bound as much as possible by careful selection of n and λ. This leads to Theorem 1.

Theorem 1. *Let $\mathcal{T}(\cdot) = \mathbb{E}[\eta(\cdot)]$ and η be a twice-differentiable function with bounded first- and second-order derivatives. Suppose Assumptions 1–3 hold. Then, $|\hat{\theta}_{n,m} - \theta| = O_{\mathbb{P}}(\Gamma^{-\kappa})$, where κ is specified as follows:*

(i) If $\nu \geq \frac{d}{2}$, then $\kappa = \frac{1}{2}$ by setting $n \asymp \Gamma$, $m \asymp 1$, and $\lambda \asymp \Gamma^{-1}$.

(ii) If $0 < \nu < \frac{d}{2}$, then $\kappa = \frac{2\nu+d}{2\nu+3d}$ by setting $n \asymp \Gamma^{\frac{2(2\nu+d)}{2\nu+3d}}$, $m \asymp \Gamma^{\frac{d-2\nu}{2\nu+3d}}$, and $\lambda \asymp \Gamma^{-\frac{2(2\nu+d)}{2\nu+3d}}$.

Theorem 1 has several implications. First, it clearly reveals a mitigating effect of ν on the curse of dimensionality on the convergence rate—the larger ν is, the faster the rate is. In particular, in the case of nested expectation with η being smooth, $\hat{\theta}_{m,n}$ achieves the square root convergence rate when $\nu \geq \frac{d}{2}$, recovering the canonical rate of Monte Carlo simulation.

Second, as $\nu \to 0$, the convergence rate of $\hat{\theta}_{m,n}$ approaches $O_{\mathbb{P}}(\Gamma^{-1/3})$, and meanwhile, the outer-level sample size becomes $n \asymp \Gamma^{2/3}$. Both recover the results for the standard nested simulation (Zhang et al. 2021). Further, in light of the fact that $\kappa > \frac{1}{3}$ for all $\nu > 0$ in Theorem 1, if η is smooth, then regardless of the dimensionality, the use of KRR will have a beneficial effect on the estimation of $\mathbb{E}[\eta(f(X))]$, at least from the perspective of convergence rates. This is in clear contrast to the effect of using kernel smoothing in the inner-level estimation. Hong et al. (2017) show that under the same assumptions on η, the use of kernel smoothing is beneficial only for low-dimensional problems and becomes detrimental when $d \geq 5$. That KRR and kernel smoothing have different effects on the convergence rate in high dimensions is mainly because the former manages to leverage the smoothness of f, providing us with a proper function space to perform function estimation.

Third, as ν increases, there exists a "phase transition" in the budget allocation rule. The inner-level sample size should remain constant ($m \asymp 1$) if ν is above a threshold ($\frac{d}{2}$ in this case), whereas it should grow as Γ increases otherwise. Intuitively, this is because if f is sufficiently smooth we do not anticipate it to vary substantially over different locations. Hence, even if each observation $f(\mathbf{x}_i)$ is highly noisy, the noises would mostly cancel one another and f would be reasonably estimated, as long as f is observed at a sufficient number of locations. However, if ν is small f may exhibit dramatic variations, and a slight change in an observation of f due to noise may lead to a significant change in the estimate of f. To reduce the impact of noise on the observation of f, we must take multiple replications at each location. Moreover, a higher proportion of the budget should be allocated to the inner-level samples the lower the smoothness of f.

Last, we consider the special case that $\eta(z) = z$, which occurs in the context of input uncertainty quantification (see Example 2). In this case, the term I_{22} in (9) vanishes and therefore Proposition 2 is no longer needed. With the same proof for Theorem 1, we have $|\hat{\theta}_{m,n} - \theta| = O_{\mathbb{P}}\big(n^{-1/2} + \lambda^{1/2} + (mn)^{-1/2}\big)$, which leads to the square root rate for all $\nu > 0$.

Corollary 1. *Let $\mathcal{T}(\cdot) = \mathbb{E}[\cdot]$. Suppose Assumptions 1–3 hold. Then, $|\hat{\theta}_{n,m} - \theta| = O_{\mathbb{P}}(\Gamma^{-1/2})$ for all $\nu > 0$, by setting $n \asymp \Gamma$, $m \asymp 1$, and $\lambda \asymp \Gamma^{-1}$.*

Numerical studies by Barton et al. (2014) and Xie et al. (2014) demonstrate that the use of stochastic kriging greatly enhances the accuracy for quantifying the impact of input uncertainty on simulation outputs. In light of the close connection between KRR and stochastic kriging as elucidated in Remark 2, Corollary 1 sheds light on this empirical success.

4.2.2. Hockey-stick Functions

Assume $\eta(z)=(z-z_0)^+$. The non-differentiability of η at z_0 implies that the magnitude of I_2 in the decomposition (8), and therefore the accuracy of $\hat{\theta}_{n,m}$, is potentially sensitive relative to the accuracy of $\hat{f}$. Namely, a slight change in $\hat{f}$ may result in a significant change in $\hat{\theta}_{n,m}$. Because the non-differentiability takes effect only when $f(\mathbf{x})$ falls in the vicinity of z_0, we impose the following assumption to characterize the likelihood of this event.

Assumption 4. *There exist positive constants C, t_0, and $\alpha\le 1$ such that*

$$\mathbb{P}(|f(X)-z_0|\le t)\le Ct^{\alpha},\quad \forall t\in(0,t_0].$$

Assumption 4 is similar to the Tsybakov margin condition (Tsybakov 2004), which is widely used in machine-learning literature to study classification algorithms. A large value of α means that $|f(\mathbf{x})-z_0|$ is bounded away from zero with a high probability. Conversely, if α is close to zero, Assumption 4 is essentially void—because it is satisfied by *any* probability distribution of X if we set $\alpha=0$ and $C=1$—and $|f(\mathbf{x})-z_0|$ can be arbitrarily close to zero, making the non-differentiability of η negatively affect nearly all $\mathbf{x}\in\Omega$. The typical case[8] in practice is $\alpha=1$. This, for example, can be easily shown via Taylor's expansion if X has a density that is bounded above and below away from zero (Assumption 2) and f has bounded first- and second-order derivatives with $\|\nabla f(\mathbf{x}_0)\|>0$ for all $\mathbf{x}_0$ such that $f(\mathbf{x}_0)=z_0$.

Due to its non-differentiability, Taylor's expansion does not apply to η, and therefore we cannot use (9) to analyze the error term I_2. Instead, we adopt the treatment in Hong et al. (2017) and construct a twice-differentiable function η_δ, which is parameterized by δ and approximates η as $\delta\to 0$. We decompose I_2 with η_δ being an intermediate step:

$$I_2\le \underbrace{\left|\frac{1}{n}\sum_{i=1}^{n}\bigl[\eta(f(\mathbf{x}_i))-\eta_\delta(f(\mathbf{x}_i))\bigr]\right|}_{J_1}+\underbrace{\left|\frac{1}{n}\sum_{i=1}^{n}\bigl[\eta_\delta(f(\mathbf{x}_i))-\eta_\delta(\hat{f}(\mathbf{x}_i))\bigr]\right|}_{J_2}+\underbrace{\left|\frac{1}{n}\sum_{i=1}^{n}\bigl[\eta_\delta(\hat{f}(\mathbf{x}_i))-\eta(\hat{f}(\mathbf{x}_i))\bigr]\right|}_{J_3}. \tag{11}$$

[8] In general, α may take values greater than one if f (not η) exhibits non-smoothness in the region $\{\mathbf{x}:f(\mathbf{x})=z_0\}$ (e.g., $|x-z_0|^{1/2}$ if $d=1$). However, as the present paper focuses on scenarios where f is smooth, we anticipate $\alpha\le 1$ in our theoretical framework. See Perchet and Rigollet (2013) for a discussion on a similar tension between the smoothness of functions in Hölder spaces and the value of α in the Tsybakov margin condition.

Assuming, without loss of generality, that $z_0=0$, the key properties of η_δ include (i) $|\eta(z)-\eta_\delta(z)|=O(\delta\,\mathbb{I}\{z\in[-\delta,\delta]\})$, (ii) $|\eta_\delta'(z)|$ is bounded uniformly for all δ, and (iii) $|\eta_\delta''(z)|=O(\delta^{-1}\,\mathbb{I}\{z\in[-\delta,\delta]\})$. First, applying property (i) to J_1, we have that for some constant $C_1>0$,

$$J_1\le\left|\frac{C_1}{n}\sum_{i=1}^n\delta\,\mathbb{I}\{f(\mathbf{x}_i)\in[-\delta,\delta]\}\right|\le C_1\delta\left|\frac{1}{n}\sum_{i=1}^n\mathbb{I}\{f(\mathbf{x}_i)\in[-\delta,\delta]\}-\mathbb{E}\,\mathbb{I}\{f(\mathbf{x}_i)\in[-\delta,\delta]\}\right|$$
$$+C_1\delta\,\mathbb{E}\big[\mathbb{I}\{f(\mathbf{x}_i)\in[-\delta,\delta]\}\big]=O_{\mathbb{P}}(\delta n^{-1/2}+\delta^{\alpha+1}),\qquad(12)$$

where the last step follows from the central limit theorem and Assumption 4.

Next, J_3 can be analyzed in the same fashion, except that the relevant event here is $\{\hat f(\mathbf{x}_i)\in[-\delta,\delta]\}$. The complication is addressed by considering the larger event $\{f(\mathbf{x}_i)\in[-\delta-\rho_n,\delta+\rho_n]\}$, where $\rho_n:=\max_{1\le i\le n}|f(\mathbf{x}_i)-\hat f(\mathbf{x}_i)|$. This leads us to $J_3=O_{\mathbb{P}}(\delta n^{-1/2}+\delta(\delta+\rho_n)^\alpha)$. The quantity ρ_n is critical and is also involved in the analysis of J_2.

At last, similar to (9), we may decompose J_2 via Taylor's expansion:

$$J_2=\underbrace{\left|\frac{1}{n}\sum_{i=1}^n\eta_\delta'(f(\mathbf{x}_i))(f(\mathbf{x}_i)-\hat f(\mathbf{x}_i))\right|}_{J_{21}}+\underbrace{\left|\frac{1}{2n}\sum_{i=1}^n\eta_\delta''(\breve z_i)(f(\mathbf{x}_i)-\hat f(\mathbf{x}_i))^2\right|}_{J_{22}},\qquad(13)$$

where $\breve z_i$ is a value between $f(\mathbf{x}_i)$ and $\hat f(\mathbf{x}_i)$. Because $|\eta_\delta'(z)|$ is bounded uniformly for all δ, J_{21} can be bounded using Proposition 1. However, unlike our treatment of I_{22} in (9), applying Proposition 2 to J_{22} would yield a loose bound because $|\eta_\delta''(z)|=O(\delta^{-1}\,\mathbb{I}\{z\in[-\delta,\delta]\})$ is not uniformly bounded as $\delta\to0$. Instead, we bound J_{22} in terms of ρ_n:

$$J_{22}\le\left|\frac{1}{n}\sum_{i=1}^n C_2\delta^{-1}\,\mathbb{I}\{\breve z_i\in[-\delta,\delta]\}\rho_n^2\right|\le\left|\frac{1}{n}\sum_{i=1}^n C_2\delta^{-1}\rho_n^2\,\mathbb{I}\{f(\mathbf{x}_i)\in[-\delta-\rho_n,\delta+\rho_n]\}\right|$$
$$=O_{\mathbb{P}}\left(\delta^{-1}\rho_n^2\big(n^{-1/2}+(\delta+\rho_n)^\alpha\big)\right),$$

for some constant $C_2>0$, where the second step holds because $\breve z_i$ is a value between $f(\mathbf{x}_i)$ and $\hat f(\mathbf{x}_i)$, and the last step can be shown with the same argument used for J_1. Putting these bounds for J_1, J_2, and J_3 together yields a bound for $|\hat\theta_{n,m}-\theta|$ that involves both δ and ρ_n. If we further set $\delta=\rho_n$, then the bound is reduced to $|\hat\theta_{n,m}-\theta|=O_{\mathbb{P}}\left(n^{-1/2}+\rho_n^{\alpha+1}\right)$. Below, we present an asymptotic property of ρ_n.

Proposition 3. *Suppose Assumptions 1–3 hold. If $\lambda\asymp\Gamma^{-1}$, then*

$$\max_{1\le i\le n}|f(\mathbf{x}_i)-\hat f(\mathbf{x}_i)|=O_{\mathbb{P}}\left(\big(n^{-\frac{\nu}{2\nu+2d}}\wedge m^{-1/2}\big)(\log n)^{1/2}\right).$$

Proof Sketch of Proposition 3. The decomposition (10) implies that it suffices to bound $\max_i|M_1(\mathbf{x}_i)|$ and $\max_i|M_2(\mathbf{x}_i)|$. Consider the former first. Note that by definition,

$$\max_{1\le i\le n}|M_1(\mathbf{x}_i)|\le\sup_{x\in\Omega}|f(\mathbf{x})-f_\dagger(\mathbf{x})|=\|f-f_\dagger\|_{\mathcal{L}_\infty(\Omega)},$$

where $f_\dagger(\mathbf{x}_i) = \mathbf{r}(\mathbf{x}_i)^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{f}$. In general, it is more difficult to derive a proper bound on the $\mathcal{L}_\infty$ norm of a function than on its $\mathcal{L}_2$ norm. The key here is to take advantage of the norm equivalence between the RKHS $\mathcal{N}_\Psi(\Omega)$ and the Sobolev space $\mathcal{H}^{\nu+d/2}$ and to apply the Gagliardo–Nirenberg interpolation inequality for functions in Sobolev spaces (Brezis and Mironescu 2019) that bounds $\|f-f_\dagger\|_{\mathcal{L}_\infty(\Omega)}$ in terms of $\|f-f_\dagger\|_{\mathcal{L}_2(\Omega)}$. Standard results in empirical process theory can then be used to bound $\|f-f_\dagger\|_{\mathcal{L}_2(\Omega)}$ in terms of $\|f-f_\dagger\|_n$, where $\|\cdot\|_n$ denotes the *empirical semi-norm* that is defined by $\|g\|_n := \left(n^{-1}\sum_{i=1}^n g^2(\mathbf{x}_i)\right)^{1/2}$. At last, to bound $\|f-f_\dagger\|_n$, note that by the representer theorem, $f_\dagger(\mathbf{x}_i)$ minimizes $n^{-1}\sum_{i=1}^n (g(\mathbf{x}_i)-f(\mathbf{x}_i))^2 + \lambda\|g\|^2_{\mathcal{N}_\Psi(\Omega)}$ over $g\in\mathcal{N}_\Psi(\Omega)$. Hence,

$$\frac{1}{n}\sum_{i=1}^n (f(\mathbf{x}_i)-f_\dagger(\mathbf{x}_i))^2 + \lambda\|f_\dagger\|^2_{\mathcal{N}_\Psi(\Omega)} \le \frac{1}{n}\sum_{i=1}^n (f(\mathbf{x}_i)-f(\mathbf{x}_i))^2 + \lambda\|f\|^2_{\mathcal{N}_\Psi(\Omega)} = \lambda\|f\|^2_{\mathcal{N}_\Psi(\Omega)},$$

which implies $\|f-f_\dagger\|_n = O_{\mathbb{P}}(\lambda^{1/2})$.

Next, consider $\max_i |M_2(\mathbf{x}_i)|$. The presence of the noise terms $\bar{\boldsymbol{\epsilon}}$ further complicates the analysis. By Assumption 1, $M_2(\mathbf{x}_i)$ is sub-Gaussian. Therefore, we can apply the union bound (Boole's inequality) to reduce the tail probability of $\max_i |M_2(\mathbf{x}_i)|$ to the asymptotic behavior of $\mathbb{V}\mathrm{ar}(M_2(\mathbf{x}_i))$. Similar to before, the approach we follow to bound $\mathbb{V}\mathrm{ar}(M_2(\mathbf{x}_i))$ also relies on the connections between the three norms $\|\cdot\|_{\mathcal{L}_\infty(\Omega)}$, $\|\cdot\|_{\mathcal{L}_2(\Omega)}$, and $\|\cdot\|_n$, although the analysis is technically more involved. □

We apply Proposition 3 to bound ρ_n in $|\hat{\theta}_{n,m}-\theta| = O_{\mathbb{P}}\left(n^{-1/2}+\rho_n^{\alpha+1}\right)$. Theorem 2 then follows from straightforward calculations.

Theorem 2. *Let $\mathcal{T}(\cdot)=\mathbb{E}[\eta(\cdot)]$ and η be a hockey-stick function. Suppose Assumptions 1–4 hold. Then, $|\hat{\theta}_{n,m}-\theta| = O_{\mathbb{P}}(\Gamma^{-\kappa}(\log\Gamma)^{\tilde{\kappa}})$, where κ and $\tilde{\kappa}$ are specified as follows:*

(i) If $\nu \ge \frac{d}{\alpha+1}$, then $\kappa = \frac{1}{2}\wedge\frac{\nu(\alpha+1)}{2(\nu+d)}$ and $\tilde{\kappa} = \frac{\alpha+1}{2}\mathbb{I}\{\nu\alpha<d\}$ by setting $n\asymp\Gamma$, $m\asymp 1$, and $\lambda\asymp\Gamma^{-1}$.

(ii) If $\nu < \frac{d}{\alpha+1}$, then $\kappa = \frac{\alpha+1}{2(\alpha+2)}$ and $\tilde{\kappa} = \frac{\alpha+1}{2}$ by setting $n\asymp\Gamma^{\frac{\alpha+1}{\alpha+2}}$, $m\asymp\Gamma^{\frac{1}{\alpha+2}}$, and $\lambda\asymp\Gamma^{-1}$.

Similar to Theorem 1, Theorem 2 manifests the mitigating effect of the smoothness on the curse of dimensionality, as well as the phase transition in the budget allocation rule. In addition, it shows that if η is a hockey-stick function, a larger value of α, through inducing a smaller probability of $f(\mathbf{X})$ falling near the point z_0 where η is non-differentiable, leads to a higher convergence rate of $\hat{\theta}_{m,n}$ for estimating $\mathbb{E}[\eta(f(X))]$.

In the typical scenario $\alpha=1$, the rate achieves $O_{\mathbb{P}}(\Gamma^{-1/2})$ if $\nu\ge d$ (and thus $\frac{\nu(\alpha+1)}{2(\nu+d)}\ge\frac{1}{2}$). Meanwhile, if $\nu<\frac{d}{2}$, the rate becomes $O_{\mathbb{P}}(\Gamma^{-1/3}(\log\Gamma))$, which is nearly (discarding the logarithmic factor) identical to that of the standard nested simulation. See Figure 1 for an illustration.

In the worst scenario $\alpha=0$, the square root rate cannot be fully recovered (because $\frac{\nu(\alpha+1)}{2(\nu+d)}\le\frac{1}{2}$ for all ν in this case), but it can be approached arbitrarily close to, as $\nu\to\infty$. Meanwhile, if $\nu<2d$, the rate is slower than $O_{\mathbb{P}}(\Gamma^{-1/3})$, and therefore the standard nested simulation is preferable to the KRR-driven method.

4.2.3. Indicator Functions Assume $\eta(z) = \mathbb{I}\{z \geq z_0\}$. The discontinuity at z_0 poses an even bigger challenge to the convergence rate of $\hat{\theta}_{n,m}$ compared to the case for hockey-stick functions. In particular, it renders the smooth approximation approach employed for hockey-stick functions ineffective. This is because for any differentiable function $\tilde{\eta}_\delta$ that converges to η uniformly as $\delta \to 0$, $|\tilde{\eta}'_\delta(z_0)|$ would blow up as $\delta \to 0$. Therefore, if we base our analysis on the decomposition (13), we would end up with an undesirable bound.

Instead, we directly work on I_2 in (8) without further decomposing it. Again, assume $z_0 = 0$ without loss of generality. Note that if $\mathbb{I}\{f(\mathbf{x}_i) \geq 0\} \neq \mathbb{I}\{\hat{f}(\mathbf{x}_i) \geq 0\}$, then we must have $f(\mathbf{x}_i) \in [-\rho_n, \rho_n]$, where $\rho_n = \max_{1\leq i \leq n} |f(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)|$. It follows that

$$\begin{aligned} I_2 &= \left| \frac{1}{n} \sum_{i=1}^n \Big(\eta(f(\mathbf{x}_i)) - \eta(\hat{f}(\mathbf{x}_i)) \Big) \mathbb{I}\{f(\mathbf{x}_i) \in [-\rho_n, \rho_n]\} \right| \\ &\leq \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{f(\mathbf{x}_i) \in [-\rho_n, \rho_n]\} = O_{\mathbb{P}}(n^{-1/2} + \rho_n^\alpha), \end{aligned}$$

where the last step follows the same argument as (12). We then can apply Proposition 3 to derive Theorem 3 below.

Theorem 3. *Let $\mathcal{T}(\cdot) = \mathbb{E}[\eta(\cdot)]$ and η be an indicator function. Suppose Assumptions 1–4 hold. Then, $|\hat{\theta}_{n,m} - \theta| = O_{\mathbb{P}}(\Gamma^{-\kappa}(\log \Gamma)^{\tilde{\kappa}})$, where κ and $\tilde{\kappa}$ are specified as follows:*

(i) If $\nu \geq \frac{d}{\alpha}$, then $\kappa = \frac{\nu\alpha}{2(\nu+d)}$ and $\tilde{\kappa} = \frac{\alpha}{2}$ by setting $n \asymp \Gamma$, $m \asymp 1$, and $\lambda \asymp \Gamma^{-1}$.

(ii) If $\nu < \frac{d}{\alpha}$, then $\kappa = \frac{\alpha}{2(\alpha+1)}$ and $\tilde{\kappa} = \frac{\alpha}{2}$ by setting $n \asymp \Gamma^{\frac{\alpha}{\alpha+1}}$, $m \asymp \Gamma^{\frac{1}{\alpha+1}}$, and $\lambda \asymp \Gamma^{-1}$.

Note that the final calculations that lead to Theorems 2 and 3 are the same, except that the bound $O_{\mathbb{P}}(n^{-1/2} + \rho_n^{\alpha+1})$ for the former is replaced with $O_{\mathbb{P}}(n^{-1/2} + \rho_n^{\alpha})$ for the latter. Therefore, if η is an indicator function and $\alpha = 1$ (the typical value), the convergence rate of $\hat{\theta}_{n,m}$ is the same as the case in which η is a hockey-stick function and $\alpha = 0$, meaning the square root rate cannot be fully recovered, but it can be approached arbitrarily close to, as $\nu \to \infty$ (see Figure 1).

In addition, if $\alpha = 0$, Theorem 3 asserts that $|\hat{\theta}_{n,m} - \theta| = O_{\mathbb{P}}(1)$ for all $\nu > 0$. Consider the following simple example. Suppose that $z_0 = 0$ and $f(\mathbf{x}) \equiv 0$ for all $\mathbf{x} \in \Omega$. Then, $\theta = \mathbb{E}[\mathbb{I}\{f(X) \geq 0\}] = 1$, and Assumption 4 is not satisfied for any $\alpha > 0$, as $\mathbb{P}(|f(X)| \leq t) = 1$ for all $t > 0$; moreover, $\hat{f}(\mathbf{x}) = \mathbf{r}(\mathbf{x})^\intercal (\mathbf{R} + n\lambda \mathbf{I})^{-1} \bar{\boldsymbol{\epsilon}}$. Suppose also that $\epsilon_{i,j}$'s are i.i.d. normal random variables. Then, given $\mathcal{D} = \{(\mathbf{x}_i, \bar{y}_i) : i = 1, \ldots, n\}$, $\hat{f}(\mathbf{x})$ has a normal distribution with a mean of zero. It follows that $\mathbb{P}(\hat{f}(\mathbf{x}) \geq 0 | \mathcal{D}) = \frac{1}{2}$ for all $\mathbf{x} \in \Omega$, and therefore

$$\mathbb{E}\big[\hat{\theta}_{n,m} \,\big|\, \mathcal{D}\big] = \mathbb{E}\big[\mathbb{I}\{\hat{f}(X) \geq z_0\} \,\big|\, \mathcal{D}\big] = \mathbb{E}\big[\mathbb{P}\big(\hat{f}(X) \geq z_0 \,\big|\, \mathcal{D}, X\big) \,\big|\, \mathcal{D}\big] = \frac{1}{2}.$$

This implies that $\mathbb{E}[\hat{\theta}_{n,m}] = \frac{1}{2}$ for all n and m, so $\hat{\theta}_{n,m}$ does not converge to θ. Recall that $\alpha = 0$ effectively nullifies Assumption 4. The preceding discussion suggests that Assumption 4 (with $\alpha > 0$) is necessary to ensure the consistency of $\hat{\theta}_{m,n}$ if $\theta = \mathbb{E}[\eta(f(\mathbf{X}))]$ and η is an indicator function.

4.3. Risk Measures

We now examine the KRR-driven method for the case that $\mathcal{T}$ represents VaR or CVaR, two popular risk measures. Let $\mathtt{G}(z) := \mathbb{P}(f(X) \le z)$ denote the cumulative distribution function (CDF) of $f(X)$ and $\mathtt{G}^{-1}(q) := \{z : \mathtt{G}(z) \ge q\}$ denote its quantile function. Fixing an arbitrary risk level $\tau \in (0,1)$, we define $\zeta_{\mathsf{VaR}} := \mathsf{VaR}_\tau(f(X)) = \mathtt{G}^{-1}(\tau)$ and let $\hat{\zeta} := \hat{f}_{(\lceil \tau n \rceil)}$ be its KRR-driven estimator defined in (7). We will analyze the convergence rate of $\hat{\zeta} - \zeta_{\mathsf{VaR}}$ by analyzing that of $\mathtt{G}(\hat{\zeta}) - \mathtt{G}(\zeta_{\mathsf{VaR}})$. To that end, we impose the following assumption to regularize the behaviors of both $\mathtt{G}$ and $\mathtt{G}^{-1}$.

Assumption 5. *There exist positive constants C_1, C_2, t_0, and $\beta \le \gamma$ such that*

$$C_2 t^\gamma \le \mathbb{P}(|f(X) - z| \le t) \le C_1 t^\beta, \quad \forall t \in (0, t_0], \, \forall z \in \{f(\mathbf{x}) : \mathbf{x} \in \Omega\}.$$

Remark 4. Note that $\mathbb{P}(|f(X) - z| \le t) = \mathtt{G}(z+t) - \mathtt{G}(z-t)$ and that $\mathtt{G}$ is a non-decreasing function by definition. Therefore, Assumption 5 is equivalent to

$$|\mathtt{G}(z) - \mathtt{G}(\tilde{z})| \le \tilde{C}_1 |z - \tilde{z}|^\beta, \tag{14}$$

$$|\mathtt{G}(z) - \mathtt{G}(\tilde{z})| \ge \tilde{C}_2 |z - \tilde{z}|^\gamma, \tag{15}$$

for all $z, \tilde{z} \in \{f(\mathbf{x}) : \mathbf{x} \in \Omega\}$, where $\tilde{C}_1$ and $\tilde{C}_2$ are some positive constants. The condition (14) means that $\mathtt{G}$ is β-Hölder continuous, which includes Lipschitz continuous ($\beta = 1$) as a special case. The condition (15) can be interpreted as follows. Let $\mathtt{G}(z) = q$ and $\mathtt{G}(\tilde{z}) = \tilde{q}$. If we further assume $\mathtt{G}^{-1}$ is continuous, then the condition (15) can be rewritten as $|q - \tilde{q}| \ge \tilde{C}_2 |\mathtt{G}^{-1}(q) - \mathtt{G}^{-1}(\tilde{q})|^\gamma$, meaning $\mathtt{G}^{-1}$ is γ^{-1}-Hölder continuous. It is known that if $\beta > 1$, then a β-Hölder continuous function on an interval is a constant. However, either $\mathtt{G}$ or $\mathtt{G}^{-1}$ is a constant by definition. Hence, Assumption 5 implicitly implies $\beta \le 1 \le \gamma$.

To analyze the convergence rate of $|\zeta_{\mathsf{VaR}} - \hat{\zeta}|$, we first note that by Assumption 5,

$$\begin{aligned} |\mathtt{G}(\hat{\zeta}) - \mathtt{G}(\zeta_{\mathsf{VaR}})| &= \mathbb{P}\Big(\min(\hat{\zeta}, \zeta_{\mathsf{VaR}}) \le f(X) \le \max(\hat{\zeta}, \zeta_{\mathsf{VaR}})\Big) \\ &= \mathbb{P}\Big(\Big|f(X) - \frac{(\hat{\zeta} + \zeta_{\mathsf{VaR}})}{2}\Big| \le \frac{|\hat{\zeta} - \zeta_{\mathsf{VaR}}|}{2}\Big) \ge C_2 |\hat{\zeta} - \zeta_{\mathsf{VaR}}|^\gamma. \end{aligned}$$

Therefore, $|\hat{\zeta} - \zeta_{\mathsf{VaR}}| = O\big(|\mathtt{G}(\hat{\zeta}) - \mathtt{G}(\zeta_{\mathsf{VaR}})|^{1/\gamma}\big)$, and we may focus on the convergence rate of $|\mathtt{G}(\hat{\zeta}) - \mathtt{G}(\zeta_{\mathsf{VaR}})|$ in the sequel. Let $\mathtt{G}_n(z) := n^{-1} \sum_{i=1}^n \mathbb{I}\{f(\mathbf{x}_i) \le z\}$ denote the empirical CDF of $f(X)$ and $\hat{\mathtt{G}}_n(z) := n^{-1} \sum_{i=1}^n \mathbb{I}\{\hat{f}(\mathbf{x}_i) \le z\}$ denote the empirical CDF of $\hat{f}(X)$. Then,

$$|\mathtt{G}(\hat{\zeta}) - \mathtt{G}(\zeta_{\mathsf{VaR}})| \le \underbrace{|\mathtt{G}(\hat{\zeta}) - \mathtt{G}_n(\hat{\zeta})|}_{V_1} + \underbrace{|\mathtt{G}_n(\hat{\zeta}) - \hat{\mathtt{G}}_n(\hat{\zeta})|}_{V_2} + \underbrace{|\hat{\mathtt{G}}_n(\hat{\zeta}) - \mathtt{G}(\zeta_{\mathsf{VaR}})|}_{V_3}.$$

The term V_1 can be bounded using the Dvoretzky–Kiefer–Wolfowitz inequality (Massart 1990), which bounds the difference between a CDF and its empirical counterpart and states that $\sup_z |\mathtt{G}(z) -$

$\mathtt{G}_n(z)| = O_{\mathbb{P}}(n^{-1/2})$. The term V_3 is also easy to handle; by definition, $\hat{\mathtt{G}}_n(\hat{\zeta}) = \frac{\lceil \tau n \rceil}{n}$ and $\mathtt{G}(\zeta_{\mathsf{VaR}}) = \tau$. Hence, $V_3 \leq n^{-1}$.

The analysis of the term V_2 is technically more involved. We will establish a (uniform) bound on $\sup_z |\mathtt{G}_n(z) - \hat{\mathtt{G}}_n(z)|$. This is done via the *chaining* method (Wainwright 2019, Chapter 5). It basically reduces the analysis of $\sup_z |\mathtt{G}_n(z) - \hat{\mathtt{G}}_n(z)|$ to that of the maximum of random variables over a finite set. We sketch the method as follows. First, we note that

$$\begin{aligned}|\mathtt{G}_n(z) - \hat{\mathtt{G}}_n(z)| &= \left| \frac{1}{n}\sum_{i=1}^n (\mathbb{I}\{f(\mathbf{x}_i) \leq z\} - \mathbb{I}\{\hat{f}(\mathbf{x}_i) \leq z\}) \right| \\ &\leq \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{f(\mathbf{x}_i) \in [z-\rho_n, z+\rho_n]\} = O_{\mathbb{P}}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{I}\{f(\mathbf{x}_i) \in [z-l_n, z+l_n]\} \right), \end{aligned} \tag{16}$$

where $\rho_n = \max_{1\leq i\leq n} |f(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)|$ and $\rho_n = O_{\mathbb{P}}(l_n)$ is the bound given by Proposition 3.

Second, we partition the domain of $\mathtt{G}(z)$ into M_n subintervals that are equally spaced and formed by the points $\{\zeta_j : j = 0, \ldots, M_n\}$, where M_n will be determined judiciously. Let $\delta_n \asymp M_n^{-1}$ denote the length of each subinterval. For any z, we define $j_*(z) = \arg\min_{0\leq j\leq M_n} |z - \zeta_j|$. Then,

$$\begin{aligned} &\mathbb{I}\{f(\mathbf{x}_i) \in [z-l_n, z+l_n]\} \\ \leq{}& \left|\mathbb{I}\{f(\mathbf{x}_i) \in [z-l_n, z+l_n]\} - \mathbb{I}\{f(\mathbf{x}_i) \in [\zeta_{j_*(z)} - l_n, \zeta_{j_*(z)} + l_n]\}\right| + \mathbb{I}\{f(\mathbf{x}_i) \in [\zeta_{j_*(z)} - l_n, \zeta_{j_*(z)} + l_n]\} \\ \leq{}& \mathbb{I}\{f(\mathbf{x}_i) \in [\zeta_{j_*(z)} - l_n - \delta_n, \zeta_{j_*(z)} + l_n + \delta_n]\} + \mathbb{I}\{f(\mathbf{x}_i) \in [\zeta_{j_*(z)} - l_n, \zeta_{j_*(z)} + l_n]\} \\ \leq{}& 2\,\mathbb{I}\{f(\mathbf{x}_i) \in [\zeta_{j_*(z)} - l_n - \delta_n, \zeta_{j_*(z)} + l_n + \delta_n]\}, \end{aligned} \tag{17}$$

where the second inequality holds because $|z - \zeta_{j_*(z)}| \leq \delta_n$. It follows from (16) and (17) that

$$\sup_z |\mathtt{G}_n(z) - \hat{\mathtt{G}}_n(z)| = O_{\mathbb{P}}\Bigg(\max_{0\leq j\leq M_n} \underbrace{\frac{2}{n}\sum_{i=1}^n \mathbb{I}\{f(\mathbf{x}_i) \in [\zeta_j - l_n - \delta_n, \zeta_j + l_n + \delta_n]\}}_{Q_{n,j}} \Bigg),$$

which provides a uniform bound through the maximum over a finite set of random variables.

The subsequent analysis is standard. We apply the union bound to the maximum in (EC.7.12), so that its tail probability can be bounded by the summation of the tail probability of each $Q_{n,j}$, which can then be handled with a Hoeffding-type inequality[9].

At last, by carefully specifying M_n, we may derive the convergence rate of $|\hat{\zeta} - \zeta_{\mathsf{VaR}}|$. Having completed the rate analysis for the case of VaR, it is not difficult to analyze the case of CVaR because the KRR-driven estimator for $\mathsf{CVaR}_\tau(f(X))$, given by (7), is calculated based on $\hat{\zeta}$.

[9] The indicator functions in $Q_{n,j}$ are i.i.d. Bernoulli random variables. However, the classic Hoefdding's inequality for bounded random variables is not sharp enough for our purpose. We use instead a refined version that is specific for Bernoulli distributions.

Theorem 4. *Let $\mathcal{T}(\cdot) = \mathsf{VaR}_\tau(\cdot)$ or $\mathcal{T}(\cdot) = \mathsf{CVaR}_\tau(\cdot)$ for some $\tau \in (0,1)$. Suppose Assumptions 1–3 and 5 hold. Then, $|\hat{\theta}_{n,m} - \theta| = O_{\mathbb{P}}(\Gamma^{-\kappa}(\log\Gamma)^{\tilde{\kappa}})$, where κ and $\tilde{\kappa}$ are specified as follows:*

(i) If $\nu \ge \frac{d}{\beta}$, then $\kappa = \frac{\nu\beta}{2\gamma(\nu+d)}$ and $\tilde{\kappa} = \frac{\beta}{2\gamma}$ by setting $n \asymp \Gamma$, $m \asymp 1$, and $\lambda \asymp \Gamma^{-1}$.

(ii) If $\nu < \frac{d}{\beta}$, then $\kappa = \frac{\beta}{2\gamma(\beta+1)}$ and $\tilde{\kappa} = \frac{\beta}{2\gamma}$ by setting $n \asymp \Gamma^{\frac{\beta}{\beta+1}}$, $m \asymp \Gamma^{\frac{1}{\beta+1}}$, and $\lambda \asymp \Gamma^{-1}$.

Theorem 4 indicates that a larger value of β or γ^{-1} leads to a faster convergence rate of $\hat{\theta}_{n,m}$ if $\mathcal{T}(\cdot) = \mathsf{VaR}_\tau(\cdot)$ or $\mathsf{CVaR}_\tau(\cdot)$. As discussed in Remark 4, β and γ^{-1} are interpreted as the parameters with which $\mathtt{G}$ and $\mathtt{G}^{-1}$ satisfy the Hölder condition, determining their degrees of smoothness. Hence, for the cases of VaR and CVaR, Theorem 4 reveals the dependence of the performance of the KRR-driven method on the smoothness of both the CDF of $f(X)$ and its quantile function, in addition to the known dependence on the smoothness of f and the dimensionality. Furthermore, because $\beta \le 1 \le \gamma$, which is implied implicitly by Assumption 5, the best scenario is $\beta = \gamma = 1$ (i.e., both $\mathtt{G}$ and $\mathtt{G}^{-1}$ are Lipschitz continuous). In this scenario, the convergence rate in Theorem 4 is the same as that in Theorem 3 with $\alpha = 1$; that is, the square root rate cannot be fully recovered, but it can be approached arbitrarily close to, as $\nu \to \infty$ (see Figure 1).

Remark 5. It can be shown via Taylor's expansion that if $\|\nabla f(\mathbf{x})\|$ is bounded below away from zero for all $\mathbf{x} \in \Omega$, then Assumption 5 is satisfied with $\beta = \gamma = 1$. However, if f has a stationary point $\mathbf{x}_0 \in \Omega$ (i.e., $\|f(\mathbf{x}_0)\| = 0$), then we may have $\beta < 1$ or $\gamma > 1$, yielding a slower convergence rate in Theorem 4. It turns out that our result can be enhanced in this scenario by virtue of "localization". For a given risk level τ and its associated VaR, $\zeta_{\mathsf{VaR}} = \mathsf{VaR}_\tau(f(X))$, if we suppose Assumption 5 holds *and* suppose there exist positive constants C_1', C_2', t_0, δ, and $\beta' \le \gamma'$ such that

$$C_2' t^{\gamma'} \le \mathbb{P}(|f(X) - z| \le t) \le C_1' t^{\beta'}, \quad \forall t \in (0, t_0], \forall z \in (\zeta_{\mathsf{VaR}} - \delta, \zeta_{\mathsf{VaR}} + \delta),$$

then a similar but refined analysis can establish the same result as in Theorem 4 with (β, γ) being replaced with (β', γ'). Therefore, even if Assumption 5 is satisfied with $\beta < 1$ or $\gamma > 1$, we may still have an improved rate result, provided that the inequalities in the assumption are satisfied with $\beta' \ge \beta$ and $\gamma' \le \gamma$ in a neighborhood of ζ_{VaR}. In particular, if $\|\nabla f(\mathbf{x})\|$ is bounded below away from zero for all $\mathbf{x}$ such that $|f(\mathbf{x}) - \zeta_{\mathsf{VaR}}| < \delta$, we have $\beta' = \gamma' = 1$.

Remark 6. As two standard risk measures that are widely used in practice, VaR and CVaR are constantly compared against each other with respect to both theoretical properties (Kou and Peng 2016) and computational efficiency (Jiang and Kou 2021). Theorem 4 contributes to the literature on the latter subject. It suggests that in the setting of KRR-driven nested simulation, both risk measures have the same convergence rates in terms of *absolute* errors. The conclusion may be different for *relative* errors, nevertheless, and we leave the investigation to future research.

Table 1 **The Convergence Rates $|\hat{\theta}_{n,m} - \theta| = O_{\mathbb{P}}(\Gamma^{-\kappa}(\log \Gamma)^{\tilde{\kappa}})$ in Theorems 1–4.**

$\mathcal{T}$	η	ν	κ	$\tilde{\kappa}$	m	λ
$\mathbb{E}[\eta(\cdot)]$	Smooth	$\nu \geq \frac{d}{2}$	$\frac{1}{2}$	0	1	Γ^{-1}
		$\nu < \frac{d}{2}$	$\frac{2\nu+d}{2\nu+3d}$	0	$\Gamma^{\frac{d-2\nu}{2\nu+3d}}$	$\Gamma^{-\frac{2(2\nu+d)}{2\nu+3d}}$
	Hockey-stick	$\nu \geq \frac{d}{\alpha+1}$	$\frac{1}{2} \wedge \frac{\nu(\alpha+1)}{2(\nu+d)}$	$\frac{\alpha+1}{2}\,\mathbb{I}\{\nu\alpha < d\}$	1	Γ^{-1}
		$\nu < \frac{d}{\alpha+1}$	$\frac{\alpha+1}{2(\alpha+2)}$	$\frac{\alpha+1}{2}$	$\Gamma^{\frac{1}{\alpha+2}}$	
	Indicator	$\nu \geq \frac{d}{\alpha}$	$\frac{\nu\alpha}{2(\nu+d)}$	$\frac{\alpha}{2}$	1	Γ^{-1}
		$\nu < \frac{d}{\alpha}$	$\frac{\alpha}{2(\alpha+1)}$	$\frac{\alpha}{2}$	$\Gamma^{\frac{1}{\alpha+1}}$	
VaR & CVaR		$\nu \geq \frac{d}{\beta}$	$\frac{\nu\beta}{2\gamma(\nu+d)}$	$\frac{\beta}{2\gamma}$	1	Γ^{-1}
		$\nu < \frac{d}{\beta}$	$\frac{\beta}{2\gamma(\beta+1)}$	$\frac{\beta}{2\gamma}$	$\Gamma^{\frac{1}{\beta+1}}$	

$0 < \alpha \leq 1$ and $0 < \beta \leq 1 \leq \gamma$.

4.4. Summary

We summarize in Table 1 the upper bounds on the convergence rates for the various forms of $\mathcal{T}$ (see Section 6.1 for a numerical examination regarding the tightness of these bounds).

First, there exist two thresholds with respect to the value of the smoothness parameter ν. One threshold determines whether the convergence rate of $\hat{\theta}_{n,m}$ exceeds $O_{\mathbb{P}}(\Gamma^{-1/3})$; that is, this threshold determines whether the use of KRR in the inner-level estimation is beneficial relative to the standard nested simulation. The other threshold determines whether the rate achieves $O_{\mathbb{P}}(\Gamma^{-1/2})$, thereby recovering the canonical rate for Monte Carlo simulation. The values of these two thresholds depend on the form of $\mathcal{T}$ and other relevant parameters. They can be easily calculated based on the results in Table 1 and are presented in Table 2 (see also Figure 1). For any given dimensionality d, the KRR-driven nested simulation enjoys a faster convergence rate than the standard nested simulation for most of the cases covered by our analysis, and in many cases it can even achieve or at least approach the square root rate, provided that ν is sufficiently large. This feature is different from previous studies in the literature that suggest the use of machine learning in nested simulation is beneficial only for low-dimensional ($d < 5$) problems. The difference stems from two facts about KRR. (i) The information about the smoothness of f allows us to postulate a proper function space to construct an estimate. (ii) It can leverage the spatial information in all the inner-level samples on a global scale.

Second, our results give refined guidelines with regard to the inner-level sample size. The existing literature suggests that when using parametric regression (Broadie et al. 2015) or kernel smoothing (Hong et al. 2017) in nested simulation, m should be fixed relative to the simulation budget Γ. We find, however, that this decision should depend on the smoothness of f. In general, m should be fixed if ν is sufficiently large; otherwise, it should grow properly without bound as Γ increases. The

Table 2 **Smoothness Thresholds for the Cubic and Square Root Convergence Rates.**

$\mathcal{T}$	η	(α,β,γ)	$\kappa \geq \frac{1}{3}$	$\kappa = \frac{1}{2}$
$\mathbb{E}[\eta(\cdot)]$	Smooth		$\nu > 0$	$\nu \geq \frac{d}{2}$
	Hockey-stick	$\alpha = 1$	$\nu > 0$	$\nu \geq d$
		$0 < \alpha < 1$	$\nu \geq \frac{2d}{3\alpha+1}$	$\nu \geq \frac{d}{\alpha}$
	Indicator	$\alpha = 1$	$\nu \geq 2d$	$\nu \to \infty$
		$\frac{2}{3} < \alpha < 1$	$\nu \geq \frac{2d}{3\alpha-2}$	n.a.
		$0 < \alpha \leq \frac{2}{3}$	n.a.	n.a.
VaR & CVaR		$\frac{\beta}{\gamma} = 1$	$\nu \geq 2d$	$\nu \to \infty$
		$\frac{2}{3} < \frac{\beta}{\gamma} < 1$	$\nu \geq \frac{2d}{3(\beta/\gamma)-2}$	n.a.
		$0 < \frac{\beta}{\gamma} \leq \frac{2}{3}$	n.a.	n.a.

$0 < \alpha \leq 1$ and $0 < \beta \leq 1 \leq \gamma$. The results hold if n, m, and λ are properly specified.

idea is that the accuracy in estimating less smooth functions is more sensitive to the sample noise, thereby requiring more replications from the inner-level simulation.

Third, the choice of the regularization parameter in all the cases that we consider is different from that when KRR is used in typical machine-learning tasks. It is known that if $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are i.i.d., $m \asymp 1$, and a Matérn kernel is used, then one should set $\lambda \asymp \Gamma^{-\frac{2\nu+d}{2(\nu+d)}}$, which is clearly different from our specifications in Table 1, in order to minimize $\mathbb{E}\,\|\hat{f} - f\|^2_{\mathcal{L}_2(\Omega)}$ (van de Geer 2000, Chapter 10). Indeed, using the standard choice of λ would lead to a significantly slower convergence rate of $\hat{\theta}_{n,m}$ in the setting of nested simulation. For example, it would lead to a rate of $\Gamma^{-\frac{2\nu+d}{4(\nu+d)}}$ if $\mathcal{T}(\cdot) = \mathbb{E}[\eta(\cdot)]$ with η being twice-differentiable with bounded first- and second-order derivatives, whereas Theorem 1 gives a rate of $\Gamma^{-\kappa}$ with $\kappa = \max(\frac{1}{2}, \frac{2\nu+d}{2\nu+3d})$. For other forms of $\mathcal{T}$, the standard choice of λ cannot even ensure the convergence of $\hat{\theta}_{n,m}$. Note that our choice of λ diminishes at a faster rate than the standard choice. Also note that with everything else the same, using a smaller value of λ in KRR results in a smaller bias but a larger variance in estimating f. Therefore, in the setting of nested simulation, it is more important to reduce the bias than to reduce the variance in the inner-level estimation to improve the estimation quality of θ.

5. $\mathcal{T}$-dependent Cross-validation

Given a simulation budget, the performance of the KRR-driven estimator depends critically on—in addition to the sample allocation rule—the selection of hyperparameters. First, the regularization parameter λ plays a significant role in light of the asymptotic analysis in Section 4. Moreover, for a given nested simulation problem we may not know precisely the smoothness of the unknown function f, and therefore it is common practice to treat ν as a hyperparameter (Salemi et al. 2019). The performance of our method may also depend on the value of ℓ that specifies the Matérn kernel (3). For notational simplicity, let $\Xi = (\lambda, \nu, \ell)$ denote the collection of these hyperparameters.

A standard approach for selecting hyperparameters of a machine-learning model is cross-validation. The basic idea is to divide the dataset $\mathcal{D} = \{(\mathbf{x}_i, \bar{y}_i) : i = 1, \ldots, n\}$ into two disjoint subsets. One subset is used, for a given value of Ξ, to train the machine-learning model $\hat{f}$ via some loss function L that measures the discrepancy between the predicted value $\hat{f}(\mathbf{x}_i)$ and the actual observation $\bar{y}_i$. (For example, $L(\hat{y}, y) = (\hat{y} - y)^2$ for many machine-learning models, including KRR.) The other subset is the validation set, and it is used to assess the said value of Ξ via the *same* loss function.

However, in the context of nested simulation, training KRR to estimate $f(\mathbf{x}) = \mathbb{E}[Y|X = \mathbf{x}]$ is merely an intermediate step. The eventual goal is to estimate $\theta = \mathcal{T}(f(X))$. The nonlinear functional $\mathcal{T}$ transforms the distribution of X to a scalar and in the process changes the relative importance of different regions of the domain of X. Thus, in order to align with the goal of estimating θ, the quality of $\hat{f}$ ought to be monitored on the validation set via a metric other than the mean squared error $|I^{\mathsf{Va}}|^{-1} \sum_{i \in I^{\mathsf{Va}}} (\hat{f}(\mathbf{x}_i) - \bar{y}_i)^2$, where I^{Va} is the set of indices for the validation set.

Instead, we measure the discrepancy between $\hat{f}(\mathbf{x}_i)$ and $\bar{y}_i$ via

$$\big(\hat{\theta}(\Xi, \mathcal{T}, I^{\mathsf{Tr}}, I^{\mathsf{Va}}) - \hat{\theta}^{\mathsf{St}}(\mathcal{T}, I^{\mathsf{Va}})\big)^2, \tag{18}$$

where I^{Tr} is the set of indices for the training set; moreover, $\hat{\theta}(\Xi, \mathcal{T}, I^{\mathsf{Tr}}, I^{\mathsf{Va}})$ and $\hat{\theta}^{\mathsf{St}}(\mathcal{T}, I^{\mathsf{Va}})$ are, respectively, the KRR-driven nested simulation estimator defined in (7) and the standard nested simulation estimator defined in (2), both of which are calculated using the data associated with I^{Va} (see Appendix A for their expressions). The notation also stresses the dependence on Ξ, $\mathcal{T}$, and I^{Tr}.

The metric (18) is interpreted as follows. When selecting Ξ, we seek to minimize the *generalization error*, meaning the expected error in estimating θ that is induced by the use of the KRR estimator $\hat{f}$ on unseen data points. Using the squared loss function to measure the discrepancy between $\hat{\theta}$ and θ—which is not the same as the discrepancy between $\hat{f}$ and f—the generalization error is

$$\mathbb{E}[(\hat{\theta} - \theta)^2 \,|\, \hat{f}] = \mathbb{E}\Big[\big(\mathcal{T}(\hat{f}(X)) - \mathcal{T}(f(X))\big)^2 \,\Big|\, \hat{f}\Big], \tag{19}$$

where the expectation is taken with respect to the distribution of X and $\hat{f}$ is taken as given. Hence, using the validation set, which is independent of the training set and thus independent of $\hat{f}$, we may estimate the generalization error (19) with its sample-based proxy, namely, the metric (18).

To reduce the variance for computing the metric (18), one may use the K-hold cross-validation setting. The dataset $\mathcal{D}$ is divided into K disjoint roughly equal-sized parts. One of them is taken as the validation set, while the remaining $K - 1$ parts are merged as the training set. The procedure is repeated K times, each time with a different part as the validation set. The performance of Ξ is evaluated via the average value of (18) over the K validation sets:

$$\mathrm{CV}(\Xi, \mathcal{T}) := \frac{1}{K} \sum_{l=1}^{K} \big(\hat{\theta}(\Xi, \mathcal{T}, \mathcal{D} \setminus I_l, I_l) - \hat{\theta}^{\mathsf{St}}(\mathcal{T}, I_l)\big)^2,$$

where I_l is the set of indices for the l-th part. At last, for a given functional $\mathcal{T}$, we determine the optimal value of Ξ by minimizing $\mathrm{CV}(\Xi,\mathcal{T})$. Namely, Ξ is selected in a $\mathcal{T}$-dependent manner.

In general, the larger K, the higher the computational cost of K-fold cross-validation because one needs to train the machine-learning model of concern K times, each on a different dataset. However, when the machine-learning method involved is KRR, the extreme case that $K=n$—that is, leave-one-out (LOO) cross-validation—can make use of a well-known trick that greatly simplifies the expression of $\mathrm{CV}(\Xi,\mathcal{T})$, thereby leading to substantial savings in computational cost relative to the case of a smaller K (see Appendix B for details). We use the LOO cross-validation setting in the numerical experiments in Section 6.

Remark 7. To see the connection between the $\mathcal{T}$-dependent cross-validation and the standard practice, let us consider the special case where $\mathcal{T}(\cdot)=\mathbb{E}[\cdot]$ and $K=n$. Then, $\theta=\mathbb{E}[Y]$ and $I_l=\{\mathbf{x}_l\}$; moreover, it is straightforward to derive from their expressions in Appendix A that

$$\hat{\theta}(\Xi,\mathcal{T},\mathcal{D}\setminus I_l,I_l)=\hat{f}^{-l}(\mathbf{x}_l;\Xi)\quad\text{and}\quad\hat{\theta}^{\mathsf{St}}(\mathcal{T},I_l)=\bar{y}_l,$$

where $\hat{f}^{-l}(\cdot;\Xi)$ denotes the KRR estimator trained with $\mathbf{x}_l$ removed from $\mathcal{D}$ and with the hyperparameters being Ξ. Therefore, $\mathrm{CV}(\Xi,\mathcal{T})$ is reduced to the LOO cross-validation with the squared loss between $\hat{f}$ and f as the measure of fit.

Remark 8. A potential benefit of the standard (i.e., $\mathcal{T}$-independent) cross-validation is that one only need to calibrate the hyperparameter once. By contrast, the $\mathcal{T}$-dependent cross-validation requires recalibration of the hyperparameter for each different $\mathcal{T}$ (e.g., different levels of VaR/CVaR), and therefore it incurs more computational cost. Nevertheless, we find in our numerical experiments that the $\mathcal{T}$-dependent cross-validation consistently outperforms the standard cross-validation by a substantial margin, which outweighs the computational overheads.

Remark 9. In typical nested simulation applications, $\mathrm{CV}(\Xi,\mathcal{T})$ lacks analytical tractability and its evaluation is computationally expensive. Moreover, the multidimensionality of Ξ renders the grid-search strategy computationally infeasible. Therefore, we view the task of minimizing $\mathrm{CV}(\Xi,\mathcal{T})$ as a black-box optimization problem and apply the Bayesian optimization framework (Frazier 2018) to find a good solution.

6. Numerical Experiments

We now numerically evaluate the KRR-driven estimator via a sequence of experiments. In Section 6.1, we study the mitigating effect of smoothness on the curse of dimensionality using test functions with known smoothness. Practical examples that arise from portfolio risk management and input uncertainty quantification are discussed in Section 6.2 and Section 6.3, respectively.

6.1. Smoothness versus Dimensionality

The asymptotic analysis in Section 4 provides upper bounds on the convergence rate of the KRR-driven estimator for various forms of the functional $\mathcal{T}$. These upper bounds reveal the mitigating effect of the smoothness on the curse of dimensionality. We numerically examine such an effect, as well as how tight these upper bounds are, using test functions with known smoothness. Let $f(\mathbf{x}) = \mathbb{E}[Y|X = \mathbf{x}]$ be the conditional expectation in the inner level. We assume

$$f(\mathbf{x}) = \sum_{i=1}^{N} c_i \Psi(\mathbf{x} - \tilde{\mathbf{x}}_i), \quad \mathbf{x} \in \Omega, \tag{20}$$

where $N = 10$, $\Omega = [-1,1]^d$, and Ψ is defined by (3) in which $\ell = 1$ and ν is specified later. Both c_i and $\tilde{\mathbf{x}}_i$ are randomly generated—with the former from $\mathsf{Uniform}[20,50]$ (i.e., the uniform distribution on $[20,50]$) and the latter from $\mathsf{Normal}(0,0.1)$ (i.e., the normal distribution with mean zero and variance 0.1)—and then fixed. According to the definition of RKHSs, f is a function in $\mathcal{N}_\Psi(\Omega)$ and has a smoothness parameter ν.

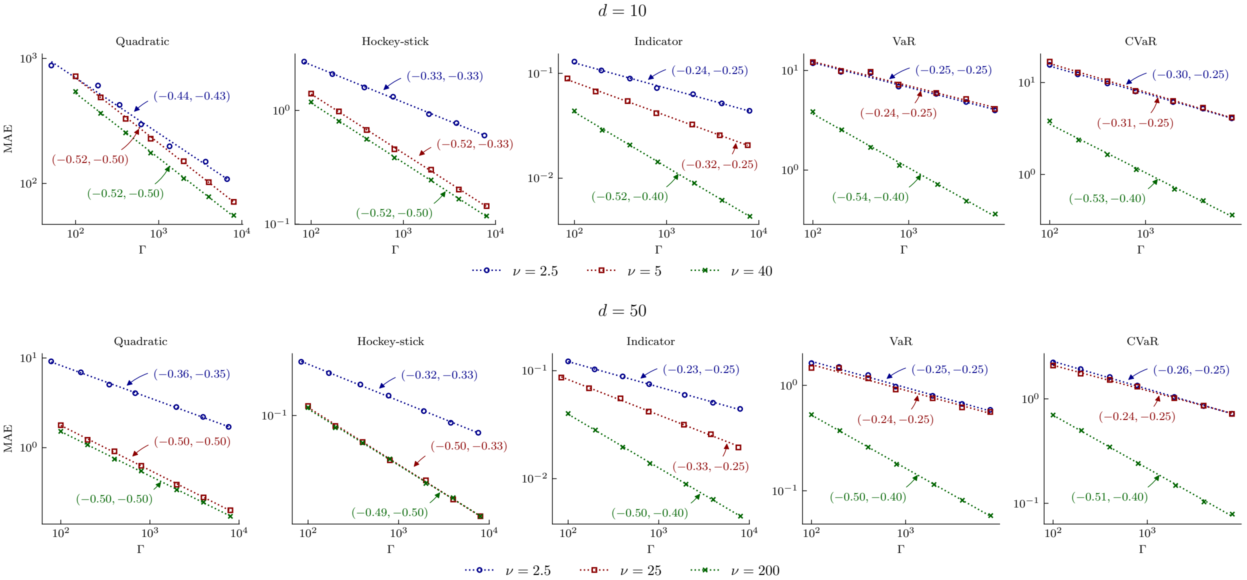
We generate each outer-level scenario $\mathbf{x} = (x_1, \ldots, x_d)$ as follows. For each $l = 1, \ldots, d$, x_l independently follows a truncated normal distribution, $\mathsf{Normal}(0,0.1)$ with truncated range $[-1,1]$. Given an outer-level scenario $\mathbf{x}_i$, the inner-level samples are simulated via $y_{ij} = f(\mathbf{x}_i) + \varepsilon_{ij}$, where ε_{ij}'s are i.i.d. $\mathsf{Normal}(0,0.1)$ random variables.

We examine five different forms of $\mathcal{T}$: $\mathcal{T}(\cdot) = \mathbb{E}[\eta(\cdot)]$ with $\eta(z) = z^2$, $\eta(z) = (z - z_0)^+$, and $\eta(z) = \mathbb{I}\{z \geq z_0\}$, $\mathcal{T}(\cdot) = \mathsf{VaR}_\tau(\cdot)$, and $\mathcal{T}(\cdot) = \mathsf{CVaR}_\tau(\cdot)$. Here, the threshold z_0 is set to be the median of the distribution of $f(\mathbf{x})$, and the risk level τ is set to be 95%. The true value of $\theta = \mathcal{T}(f(X))$ is estimated based on 10^8 i.i.d. copies of X. The budget allocation (n,m) and the regularization parameter λ are specified exactly according to our asymptotic results in Table 1 *without tuning*. For each $d \in \{10, 50\}$, $\nu \in \{5/2, d/2, 4d\}$, and a budget Γ ranging from 10^2 to 10^4, we estimate the mean absolute error (MAE), defined as $|\hat{\theta}_{n,m} - \theta|$, of the KRR-driven method using 1,000 macro-replications. In Figure 4, we plot the MAE against the budget on a logarithmic scale and compare the slope of the resulting line with the parameter κ in the upper bound $|\hat{\theta}_{n,m} - \theta| = O_{\mathbb{P}}(\Gamma^{-\kappa}(\log\Gamma)^{\tilde{\kappa}})$, which is specified in Table 1 (it is easy to show that Assumptions 1–3 hold, Assumption 4 holds with $\alpha = 1$, and Assumption 5 holds with $\beta = \gamma = 1$).

Looking at Figure 4, we have several observations. First, the mitigating effect of the smoothness on the curse of dimensionality is clearly demonstrated. Everything being equal, a higher smoothness yields a faster decay rate of the MAE. For example, in the case that $\mathcal{T} = \mathsf{VaR}$, the actual decay rate of the MAE is approximately $\Gamma^{-1/4}$ if $\nu = 5/2$ or $\nu = d/2$, whereas it is approximately $\Gamma^{-1/2}$ if $\nu = 4d$.

Second, the slope of each dashed line, which describes the trend (on a logarithmic scale) of the MAE as Γ grows, is no greater than to $-\kappa$. (In many cases, the slope is basically identical to

Figure 4 MAEs for the KRR-driven Nested Simulation.



Note. The data are plotted on a logarithmic scale. Each dashed line is drawn via linear regression between log(MAE) and log(Γ). The numbers in each set of parentheses are, respectively, the estimated slope and the value of $-\kappa$.

$-\kappa$.) This is consistent with our theoretical results in Theorems 1–4 asserting that the MAE is asymptotically upper bounded by $\Gamma^{-\kappa}(\log\Gamma)^{\tilde{\kappa}}$.

Third, in some cases (e.g., $d=50$, $\nu=d/2$, and $\mathcal{T}(\cdot)=\mathbb{E}[\eta(\cdot)]$ with η being a hockey-stick function), the slope of a dashed line does not match $-\kappa$. Nevertheless, we emphasize that our upper bounds hold for *all* functions f in the RKHS $\mathcal{N}_\Psi(\Omega)$, while the results in Figure 4 are computed for a *particular* function f defined by (20). Therefore, although it suggests that there *may* be room for improvement in our upper bounds, the mismatch is not conclusive in itself. To fully address the question regarding the tightness of our upper bounds may require minimax lower bounds (Györfi et al. 2002, Chapter 3) on the convergence rate. We leave that investigation to future research.

6.2. Portfolio Risk Management

Consider a portfolio consisting of financial derivatives that derive their values from the performance of q underlying assets. Suppose that the constituent derivatives share the same expiration date T. The manager of the portfolio is interested in assessing the risk at some future time $T_0 < T$ in terms of five metrics: expected quadratic loss $\mathbb{E}[Z^2]$, expected excess loss $\mathbb{E}[(Z-z_0)^+]$, probability of a large loss $\mathbb{P}(Z \ge z_0)$, and VaR and CVaR of Z at some risk level τ, where Z is the portfolio loss at time T_0 and z_0 is some threshold. The first three of these metrics correspond to $\mathcal{T}(\cdot)=\mathbb{E}[\eta(\cdot)]$ with $\eta(z)=z^2$, $\eta(z)=(z-z_0)^+$, and $\eta(z)=\mathbb{I}\{z\ge z_0\}$, respectively.

Let $\mathbf{S}(t)=(S_1(t),\dots,S_q(t))$ denote the vector of the asset prices at time t. In the framework of nested simulation, we simulate $\mathbf{S}(t)$ up to time T_0 and the sample paths constitute outer-level scenarios. Then, in the inner-level simulation, we simulate $\mathbf{S}(t)$ from T_0 to T to estimate the prices of the options in the portfolio. A subtlety here is that the probability measures for the outer- and inner-level simulations are different; the former uses the real-world measure, whereas the latter uses a risk-neutral measure. Suppose $\mathbf{S}(t)$ follows a q-dimensional geometric Brownian motion (GBM):

$$\frac{\mathrm{d}S_i(t)}{S_i(t)} = \mu_i \mathrm{d}t + \sum_{i=1}^{q} \sigma_{ij} \mathrm{d}B_j(t), \quad i = 1, \dots, q,$$

where $B_1(t), \dots, B_q(t)$ are independent standard one-dimensional Brownian motions, the matrix $(\sigma_{ij})_{i,j=1}^{q}$ is lower-triangular (i.e., $\sigma_{ij} = 0$ for all $i > j$), and μ_i takes different values depending on which probability measure is used. We assume, for simplicity, that each underlying asset has the same return μ (i.e., $\mu_i = \mu$ for all i) under the real-world measure. However, under the risk-neutral measure, it should be identical to the risk-free interest rate r (i.e., $\mu_i = r$ for all i).

In the portfolio, there are three geometric Asian call options with discrete monitoring and three up-and-out barrier call options written on each underlying asset (so the portfolio consists of $6q$ options in total). Note that both types of options have path-dependent payoffs. For an underlying asset with price $S(t)$, the payoff of a geometric Asian call option with monitoring at the times $0 = t_0 < t_1 < \cdots < t_M = T$ is $\big(\big(\prod_{k=1}^{M} S(t_k)\big)^{1/M} - K\big)^+$, where K is the strike price, and the payoff of an up-and-out barrier option with barrier H is $(S(T) - K)^+ \mathbb{I}\{\max_{0 \le t \le T} S(t) \le H\}$. We assume that the strike prices of the three Asian options written on each asset are K_1, K_2, and K_3, and the same holds for the three barrier options; moreover, these barrier options have the same barrier H. We also assume that the risk horizon $T_0 = t_{M_0}$ for some $M_0 = 1, \dots, M$. According to the theory of derivatives pricing (Glasserman 2003, Chapter 1), the value of each option at time T_0 equals its expected discounted payoff. Therefore, the value of the portfolio at time T_0 is

$$V_{T_0}(X) = \mathbb{E}\Bigg[\underbrace{e^{-r(T-T_0)} \sum_{i=1}^{q} \sum_{l=1}^{3} \Bigg(\Big(\prod_{k=1}^{M} S_i(t_k)\Big)^{1/M} - K_l\Bigg)^+ + (S_i(T) - K_l)^+ \mathbb{I}\Big\{\max_{0 \le t \le T} S_i(t) \le H\Big\}}_{W} \Bigg| X\Bigg],$$

where $X \in \mathbb{R}^{3q}$ is a vector of risk factors defined as

$$X = \Bigg(S_1(T_0), \dots, S_q(T_0), \Big(\prod_{k=1}^{M_0} S_1(t_k)\Big)^{1/M_0}, \dots, \Big(\prod_{k=1}^{M_0} S_q(t_k)\Big)^{1/M_0}, \max_{0 \le t \le T_0} S_1(t), \dots, \max_{0 \le t \le T_0} S_q(t)\Bigg).$$

The portfolio loss at time T_0 is then $Z := V_0 - V_{T_0}(X)$, where V_0 is the value of the portfolio at time 0 and is calculated in the same way as V_{T_0}, except with $T_0 = 0$. Further, it can be expressed as $Z = \mathbb{E}[Y|X]$, where $Y = V_0 - W$.

To assess the performance of a nested simulation method, we need to compute the true value of θ, which is done as follows. Under the assumption that $\mathbf{S}(t)$ follows a GBM, $V_{T_0}(X)$ can be calculated in closed form (see, e.g., Haug 2007, Chapter 4). We generate 10^8 i.i.d. copies of X and calculate the corresponding values of $V_0 - V_{T_0}(X)$, which are i.i.d. copies of Z and can be used to accurately approximate $\theta = \mathcal{T}(Z)$. We compare different methods based on the relative root mean squared error (RRMSE) defined as the ratio of the MSE to the true value of θ. For each problem instance, the RRMSE is estimated via 1,000 macro-replications.

The other parameters involved are specified as follows. The expiration date of each option in the portfolio is $T = 1$, and the risk horizon is $T_0 = 3/50$. In the GBM model, the initial price of each asset is $S_i(0) = 100$ for all $i = 1, \ldots, q$, the return of each asset under the real-world measure is $\mu = 8\%$, the risk-free rate is $r = 5\%$, and the volatility term σ_{ij} is generated randomly (and then fixed) if $i \le j$ and is $\sigma_{ij} = 0$ otherwise. We vary $q \in \{10, 20, 50, 100\}$ (note that the dimensionality of the conditioning variable X is $d = 3q$)[10]. For the options in the portfolio, the three different strike prices are $K_1 = 90$, $K_2 = 100$, and $K_3 = 110$. Moreover, the monitoring times of each Asian option are $\{t_k = kT/M : k = 1, \ldots, M\}$ with $M = 50$, and the barrier of each barrier option is $H = 150$.

Because exact simulation of the running maximum $\max_{0 \le t \le T} S_i(t)$ is unavailable, the payoffs of the barrier options are simulated as follows. We first perform the Euler discretization scheme with 200 time steps, evenly spaced on $[0, T]$, to generate sample paths of $S_i(t)$. We then adopt the commonly used Brownian interpolation approach to correct for the barrier crossing-probabilities (see Glasserman 2003, Chapter 6.4 for details).

Given a simulation budget $\Gamma = 10^4$ or $\Gamma = 10^5$, we compare three methods for nested simulation:

(i) The standard method. We try 10 different values of the inner-level sample size m ranging from 5 to 2,000 and report the best performance among them[11].

(ii) The KRR-driven method. Similarly, we try different values of m and report the best performance. Given Γ and m, we apply the LOO $\mathcal{T}$-dependent cross-validation to select the hyperparameters $\lambda \in (0, 0.1)$, $\nu \in (1/2, 4d)$, and $\ell \in \{10^k : -3 \le k \le 3\}$.

(iii) The regression-driven method. This method is similar to the KRR-driven method, except that the conditional expectation $\mathbb{E}[Y|X]$ is modeled as a linear combination of basis functions in X. In addition to different values of m, we also try different sets of basis functions, including polynomials and orthogonal polynomials (the Legendre, Laguerre, Hermite, and Chebyshev polynomials) up to order 5 and without interaction terms. We report the best performance among all the combinations of different values of m and sets of basis functions.

[10] As pointed out by Hong et al. (2017), taking advantage of the additive structure of $V_{T_0}(X)$ and the fact that each option is written on a single underlying asset, one may decompose this $3q$-dimensional problem into $6q$ two-dimensional sub-problems—because the portfolio is comprised of $6q$ options, each having two risk factors—and apply the KRR-driven method to each sub-problem separately. We do not pursue this idea in this paper, relegating the analysis under both the additivity and smoothness assumptions to future research.

[11] The optimal value of m is generally unknown and problem-dependent. Zhang et al. (2021) develop a bootstrap-based approach to estimate the optimal value of m using a small proportion of the total budget. They report (and we confirm in our experiments) that the performance of the bootstrap-based approach is often close to, but slightly worse than, the best performance among those associated with a wide range of values of m.

Table 3 RRMSE (%) for the Portfolio Risk Management Problem.

	$\mathcal{T}$	KRR		Regression		Standard	
		$\Gamma = 10^4$	$\Gamma = 10^5$	$\Gamma = 10^4$	$\Gamma = 10^5$	$\Gamma = 10^4$	$\Gamma = 10^5$
$q = 10$	Quadratic	3.76	1.26	7.29	2.58	19.24	8.43
	Hockey-stick	3.27	1.10	6.28	1.99	15.64	6.91
	Indicator	2.47	0.77	4.05	2.45	5.73	2.63
	VaR	2.57	0.87	8.23	5.43	10.11	5.37
	CVaR	2.63	0.87	9.43	6.61	11.27	5.56
$q = 20$	Quadratic	3.36	1.15	8.36	2.61	18.81	8.37
	Hockey-stick	3.88	1.19	8.56	2.30	19.87	9.01
	Indicator	2.84	0.89	4.33	1.91	8.17	3.90
	VaR	3.21	1.05	7.03	2.82	11.95	6.48
	CVaR	3.33	1.09	7.75	3.35	13.64	6.84
$q = 50$	Quadratic	3.33	1.04	13.45	2.19	18.57	7.78
	Hockey-stick	4.03	1.29	13.15	2.65	21.54	9.72
	Indicator	3.02	0.98	4.46	1.65	10.10	4.93
	VaR	3.34	1.06	9.57	2.83	12.09	6.23
	CVaR	3.42	1.08	10.14	3.24	14.21	6.50
$q = 100$	Quadratic	3.26	1.05	24.83	2.54	18.26	8.08
	Hockey-stick	4.87	1.59	27.75	3.87	25.02	10.95
	Indicator	3.74	1.19	9.29	1.75	13.43	6.11
	VaR	3.41	1.16	15.33	2.89	12.21	6.27
	CVaR	3.56	1.19	15.56	3.21	13.40	6.82

The dimensionality of the conditioning variable is $d = 3q$. For hockey-stick and indicator functions, the threshold is $z_0 = 0.02V_0$. For VaR and CVaR, the risk level is $\tau = 99\%$.

The numerical results are presented in Table 3. First, the KRR-driven method consistently outperforms the other two methods by a significant margin. In some cases, the RRMSEs for the former with a budget $\Gamma = 10^4$ are even comparable to those for the latter with a budget $\Gamma = 10^5$. For example, when $q = 10$ and $\mathcal{T}(Z) = \mathbb{E}[\mathbb{I}\{Z \geq z_0\}]$, the RRMSE for the KRR-driven method with $\Gamma = 10^4$ is 2.47%, whereas it is 2.45% for the regression-driven method and 2.63% for the standard method even a 10-times larger budget ($\Gamma = 10^5$).

Second, the dimensionality does not significantly affect the performance of the KRR-driven method. As q increases from 10 to 100 (i.e., d increases from 30 to 300), the RRMSEs that the KRR-driven method achieves and do not grow rapidly, remaining at a relatively low level compared to the other two methods. For example, when $q = 100$, $\mathcal{T}$ is VaR, and $\Gamma = 10^5$, the RRMSEs for the three methods are 1.16% (KRR-driven), 2.89% (regression-driven), and 6.27% (standard), respectively. This may be attributed to the fact that the estimated value of ν is large, indicating that $\mathbb{E}[Y|X]$ is highly smooth with respect to X.

Last, although it is not the focus of the present paper, we stress that the superior performance of the KRR-driven method is achieved with a higher computational cost than the other two methods. This is because (i) it requires the selection of hyperparameters, and (ii) its computation involves numerical inversion of $n \times n$ matrices, an intensive computational task for large n. This makes the KRR-driven method—in its current form—possibly unfit for situations where more than $n = 10^5$ outer-level scenarios are simulated. Kernel approximation methods (Rasmussen and Williams 2006, Chapter 8) can be used to alleviate the computational burden, but they require a careful design to balance computational speed and estimation accuracy, which is beyond the scope of this paper.

6.3. Input Uncertainty Quantification

Consider a newsvendor problem with multiple products. For product $i = 1, \dots, d$, let p_i denote its unit selling price, c_i its unit procurement cost, and D_i its demand. Given a price vector $(p_1, \dots, p_d)$, the demand of these products is driven by a multinomial logit choice model (Aydin and Porteus 2008): $D_i = v_i \varepsilon_i$, where $v_i = e^{\alpha_i - p_i}/(1 + \sum_{i=1}^d e^{\alpha_i - p_i})$, and ε_i is an independent $\mathsf{Uniform}[a, b]$ random variable. The parameter α_i can be interpreted as a customer's expected utility for product i, and it needs to be estimated from real data, which gives rise to the issue of input uncertainty. We assume that the distribution of α_i is $\mathsf{Normal}(\mu_i, \sigma_i^2)$. Denote $X = (\alpha_1, \dots, \alpha_d)$.

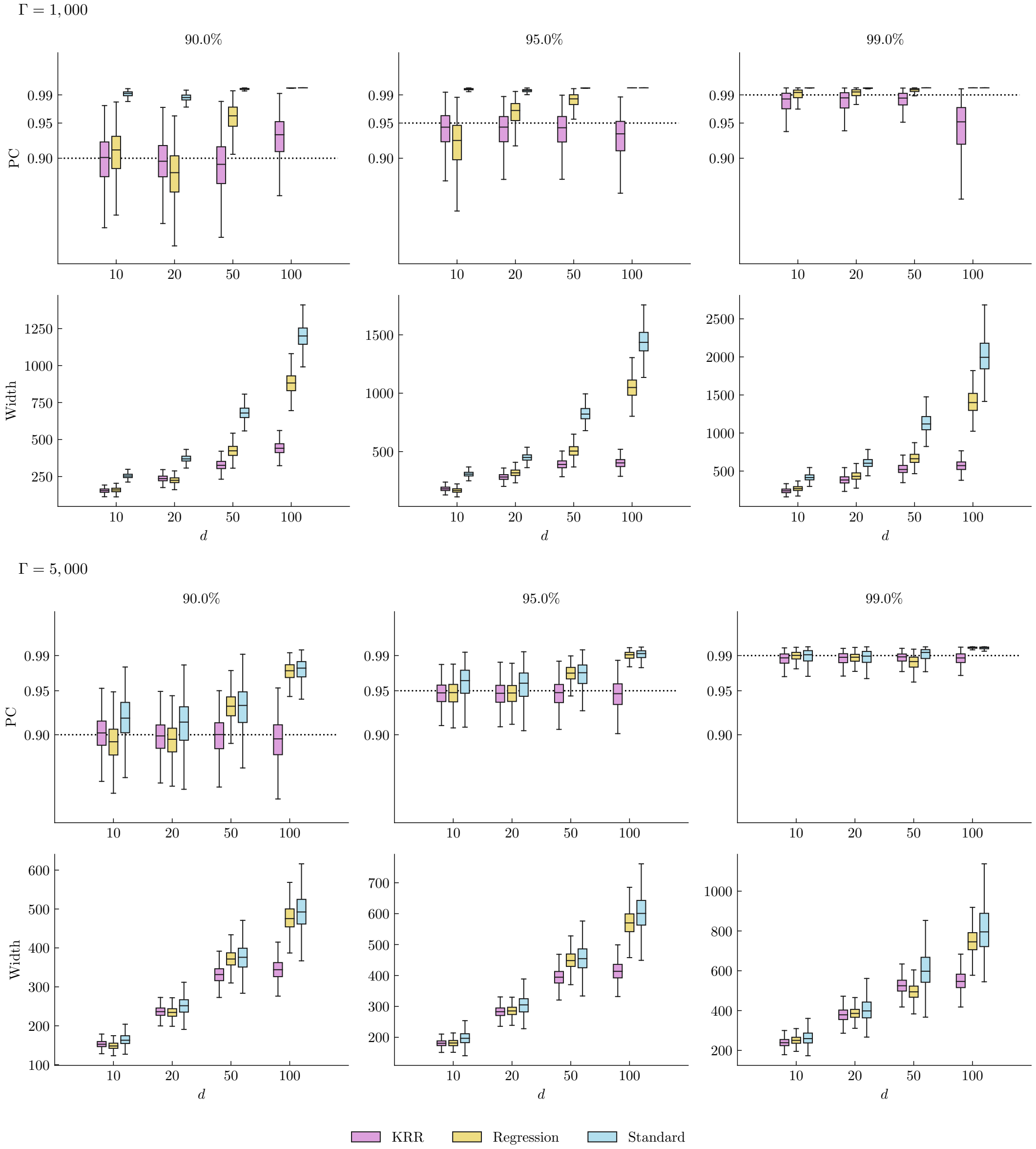
Let q_i denote the order quantity of product i. Given a realized value of X, the newsvendor's objective is to maximize its expected profit $\mathbb{E}[\sum_{i=1}^d p_i(D_i \wedge q_i) - c_i q_i]$ by optimizing the order quantities. It is easy to show that the optimal order quantities are $q_i^* = F_{\varepsilon_i}^{-1}\left(\frac{p_i - c_i}{p_i}\right) v_i = \left((b-a)\left(\frac{p_i - c_i}{p_i}\right) + a\right) v_i$, where $F_{\varepsilon_i}^{-1}(\cdot)$ is the inverse CDF (i.e., the quantile function) of ε_i. Thus, the optimal profit is

$$Z := \mathbb{E}\left[\underbrace{\sum_{i=1}^d p_i(D_i \wedge q_i^*) - c_i q_i^*}_{Y} \,\middle|\, X\right].$$

To quantify the impact of input uncertainty (i.e., the uncertainty about α_i's) on the calculation of the newsvendor's optimal profit, we use nested simulation to construct $100(1-\tau)\%$ credible intervals (CrIs) of the form $[\mathsf{VaR}_{\tau/2}(Z), \mathsf{VaR}_{1-\tau/2}(Z)]$ for the distribution of Z. We assess the performance of a nested simulation method using both the probability content (PC) and the width of the CrI that the method constructs. Both of them are estimated based on 1,000 macro-replications. In each macro-replication, we construct a CrI, say (l, u), for a particular method and estimate its PC as follows. We generate 10^8 i.i.d. copies of X and calculate the corresponding values of Z (which can be done in closed form). We then use these copies of Z to accurately approximate $\mathbb{P}(Z \in (l, u))$. We vary $d \in \{10, 20, 50, 100\}$ and $\tau \in \{0.1, 0.05, 0.01\}$. The other parameters are specified as follows: $a = 100$, $b = 500$, $p_i = 0.2i + 3$, $c_i = 2$, $\mu_i = 0.3i + 5$, and $\sigma_i = 1$ for $i = 1, \dots, d$.

Given a simulation budget $\Gamma = 1{,}000$ or $\Gamma = 5{,}000$, similar to Section 6.2, for each of the three methods (standard, KRR-driven, and regression-driven) we report the best performance among different values of the inner-level sample size m (and different sets of basis functions in the case of the regression-driven method). The numerical results are presented in Figure 5.

Figure 5 Statistics (PC and Width) of CrIs for the Input Uncertainty Quantification Problem.



Note. The boxplots are made based on 1,000 macro-replications. The percentages are the CrIs' credible levels.

First, if the budget is small ($\Gamma = 1,000$), the standard method basically fails; the CrIs it constructed are all excessively wide and have significant over-coverage[12]. The regression-driven method produces accurate CrIs, having a PC close to the corresponding nominal level for $d = 10, 20$; however, it suffers from a severe over-coverage issue in higher dimensions. Compared to the regression-driven method, the performance of the KRR-driven method is similar for $d = 10, 20$ and somewhat better for $d = 50$. For $d = 100$, albeit showing clear under-coverage, the CrIs constructed by the KRR-driven method are significantly better—in terms of width—than those constructed by the other two methods.

[12] The notion of "over-coverage" here means that the PC of a CrI is larger than the nominal value $100(1-\tau)\%$. This is different from the over-coverage of a confidence interval from a frequentist perspective.

Second, if the budget is large ($\Gamma = 5,000$), the standard and the regression-driven methods have comparable performances. Both can construct accurate CrIs for $d = 10, 20$. However, their performances begin to deteriorate for $d = 50$ and become unacceptable for $d = 100$. By contrast, the KRR-driven method consistently produce accurate CrIs. For example, for $d = 100$, the 95% CrIs constructed by the KRR-driven, regression-driven, and standard methods have an estimated PC of 94.49%, 98.98%, and 99.05%, respectively. This shows that the KRR-driven method is a viable option for high-dimensional settings, which aligns with the experiment results in Section 6.2.

7. Concluding Remarks

In this paper, we develop a new method based on KRR for nested simulation. We show that for various forms of nested simulation, the new method may recover (or at least approach) the square root convergence rate—that is, the canonical rate for the standard Monte Carlo simulation—provided that the conditional expectation as a function of the conditioning variable is sufficiently smooth with respect to its dimensionality. In other words, the new method can substantially alleviate the curse of dimensionality by exploiting the smoothness. Our theoretical framework for convergence rate analysis is general. Not only does it incorporate different forms of nested simulation, but it may also be applied to examine the use of kernels other than the Matérn class in KRR or even the use of machine-learning methods other than KRR in nested simulation.

Our work can be extended in several ways. First, the numerical experiments in Section 6.1 suggest that there may exist room for improvement in our upper bounds on the convergence rates. It is interesting, albeit challenging, to identify lower bounds on the convergence rates. They would address the question as to whether our upper bounds are optimal in a minimax sense.

Second, our theory prescribes rules for budget allocation (n, m) in an asymptotic sense. Although our numerical experiments suggest that these asymptotic rules indeed serve as a good guideline, a more refined rule may further facilitate the use of the KRR-driven method. Given the asymptotic orders of magnitude of m and n, one might adopt the idea of Zhang et al. (2021) to use a small proportion of the total simulation budget to compute a bootstrap-based estimate of the leading constants in the asymptotic orders. This would conceivably yield a reasonable choice of (n, m).

Last, the computation of KRR involves matrix inversion, and it may become computationally challenging as the matrix size—which equals the outer-level sample size—grows. Numerous approximation methods (Zhang et al. 2015, Lu et al. 2020) have been developed to address this issue. These methods are designed to strike a balance between the approximation's computational efficiency and KRR's prediction accuracy. To use them in the KRR-driven nested simulation, however, this kind of balance needs to be adjusted because KRR's prediction accuracy is measured differently in nested simulation relative to typical machine-learning tasks. The adjustment would make the KRR-driven method applicable for large-scale problems.

Appendix. LOO $\mathcal{T}$-dependent Cross-validation

A. Discrepancy Metric

We present the expressions of $\hat{\theta}^{\mathsf{St}}(\mathcal{T}, I^{\mathsf{Va}})$ and $\hat{\theta}(\Xi, \mathcal{T}, I^{\mathsf{Tr}}, I^{\mathsf{Va}})$ in the metric (18) as follows. Let $n^{\mathsf{Va}} = |I^{\mathsf{Va}}|$ denote the size of the set I^{Va}. Then,

$$\hat{\theta}^{\mathsf{St}}(\mathcal{T}, I^{\mathsf{Va}}) := \begin{cases} \dfrac{1}{n^{\mathsf{Va}}} \displaystyle\sum_{i \in I^{\mathsf{Va}}} \eta(\bar{y}_i), & \text{if } \mathcal{T}(\cdot) = \mathbb{E}[\eta(\cdot)], \\ \bar{y}_{(\lceil \tau n^{\mathsf{Va}} \rceil)}, & \text{if } \mathcal{T}(\cdot) = \mathsf{VaR}_\tau(\cdot), \\ \bar{y}_{(\lceil \tau n^{\mathsf{Va}} \rceil)} + \dfrac{1}{(1-\tau)n^{\mathsf{Va}}} \displaystyle\sum_{i \in I^{\mathsf{Va}}} (\bar{y}_i - \bar{y}_{(\lceil \tau n^{\mathsf{Va}} \rceil)})^+, & \text{if } \mathcal{T}(\cdot) = \mathsf{CVaR}_\tau(\cdot), \end{cases} \tag{21}$$

where $\bar{y}_{(1)} \le \cdots \le \bar{y}_{(n^{\mathsf{Va}})}$ denote the order statistics of $\{\bar{y}_i : i \in I^{\mathsf{Va}}\}$.

In addition, let $\hat{f}^{\Xi, I^{\mathsf{Tr}}}$ denote the KRR estimator trained using the hyperparameters Ξ and the data associated with I^{Tr}. Then,

$$\hat{\theta}(\Xi, \mathcal{T}, I^{\mathsf{Tr}}, I^{\mathsf{Va}}) := \begin{cases} \dfrac{1}{n^{\mathsf{Va}}} \displaystyle\sum_{i \in I^{\mathsf{Va}}} \eta\Big(\hat{f}^{\Xi, I^{\mathsf{Tr}}}(\mathbf{x}_i)\Big), & \text{if } \mathcal{T}(\cdot) = \mathbb{E}[\eta(\cdot)], \\ \hat{f}^{\Xi, I^{\mathsf{Tr}}}_{(\lceil \tau n^{\mathsf{Va}} \rceil)}, & \text{if } \mathcal{T}(\cdot) = \mathsf{VaR}_\tau(\cdot), \\ \hat{f}^{\Xi, I^{\mathsf{Tr}}}_{(\lceil \tau n^{\mathsf{Va}} \rceil)} + \dfrac{1}{(1-\tau)n^{\mathsf{Va}}} \displaystyle\sum_{i \in I^{\mathsf{Va}}} \Big(\hat{f}^{\Xi, I^{\mathsf{Tr}}}(\mathbf{x}_i) - \hat{f}^{\Xi, I^{\mathsf{Tr}}}_{(\lceil \tau n^{\mathsf{Va}} \rceil)}\Big)^+, & \text{if } \mathcal{T}(\cdot) = \mathsf{CVaR}_\tau(\cdot), \end{cases} \tag{22}$$

where $\hat{f}^{\Xi, I^{\mathsf{Tr}}}_{(1)} \le \cdots \le \hat{f}^{\Xi, I^{\mathsf{Tr}}}_{(n^{\mathsf{Va}})}$ denote the order statistics of $\{\hat{f}^{\Xi, I^{\mathsf{Tr}}}(\mathbf{x}_i) : i \in I^{\mathsf{Va}}\}$.

B. LOO

In the LOO cross-validation setting, the dataset $\mathcal{D} = \{(\mathbf{x}_i, \bar{y}_i : i = 1, \ldots, n\}$ is divided into n parts, each having exactly one point. In the l-th iteration, KRR is trained using the data $\mathcal{D} \setminus \{\mathbf{x}_l\}$. The discrepancy metric of the LOO $\mathcal{T}$-dependent cross-validation is

$$\mathrm{CV}^{\mathsf{LOO}}(\Xi, \mathcal{T}) := \frac{1}{n} \sum_{l=1}^{n} \big(\hat{\theta}(\Xi, \mathcal{T}, \mathcal{D} \setminus \{\mathbf{x}_l\}, \{\mathbf{x}_l\}) - \hat{\theta}^{\mathsf{St}}(\mathcal{T}, \{\mathbf{x}_l\})\big)^2,$$

where, according to (21) and (22),

$$\hat{\theta}^{\mathsf{St}}(\mathcal{T}, \{\mathbf{x}_l\}) = \begin{cases} \eta(\bar{y}_l), & \text{if } \mathcal{T}(\cdot) = \mathbb{E}[\eta(\cdot)], \\ \bar{y}_l, & \text{if } \mathcal{T}(\cdot) = \mathsf{VaR}_\tau(\cdot) \text{ or } \mathsf{CVaR}_\tau(\cdot), \end{cases}$$

and

$$\hat{\theta}(\Xi, \mathcal{T}, \mathcal{D} \setminus \{\mathbf{x}_l\}, \{\mathbf{x}_l\}) = \begin{cases} \eta(\hat{f}^{-l}(\mathbf{x}_l; \Xi)), & \text{if } \mathcal{T}(\cdot) = \mathbb{E}[\eta(\cdot)], \\ \hat{f}^{-l}(\mathbf{x}_l; \Xi), & \text{if } \mathcal{T}(\cdot) = \mathsf{VaR}_\tau(\cdot) \text{ or } \mathsf{CVaR}_\tau(\cdot). \end{cases}$$

Here, $\hat{f}^{-l}(\cdot; \Xi)$ denotes the KRR estimator trained with $\mathcal{D} \setminus \{\mathbf{x}_l\}$ and with the hyperparameters being Ξ.

For each l, let $\bar{\mathbf{y}}^{-l} \in \mathbb{R}^d$ denote the vector of which the l-th entry is $\hat{f}^{-l}(\mathbf{x}_l; \Xi)$ and the i-th entry is $\bar{y}_i$ for all $i \ne l$. It can be shown (see, e.g., Rifkin and Lippert 2007) that

$$\hat{f}^{-l}(\mathbf{x}_l; \Xi) = \frac{\big(\mathbf{R}\,(\mathbf{R} + n\lambda\mathbf{I})^{-1}\,\bar{\mathbf{y}}\big)_l - \big(\mathbf{R}\,(\mathbf{R} + n\lambda\mathbf{I})^{-1}\big)_{ll}\,\bar{y}_l}{1 - \big(\mathbf{R}\,(\mathbf{R} + n\lambda\mathbf{I})^{-1}\big)_{ll}}. \tag{23}$$

The key point here is that according to the expression (23), to compute $\mathrm{CV}^{\mathsf{LOO}}(\Xi, \mathcal{T})$ for given Ξ and $\mathcal{T}$, we need to compute the inverse matrix $\mathbf{R}\,(\mathbf{R} + n\lambda\mathbf{I})^{-1}$ only *once*, the computational time complexity of which is $O(n^3)$. By contrast, if we use K-fold cross-validation (with K usually set to be 5–20), then the computational trick (23) is not applicable, and therefore we would need to train KRR separately on each training set of size $(1 - 1/K)n$ from scratch. The computational cost involved in matrix inversion in the training process would amount to $O(Kn^3)$, which is substantially higher than LOO cross-validation.

References

Adams RA, Fournier JJ (2003) *Sobolev Spaces* (Academic Press), 2nd edition.

Andradóttir S, Glynn PW (2016) Computing Bayesian means using simulation. *ACM Trans. Model. Comput. Simul.* 26(2):Article 10.

Ankenman B, Nelson BL, Staum J (2010) Stochastic kriging for simulation metamodeling. *Oper. Res.* 58(2):371–382.

Asmussen S, Glynn PW (2007) *Stochastic Simulation: Algorithm and Analysis* (Springer).

Aydin G, Porteus EL (2008) Joint inventory and pricing decisions for an assortment. *Oper. Res.* 56(5):1247–1255.

Barton RR (2012) Tutorial: Input uncertainty in output analysis. *Proceedings of the 2012 Winter Simulation Conference*, 67–78.

Barton RR, Nelson BL, Xie W (2014) Quantifying input uncertainty via simulation confidence intervals. *INFORMS J. Comput.* 26(1):74–87.

Berlinet A, Thomas-Agnan C (2004) *Reproducing Kernel Hilbert Spaces in Probability and Statistics* (Springer).

Brezis H, Mironescu P (2019) Where Sobolev interacts with Gagliardo–Nirenberg. *Journal of Functional Analysis* 277(8):2839–2864.

Broadie M, Du Y, Moallemi CC (2011) Efficient risk estimation via nested sequential simulation. *Manag. Sci.* 57(6):1172–1194.

Broadie M, Du Y, Moallemi CC (2015) Risk estimation via regression. *Oper. Res.* 63(5):1077–1097.

Caponnetto A, De Vito E (2007) Optimal rates for the regularized least-squares algorithm. *Found. Comput. Math.* 7(3):331–368.

Chick SE (2001) Input distribution selection for simulation experiments: Accounting for input uncertainty. *Oper. Res.* 49(5):744–758.

Chick SE (2006) Subjective probability and Bayesian methodology. Henderson SG, Nelson BL, eds., *Handbooks in Operations Research and Management Science*, volume 13, chapter 9, 225–257 (Elsevier).

Dang O, Feng M, Hardy MR (2020) Efficient nested simulation for conditional tail expectation of variable annuities. *North American Actuarial Journal* 24(2):187–210.

Feng MB, Song E (2021) Optimal nested simulation experiment design via likelihood ratio method. Preprint available at `https://arxiv.org/abs/2008.13087`.

Frazier PI (2018) Bayesian optimization. *Recent Advances in Optimization and Modeling of Contemporary Problems*, 255–278, INFORMS TutORials in Operations Research (INFORMS).

Fu MC, Hong LJ, Hu JQ (2009) Conditional Monte Carlo estimation of quantile sensitivities. *Manag. Sci.* 55(12):2019–2027.

Fu MC, Hu JQ (1997) *Conditional Monte Carlo: Gradient Estimation and Optimization Applications* (Springer).

Fuh CD, Hu I, Hsu YH, Wang RH (2011) Efficient simulation of value at risk with heavy-tailed risk factors. *Oper. Res.* 59(6):1395–1406.

Glasserman P (2003) *Monte Carlo Methods in Financial Engineering* (Springer).

Glasserman P, Heidelberger P, Shahabuddin P (2000) Variance reduction techniques for estimating value-at-risk. *Manag. Sci.* 46(10):1349–1364.

Gordy MB, Juneja S (2010) Nested simulation in portfolio risk measurement. *Manag. Sci.* 56(10):1833–1848.

Györfi L, Kohler M, Krzyżak A, Walk H (2002) *A Distribution-Free Theory of Nonparametric Regression* (Springer).

Han Q, Wellner JA (2019) Convergence rates of least squares regression estimators with heavy-tailed errors. *Ann. Stat.* 47(4):2286–2319.

Hastie T (2020) Ridge regularization: An essential concept in data science. *Technometrics* 62(4):426–433.

Haug EG (2007) *The Complete Guide to Option Pricing Formulas* (McGraw-Hill), 2nd edition.

Hong LJ, Hu Z, Liu G (2014) Monte Carlo methods for value-at-risk and conditional value-at-risk: A review. *ACM Trans. Model. Comput. Simul.* 24(4):Article 22.

Hong LJ, Juneja S, Liu G (2017) Kernel smoothing for nested estimation with application to portfolio risk measurement. *Oper. Res.* 65(3):657–673.

Jiang W, Kou S (2021) Simulating risk measures via asymptotic expansions for relative errors. *Math. Finance* 31(3):907–942.

Jin X, Fu MC, Xiong X (2003) Probabilistic error bounds for simulation quantile estimators. *Manag. Sci.* 49(2):230–246.

Kanagawa M, Hennig P, Sejdinovic D, Sriperumbudur BK (2018) Gaussian processes and kernel methods: A review on connections and equivalences. Preprint available at `https://arxiv.org/abs/1807.02582`.

Kou S, Peng X (2016) On the measurement of economic tail risk. *Oper. Res.* 64(5):1056–1072.

Lan H (2010) *Two-level simulation of expected shortfall: Confidence intervals, efficient simulation procedures, and high-performance computing*. Ph.D. thesis, Northwestern University, Evanston, IL.

Lan H, Nelson BL, Staum J (2010) A confidence interval procedure for expected shortfall risk measurement via two-level simulation. *Oper. Res.* 58(5):1481–1490.

Lee SH, Glynn PW (2003) Computing the distribution function of a conditional expectation via Monte Carlo: Discrete conditioning spaces. *ACM Trans. Model. Comput. Simul.* 13(3):238–258.

Lesnevski V, Nelson BL, Staum J (2007) Simulation of coherent risk measures based on generalized scenarios. *Manag. Sci.* 53(11):1756–1769.

Liu M, Staum J (2010) Stochastic kriging for efficient nested simulation of expected shortfall. *J. Risk* 12(3):3–27.

Lu X, Rudi A, Borgonovo E, Rosasco L (2020) Faster kriging: Facing high-dimensional simulators. *Oper. Res.* 68(1):233–249.

Massart P (1990) The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.* 18(3):1269–1283.

Perchet V, Rigollet P (2013) The multi-armed bandit problem with covariates. *Ann. Stat.* 41(2):693–721.

Rasmussen CE, Williams CKI (2006) *Gaussian Processes for Machine Learning* (MIT Press).

Rifkin RM, Lippert RA (2007) Notes on regularized least-squares. Technical report, MIT, Computer Science and Artificial Intelligence Laboratory, URL `https://dspace.mit.edu/handle/1721.1/37318`.

Salemi P, Staum J, Nelson BL (2019) Generalized integrated Brownian fields for simulation metamodeling. *Oper. Res.* 67(3):874–891.

Schölkopf B, Smola AJ (2002) *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (MIT Press).

Steinwart I, Hush D, Scovel C (2009) Optimal rates for regularized least squares regression. *Proceedings of the 22nd Annual Conference on Learning Theory*, 79–93.

Sun Y, Apley DW, Staum J (2011) Efficient nested simulation for estimating the variance of a conditional expectation. *Oper. Res.* 59(4):998–1007.

Tsybakov AB (2004) Optimal aggregation of classifiers in statistical learning. *Ann. Stat.* 32(1):135–166.

Tuo R, Wang Y, Wu CFJ (2020) On the improved rates of convergence for Matérn-type kernel ridge regression with application to calibration of computer models. *SIAM/ASA J. Uncertainty Quantification* 8(4):1522–1547.

van de Geer S (2000) *Empirical Processes in M-Estimation* (Cambridge University Press).

Wainwright MJ (2019) *High-Dimensional Statistics: A Non-Asymptotic Viewpoint* (Cambridge University Press).

Xie W, Nelson BL, Barton RR (2014) A Bayesian framework for quantifying uncertainty in stochastic simulation. *Oper. Res.* 62(6):1439–1452.

Zhang K, Liu G, Wang S (2021) Bootstrap-based budget allocation for nested simulation. *Oper. Res.*, forthcoming.

Zhang Y, Duchi J, Wainwright M (2015) Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *J. Mach. Learn. Res.* 16(102):3299–3340.

Zhu H, Liu T, Zhou E (2020) Risk quantification in stochastic simulation under input uncertainty. *ACM Trans. Model. Comput. Simul.* 30(1):Article 1.

Supplemental Material

EC.1. Proof of Proposition 1

Lemma EC.1 **(Representer Theorem (Theorem 4.2 in Schölkopf and Smola 2002))**. *Let $k:\Omega\times\Omega\mapsto\mathbb{R}$ be a positive definite kernel, $\mathcal{N}_k(\Omega)$ be the RKHS of k, $L:\mathbb{R}\times\mathbb{R}\mapsto\mathbb{R}_+$ be an arbitrary loss function, $Q:\mathbb{R}_+\mapsto\mathbb{R}$ be a strictly increasing function, and $\{(\mathbf{x}_i,y_i)\}_{i=1}^n$ be a set of training data with $\mathbf{x}_i\in\Omega$ and $y_i\in\mathbb{R}$. Then, each optimal solution to the optimization problem*

$$\min_{g\in\mathcal{N}_k(\Omega)}\frac{1}{n}\sum_{i=1}^n L(y_i,g(\mathbf{x}_i))+Q(\|g\|_{\mathcal{N}_k(\Omega)}),\tag{EC.1.1}$$

admits a representation of the form $g^*=\sum_{i=1}^n\beta_i^*k(\mathbf{x}_i,\cdot)$ for some constants $\beta_i^*\in\mathbb{R}$, $i=1,\dots,n$.

Lemma EC.2 **(Theorem 14.4-1 in Bishop et al. 2007)**. *Let $Z_1,\dots,Z_n$ be zero-mean random variables such that $\mathbb{V}\mathrm{ar}(Z_i)<\infty$ for all $i=1,\dots,n$. Then, $Z_n=O_{\mathbb{P}}(\sqrt{\mathbb{V}\mathrm{ar}(Z_n)})$.*

Proof of Proposition 1. Let $\mathbf{f}=(f(\mathbf{x}_1),\dots,f(\mathbf{x}_n))^\intercal$ and $\bar{\boldsymbol{\epsilon}}=(\bar\epsilon_1,\dots,\bar\epsilon_n)^\intercal$, where $\bar\epsilon_i=\frac{1}{m}\sum_{j=1}^m\epsilon_{ij}$. It follows from the expression of $\hat f$ that

$$\hat f(\mathbf{x}_i)=\mathbf{r}(\mathbf{x}_i)^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\bar{\mathbf{y}}=\mathbf{r}(\mathbf{x}_i)^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{f}+\mathbf{r}(\mathbf{x}_i)^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\bar{\boldsymbol{\epsilon}}.$$

Thus,

$$\begin{aligned}&\left|\frac{1}{n}\sum_{i=1}^n\varphi(f(\mathbf{x}_i))(f(\mathbf{x}_i)-\hat f(\mathbf{x}_i))\right|\\ \le&\underbrace{\left|\frac{1}{n}\sum_{i=1}^n\varphi(f(\mathbf{x}_i))(f(\mathbf{x}_i)-\mathbf{r}(\mathbf{x}_i)^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{f})\right|}_{H_1}+\underbrace{\left|\frac{1}{n}\sum_{i=1}^n\varphi(f(\mathbf{x}_i))\mathbf{r}(\mathbf{x}_i)^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\bar{\boldsymbol{\epsilon}}\right|}_{H_2}.\end{aligned}\tag{EC.1.2}$$

To bound H_1, we apply the Cauchy–Schwarz inequality:

$$\begin{aligned}|H_1|&\le\left(\frac{1}{n}\sum_{i=1}^n\varphi(f(\mathbf{x}_i))^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^n(f(\mathbf{x}_i)-\mathbf{r}(\mathbf{x}_i)^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{f})^2\right)^{1/2}\\&\le C\left(\frac{1}{n}\sum_{i=1}^n(f(\mathbf{x}_i)-\mathbf{r}(\mathbf{x}_i)^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{f})^2\right)^{1/2},\end{aligned}\tag{EC.1.3}$$

where $C=\sup_{z\in\{f(\mathbf{x}):\mathbf{x}\in\Omega\}}|g(z)|<\infty$.

Let $f_\dagger(\mathbf{x}):=\mathbf{r}(\mathbf{x})^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{f}$. By Lemma EC.1, it can be checked that $f_\dagger$ is the solution to the optimization problem:

$$f_\dagger=\operatorname*{arg\,min}_{g\in\mathcal{N}_\Psi(\Omega)}\frac{1}{n}\sum_{i=1}^n(g(\mathbf{x}_i)-f(\mathbf{x}_i))^2+\lambda\|g\|^2_{\mathcal{N}_\Psi(\Omega)}.\tag{EC.1.4}$$

Clearly, (EC.1.4) implies

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i)-\mathbf{r}(\mathbf{x}_i)^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{f})^2 &= \frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i)-f_\dagger(\mathbf{x}_i))^2\\ &\le \frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i)-f_\dagger(\mathbf{x}_i))^2+\lambda\|f_\dagger\|^2_{\mathcal{N}_\Psi(\Omega)}\\ &\le \frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i)-f(\mathbf{x}_i))^2+\lambda\|f\|^2_{\mathcal{N}_\Psi(\Omega)}=\lambda\|f\|^2_{\mathcal{N}_\Psi(\Omega)}.\end{aligned}$$

Thus,

$$|H_1| = O(\lambda^{1/2}). \tag{EC.1.5}$$

Now consider H_2. Because ϵ_{ij}'s are zero-mean sub-Gaussian random variables, H_2 is also sub-Gaussian and has mean zero, which implies that H_2 has a finite variance. By Lemma EC.2,

$$|H_2| = O_{\mathbb{P}}(\sqrt{\mathbb{V}\mathrm{ar}(H_2)}). \tag{EC.1.6}$$

Let $\mathbf{u}_i=(u_1(\mathbf{x}_i),\dots,u_n(\mathbf{x}_i))^\intercal=(\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{r}(\mathbf{x}_i)$. Since $|\varphi(f(\mathbf{x}_i))|\le C$ for all $i=1,\dots,n$, by Assumption 1, we have that

$$\begin{aligned}\mathbb{V}\mathrm{ar}(H_2) &= \mathbb{V}\mathrm{ar}\left(\frac{1}{n}\sum_{i=1}^{n}\varphi(f(\mathbf{x}_i))\mathbf{r}(\mathbf{x}_i)^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\bar{\boldsymbol{\epsilon}}\right)\\ &= \mathbb{V}\mathrm{ar}\left(\frac{1}{n}\sum_{i=1}^{n}\varphi(f(\mathbf{x}_i))\sum_{j=1}^{n}u_j(\mathbf{x}_i)\bar{\epsilon}_j\right)\\ &\le \frac{\sigma^2}{m}\sum_{j=1}^{n}\left(\frac{1}{n}\sum_{i=1}^{n}\varphi(f(\mathbf{x}_i))u_j(\mathbf{x}_i)\right)^2.\end{aligned} \tag{EC.1.7}$$

Let $\mathbf{e}_j\in\mathbb{R}^d$ be a vector of zeros except the j-th entry being one. Direct computation shows that

$$\begin{aligned}&\sum_{j=1}^{n}\left(\frac{1}{n}\sum_{i=1}^{n}\varphi(f(\mathbf{x}_i))u_j(\mathbf{x}_i)\right)^2\\ =&\sum_{j=1}^{n}\left(\frac{1}{n}\sum_{i=1}^{n}\varphi(f(\mathbf{x}_i))\mathbf{e}_j^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{r}(\mathbf{x}_i)\right)^2\\ =&\sum_{j=1}^{n}\left(\mathbf{e}_j^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\varphi(f(\mathbf{x}_i))\mathbf{r}(\mathbf{x}_i)\right)\right)^2\\ =&\sum_{j=1}^{n}\left(\frac{1}{n}\sum_{i=1}^{n}\varphi(f(\mathbf{x}_i))\mathbf{r}(\mathbf{x}_i)\right)^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{e}_j\mathbf{e}_j^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\varphi(f(\mathbf{x}_i))\mathbf{r}(\mathbf{x}_i)\right)\\ =&\left(\frac{1}{n}\sum_{i=1}^{n}\varphi(f(\mathbf{x}_i))\mathbf{r}(\mathbf{x}_i)\right)^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-2}\left(\frac{1}{n}\sum_{i=1}^{n}\varphi(f(\mathbf{x}_i))\mathbf{r}(\mathbf{x}_i)\right)\end{aligned}$$

$$
\begin{aligned}
&\leq \left(\frac{1}{n}\sum_{i=1}^{n}\varphi(f(\mathbf{x}_i))\mathbf{r}(\mathbf{x}_i)\right)^{\intercal}\mathbf{R}^{-2}\left(\frac{1}{n}\sum_{i=1}^{n}\varphi(f(\mathbf{x}_i))\mathbf{r}(\mathbf{x}_i)\right)\\
&=\frac{1}{n^2}\sum_{i,j=1}^{n}\varphi(f(\mathbf{x}_i))g(f(\mathbf{x}_j))\mathbf{r}(\mathbf{x}_i)^{\intercal}\mathbf{R}^{-2}r(\mathbf{x}_j)\\
&=\frac{1}{n^2}\sum_{i,j=1}^{n}\varphi(f(\mathbf{x}_i))g(f(\mathbf{x}_j))\mathbf{e}_i^{\intercal}\mathbf{e}_j\leq\frac{C^2}{n}, \qquad \text{(EC.1.8)}
\end{aligned}
$$

where the fourth equality holds because $\sum_{j=1}^{n}\mathbf{e}_j\mathbf{e}_j^{\intercal}=\mathbf{I}$, and the last step because $\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}_i)=\mathbf{e}_i$. Then, plugging (EC.1.7) and (EC.1.8) into (EC.1.6) yields

$$
|H_2| = O_{\mathbb{P}}((mn)^{-1/2}). \qquad \text{(EC.1.9)}
$$

The proof is completed by combining (EC.1.2), (EC.1.5), and (EC.1.9). □

EC.2. Proof of Proposition 2

Lemma EC.3 **(Lemma A.1 in Tuo et al. 2020)**. *Let Ω be a bounded convex subset of $\mathbb{R}^d$ and $\{\epsilon_1,\dots,\epsilon_n\}$ be independent, zero-mean sub-Gaussian random variables. Then, there exist positive constants C_1 and C_2 such that for all t large enough,*

$$
\Pr\left(\sup_{g\in\mathcal{N}_\Psi(\Omega)}\frac{\left|\frac{1}{n}\sum_{j=1}^{n}\epsilon_j g(\mathbf{x}_j)\right|}{\|g\|_n^{1-\frac{d}{2\nu+d}}\|g\|_{\mathcal{N}_\Psi(\Omega)}^{\frac{d}{2\nu+d}}}\geq tn^{-1/2}\right)\leq C_1\exp\bigl(-C_2t^2\bigr),
$$

where $\|g\|_n=\bigl(n^{-1}\sum_{i=1}^{n}g^2(\mathbf{x}_i)\bigr)^{1/2}$ denotes the empirical semi-norm of g

Proof of Proposition 2. Note that $\hat{f}$ is the solution to

$$
\min_{g\in\mathcal{N}_\Psi(\Omega)}\left(\frac{1}{n}\sum_{i=1}^{n}(\bar{y}_i-g(\mathbf{x}_i))^2+\lambda\|g\|_{\mathcal{N}_\Psi(\Omega)}^2\right).
$$

Hence,

$$
\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(\mathbf{x}_i)-\bar{y}_i)^2+\lambda\|\hat{f}\|_{\mathcal{N}_\Psi(\Omega)}^2\leq\frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i)-\bar{y}_i)^2+\lambda\|f\|_{\mathcal{N}_\Psi(\Omega)}^2. \qquad \text{(EC.2.1)}
$$

Using the notation $\|\hat{f}-f\|_n^2=\frac{1}{n}\sum_{i=1}^{n}(f(\mathbf{x}_i)-\hat{f}(\mathbf{x}_i))^2$, and plugging $\bar{y}_i=f(\mathbf{x}_i)+\bar{\epsilon}_i$ into (EC.2.1), we have that

$$
\|\hat{f}-f\|_n^2+\lambda\|\hat{f}\|_{\mathcal{N}_\Psi(\Omega)}^2\leq\frac{2}{n}\sum_{i=1}^{n}\bar{\epsilon}_i(\hat{f}-f)(\mathbf{x}_i)+\lambda\|f\|_{\mathcal{N}_\Psi(\Omega)}^2.
$$

Moreover, note that $\bar{\epsilon}_i$ is $\mathsf{subG}(\sigma^2/m)$. It then follows from Lemma EC.3 that

$$
\frac{1}{n}\sum_{i=1}^{n}\bar{\epsilon}_i(\hat{f}-f)(\mathbf{x}_i)=O_{\mathbb{P}}((mn)^{-1/2})\|\hat{f}-f\|_n^{1-\frac{d}{2\nu+d}}\|\hat{f}-f\|_{\mathcal{N}_\Psi(\Omega)}^{\frac{d}{2\nu+d}}.
$$

Hence,

$$\|\hat{f}-f\|_n^2+\lambda\|\hat{f}\|^2_{\mathcal{N}_\Psi(\Omega)} \le O_{\mathbb{P}}((mn)^{-1/2})\|\hat{f}-f\|_n^{1-\frac{d}{2\nu+d}}\|\hat{f}-f\|^{\frac{d}{2\nu+d}}_{\mathcal{N}_\Psi(\Omega)}+\lambda\|f\|^2_{\mathcal{N}_\Psi(\Omega)}. \tag{EC.2.2}$$

It can be seen that (EC.2.2) implies either

$$\|\hat{f}-f\|_n^2+\lambda\|\hat{f}\|^2_{\mathcal{N}_\Psi(\Omega)} \le O_{\mathbb{P}}((mn)^{-1/2})\|\hat{f}-f\|_n^{1-\frac{d}{2\nu+d}}\|\hat{f}-f\|^{\frac{d}{2\nu+d}}_{\mathcal{N}_\Psi(\Omega)}, \tag{EC.2.3}$$

or

$$\|\hat{f}-f\|_n^2+\lambda\|\hat{f}\|^2_{\mathcal{N}_\Psi(\Omega)} \le 2\lambda\|f\|^2_{\mathcal{N}_\Psi(\Omega)}. \tag{EC.2.4}$$

Case 1: Assume (EC.2.3) holds. Then, by the triangle inequality,

$$\|\hat{f}-f\|_n^2+\lambda\|\hat{f}\|^2_{\mathcal{N}_\Psi(\Omega)} \le O_{\mathbb{P}}((mn)^{-1/2})\|\hat{f}-f\|_n^{1-\frac{d}{2\nu+d}}\left(\|\hat{f}\|_{\mathcal{N}_\Psi(\Omega)}+\|f\|_{\mathcal{N}_\Psi(\Omega)}\right)^{\frac{d}{2\nu+d}}. \tag{EC.2.5}$$

Next, we analyze (EC.2.5) separately depending on whether $\|f\|_{\mathcal{N}_\Psi(\Omega)} \le \|\hat{f}\|_{\mathcal{N}_\Psi(\Omega)}$.

If $\|f\|_{\mathcal{N}_\Psi(\Omega)} \le \|\hat{f}\|_{\mathcal{N}_\Psi(\Omega)}$, then (EC.2.5) implies that

$$\|\hat{f}-f\|_n^2+\lambda\|\hat{f}\|^2_{\mathcal{N}_\Psi(\Omega)} \le O_{\mathbb{P}}((mn)^{-1/2})\|\hat{f}-f\|_n^{1-\frac{d}{2\nu+d}}\|\hat{f}\|^{\frac{d}{2\nu+d}}_{\mathcal{N}_\Psi(\Omega)}.$$

Therefore,

$$\|\hat{f}-f\|_n^2 \le O_{\mathbb{P}}((mn)^{-1/2})\|\hat{f}-f\|_n^{1-\frac{d}{2\nu+d}}\|\hat{f}\|^{\frac{d}{2\nu+d}}_{\mathcal{N}_\Psi(\Omega)}, \tag{EC.2.6}$$

$$\lambda\|\hat{f}\|^2_{\mathcal{N}_\Psi(\Omega)} \le O_{\mathbb{P}}((mn)^{-1/2})\|\hat{f}-f\|_n^{1-\frac{d}{2\nu+d}}\|\hat{f}\|^{\frac{d}{2\nu+d}}_{\mathcal{N}_\Psi(\Omega)}. \tag{EC.2.7}$$

Solving the system of inequalities (EC.2.6)–(EC.2.7) yields

$$\|\hat{f}-f\|_n^2 = O_{\mathbb{P}}\left((mn)^{-1}\lambda^{-\frac{d}{2\nu+d}}\right), \tag{EC.2.8}$$

$$\|\hat{f}\|^2_{\mathcal{N}_\Psi(\Omega)} = O_{\mathbb{P}}\left((mn)^{-1}\lambda^{-\frac{2(\nu+d)}{2\nu+d}}\right). \tag{EC.2.9}$$

If $\|f\|_{\mathcal{N}_\Psi(\Omega)} > \|\hat{f}\|_{\mathcal{N}_\Psi(\Omega)}$, then (EC.2.5) implies that

$$\|\hat{f}-f\|_n^2+\lambda\|\hat{f}\|^2_{\mathcal{N}_\Psi(\Omega)} \le O_{\mathbb{P}}((mn)^{-1/2})\|\hat{f}-f\|_n^{1-\frac{d}{2\nu+d}}\|f\|^{\frac{d}{2\nu+d}}_{\mathcal{N}_\Psi(\Omega)}.$$

Hence, noting that $\|f\|_{\mathcal{N}_\Psi(\Omega)}$ is a constant, we have

$$\|\hat{f}-f\|_n^2 \le O_{\mathbb{P}}((mn)^{-1/2})\|\hat{f}-f\|_n^{1-\frac{d}{2\nu+d}}. \tag{EC.2.10}$$

Solving (EC.2.10) yields

$$\|\hat{f}-f\|_n^2 = O_{\mathbb{P}}\left((mn)^{-\frac{2\nu+d}{2\nu+2d}}\right). \tag{EC.2.11}$$

Case 2: Assume (EC.2.4) holds. Then,

$$\|\hat{f}-f\|_n^2 \le 2\lambda\|f\|^2_{\mathcal{N}_\Psi(\Omega)} = O_{\mathbb{P}}(\lambda). \tag{EC.2.12}$$

Combining the two cases, i.e., combining (EC.2.8), (EC.2.11), and (EC.2.12), we conclude that

$$\|\hat{f}-f\|_n^2 = O_{\mathbb{P}}\left((mn)^{-1}\lambda^{-\frac{d}{2\nu+d}}+(mn)^{-\frac{2\nu+d}{2\nu+2d}}+\lambda\right). \quad \square$$

EC.3. Proof of Theorem 1

Proof of Theorem 1. By the triangle inequality, we have

$$\begin{aligned}|\hat{\theta}_{n,m}-\theta| &= \left|\mathbb{E}[\eta(f(X))]-\frac{1}{n}\sum_{i=1}^{n}\eta(f(\mathbf{x}_i))+\frac{1}{n}\sum_{i=1}^{n}\eta(f(\mathbf{x}_i))-\frac{1}{n}\sum_{i=1}^{n}\eta(\hat{f}(\mathbf{x}_i))\right| \\ &\leq \underbrace{\left|\mathbb{E}[\eta(f(X))]-\frac{1}{n}\sum_{i=1}^{n}\eta(f(\mathbf{x}_i))\right|}_{I_1}+\underbrace{\left|\frac{1}{n}\sum_{i=1}^{n}\left[\eta(f(\mathbf{x}_i))-\eta(\hat{f}(\mathbf{x}_i))\right]\right|}_{I_2}. \end{aligned} \tag{EC.3.1}$$

The reproducing property of RKHSs implies that for any $\mathbf{x}\in\Omega$,

$$\begin{aligned}|f(\mathbf{x})| = |\langle f,\Psi(\mathbf{x}-\cdot)\rangle_{\mathcal{N}_\Psi(\Omega)}| &\leq \|f\|_{\mathcal{N}_\Psi(\Omega)}\|\Psi(\mathbf{x}-\cdot)\|_{\mathcal{N}_\Psi(\Omega)} \\ &= \|f\|_{\mathcal{N}_\Psi(\Omega)}\Psi(\mathbf{x}-\mathbf{x}) = \|f\|_{\mathcal{N}_\Psi(\Omega)}\Psi(\mathbf{0}),\end{aligned} \tag{EC.3.2}$$

which implies both $\|f\|_{\mathcal{L}_\infty(\Omega)}$ and $\|\eta(f(\cdot))\|_{\mathcal{L}_\infty(\Omega)}$ are finite. Therefore, $\eta(f(\mathbf{x}_i))$'s are bounded random variables, thereby being sub-Gaussian. Then, the central limit theorem implies that

$$I_1 = O_{\mathbb{P}}(n^{-1/2}). \tag{EC.3.3}$$

It remains to bound I_2. It follows from Taylor's expansion and the triangle inequality that

$$\begin{aligned} I_2 &= \left|\frac{1}{n}\sum_{i=1}^{n}\eta'(f(\mathbf{x}_i))(f(\mathbf{x}_i)-\hat{f}(\mathbf{x}_i))+\frac{1}{2n}\sum_{i=1}^{n}\eta''(\tilde{z}_i)(f(\mathbf{x}_i)-\hat{f}(\mathbf{x}_i))^2\right| \\ &\leq \underbrace{\left|\frac{1}{n}\sum_{i=1}^{n}\eta'(f(\mathbf{x}_i))(f(\mathbf{x}_i)-\hat{f}(\mathbf{x}_i))\right|}_{I_{21}}+\underbrace{\left|\frac{1}{2n}\sum_{i=1}^{n}\eta''(\tilde{z}_i)(f(\mathbf{x}_i)-\hat{f}(\mathbf{x}_i))^2\right|}_{I_{22}}, \end{aligned} \tag{EC.3.4}$$

where $\tilde{z}_i$ is a value between $f(\mathbf{x}_i)$ and $\hat{f}(\mathbf{x}_i)$.

Because $|\eta'(f(\mathbf{x}))|$ is bounded for all $\mathbf{x}\in\Omega$, it follows from Proposition 1 that

$$I_{21} = O_{\mathbb{P}}(\lambda^{1/2}+(mn)^{-1/2}). \tag{EC.3.5}$$

Let $C=\sup_{z\in\{f(\mathbf{x}):\mathbf{x}\in\Omega\}}|\eta''(z)|<\infty$. The term I_{22} can be bounded using Proposition 2:

$$I_{22} \leq \frac{C}{2n}\sum_{i=1}^{n}(f(\mathbf{x}_i)-\hat{f}(\mathbf{x}_i))^2 = O_{\mathbb{P}}\left((mn)^{-1}\lambda^{-\frac{d}{2\nu+d}}+(mn)^{-\frac{2\nu+d}{2\nu+2d}}+\lambda\right). \tag{EC.3.6}$$

Then, we combine (EC.3.1), (EC.3.3), (EC.3.4), (EC.3.5), and (EC.3.6). This yields

$$\begin{aligned}|\hat{\theta}_{n,m}-\theta| &= O_{\mathbb{P}}\left(n^{-1/2}+\lambda^{1/2}+(mn)^{-1/2}+(mn)^{-1}\lambda^{-\frac{d}{2\nu+d}}+(mn)^{-\frac{2\nu+d}{2\nu+2d}}+\lambda\right) \\ &= O_{\mathbb{P}}\left(n^{-1/2}+\lambda^{1/2}+(mn)^{-1}\lambda^{-\frac{d}{2\nu+d}}\right),\end{aligned} \tag{EC.3.7}$$

where the second inequality holds because $\lambda=O(1)$, $(mn)^{-1/2}\leq n^{-1/2}$, and $(mn)^{-\frac{2\nu+d}{2\nu+2d}}\leq n^{-1/2}$.

If $\nu \geq d/2$, then we set $n \asymp \Gamma$ and $\lambda \asymp \Gamma^{-1}$. Because $\frac{2\nu}{2\nu+d} \geq \frac{1}{2}$ in this case, (EC.3.7) becomes

$$|\hat{\theta}_{n,m} - \theta| = O_{\mathbb{P}}\left(\Gamma^{-1/2} + \Gamma^{-1/2} + \Gamma^{-\frac{2\nu}{2\nu+d}}\right) = O_{\mathbb{P}}\left(\Gamma^{-1/2}\right).$$

If $0 < \nu < d/2$, then we set $n \asymp \Gamma^{\frac{2(2\nu+d)}{2\nu+3d}}$ and $\lambda \asymp \Gamma^{-\frac{2(2\nu+d)}{2\nu+3d}}$. (This choice of n is possible because $\frac{2(2\nu+d)}{2\nu+3d} < 1$ for $\nu < d/2$.) This makes (EC.3.7) become

$$|\hat{\theta}_{n,m} - \theta| = O_{\mathbb{P}}\left(\Gamma^{-\frac{2\nu+d}{2\nu+3d}} + \Gamma^{-\frac{2\nu+d}{2\nu+3d}} + \Gamma^{-\frac{2\nu+d}{2\nu+3d}}\right) = O_{\mathbb{P}}\left(\Gamma^{-\frac{2\nu+d}{2\nu+3d}}\right). \quad \square$$

EC.4. Proof of Proposition 3

Lemma EC.4. *Let Ω be a bounded convex subset of $\mathbb{R}^d$ and $0 < s_1 < s_2$. Then, there exists a constant $C > 0$ such that for any $g \in \mathcal{H}^{s_2}(\Omega)$,*

$$\|g\|_{\mathcal{L}_\infty(\Omega)} \leq C\|g\|_{\mathcal{L}_2(\Omega)}^{1-\frac{s_1}{s_2}} \|g\|_{\mathcal{H}^{s_2}(\Omega)}^{\frac{s_1}{s_2}}.$$

Proof of Lemma EC.4. This is a direct result of applying the Gagliardo–Nirenberg interpolation inequality to the Sobolev spaces (see, e.g., Brezis and Mironescu 2019). □

In order to quantify the capacity of a function class, we need the following two definitions of entropy number and bracket entropy number (see van de Geer 2000 for more discussions).

Definition EC.1 (Entropy Number). Let $\mathcal{G}$ be a function space equipped with a norm $\|\cdot\|$, and $\mathcal{G}_0 \subset \mathcal{G}$ be a function class. For any $\varepsilon > 0$, let $\mathcal{B}_\varepsilon(h, \|\cdot\|) := \{g \in \mathcal{G} : \|g - h\| \leq \varepsilon\}$ be an ε-ball that is centered at $h \in \mathcal{G}$. The *covering number* $\mathcal{N}(\varepsilon, \mathcal{G}_0, \|\cdot\|)$ is defined as

$$\mathcal{N}(\varepsilon, \mathcal{G}_0, \|\cdot\|) := \min\left\{n : \text{ There exist } g_1, \ldots, g_n \in \mathcal{G}_0 \text{ such that } \mathcal{G}_0 \subseteq \bigcup_{i=1}^{n} \mathcal{B}_\varepsilon(g_i, \|\cdot\|)\right\}.$$

Then, $\mathcal{H}(\varepsilon, \mathcal{G}_0, \|\cdot\|) := \log_2 \mathcal{N}(\varepsilon, \mathcal{G}_0, \|\cdot\|)$ is called the *entropy number* of $\mathcal{G}_0$.

Definition EC.2 (Bracket Entropy Number). Let $\mathcal{G}$ and $\mathcal{G}_0 \subset \mathcal{G}$ be in Definition EC.1. For any $\varepsilon > 0$, let $\mathcal{N}_{[\,]}(\varepsilon, \mathcal{G}_0, \|\cdot\|)$ be the smallest value of n for which there exist pairs of functions $\{g_j^L, g_j^U\} \subset \mathcal{G}_0$ such that $\|g_j^U - g_j^L\| \leq \varepsilon$ for all $j = 1, ..., n$, and such that for each $g \in \mathcal{G}_0$, $g_j^L \leq g \leq g_j^U$. Then, $\mathcal{H}_{[\,]}(\varepsilon, \mathcal{G}_0, \|\cdot\|) := \log_2 \mathcal{N}_{[\,]}(\varepsilon, \mathcal{G}_0, \|\cdot\|)$ is called the *bracket entropy number* of $\mathcal{G}_0$.

Lemma EC.5 **(Lemma 2.1 in van de Geer 2000)**. *Let Ω be a bounded convex subset of $\mathbb{R}^d$, $\mathcal{G} = \{g : \Omega \mapsto \mathbb{R} : \|g\|_{\mathcal{L}_\infty} < \infty\}$, and $\mathcal{G}_0 \subset \mathcal{G}$. Then, for $p = 1, 2$ and any $\varepsilon > 0$, there exists a constant $C > 0$ such that*

$$\mathcal{H}(\varepsilon, \mathcal{G}_0, \|\cdot\|_{\mathcal{L}_p}) \leq \mathcal{H}_{[\,]}(\varepsilon, \mathcal{G}_0, \|\cdot\|_{\mathcal{L}_p}) \quad anbd \quad \mathcal{H}_{[\,]}(\varepsilon, \mathcal{G}_0, \|\cdot\|_{\mathcal{L}_p}) \leq C\mathcal{H}(\varepsilon/2, \mathcal{G}_0, \|\cdot\|_{\mathcal{L}_\infty}).$$

Lemma EC.6 **(Lemma 5.16 in van de Geer 2000)**. *Let Ω be a bounded convex subset of $\mathbb{R}^d$, $\mathcal{G} = \{g : \Omega \mapsto \mathbb{R} : \|g\|_{\mathcal{L}_2} < \infty\}$, and $\mathcal{G}_0 \subset \mathcal{G}$. Suppose that the sequence $\{\delta_n > 0 : n \geq 1\}$ satisfies $n\delta_n^2 \geq \mathcal{H}_{[\,]}(\delta_n, \mathcal{G}_0, \|\cdot\|_{\mathcal{L}_2})$ for all $n \geq 1$, and $n\delta_n^2 \to \infty$ as $n \to \infty$. Then, for any $C > 0$ and $t \in (0,1)$,*

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{g \in \mathcal{G}_0, \|g\|_{\mathcal{L}_2} > Ct^{-1}\delta_n} \left|\frac{\|g\|_n}{\|g\|_{\mathcal{L}_2}} - 1\right| > t\right) = 0.$$

Lemma EC.7. *Let Ω be a bounded convex subset of $\mathbb{R}^d$ and $\mathcal{B} = \{g \in \mathcal{H}^s(\Omega) : \|g\|_{\mathcal{H}^s(\Omega)} \le 1\}$. Then, $\|g\|_{\mathcal{L}_2(\Omega)} = O_{\mathbb{P}}(n^{-\frac{s}{2s+d}} + \|g\|_n)$ for all $g \in \mathcal{B}$.*

Proof of Lemma EC.7. Note that $\mathcal{B}$ is the unit ball in the Sobolev space $\mathcal{H}^s(\Omega)$. It can be bounded by Lemma EC.5

$$\mathcal{H}_{[\,]}(\delta_n, \mathcal{B}, \|\cdot\|_{\mathcal{L}_2}) \le C\mathcal{H}(\delta_n/2, \mathcal{B}, \|\cdot\|_{\mathcal{L}_\infty(\Omega)}) \le C_1 \delta_n^{-s/d},$$

for some constants $C, C_1 > 0$, where the second inequality follows from the theorem on page 105 of Edmunds and Triebel (1996). Now, we take $\delta_n = C_1 n^{-\frac{s}{2s+d}}$ such that

$$n\delta_n^2 \ge \mathcal{H}_{[\,]}(\delta_n, \mathcal{B}, \|\cdot\|_{\mathcal{L}_2}),$$

and then apply Lemma EC.6. This leads to

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{g\in\mathcal{B}, \|g\|_{\mathcal{L}_2} > C_1 t^{-1} n^{-\frac{s}{2s+d}}} \left| \frac{\|g\|_n}{\|g\|_{\mathcal{L}_2}} - 1 \right| > t \right) = 0.$$

Hence, $\|g\|_{\mathcal{L}_2} = O_{\mathbb{P}}(\max\{n^{-\frac{s}{2s+d}}, \|g\|_n\}) = O_{\mathbb{P}}(n^{-\frac{s}{2s+d}} + \|g\|_n)$. □

To prove Proposition 3, note that, by the expression of $\hat{f}$,

$$|f(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)| \le \Big|\underbrace{f(\mathbf{x}_i) - \mathbf{r}(\mathbf{x}_i)^\intercal(\mathbf{R} + n\lambda\mathbf{I})^{-1}\mathbf{f}}_{M_1(\mathbf{x}_i)}\Big| + \Big|\underbrace{\mathbf{r}(\mathbf{x}_i)^\intercal(\mathbf{R} + n\lambda\mathbf{I})^{-1}\bar{\boldsymbol{\epsilon}}}_{M_2(\mathbf{x}_i)}\Big|. \tag{EC.4.1}$$

Hence, it suffices to bound $\max_{1\le i\le n}|M_1(\mathbf{x}_i)|$ and $\max_{1\le i\le n}|M_2(\mathbf{x}_i)|$, respectively.

Lemma EC.8. *Suppose $f \in \mathcal{N}_\Psi(\Omega)$ and Assumption 2 holds. Then,*

$$\max_{1\le i\le n} |M_1(\mathbf{x}_i)| = O_{\mathbb{P}}\left(\left(n^{-\frac{\nu}{2\nu+2d}} + \lambda^{\frac{\nu}{2\nu+d}}\right) \wedge (n\lambda)^{1/2} \right).$$

Proof of Lemma EC.8. Let $f_\dagger(\mathbf{x}) := \mathbf{r}(\mathbf{x})^\intercal(\mathbf{R} + n\lambda\mathbf{I})^{-1}\mathbf{f}$ be the solution to (EC.1.4). The term $M_1(\mathbf{x}_i)$ can be bounded in two different ways, and we will take the smaller one as the upper bound.

First way to bound $M_1(\mathbf{x}_i)$**.** Note that $|f(\mathbf{x}) - f_\dagger(\mathbf{x})| \le \|f - f_\dagger\|_{\mathcal{L}_\infty(\Omega)}$ for all $\mathbf{x} \in \Omega$, so $\max_{1\le i\le n}|M_1(\mathbf{x}_i)| \le \|f - f_\dagger\|_{\mathcal{L}_\infty(\Omega)}$. Because $f_\dagger(\mathbf{x})$ is the solution to (EC.1.4), we have that

$$\frac{1}{n}\sum_{i=1}^n (f(\mathbf{x}_i) - f_\dagger(\mathbf{x}_i))^2 + \lambda\|f_\dagger\|^2_{\mathcal{N}_\Psi(\Omega)} \le \frac{1}{n}\sum_{i=1}^n (f(\mathbf{x}_i) - f(\mathbf{x}_i))^2 + \lambda\|f\|^2_{\mathcal{N}_\Psi(\Omega)} = \lambda\|f\|^2_{\mathcal{N}_\Psi(\Omega)},$$

which implies

$$\|f - f_\dagger\|_n = O_{\mathbb{P}}(\lambda^{1/2}), \tag{EC.4.2}$$

$$\|f_\dagger\|_{\mathcal{N}_\Psi(\Omega)} \le \|f\|_{\mathcal{N}_\Psi(\Omega)}. \tag{EC.4.3}$$

Because of the norm equivalence between $\mathcal{N}_\Psi(\Omega)$ and $\mathcal{H}^{\nu+d/2}(\Omega)$, there exists a constant C_1 such that

$$\|f-f_\dagger\|_{\mathcal{H}^{\nu+d/2}(\Omega)} \le C_1\|f-f_\dagger\|_{\mathcal{N}_\Psi(\Omega)} \le C_1(\|f\|_{\mathcal{N}_\Psi(\Omega)}+\|f_\dagger\|_{\mathcal{N}_\Psi(\Omega)}) \le 2C_1\|f\|_{\mathcal{N}_\Psi(\Omega)}, \qquad \text{(EC.4.4)}$$

where the last equality is by (EC.4.3).

Let $g=\frac{f-f_\dagger}{2C_1\|f\|_{\mathcal{N}_\Psi(\Omega)}}$. Then, $\|g\|_{\mathcal{H}^{\nu+d/2}(\Omega)}\le 1$. It follows immediately from Lemma EC.7 that

$$\|g\|_{\mathcal{L}_2(\Omega)} = O_{\mathbb{P}}\left(n^{-\frac{\nu+d/2}{2\nu+2d}}+\|g\|_n\right),$$

which, together with (EC.4.2), implies

$$\|f-f_\dagger\|_{\mathcal{L}_2(\Omega)} = O_{\mathbb{P}}\left(n^{-\frac{\nu+d/2}{2\nu+2d}}+\|f-f_\dagger\|_n\right) = O_{\mathbb{P}}\left(n^{-\frac{\nu+d/2}{2\nu+2d}}+\lambda^{1/2}\right). \qquad \text{(EC.4.5)}$$

By Lemma EC.4, it follows from (EC.4.4) and (EC.4.5) that for there exists a constant $C_2>0$ such that

$$\begin{aligned}\max_{1\le i\le n}|M_1(\mathbf{x}_i)| \le \|f-f_\dagger\|_{\mathcal{L}_\infty(\Omega)} &\le C_2\|f-f_\dagger\|_{\mathcal{L}_2(\Omega)}^{1-\frac{d}{2\nu+d}}\|f-f_\dagger\|_{\mathcal{N}_\Psi(\Omega)}^{\frac{d}{2\nu+d}}\\ &= O_{\mathbb{P}}\left(n^{-\frac{\nu}{2\nu+2d}}+\lambda^{\frac{\nu}{2\nu+d}}\right). \qquad \text{(EC.4.6)}\end{aligned}$$

Second way to bound $M_1(\mathbf{x}_i)$. A second way to bound $M_1(\mathbf{x}_i)$ is to work on $M_1(\mathbf{x}_i)$ directly without using the $\mathcal{L}_\infty$ bound. For $M_1(\mathbf{x}_i)$, Lemma F.8 in Wang (2020) asserts that

$$(f(\mathbf{x})-\mathbf{r}(\mathbf{x})^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{f})^2 \le (\Psi(\mathbf{0})-\mathbf{r}(\mathbf{x})^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{r}(\mathbf{x}))\|f\|^2_{\mathcal{N}_\Psi(\Omega)}.$$

Moreover, note that

$$\begin{aligned}\Psi(\mathbf{0})-\mathbf{r}(\mathbf{x}_i)^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{r}(\mathbf{x}_i) &= \mathbf{r}(\mathbf{x}_i)^\intercal\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}_i)-\mathbf{r}(\mathbf{x}_i)^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{r}(\mathbf{x}_i)\\ &= \mathbf{r}(\mathbf{x}_i)^\intercal\mathbf{R}^{-1}[(\mathbf{R}+n\lambda)\mathbf{I}-\mathbf{R}](\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{r}(\mathbf{x}_i)\\ &= n\lambda\mathbf{e}_i^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{r}(\mathbf{x}_i)\\ &\le n\lambda\sqrt{\mathbf{e}_i^\intercal\mathbf{e}_i}\sqrt{\mathbf{r}(\mathbf{x}_i)^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-2}\mathbf{r}(\mathbf{x}_i)}\\ &\le n\lambda\sqrt{\mathbf{r}(\mathbf{x}_i)^\intercal\mathbf{R}^{-2}\mathbf{r}(\mathbf{x}_i)} = n\lambda,\end{aligned}$$

where the first inequality follows from the Cauchy–Schwarz inequality. Hence,

$$|M_1(\mathbf{x}_i)|^2 \le (\Psi(\mathbf{0})-\mathbf{r}(\mathbf{x}_i)^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{r}(\mathbf{x}_i))\|f\|^2_{\mathcal{N}_\Psi(\Omega)} \le n\lambda\|f\|^2_{\mathcal{N}_\Psi(\Omega)}.$$

Because the above bound is uniform for all $\mathbf{x}_i$, we have

$$\max_{1\le i\le n}|M_1(\mathbf{x}_i)| = O_{\mathbb{P}}\left((n\lambda)^{1/2}\right). \qquad \text{(EC.4.7)}$$

Therefore, combining (EC.4.6) and (EC.4.7) completes the proof. □

LEMMA EC.9. *Suppose $f \in \mathcal{N}_\Psi(\Omega)$ and Assumptions 1 and 2 hold. Then,*

$$\max_{1\le i\le n}|M_2(\mathbf{x}_i)| = O_{\mathbb{P}}\left(\left(r_n^{1/2}\wedge 1\right)m^{-1/2}(\log n)^{1/2}\right),$$

where $r_n = \lambda^{-1}n^{-\frac{2\nu+d}{\nu+d}} + n^{-1}\lambda^{-\frac{d}{2\nu+d}}$.

Proof of Lemma EC.9. We begin with bounding $\mathbb{V}\mathrm{ar}(M_2(\mathbf{x}_i))$. We do it in two different ways and take the smaller one as the upper bound.

First way to bound $\mathbb{V}\mathrm{ar}(M_2(\mathbf{x}_i))$. Note that

$$\mathbb{V}\mathrm{ar}(M_2(\mathbf{x}_i)) \le \frac{\sigma^2}{m}\mathbf{r}(\mathbf{x}_i)^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-2}\mathbf{r}(\mathbf{x}_i). \tag{EC.4.8}$$

Fix $\mathbf{x}$. Consider the quadratic function

$$\mathcal{P}(\mathbf{u}) = \Psi(\mathbf{0}) - 2\sum_{i=1}^n \Psi(\mathbf{x}-\mathbf{x}_i)u_i + \sum_{i,j=1}^n u_iu_j\Psi(\mathbf{x}_i-\mathbf{x}_j) + n\lambda\|\mathbf{u}\|_2^2,$$

for $\mathbf{u} = (u_1,\dots,u_n)\in\mathbb{R}^n$. Clearly, $\mathbf{u}_* := (\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{r}(\mathbf{x})$ minimizes $\mathcal{P}(\mathbf{u})$. Since Ψ is positive definite,

$$\Psi(\mathbf{0}) - 2\sum_{i=1}^n \Psi(\mathbf{x}-\mathbf{x}_i)u_i + \sum_{i,j=1}^n u_iu_j\Psi(\mathbf{x}_i-\mathbf{x}_j) \ge 0$$

for all $\mathbf{u} = (u_1,\dots,u_n)\in\mathbb{R}^n$, which implies

$$\mathcal{P}(\mathbf{u}_*) \ge n\lambda\|\mathbf{u}_*\|_2^2 = n\lambda\mathbf{r}(\mathbf{x})^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-2}\mathbf{r}(\mathbf{x}). \tag{EC.4.9}$$

Direct calculation shows

$$\mathcal{P}(\mathbf{u}_*) = \Psi(\mathbf{0}) - \mathbf{r}(\mathbf{x})^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{r}(\mathbf{x}).$$

In order to obtain an upper bound of $\mathcal{P}(\mathbf{u}_*)$, we follow the idea from the proof of Lemma F.8 in Wang (2020).

For a fixed $\mathbf{x}$, define $h(\mathbf{t}) = \Psi(\mathbf{x}-\mathbf{t})$. Let $h_\dagger(\mathbf{t}) = \mathbf{r}(\mathbf{t})^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{r}(\mathbf{x})$. Clearly,

$$\Psi(\mathbf{0}) - \mathbf{r}(\mathbf{x})^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{r}(\mathbf{x}) \le \|h-h_\dagger\|_{\mathcal{L}_\infty(\Omega)}. \tag{EC.4.10}$$

By Lemma EC.1, it can be checked that $h_\dagger$ is the solution to the optimization problem:

$$h_\dagger = \operatorname*{arg\,min}_{g\in\mathcal{N}_\Psi(\Omega)} \frac{1}{n}\sum_{i=1}^n (g(\mathbf{x}_i)-h(\mathbf{x}_i))^2 + \lambda\|g\|^2_{\mathcal{N}_\Psi(\Omega)}. \tag{EC.4.11}$$

Note that $\|h-h_\dagger\|^2_{\mathcal{N}_\Psi(\Omega)}$ can be bounded by

$$\begin{aligned}\|h-h_\dagger\|^2_{\mathcal{N}_\Psi(\Omega)} &= \Psi(\mathbf{x}-\mathbf{x}) - 2\mathbf{r}(\mathbf{x})^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{r}(\mathbf{x}) + \mathbf{r}(\mathbf{x})^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{R}(\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{r}(\mathbf{x})\\ &\le \Psi(\mathbf{0}) - \mathbf{r}(\mathbf{x})^\intercal(\mathbf{R}+n\lambda\mathbf{I})^{-1}\mathbf{r}(\mathbf{x})\end{aligned}$$

$$\leq \|h - h_\dagger\|_{\mathcal{L}_\infty(\Omega)}. \tag{EC.4.12}$$

An upper bound on $\|h - h_\dagger\|_n$ can be obtained by

$$\begin{aligned}
\|h - h_\dagger\|_n^2 &= \|h - h_\dagger\|_n^2 + \lambda \|h_\dagger\|^2_{\mathcal{N}_\Psi(\Omega)} - \lambda \|h_\dagger\|^2_{\mathcal{N}_\Psi(\Omega)} \\
&\leq \|h - h\|_n^2 + \lambda \|h\|^2_{\mathcal{N}_\Psi(\Omega)} - \lambda \mathbf{r}(\mathbf{x})^\intercal (\mathbf{R} + n\lambda \mathbf{I})^{-1} \mathbf{R} (\mathbf{R} + n\lambda \mathbf{I})^{-1} \mathbf{r}(\mathbf{x}) \\
&= \lambda(\Psi(\mathbf{x} - \mathbf{x}) - \mathbf{r}(\mathbf{x})^\intercal (\mathbf{R} + n\lambda \mathbf{I})^{-1} \mathbf{R} (\mathbf{R} + n\lambda \mathbf{I})^{-1} \mathbf{r}(\mathbf{x})) \\
&= \lambda(\Psi(\mathbf{x} - \mathbf{x}) - \mathbf{r}(\mathbf{x})^\intercal (\mathbf{R} + n\lambda \mathbf{I})^{-1} \mathbf{r}(\mathbf{x}) \\
&\quad + \mathbf{r}(\mathbf{x})^\intercal (\mathbf{R} + n\lambda \mathbf{I})^{-1} \mathbf{r}(\mathbf{x}) - \mathbf{r}(\mathbf{x})^\intercal (\mathbf{R} + n\lambda \mathbf{I})^{-1} \mathbf{R} (\mathbf{R} + n\lambda \mathbf{I})^{-1} \mathbf{r}(\mathbf{x})) \\
&= \lambda(\Psi(\mathbf{x} - \mathbf{x}) - \mathbf{r}(\mathbf{x})^\intercal (\mathbf{R} + n\lambda \mathbf{I})^{-1} \mathbf{r}(\mathbf{x}) + n\lambda \mathbf{r}(\mathbf{x})^\intercal (\mathbf{R} + n\lambda \mathbf{I})^{-2} \mathbf{r}(\mathbf{x})) \\
&\leq 2\lambda(\Psi(\mathbf{x} - \mathbf{x}) - \mathbf{r}(\mathbf{x})^\intercal (\mathbf{R} + n\lambda \mathbf{I})^{-1} \mathbf{r}(\mathbf{x})) \\
&\leq 2\lambda \|h - h_\dagger\|_{\mathcal{L}_\infty(\Omega)},
\end{aligned}$$

where the first inequality is because $h_\dagger$ is the solution to the optimization problem (EC.4.11); the second inequality is by (EC.4.9); the last inequality is by (EC.4.10). Thus,

$$\|h - h_\dagger\|_n = O_{\mathbb{P}}(\lambda^{1/2} \|h - h_\dagger\|^{1/2}_{\mathcal{L}_\infty(\Omega)}). \tag{EC.4.13}$$

Because of the norm equivalence between $\mathcal{N}_\Psi(\Omega)$ and $\mathcal{H}^{\nu+d/2}(\Omega)$, there exists a constant C_3 such that

$$\|h - h_\dagger\|_{\mathcal{H}^{\nu+d/2}(\Omega)} \leq C_3 \|h - h_\dagger\|_{\mathcal{N}_\Psi(\Omega)} \leq C_3 \|h - h_\dagger\|^{1/2}_{\mathcal{L}_\infty(\Omega)}, \tag{EC.4.14}$$

where the last inequality is because of (EC.4.12). Let $g_1 = \frac{h - h_\dagger}{C_3 \|h - h_\dagger\|^{1/2}_{\mathcal{L}_\infty(\Omega)}}$. Thus, $\|g_1\|_{\mathcal{H}^{\nu+d/2}(\Omega)} \leq 1$. It follows from Lemma EC.7 that

$$\|g_1\|_{\mathcal{L}_2(\Omega)} = O_{\mathbb{P}}\left(n^{-\frac{\nu+d/2}{2\nu+2d}} + \|g_1\|_n \right),$$

which, by (EC.4.13), implies that

$$\begin{aligned}
\|h - h_\dagger\|_{\mathcal{L}_2(\Omega)} &= O_{\mathbb{P}}\left(n^{-\frac{\nu+d/2}{2\nu+2d}} \|h - h_\dagger\|^{1/2}_{\mathcal{L}_\infty(\Omega)} + \|h - h_\dagger\|_n \right) \\
&= O_{\mathbb{P}}\left(n^{-\frac{\nu+d/2}{2\nu+2d}} \|h - h_\dagger\|^{1/2}_{\mathcal{L}_\infty(\Omega)} + \lambda^{1/2} \|h - h_\dagger\|^{1/2}_{\mathcal{L}_\infty(\Omega)} \right).
\end{aligned} \tag{EC.4.15}$$

By Lemma EC.4, it follows from (EC.4.12) that

$$\begin{aligned}
\|h - h_\dagger\|_{\mathcal{L}_\infty(\Omega)} &\leq C_4 \|h - h_\dagger\|^{1 - \frac{d}{2\nu+d}}_{\mathcal{L}_2(\Omega)} \|h - h_\dagger\|^{\frac{d}{2\nu+d}}_{\mathcal{N}_\Psi(\Omega)} \leq C_4 \|h - h_\dagger\|^{1 - \frac{d}{2\nu+d}}_{\mathcal{L}_2(\Omega)} \|h - h_\dagger\|^{\frac{d}{2(2\nu+d)}}_{\mathcal{L}_\infty(\Omega)} \\
&= \|h - h_\dagger\|^{1/2}_{\mathcal{L}_\infty(\Omega)} O_{\mathbb{P}}\left(\left(n^{-\frac{\nu+d/2}{2\nu+2d}} + \lambda^{1/2} \right)^{\frac{2\nu}{2\nu+d}} \right)
\end{aligned}$$

which implies

$$\|h-h_{\dagger}\|_{\mathcal{L}_\infty(\Omega)} = O_{\mathbb{P}}\left(n^{-\frac{\nu}{\nu+d}} + \lambda^{\frac{2\nu}{2\nu+d}}\right). \tag{EC.4.16}$$

Recall that $\mathcal{P}(u_*) \le \|h-h_{\dagger}\|_{\mathcal{L}_\infty(\Omega)}$, which, together with (EC.4.9) and (EC.4.16), leads to

$$\begin{aligned}\mathbf{r}(\mathbf{x})^{\intercal}(\mathbf{R}+n\lambda\mathbf{I})^{-2}\mathbf{r}(\mathbf{x}) &= O_{\mathbb{P}}\left((n\lambda)^{-1}n^{-\frac{\nu}{\nu+d}} + (n\lambda)^{-1}\lambda^{\frac{2\nu}{2\nu+d}}\right)\\ &= O_{\mathbb{P}}\left(\lambda^{-1}n^{-\frac{2\nu+d}{\nu+d}} + n^{-1}\lambda^{-\frac{d}{2\nu+d}}\right),\end{aligned}$$

which, together with (EC.4.8), implies

$$\mathbb{V}\mathrm{ar}(M_2(\mathbf{x}_i)) = O_{\mathbb{P}}\left(\lambda^{-1}n^{-\frac{2\nu+d}{\nu+d}} + n^{-1}\lambda^{-\frac{d}{2\nu+d}}\right). \tag{EC.4.17}$$

Second way to bound $\mathbb{V}\mathrm{ar}(M_2(\mathbf{x}_i))$. With an argument similar to (EC.1.7), we have

$$\mathbb{V}\mathrm{ar}(M_2(\mathbf{x}_i)) \le \frac{\sigma^2}{m}\mathbf{r}(\mathbf{x}_i)^{\intercal}(\mathbf{R}+n\lambda\mathbf{I})^{-2}\mathbf{r}(\mathbf{x}_i) \le \frac{\sigma^2}{m}\mathbf{r}(\mathbf{x}_i)^{\intercal}\mathbf{R}^{-2}\mathbf{r}(\mathbf{x}_i) = \frac{\sigma^2}{m}. \tag{EC.4.18}$$

Let $r_n = \lambda^{-1}n^{-\frac{2\nu+d}{\nu+d}} + n^{-1}\lambda^{-\frac{d}{2\nu+d}}$ and $s_n = r_n \wedge 1$. Combining (EC.4.17) and (EC.4.18) yields

$$\mathbb{V}\mathrm{ar}(M_2(\mathbf{x}_i)) = m^{-1}O_{\mathbb{P}}(s_n).$$

Therefore, for any $\epsilon > 0$, there exist M_ϵ and N_ϵ such that $\mathbb{P}(\mathbb{V}\mathrm{ar}(M_2(\mathbf{x}_i)) > M_\epsilon m^{-1}s_n) < \epsilon$, when $n > N_\epsilon$. Take $N_0 = \max\{N_\epsilon, \epsilon^{-1}\}$. Note that $M_2(\mathbf{x}_i)$ is sub-Gaussian by Assumption 2. It follows that for all $n \ge N_0$,

$$\begin{aligned}&\mathbb{P}\left(\max_{1\le i\le n}|M_2(\mathbf{x}_i)| > 2\sqrt{\sigma^2 M_\epsilon m^{-1}s_n\log n}\right)\\ &= \mathbb{P}\left(\max_{1\le i\le n}|M_2(\mathbf{x}_i)| > 2\sqrt{\sigma^2 M_\epsilon m^{-1}s_n\log n}, \mathbb{V}\mathrm{ar}(M_2(\mathbf{x}_i)) \le M_\epsilon m^{-1}s_n\right)\\ &\quad + \mathbb{P}\left(\max_{1\le i\le n}|M_2(\mathbf{x}_i)| > 2\sqrt{\sigma^2 M_\epsilon m^{-1}s_n\log n}, \mathbb{V}\mathrm{ar}(M_2(\mathbf{x}_i)) > M_\epsilon m^{-1}s_n\right)\\ &\le \mathbb{P}\left(\text{There exists } i=1,\ldots,n \text{ such that } |M_2(\mathbf{x}_i)| > 2\sqrt{\sigma^2\,\mathbb{V}\mathrm{ar}(M_2(\mathbf{x}_i))\log n}\right) + \epsilon\\ &\le \sum_{i=1}^n \mathbb{P}\left(|M_2(\mathbf{x}_i)| > 2\sqrt{\sigma^2\,\mathbb{V}\mathrm{ar}(M_2(\mathbf{x}_i))\log n}\right) + \epsilon\\ &\le \sum_{i=1}^n 2\exp\left(-\frac{4\sigma^2\log n}{2\sigma^2}\right) + \epsilon = \frac{2}{n} + \epsilon \le 3\epsilon,\end{aligned}$$

where the first inequality follows from the probability union bound. Hence,

$$\max_{1\le i\le n}|M_2(\mathbf{x}_i)| = O_{\mathbb{P}}\left((r_n^{1/2}\wedge 1)m^{-1/2}(\log n)^{1/2}\right). \quad \square$$

Proof of Proposition 3. Applying Lemmas EC.8 and EC.9 to (EC.4.1) leads to

$$\rho_n = O_{\mathbb{P}}\left(\left(n^{-\frac{\nu}{2\nu+2d}} + \lambda^{\frac{\nu}{2\nu+d}}\right) \wedge (n\lambda)^{1/2} + \left(r_n^{1/2} \wedge 1\right) m^{-1/2} (\log n)^{1/2}\right). \tag{EC.4.19}$$

If we set $\lambda \asymp 1/(mn)$, then

$$r_n = mn^{-\frac{\nu}{\nu+d}} + n^{-\frac{2\nu}{2\nu+d}} m^{\frac{2\nu+d}{4\nu+d}} \leq 2mn^{-\frac{\nu}{\nu+d}}.$$

Noting $r_n^{1/2} \wedge 1 \leq r_n^{1/2} \wedge \sqrt{2}$, we have

$$\begin{aligned}
\rho_n &= O_{\mathbb{P}}\left(\left(n^{-\frac{\nu}{2\nu+2d}} + (mn)^{-\frac{\nu}{2\nu+d}}\right) \wedge m^{-1/2} + \left(r_n^{1/2} \wedge \sqrt{2}\right) m^{-1/2} (\log n)^{1/2}\right) \\
&= O_{\mathbb{P}}\left(\left(n^{-\frac{\nu}{2\nu+2d}} \wedge m^{-1/2}\right) + \left(m^{1/2} n^{-\frac{\nu}{2\nu+2d}} \wedge 1\right) m^{-1/2} (\log n)^{1/2}\right) \\
&= O_{\mathbb{P}}\left(\left(n^{-\frac{\nu}{2\nu+2d}} \wedge m^{-1/2}\right) + \left(n^{-\frac{\nu}{2\nu+2d}} \wedge m^{-1/2}\right) (\log n)^{1/2}\right) \\
&= O_{\mathbb{P}}\left(\left(n^{-\frac{\nu}{2\nu+2d}} \wedge m^{-1/2}\right) (\log n)^{1/2}\right). \quad \square
\end{aligned}$$

EC.5. Proof of Theorem 2

Proof of Theorem 2. Without loss of generality, we assume $z_0 = 0$ so that $\eta(z) = z^+$. To handle the term I_2 in the decomposition (EC.3.1), we use a smooth approximation of η as in Hong et al. (2017). Let

$$\eta_\delta(z) = \left(\frac{1}{2}(z+\delta) - \frac{\delta}{\pi}\cos\left(\frac{\pi}{2\delta} z\right)\right) \mathbb{I}\{-\delta \leq z \leq \delta\} + z\,\mathbb{I}\{z \geq \delta\}, \tag{EC.5.1}$$

where $\delta > 0$ is a parameter to be determined later. It can be verified that

$$\begin{aligned}
\eta_\delta'(z) &= \left(\frac{1}{2} + \frac{1}{2}\sin\left(\frac{\pi}{2\delta} z\right)\right) \mathbb{I}\{-\delta \leq z \leq \delta\} + \mathbb{I}\{z \geq \delta\}, \\
\eta_\delta''(z) &= \frac{\pi}{4\delta}\cos\left(\frac{\pi}{2\delta} z\right) \mathbb{I}\{-\delta \leq z \leq \delta\}.
\end{aligned} \tag{EC.5.2}$$

Moreover, $|\eta_\delta'(z)| \leq 1$, $|\eta_\delta''(z)| \leq \frac{\pi}{4\delta}$, and there exists a constant $C > 0$ such that

$$\begin{cases} |\eta(z) - \eta_\delta(z)| \leq C\delta, & \text{if } z \in [-\delta, \delta], \\ |\eta(z) - \eta_\delta(z)| = 0, & \text{otherwise.} \end{cases} \tag{EC.5.3}$$

By the triangle inequality, we have

$$I_2 \leq \underbrace{\left|\frac{1}{n}\sum_{i=1}^n \left[\eta(f(\mathbf{x}_i)) - \eta_\delta(f(\mathbf{x}_i))\right]\right|}_{J_1} + \underbrace{\left|\frac{1}{n}\sum_{i=1}^n \left[\eta_\delta(f(\mathbf{x}_i)) - \eta_\delta(\hat{f}(\mathbf{x}_i))\right]\right|}_{J_2} + \underbrace{\left|\frac{1}{n}\sum_{i=1}^n \left[\eta_\delta(\hat{f}(\mathbf{x}_i)) - \eta(\hat{f}(\mathbf{x}_i))\right]\right|}_{J_3}. \tag{EC.5.4}$$

Bound for J_1. It follows from (EC.5.3) that

$$J_1 \leq \left|\frac{C}{n}\sum_{i=1}^n \delta\,\mathbb{I}\{f(\mathbf{x}_i) \in [-\delta, \delta]\}\right| \leq C\delta \left|\frac{1}{n}\sum_{i=1}^n \mathbb{I}\{f(\mathbf{x}_i) \in [-\delta, \delta]\} - \mathbb{E}\left[\mathbb{I}\{f(\mathbf{x}_i) \in [-\delta, \delta]\}\right]\right|$$

$$+C\delta\,\mathbb{E}\,\mathbb{I}\{f(\mathbf{x}_i)\in[-\delta,\delta]\}=O_{\mathbb{P}}(\delta n^{-1/2})+O_{\mathbb{P}}(\delta^{\alpha+1}), \quad \text{(EC.5.5)}$$

for some constant $C>0$, where the last step follows from the central limit theorem and the fact that $\mathbb{E}\,\mathbb{I}\{f(\mathbf{x}_i)\in[-\delta,\delta]\}=\mathbb{P}(|f(\mathbf{x}_i)|\le\delta))=O(\delta^\alpha)$ by Assumption 4.

Bound for J_2**.** Applying Taylor's expansion to η_δ, we have

$$J_2=\underbrace{\left|\frac{1}{n}\sum_{i=1}^n\eta_\delta'(f(\mathbf{x}_i))(f(\mathbf{x}_i)-\hat f(\mathbf{x}_i))\right|}_{J_{21}}+\underbrace{\left|\frac{1}{2n}\sum_{i=1}^n\eta_\delta''(\tilde z_i)(f(\mathbf{x}_i)-\hat f(\mathbf{x}_i))^2\right|}_{J_{22}}, \quad \text{(EC.5.6)}$$

where $\tilde z_i$ is a value between $f(\mathbf{x}_i)$ and $\hat f(\mathbf{x}_i)$.

The first term J_{21} can be bounded using Proposition 1, which gives us

$$J_{21}=O_{\mathbb{P}}(\lambda^{1/2}+(mn)^{-1/2}). \quad \text{(EC.5.7)}$$

For J_{22}, by (EC.5.2), we find that

$$\begin{aligned}J_{22}&=\left|\frac{1}{2n}\sum_{i=1}^n\frac{\pi}{4\delta}\cos\left(\frac{\tilde z_i}{2\delta}\pi\right)\mathbb{I}\{-\delta\le\tilde z_i\le\delta\}(f(\mathbf{x}_i)-\hat f(\mathbf{x}_i))^2\right|\\&\le\frac{\pi}{8\delta}\left|\frac{1}{n}\sum_{i=1}^n\mathbb{I}\{-\delta\le\tilde z_i\le\delta\}(f(\mathbf{x}_i)-\hat f(\mathbf{x}_i))^2\right|. \quad \text{(EC.5.8)}\end{aligned}$$

Since $\tilde z_i$ is a value between $f(\mathbf{x}_i)$ and $\hat f(\mathbf{x}_i)$, $\tilde z_i\in[-\delta,\delta]$ implies that $f(\mathbf{x}_i)\in[-\delta-\rho_n,\delta+\rho_n]$, where $\rho_n=\max_{1\le i\le n}|f(\mathbf{x}_i)-\hat f(\mathbf{x}_i)|$. By (EC.5.8), J_{22} can be further bounded by

$$\begin{aligned}J_{22}&=O_{\mathbb{P}}\left(\frac{\pi}{8\delta}\left|\frac{1}{n}\sum_{i=1}^n\mathbb{I}\{-\delta-\rho_n\le f(\mathbf{x}_i)\le\delta+\rho_n\}(f(\mathbf{x}_i)-\hat f(\mathbf{x}_i))^2\right|\right)\\&=O_{\mathbb{P}}\left(\frac{1}{\delta}\left|\frac{\rho_n^2}{n}\sum_{i=1}^n\mathbb{I}\{-\delta-\rho_n\le f(\mathbf{x}_i)\le\delta+\rho_n\}\right|\right)\\&=O_{\mathbb{P}}\left(\frac{\rho_n^2}{\delta}(n^{-1/2}+(\delta+\rho_n)^\alpha)\right), \quad \text{(EC.5.9)}\end{aligned}$$

where the last step can be shown similarly for (EC.5.5). Plugging (EC.5.7) and (EC.5.9) in (EC.5.6), we have

$$J_2=O_{\mathbb{P}}\left(\lambda^{1/2}+(mn)^{-1/2}+\frac{\rho_n^2}{\delta}n^{-1/2}+\frac{\rho_n^2}{\delta}(\delta+\rho_n)^\alpha\right). \quad \text{(EC.5.10)}$$

Bound for J_3**.** Following the same argument for (EC.5.5), it can be seen that

$$J_3\le\left|\frac{1}{n}\sum_{i=1}^n\delta\,\mathbb{I}\{-\delta\le\hat f(\mathbf{x}_i)\le\delta\}\right|\le\left|\frac{1}{n}\sum_{i=1}^n\delta\,\mathbb{I}\{-\delta-\rho_n\le f(\mathbf{x}_i)\le\delta+\rho_n\}\right|=O_{\mathbb{P}}(\delta n^{-1/2}+\delta(\delta+\rho_n)^\alpha). \quad \text{(EC.5.11)}$$

In (EC.5.11), the first inequality holds because (EC.5.3), and the second inequality holds because $\hat{f}(\mathbf{x}_i) \in [-\delta, \delta]$ implies $f(\mathbf{x}_i) \in [-\delta - \rho_n, \delta + \rho_n]$.

Putting the bounds together. Combining (EC.3.1), (EC.3.3), (EC.5.4), (EC.5.5), (EC.5.10), and (EC.5.11), we have

$$\begin{aligned}|\hat{\theta}_{n,m} - \theta| &= O_{\mathbb{P}}\left(n^{-1/2} + \delta n^{-1/2} + \delta^{\alpha+1} + \frac{\rho_n^2}{\delta} n^{-1/2} + \frac{\rho_n^2}{\delta}(\delta + \rho_n)^{\alpha} + \lambda^{1/2} + (mn)^{-1/2} + \delta(\delta+\rho_n)^{\alpha}\right) \\ &= O_{\mathbb{P}}\left(n^{-1/2} + \delta^{\alpha+1} + \frac{\rho_n^2}{\delta} n^{-1/2} + \frac{\rho_n^2}{\delta}(\delta+\rho_n)^{\alpha} + \lambda^{1/2} + \delta\rho_n^{\alpha}\right). \qquad \text{(EC.5.12)}\end{aligned}$$

Then, by setting $\delta = \rho_n$ and $\lambda \asymp 1/(mn) = \Gamma^{-1}$, (EC.5.12) becomes $|\hat{\theta}_{n,m} - \theta| = O_{\mathbb{P}}\left(n^{-1/2} + \rho_n^{\alpha+1}\right)$.

At last, we apply Proposition 3 to conclude that: if $\nu \geq \frac{d}{\alpha+1}$, we set $n = \Gamma$ to obtain the rate $\max\{\Gamma^{-1/2}, \Gamma^{-\frac{\nu(\alpha+1)}{2\nu+2d}}(\log\Gamma)^{\frac{\alpha+1}{2}}\}$; if $\nu < \frac{d}{\alpha+1}$, we set $n \asymp \Gamma^{\frac{\alpha+1}{\alpha+2}}$ to obtain the rate $\Gamma^{-\frac{\alpha+1}{2(\alpha+2)}}(\log\Gamma)^{\frac{\alpha+1}{2}}$. □

EC.6. Proof of Theorem 3

Proof of Theorem 3. Without loss of generality, we assume $z_0 = 0$ so that $\eta(z) = \mathbb{I}\{z \geq 0\}$. Again, we work on the decomposition (EC.3.1). For the term I_2, note that if $\eta(f(\mathbf{x}_i)) \neq \eta(\hat{f}(\mathbf{x}_i))$, then we must have $f(\mathbf{x}_i) \in [-\rho_n, \rho_n]$, where $\rho_n = \max_{1\leq i\leq n} |f(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)|$. Therefore,

$$\begin{aligned}I_2 &= \left|\frac{1}{n}\sum_{i=1}^{n}\Big(\eta(f(\mathbf{x}_i)) - \eta(\hat{f}(\mathbf{x}_i))\Big)\mathbb{I}\{f(\mathbf{x}_i) \in [-\rho_n, \rho_n]\}\right| \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{f(\mathbf{x}_i) \in [-\rho_n, \rho_n]\} \\ &= O_{\mathbb{P}}(n^{-1/2} + \rho_n^{\alpha}), \qquad \text{(EC.6.1)}\end{aligned}$$

where the inequality holds because $|\eta(a) - \eta(b)| \leq 1$ for all $a, b \in \mathbb{R}$, and the last step from (EC.5.5).

Combining (EC.3.1), (EC.3.3), and (EC.6.1), we conclude that $|\hat{\theta}_{n,m} - \theta| = O_{\mathbb{P}}(n^{-1/2} + \rho_n^{\alpha})$. At last, we apply Proposition 3 to conclude that: if $\nu \geq \frac{d}{\alpha}$, we set $n = \Gamma$ to obtain the rate $\max\{\Gamma^{-1/2}, \Gamma^{-\frac{\nu\alpha}{2\nu+2d}}(\log\Gamma)^{\frac{\alpha}{2}}\} = \Gamma^{-\frac{\nu\alpha}{2\nu+2d}}(\log\Gamma)^{\frac{\alpha}{2}}$ because $\alpha \leq 1$; if $\nu < \frac{d}{\alpha}$, we set $n \asymp \Gamma^{\frac{\alpha}{\alpha+1}}$ to obtain the rate $\Gamma^{-\frac{\alpha}{2(\alpha+1)}}(\log\Gamma)^{\frac{\alpha}{2}}$. □

EC.7. Proof of Theorem 4

Lemma EC.10 **(Refined Hoeffding's Inequality for Bernoulli Random Variables)**. *Let $Z_1, \ldots, Z_n$ be independent Bernoulli random variables with parameter $p \in (0,1)$ such that $\mathbb{P}(Z_i = 1) = 1 - \mathbb{P}(Z_i = 0) = p$, $i = 1, \ldots, n$. Then,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(Z_i - p)\right| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{2p}\right), \quad \forall t > 0.$$

Proof of Lemma EC.10. Note that $\mathbb{E}[Z_i] = p$ and for any $s \in \mathbb{R}$,

$$\log\mathbb{E}[\exp(s(Z_i - p))] = \log\left(pe^{s(1-p)} + (1-p)e^{-sp}\right) = -sp + \log\left(1 + (e^s - 1)p\right)$$

$$\leq -sp + (e^s - 1)p \leq -sp + \left(s + \frac{s^2}{2}\right)p = \frac{s^2 p}{2},$$

where the first inequality holds because $\log(1+x) \leq x$ for all $x > -1$ and $(e^s - 1)p > -1$ for all $s \in \mathbb{R}$, and the second inequality holds because $e^x \geq 1 + x + \frac{x^2}{2}$ for all $x \in \mathbb{R}$. Hence, $Z_i \sim \mathsf{subG}(p)$ by definition. The proof is completed by applying Hoeffding's inequality for sub-Gaussian random variables (Wainwright 2019, Proposition 2.5). □

Proof of Theorem 4: The Case of VaR. Let $\zeta_{\mathsf{VaR}} := \mathsf{VaR}_\tau(f(X))$ and $\hat{\zeta} := \hat{f}_{(\lceil \tau n \rceil)}$ be its KRR-driven estimator. Moreover, let $\mathtt{G}(z)$, $\mathtt{G}_n(z)$, and $\hat{\mathtt{G}}_n(z)$ denote the cumulative distribution function (CDF) of $f(X)$, the empirical CDF of $f(X)$, and the empirical CDF of $\hat{f}(X)$, respectively:

$$\mathtt{G}(z) := \mathbb{P}(f(X) \leq z), \quad \mathtt{G}_n(z) := \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{f(\mathbf{x}_i) \leq z\}, \quad \text{and} \quad \hat{\mathtt{G}}_n(z) := \frac{1}{n}\sum_{i=1}^n \mathbb{I}\{\hat{f}(\mathbf{x}_i) \leq z\}.$$

Then, $\mathtt{G}(\zeta_{\mathsf{VaR}}) = \tau$ and $\hat{\mathtt{G}}_n(\hat{\zeta}) = \frac{\lceil \tau n \rceil}{n}$.

Note that

$$\begin{aligned}
|\mathtt{G}(\hat{\zeta}) - \mathtt{G}(\zeta_{\mathsf{VaR}})| &= \left|\mathbb{E}\big[\mathbb{I}\{f(X) \leq \hat{\zeta}\}\big] - \mathbb{E}\big[\mathbb{I}\{f(X) \leq \zeta_{\mathsf{VaR}}\}\big]\right| \\
&= \mathbb{E}\big[\mathbb{I}\{f(X) \in [\min(\hat{\zeta}, \zeta_{\mathsf{VaR}}), \max(\hat{\zeta}, \zeta_{\mathsf{VaR}})]\}\big] \\
&= \mathbb{P}\left(\left|f(X) - \frac{(\hat{\zeta} + \zeta_{\mathsf{VaR}})}{2}\right| \leq \frac{|\hat{\zeta} - \zeta_{\mathsf{VaR}}|}{2}\right) \\
&\geq C_2 |\hat{\zeta} - \zeta_{\mathsf{VaR}}|^\gamma,
\end{aligned}$$

for some constant $C_2 > 0$, where the last step follows from Assumption 5. Thus,

$$|\hat{\zeta} - \zeta_{\mathsf{VaR}}| = O\left(|\mathtt{G}(\hat{\zeta}) - \mathtt{G}(\zeta_{\mathsf{VaR}})|^{1/\gamma}\right). \tag{EC.7.1}$$

Next, we bound $|\mathtt{G}(\hat{\zeta}) - \mathtt{G}(\zeta_{\mathsf{VaR}})|$. Note that

$$|\mathtt{G}(\hat{\zeta}) - \mathtt{G}(\zeta_{\mathsf{VaR}})| \leq \underbrace{|\mathtt{G}(\hat{\zeta}) - \mathtt{G}_n(\hat{\zeta})|}_{V_1} + \underbrace{|\mathtt{G}_n(\hat{\zeta}) - \hat{\mathtt{G}}_n(\hat{\zeta})|}_{V_2} + \underbrace{|\hat{\mathtt{G}}_n(\hat{\zeta}) - \mathtt{G}(\zeta_{\mathsf{VaR}})|}_{V_3}. \tag{EC.7.2}$$

Let $\rho_n = \max_{1 \leq i \leq n} |f(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)|$. Because $\mathtt{G}(-\|f\|_{\mathcal{L}_\infty}) = 0$ and $\mathtt{G}(\|f\|_{\mathcal{L}_\infty}) = 1$, and $|\hat{\zeta}| = |\hat{f}_{(\lceil \tau n \rceil)}| \leq \|f\|_{\mathcal{L}_\infty} + \rho_n$ with $\rho_n \to 0$ in probability by Proposition 3, we can focus on the case $z \in [-2\|f\|_{\mathcal{L}_\infty}, 2\|f\|_{\mathcal{L}_\infty}]$ when analyzing V_i, $i = 1, 2, 3$. For simplicity, define $\mathcal{I} := [-2\|f\|_{\mathcal{L}_\infty}, 2\|f\|_{\mathcal{L}_\infty}]$.

Bound for V_1. By the Dvoretzky–Kiefer–Wolfowitz inequality (Massart 1990),

$$\mathbb{P}\left(\sup_{z \in \mathcal{I}} |\mathtt{G}_n(z) - \mathtt{G}(z)| > t\right) \leq \mathbb{P}\left(\sup_{z \in \mathbb{R}} |\mathtt{G}_n(z) - \mathtt{G}(z)| > t\right) \leq 2e^{-2nt^2}, \quad \forall t > 0.$$

This yields $\|\mathtt{G} - \mathtt{G}_n\|_{\mathcal{L}_\infty(\mathcal{I})} = O_{\mathbb{P}}(n^{-1/2})$. Hence,

$$V_1 = O_{\mathbb{P}}(\|\mathtt{G} - \mathtt{G}_n\|_{\mathcal{L}_\infty(\mathcal{I})}) = O_{\mathbb{P}}(n^{-1/2}). \tag{EC.7.3}$$

Bound for V_2. Let $\rho_n = \max_{1\le i\le n}|f(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)|$. Note that

$$\begin{aligned}|\mathtt{G}_n(z) - \hat{\mathtt{G}}_n(z)| &= \left|\frac{1}{n}\sum_{i=1}^{n}(\mathbb{I}\{f(\mathbf{x}_i)\le z\} - \mathbb{I}\{\hat{f}(\mathbf{x}_i)\le z\})\right| \\ &\le \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{f(\mathbf{x}_i)\in[z-\rho_n, z+\rho_n]\} \\ &= O_{\mathbb{P}}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{f(\mathbf{x}_i)\in[z-l_n, z+l_n]\}\right), \end{aligned} \tag{EC.7.4}$$

where the last step follows from (EC.4.19) and

$$l_n = \big(n^{-\frac{\nu}{2\nu+2d}} + \lambda^{\frac{\nu}{2\nu+d}}\big)\wedge(n\lambda)^{1/2} + \big(r_n^{1/2}\wedge 1\big)m^{-1/2}(\log n)^{1/2}, \tag{EC.7.5}$$

with $r_n = \lambda^{-1}n^{-\frac{2\nu+d}{\nu+d}} + n^{-1}\lambda^{-\frac{2\nu+d}{4\nu+d}}$.

Let $\delta_n = \max\{n^{-1/(2\beta)}, l_n\}$ and $M_n = \lceil 4\|f\|_{\mathcal{L}_\infty}/\delta_n\rceil$. Consider a partition of $\mathcal{I}$ as follows: let $-2\|f\|_{\mathcal{L}_\infty} = \zeta_0 < \zeta_1 < \cdots < \zeta_{M_n} = 2\|f\|_{\mathcal{L}_\infty}$ with $\zeta_{j+1} - \zeta_j = 4\|f\|_{\mathcal{L}_\infty}/M_n$, $j = 0,\ldots,M_n - 1$. Clearly, $\zeta_{j+1} - \zeta_j \le \delta_n$.

For any $z\in\mathcal{I}$, we take $j_* = \arg\min_{0\le j\le M_n}|z-\zeta_j|$ and note that $|z - \zeta_{j_*(z)}|\le\delta_n$. Thus,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{f(\mathbf{x}_i)\in[z-l_n, z+l_n]\} \\ &\le \left|\frac{1}{n}\sum_{i=1}^{n}(\mathbb{I}\{f(\mathbf{x}_i)\in[z-l_n, z+l_n]\} - \mathbb{I}\{f(\mathbf{x}_i)\in[\zeta_{j_*(z)}-l_n, \zeta_{j_*(z)}+l_n]\})\right| \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{f(\mathbf{x}_i)\in[\zeta_{j_*(z)}-l_n, \zeta_{j_*(z)}+l_n]\} \\ &\le \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{f(\mathbf{x}_i)\in[\zeta_{j_*(z)}-l_n-\delta_n, \zeta_{j_*(z)}+l_n+\delta_n]\} + \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{f(\mathbf{x}_i)\in[\zeta_{j_*(z)}-l_n, \zeta_{j_*(z)}+l_n]\} \\ &\le \frac{2}{n}\sum_{i=1}^{n}\mathbb{I}\{f(\mathbf{x}_i)\in[\zeta_{j_*(z)}-l_n-\delta_n, \zeta_{j_*(z)}+l_n+\delta_n]\}, \end{aligned} \tag{EC.7.6}$$

where the first inequality holds because of the triangle inequality, and the second because $|z-\zeta_{j_*(z)}|\le\delta_n$. It follows from (EC.7.4) and (EC.7.6) that

$$\sup_{z\in\mathcal{I}}|\mathtt{G}_n(z) - \hat{\mathtt{G}}_n(z)| = O_{\mathbb{P}}\Bigg(\max_{0\le j\le M_n}\underbrace{\frac{2}{n}\sum_{i=1}^{n}\mathbb{I}\{f(\mathbf{x}_i)\in[\zeta_j - l_n - \delta_n, \zeta_j + l_n + \delta_n]\}}_{Q_{n,j}}\Bigg). \tag{EC.7.7}$$

Let $p_n = \mathbb{P}(|f(X) - \zeta_j|\le l_n + \delta_n)$. Then, Assumption 5 implies that

$$p_n \le C_1(l_n+\delta_n)^\beta \le C_1\left(l_n + (n^{-1/(2\beta)} + l_n)\right)^\beta = O\left(l_n^\beta + n^{-1/2}\right), \tag{EC.7.8}$$

for some constant $C_1 > 0$, where second inequality follows from the definition of δ_n.

By the definition of l_n in (EC.7.5), if $\nu \geq \frac{d}{\beta}$, by Proposition 3, setting $n = \Gamma$ (so that $m = 1$) yields

$$l_n = O\left(n^{-\frac{\nu}{(2\nu+2d)}} (\log n)^{1/2}\right),$$

which, together with (EC.7.8), implies that

$$p_n = O\left(n^{-\frac{\nu}{(2\nu+2d)}} (\log n)^{1/2} + n^{-1/2}\right) = O\left(n^{-\frac{\nu}{(2\nu+2d)}} (\log n)^{1/2}\right). \tag{EC.7.9}$$

Otherwise, if $\nu < \frac{d}{\beta}$, then setting $n \asymp \Gamma^{\frac{\beta}{\beta+1}}$ (so that $m \asymp \Gamma^{\frac{1}{\beta+1}} \asymp n^{1/\beta}$) yields

$$l_n = O\left(n^{-1/(2\beta)} (\log n)^{1/2}\right),$$

which, together with (EC.7.8), implies that

$$p_n = O\left(n^{-1/2} (\log n)^{\beta/2} + n^{-1/2}\right) = O\left(n^{-1/2} (\log n)^{\beta/2}\right). \tag{EC.7.10}$$

Because $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are i.i.d., we have

$$\mathbb{E}\, Q_{n,j} = \frac{2}{n} \sum_{i=1}^{n} \mathbb{E}\, \mathbb{I}\left\{f(\mathbf{x}_i) \in [\zeta_j - l_n - \delta_n, \zeta_j + l_n + \delta_n]\right\} = 2p_n. \tag{EC.7.11}$$

Applying the probability union bound yields

$$\begin{aligned}
&\mathbb{P}\left(\max_{0\leq j\leq M_n} (Q_{n,j} - 2p_n) > 2n^{-1/2}\right) \\
&= \mathbb{P}\left(\text{There exists } j = 0, \ldots, M_n \text{ such that } Q_{n,j} - p_n > 2n^{-1/2}\right) \\
&\leq \sum_{0\leq j\leq M_n} \mathbb{P}\left(Q_{n,j} - 2p_n > 2n^{-1/2}\right) \\
&= \sum_{0\leq j\leq M_n} \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{I}\left\{f(\mathbf{x}_i) \in [\zeta_j - l_n - \delta_n, \zeta_j + l_n + \delta_n]\right\} - p_n\right) > n^{-1/2}\right) \\
&\leq 2(M_n + 1) \exp\left(-\frac{1}{2p_n}\right) \\
&\leq 2\left(4\|f\|_{\mathcal{L}_\infty} n^{1/(2\beta)} + 1\right) \exp\left(-\frac{1}{2p_n}\right) \to 0,
\end{aligned} \tag{EC.7.12}$$

as $n \to \infty$, where the second inequality follows from Lemma EC.10, the third from the definitions of M_n and δ_n, and the convergence to zero from (EC.7.9) and (EC.7.10).

It follows from (EC.7.7) and (EC.7.12) that

$$\begin{aligned}
\mathbb{P}\left(\|\mathtt{G}_n - \hat{\mathtt{G}}_n\|_{\mathcal{L}_\infty(\mathcal{I})} > 2p_n + 2n^{-1/2}\right) &\leq \mathbb{P}\left(\max_{0\leq j\leq M_n} Q_j > 2p_n + 2n^{-1/2}\right) \\
&= \mathbb{P}\left(\max_{0\leq j\leq M_n} (Q_{n,j} - 2p_n) > 2n^{-1/2}\right) \to 0,
\end{aligned}$$

as $n \to \infty$. Hence, $\|\mathtt{G}_n - \hat{\mathtt{G}}_n\|_{\mathcal{L}_\infty(\mathcal{I})} = O_{\mathbb{P}}(p_n + n^{-1/2}) = O_{\mathbb{P}}(l_n^\beta + n^{-1/2})$, implies that

$$V_2 = O_{\mathbb{P}}\left(\|\mathtt{G}_n - \hat{\mathtt{G}}_n\|_{\mathcal{L}_\infty(\mathcal{I})}\right) = O_{\mathbb{P}}(l_n^\beta + n^{-1/2}). \tag{EC.7.13}$$

Bound for V_3.

$$V_3 = |\hat{\mathtt{G}}_n(\hat{\zeta}) - \mathtt{G}(\zeta_{\mathsf{VaR}})| = \left|\frac{\lceil \tau n\rceil}{n} - \tau\right| \le n^{-1}. \tag{EC.7.14}$$

Putting the bounds together. Plugging (EC.7.3), (EC.7.13), and (EC.7.14) into (EC.7.2) leads to

$$\begin{aligned}|\mathtt{G}(\hat{\zeta}) - \mathtt{G}(\zeta_{\mathsf{VaR}})| &= O_{\mathbb{P}}(n^{-1/2}) + O_{\mathbb{P}}(l_n^\beta + n^{-1/2}) + O(n^{-1}) \\ &= O_{\mathbb{P}}(l_n^\beta + n^{-1/2}),\end{aligned}$$

which, together with (EC.7.1), implies that

$$|\hat{\zeta} - \zeta_{\mathsf{VaR}}| = O_{\mathbb{P}}(l_n^{\beta/\gamma} + n^{-1/(2\gamma)}). \tag{EC.7.15}$$

At last, we apply Proposition 3 to conclude that: if $\nu \ge \frac{d}{\beta}$, we set $n = \Gamma$ to obtain the rate $\Gamma^{-\frac{\nu\beta}{(2\nu+2d)\gamma}}(\log\Gamma)^{\frac{\beta}{2\gamma}}$; if $\nu < \frac{d}{\beta}$, we set $n \asymp \Gamma^{\frac{\beta}{\beta+1}}$ to obtain the rate $\Gamma^{-\frac{\beta}{2\gamma(\beta+1)}}(\log\Gamma)^{\frac{\beta}{2\gamma}}$. □

Proof of Theorem 4: The Case of CVaR. Let $\hat{z} = n^{-1}\sum_{i=1}^n (\hat{f}(\mathbf{x}_i) - \hat{f}_{(\lceil \tau n\rceil)})^+$. By the triangle inequality, we have

$$|\hat{\theta}_{n,m} - \mathsf{CVaR}_\tau(f(X))| \le |\hat{\zeta} - \zeta_{\mathsf{VaR}}| + (1-\tau)^{-1}|\hat{z} - \mathbb{E}[(f(X) - \mathsf{VaR}_\tau(f(X)))^+]|. \tag{EC.7.16}$$

Note that

$$\begin{aligned}|\hat{z} - \mathbb{E}[f(X) - \mathsf{VaR}_\tau(f(X)))^+]| \le{}& \underbrace{\left|\frac{1}{n}\sum_{i=1}^n (\hat{f}(\mathbf{x}_i) - \hat{f}_{(\lceil \tau n\rceil)})^+ - \frac{1}{n}\sum_{i=1}^n (f(\mathbf{x}_i) - \zeta_{\mathsf{VaR}})^+\right|}_{W_1} \\ &+ \underbrace{\left|\frac{1}{n}\sum_{i=1}^n (f(\mathbf{x}_i) - \zeta_{\mathsf{VaR}})^+ - \mathbb{E}[(f(X) - \mathsf{VaR}_\tau(f(X)))^+]\right|}_{W_2}.\end{aligned} \tag{EC.7.17}$$

For W_1, applying the basic inequality $|\max(a,0) - \max(b,0)| \le |a-b|$ yields

$$\begin{aligned}W_1 &\le \frac{1}{n}\sum_{i=1}^n \left|(\hat{f}(\mathbf{x}_i) - \hat{f}_{(\lceil \tau n\rceil)}) - (f(\mathbf{x}_i) - \zeta_{\mathsf{VaR}})\right| \le \frac{1}{n}\sum_{i=1}^n \Big(|\hat{f}(\mathbf{x}_i) - f(\mathbf{x}_i)| + |\hat{\zeta} - \zeta_{\mathsf{VaR}}|\Big) \\ &\le \rho_n + |\hat{\zeta} - \zeta_{\mathsf{VaR}}| = O_{\mathbb{P}}(l_n) + |\hat{\zeta} - \zeta_{\mathsf{VaR}}|,\end{aligned} \tag{EC.7.18}$$

where $\rho_n = \max_{1\le i\le n}|f(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)|$, l_n is given by (EC.7.5), and the last step follows from Proposition 3.

For W_2, the central limit theorem implies that

$$W_2 = O_{\mathbb{P}}(n^{-1/2}). \tag{EC.7.19}$$

Combining (EC.7.16)–(EC.7.19) yields

$$|\hat{\theta}_{n,m} - \mathsf{CVaR}_\tau(f(X))| = O_{\mathbb{P}}(l_n + |\hat{\zeta} - \zeta_{\mathsf{VaR}}| + n^{-1/2}). \tag{EC.7.20}$$

Plugging (EC.7.15) into (EC.7.20), we have

$$\begin{aligned}|\hat{\theta}_{n,m} - \mathsf{CVaR}_\tau(f(X))| &= O_{\mathbb{P}}(l_n + l_n^{\beta/\gamma} + n^{-1/(2\gamma)} + n^{-1/2}) \\ &= O_{\mathbb{P}}(l_n^{\beta/\gamma} + n^{-1/(2\gamma)} + n^{-1/2}),\end{aligned} \tag{EC.7.21}$$

where the second equality holds because $\beta \le \gamma$ by Assumption 5. Therefore, since $\gamma \ge 1$, the convergence rate is $O_{\mathbb{P}}(l_n^{\beta/\gamma} + n^{-1/(2\gamma)})$, which is the same as that of $|\hat{\zeta} - \zeta_{\mathsf{VaR}}|$, given by the case of VaR in Theorem 4. □

References

Bishop YM, Fienberg SE, Holland PW (2007) *Discrete Multivariate Analysis: Theory and Practice.* (Springer)

Brezis H, Mironescu P (2019) Where Sobolev interacts with Gagliardo-Nirenberg. *Journal of Functional Analysis* 277(8):2839–2864.

Edmunds DE, Triebel H (1996) *Function Spaces, Entropy Numbers, Differential Operators* (Cambridge University Press).

Massart P (1990) The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.* 18(3):1269–1283.

Schölkopf B, Smola AJ (2002) *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (MIT Press).

Tuo R, Wang Y, Wu CFJ (2020) On the improved rates of convergence for Matérn-type kernel ridge regression with application to calibration of computer models. *SIAM/ASA J. Uncertainty Quantification* 8(4):1522–1547.

van de Geer S (2000) *Empirical Processes in M-Estimation* (Cambridge University Press).

Wainwright MJ (2019) *High-Dimensional Statistics: A Non-Asymptotic Viewpoint* (Cambridge University Press).

Wang W (2020) On the inference of applying Gaussian process modeling to a deterministic function. Preprint available at `https://arxiv.org/abs/2002.01381`.