When can Regression-Adjusted Control Variates Help?
Rare Events, Sobolev Embedding and Minimax Optimality

Jose Blanchet
Department of MS&E and ICME
Stanford University
Stanford, CA 94305
jose.blanchet@stanford.edu

Haoxuan Chen
ICME
Stanford University
Stanford, CA 94305
haoxuanc@stanford.edu

Yiping Lu
Courant Institute of Mathematical Sciences
New York University
New York, NY 10012
yiping.lu@nyu.edu

Lexing Ying
Department of Mathematics and ICME
Stanford University
Stanford, CA 94305
lexing@stanford.edu

Abstract

This paper studies the use of a machine learning-based estimator as a control variate for mitigating the variance of Monte Carlo sampling. Specifically, we seek to uncover the key factors that influence the efficiency of control variates in reducing variance. We examine a prototype estimation problem that involves simulating the moments of a Sobolev function based on observations obtained from (random) quadrature nodes. Firstly, we establish an information-theoretic lower bound for the problem. We then study a specific quadrature rule that employs a nonparametric regression-adjusted control variate to reduce the variance of the Monte Carlo simulation. We demonstrate that this kind of quadrature rule can improve the Monte Carlo rate and achieve the minimax optimal rate under a sufficient smoothness assumption. Due to the Sobolev Embedding Theorem, the sufficient smoothness assumption eliminates the existence of rare and extreme events. Finally, we show that, in the presence of rare and extreme events, a truncated version of the Monte Carlo algorithm can achieve the minimax optimal rate while the control variate cannot improve the convergence rate.

1 Introduction

In this paper, we consider a nonparametric quadrature rule on (random) quadrature points based on regression-adjusted control variate [1, 2, 3, 4]. To construct the quadrature rule, we partition our available data into two halves. The first half is used to construct a nonparametric estimator, which is then utilized as a control variate to reduce the variance of the Monte Carlo algorithm implemented over the second half of our data. Traditional and well-known results [1, Chapter 5.2] show that the optimal linear control variate can be obtained via Ordinary Least Squares regression. In this paper, we investigate a similar idea for constructing a quadrature rule [3, 5, 6, 7, 8, 9, 10], which uses a non-parametric machine learning-based estimator as a regression-adjusted control variate. We aim to answer the following two questions:

> *Is using optimal nonparametric machine learning algorithms to construct control variates an optimal way to improve Monte Carlo methods? What are the factors that determine the effectiveness of the control variate?*

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

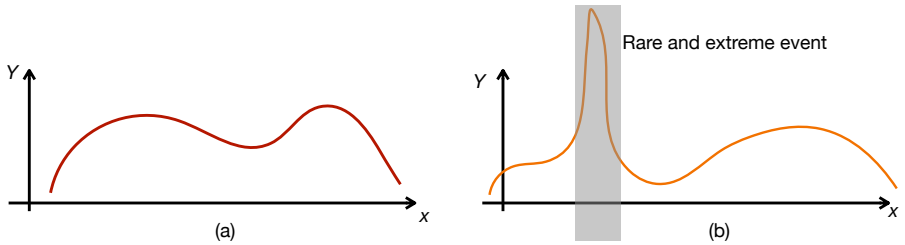

Figure 1: According to the Sobolev Embedding Theorem [11], the Sobolev space $W^{s,p}$ can be embedded in L^{p^*}, where $\frac{1}{p^*} = \frac{1}{p} - \frac{s}{d}$. When s is large enough, as shown in (a), the smoothness assumption can rule out the existence of rare and extreme events. When s is not sufficiently large, specifically $s < \frac{2dq-dp}{2pq}$, there may exist a peak (*a.k.a* rare and extreme event) that makes the Monte Carlo simulation hard. Under such circumstances, the function's $2q$-th moment is unbounded.

To understand the two questions, we consider a basic but fundamental prototype problem of estimating moments of a Sobolev function from its values observed on (random) quadrature nodes, which has a wide range of applications in Bayesian inference, the study of complex systems, computational physics, and financial risk management [1]. Specifically, we estimate the q-th moment $\int_\Omega f(x)^q dx$ of f based on values $f(x_1), \cdots, f(x_n)$ observed on n (random) quadrature nodes $x_1, \cdots, x_n \in \Omega$ for a function f in the Sobolev space $W^{s,p}(\Omega)$, where $\Omega \subset \mathbb{R}^d$. The parameter q here is introduced to characterize the rare events' extremeness for estimation. To verify the effectiveness of the non-parametric regression adjusted quadrature rule, we first study the statistical limit of the problem by providing a minimax information-theoretic lower bound of magnitude $n^{\max\{(\frac{1}{p}-\frac{s}{d})q-1, -\frac{s}{d}-\frac{1}{2}\}}$.

We also provide matching upper bounds for different levels of function smoothness. Under the sufficient smoothness assumption that $s > \frac{d(2q-p)}{2pq}$, we find that the non-parametric regression adjusted control variate $\hat{f}$ can improve the rate of classical Monte Carlo algorithm and help us attain a minimax optimal upper bound. In (3.4) below, we bound variance $\int_\Omega (f^q - \hat{f}^q)^2$ of the Monte Carlo target by the sum of the semi-parametric influence part $\int_\Omega f^{2q-2}(f-\hat{f})^2$ and the propagated estimation error $\int_\Omega (f-\hat{f})^{2q}$. Although the optimal algorithm in this regime remains the same, we need to consider three different cases to derive an upper bound on the semi-parametric influence part, which is the main contribution of our proof. We propose a new proof technique that embeds the square of the influence function $(qf^{q-1})^2$ and estimation error $(f-\hat{f})^2$ in appropriate spaces via the Sobolev Embedding Theorem [11]. The two norms used for evaluating $(f^{q-1})^2$ and $(f-\hat{f})^2$ should be dual norms of each other. Also, we should select the norm for evaluating $(f-\hat{f})^2$ in a way that it's easy to estimate f under the selected norm, which helps us control the error induced by $(f-\hat{f})^2$. A detailed explanation of how to select the proper norms in different cases via the Sobolev Embedding Theorem is exhibited in Figure 2. In the first regime when $s > \frac{d}{p}$, we can directly embed f in $L^\infty(\Omega)$ and attain a final convergence rate of magnitude $n^{-\frac{s}{d}-\frac{1}{2}}$. For the second regime when $\frac{d(2q-p)}{p(2q-2)} < s < \frac{d}{p}$, the smoothness parameter s is not large enough to ensure that $f \in L^\infty(\Omega)$. Thus, we evaluate the estimation error $(f-\hat{f})^2$ under the $L^{\frac{p}{2}}$ norm and embed the square of the influence function $(qf^{q-1})^2$ in the dual space of $L^{\frac{p}{2}}(\Omega)$. Here the validity of such embedding is ensured by the lower bound $\frac{d(2q-p)}{p(2q-2)}$ on s. Moreover, the semi-parametric influence part is still dominant in the second regime, so the final convergence rate is the same as that of the first case. In the third regime, when $\frac{d(2q-p)}{2pq} < s < \frac{d(2q-p)}{p(2q-2)}$, the semi-parametric influence no longer dominates and the final converge rate transits from $n^{-\frac{s}{d}-\frac{1}{2}}$ to $n^{q(\frac{1}{p}-\frac{s}{d})-1}$.

When the sufficient smoothness assumption breaks, *i.e.* $s < \frac{d(2q-p)}{2pq}$, according to the Sobolev Embedding Theorem [11], the Sobolev space $W^{s,p}$ is embedded in $L^{\frac{dp}{d-sp}}$ and $\frac{dp}{d-sp} < 2q$. This indicates that rare and extreme events might be present, and they are not even guaranteed to have bounded L^{2q} norm, which makes the Monte Carlo estimate of the q-th moment have infinite variance. Under this scenario, we consider a truncated version of the Monte Carlo algorithm, which can be proved to attain the minimax optimal rate of magnitude $n^{q(\frac{1}{p}-\frac{s}{d})-1}$. In contrast, the usage of

regression-adjusted control variates does not improve the convergence rate under this scenario. Our results reveal how the existence of rare events will change answers to the questions raised at the beginning of the section.

We also use the estimation of a linear functional as an example to investigate the algorithm's adaptivity to the noise level. In this paper, we provide minimax lower bounds for estimating the integral of a fixed function with a general assumption on the noise level. Specifically, we consider all estimators that have access to observations $\{x_i, f(x_i) + \epsilon_i\}_{i=1}^n$ of some function f that is s-Hölder smooth, where $x_i \overset{\text{i.i.d}}{\sim} \text{Uniform}([0,1]^d)$ and $\epsilon_i \overset{\text{i.i.d}}{\sim} n^{-\gamma}\mathcal{N}(0,1)$ for some $\gamma > 0$. Based on the method of two fuzzy hypotheses, we present a lower bound of magnitude $n^{\max\{-\frac{1}{2}-\gamma, -\frac{1}{2}-\frac{s}{d}\}}$, which exhibits a smooth transition from the Monte Carlo rate to the Quasi-Monte Carlo rate. At the same time, our information-theoretic lower bound also matches the upper bound built for quadrature rules taking use of non-parametric regression-adjusted control variates.

1.1 Related Work

Regression-Adjusted Control Variate The control variate method is a technique used for variance reduction in Monte-Carlo simulation. Consider the task of estimating the expectation $\mathbb{E}X$ for some random variable X. The idea of control variate method is to introduce another random variable Y correlated with the random variable X, such that the random variable $X - Y$ has smaller variance than X. Since $\mathbb{E}X = \mathbb{E}[X-Y] + \mathbb{E}[Y]$ and $\mathbb{E}[Y]$ is deterministic, one may obtain a variance reduced estimator of $\mathbb{E}[X]$ by summing up $\mathbb{E}[Y]$ and an empirical estimate of $\mathbb{E}[X - Y]$. Such a random variable Y is called a control variate. Regression-adjusted control variate, in particular, refers to the case when Y is obtained by applying regression methods to observed data samples of X.

Regression-adjusted control variates have shown both theoretical and empirical improvements in a wide range of applications, including the construction of confidence intervals [12, 13], randomized trace-estimation [14, 15], dimension reduction [16], causal inference [17], light transport simulation [18], MCMC simulation [19], estimation of the normalizing factor [10] and gradient estimation [20, 21]. It is also used as a technique for proving the approximation bounds on two-layer neural networks in the Barron space [22].

Regarding literature most related to our work, we mention [3, 7, 8, 10], which also study the theoretical properties of nonparametric control variate estimator. However, the theoretical analysis in [3, 7] does not provide a specific convergence rate in the Reproducing Kernel Hilbert Space, which requires a high level of smoothness for the underlying function. In contrast to prior work, our research delves into the effectiveness of a non-parametric regression-adjusted control variate in boosting convergence rates across various degrees of smoothness assumptions and identifies the key factor that determines the efficacy of these control variates.

Quadrature Rule There is a long literature on building quadrature rules in the Reproducing Kernel Hilbert Space, including Bayes–Hermite quadrature [23, 24, 25, 26, 27], determinantal point processes [28, 29, 30, 31], Nyström approximation [32, 33], kernel herding[34, 35, 36] and kernel thinning [37, 38, 39]. Nevertheless, the quadrature points chosen in these studies all have the ability to reconstruct the function's information, which results in a suboptimal rate for estimating the moments.

Functional Estimation There are also lines of research that investigated the optimal rates of estimating both linear [8, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50] and nonlinear [51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64] functionals, such as integrals and the L^q norm. However, as far as the authors know, previous works on this topic have assumed sufficient smoothness, which rules out the existence of rare and extreme events that are hard to simulate. Additionally, existing proof techniques are only applicable in scenarios where there is either no noise or a constant level of noise present. We have developed a novel and unified proof technique that leverages the method of two fuzzy hypotheses, which allows us to account for not only rare and extreme events but also different levels of noise.

1.2 Contribution

- We determine all the regimes when a quadrature rule utilizing a nonparametric estimator as a control variate to reduce the Monte Carlo estimate's variance can boost the convergence rate of estimating the moments of a Sobolev function. Under sufficient smoothness assumption,

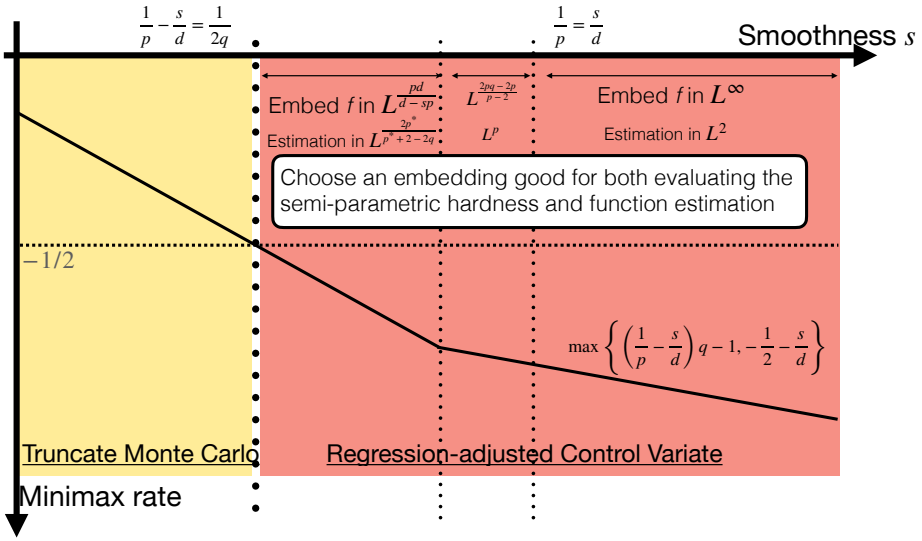

Figure 2: We summarize the minimax optimal rates and the corresponding optimal algorithms with respect to the function smoothness here. When the function is smooth enough, regression-adjusted control variates can improve the Monte Carlo rate. However, when there exist rare and extreme events that are hard to simulate, truncating the Monte Carlo estimate directly yields a minimax optimal algorithm. Above the transition point of algorithm selection is $s = \frac{d(2q-p)}{2pq}$, while the transition point of the optimal convergence rate is $s = \frac{d(2q-p)}{p(2q-2)}$. To build the optimal convergence guarantee for any algorithm that utilizes a regression-adjusted control variate $\hat{f}$, we need to embed the square of the influence function $(qf^{q-1})^2$ in an appropriate space via the Sobolev Embedding Theorem and evaluate the estimation error $(f-\hat{f})^2$ under the dual norm of the norm associated with the chosen space, which allows us to achieve optimal semi-parametric efficiency. Our selections of the metrics in different regimes are shown in this figure.

which rules out the existence of rare and extreme events due to the Sobolev Embedding Theorem, the regression-adjusted control variate improves the convergence rate and achieves the minimax optimal rate. Without the sufficient smoothness assumption, however, there may exist rare and extreme events that are hard to simulate. In this circumstance, we discover that a truncated version of the Monte Carlo method is minimax optimal, while regression-adjusted control variate can't improve the convergence rate.

- As far as the authors know, our paper is the first work considering this problem without assuming that the underlying function f is uniformly bounded. All previous work assumed that $s > \frac{d}{p}$, which implies $f \in L^\infty(\Omega)$ and neglects the possibility of spike functions. As a result, they were unable to discover the transition between the two regimes described above. Under the assumption that $s > \frac{d(2q-p)}{2pq}$, the main difficulty in establishing the convergence guarantee lies in determining the right evaluation metric for function estimation. To select a suitable metric, we introduce a new proof technique by embedding the influence function into an appropriate space via the Sobolev Embedding Theorem and evaluating the function estimation in the corresponding dual norm to achieve optimal semi-parametric efficiency. Our selection of the proper embedding metrics is shown in Figure 2.
- To study how the regression adjusted control variate adapts to the noise level, we examine the linear functionals, *i.e.* the definite integral. We prove that this method is minimax optimal regardless of the level of noise present in the observed data.

1.3 Notations

Let $\|\cdot\|$ be the standard Euclidean norm and $\Omega = [0,1]^d$ be the unit cube in $\mathbb{R}^d$ for any fixed $d \in \mathbb{N}$. Also, let $\mathbb{1} = \mathbb{1}\{\cdot\}$ denote the indicator function, *i.e*, for any event A we have $\mathbb{1}\{A\} = 1$ if A is

true and $\mathbb{1}\{A\} = 0$ otherwise. For any region $R \subseteq \Omega$, we use $V(R) := \int_\Omega \mathbb{1}\{x \in R\}dx$ to denote the volume of R. Let $C(\Omega)$ denote the space of all continuous functions $f : \Omega \to \mathbb{R}$ and $\lfloor \cdot \rfloor$ be the rounding function. For any $s > 0$ and $f \in C(\Omega)$, we define the Hölder norm $\|\cdot\|_{C^s(\Omega)}$ by

$$\|f\|_{C^s(\Omega)} := \max_{|k| \le \lfloor s \rfloor} \|D^k f\|_{L^\infty(\Omega)} + \max_{|k| = \lfloor s \rfloor} \sup_{x,y \in \Omega, x \neq y} \frac{|D^k f(x) - D^k f(y)|}{\|x - y\|^{s - \lfloor s \rfloor}}. \tag{1.1}$$

The corresponding Hölder space is defined as $C^s(\Omega) := \Big\{ f \in C(\Omega) : \|f\|_{C^s(\Omega)} < \infty \Big\}$. When $s = 0$, we have that the two norms $\|\cdot\|_{C^0(\Omega)}$ and $\|\cdot\|_{L^\infty(\Omega)}$ are equivalent and $C^0(\Omega) = L^\infty(\Omega)$. Let $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$ be the set of all non-negative integers. For any $s \in \mathbb{N}_0$ and $1 \le p \le \infty$, we define the Sobolev space $W^{s,p}(\Omega)$ by

$$W^{s,p}(\Omega) := \Big\{ f \in L^p(\Omega) : D^\alpha f \in L^p(\Omega), \forall\, \alpha \in \mathbb{N}_0^d \text{ satisfying } |\alpha| \le s \Big\}. \tag{1.2}$$

Let $(c)_+$ denote $\max\{c, 0\}$ for any $c \in \mathbb{R}$. Fix any two non-negative sequences $\{a_n\}_{n=1}^\infty$ and $\{b_n\}_{n=1}^\infty$. We write $a_n \lesssim b_n$, or $a_n = O(b_n)$, to denote that $a_n \le C b_n$ for some constant C independent of n. Similarly, we write $a_n \gtrsim b_n$, or $a_n = \omega(b_n)$, to denote that $a_n \ge c b_n$ for some constant c independent of n. We use $a_n = \Theta(b_n)$ to denote that $a_n = O(b_n)$ and $a_n = \omega(b_n)$.

2 Information-Theoretic Lower Bound on Moment Estimation

Problem Setup To understand how the non-parametric regression-adjusted control variate improves the Monte Carlo estimator's convergence rate, we consider a prototype problem that estimates a function's q-th moment. For any fixed $q \in \mathbb{N}$ and $f \in W^{s,p}(\Omega)$, we want to estimate the q-th moment $I_f^q := \int_\Omega f^q(x)dx$ with n random quadrature points $\{x_i\}_{i=1}^n \subset \Omega$. On each quadrature point x_i $(i = 1, \cdots, n)$, we can observe the function value $y_i := f(x_i)$.

In this section, we study the information-theoretic limit for the problem above via the method of two fuzzy hypotheses [65]. We have the following information-theoretic lower bound on the class $\mathcal{H}_n^{f,q}$ that contains all estimators $\hat{H}^q : \Omega^n \times \mathbb{R}^n \to \mathbb{R}$ of the q-th moment I_f^q.

Theorem 2.1 (Lower Bound on Estimating the Moment) *When $p > 2$ and $q < p < 2q$, let $\mathcal{H}_n^f$ denote the class of all the estimators that use n quadrature points $\{x_i\}_{i=1}^n$ and observed function values $\{y_i = f(x_i)\}_{i=1}^n$ to estimate the q-th moment of f, where $\{x_i\}_{i=1}^n$ are independently and identically sampled from the uniform distribution on Ω. Then we have*

$$\inf_{\hat{H}^q \in \mathcal{H}_n^{f,q}} \sup_{f \in W^{s,p}(\Omega)} \mathbb{E}_{\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n} \left[\left| \hat{H}^q\Big(\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n\Big) - I_f^q \right| \right] \gtrsim n^{\max\left\{-q\left(\frac{s}{d} - \frac{1}{p}\right) - 1, -\frac{1}{2} - \frac{s}{d}\right\}}. \tag{2.1}$$

Proof Sketch Here we give a sketch for our proof of Theorem 2.1. Our proof is based on the method of two fuzzy hypotheses, which is a generalization of the traditional Le Cam's two-point method. In fact, each hypothesis in the generalized method is constructed via a prior distribution. In order to attain a lower bound of magnitude Δ via the method of two fuzzy hypotheses, one needs to pick two prior distributions μ_0, μ_1 on the Sobolev space $W^{s,p}(\Omega)$ such that the following two conditions hold. Firstly, the estimators I_f^q differ by Δ with constant probability under the two priors. Secondly, the TV distance between the two corresponding distributions $\mathbb{P}_0$ and $\mathbb{P}_1$ of data generated by μ_0 and μ_1 is of constant magnitude. In order to prove the two lower bounds given in (2.1), we pick two different pairs of prior distributions as follows:

Below we set $m = \Theta(n^{\frac{1}{d}})$ and divide the domain Ω into m^d small cubes $\Omega_1, \Omega_2, \cdots, \Omega_{m^d}$, each of which has side length m^{-1}. For any $p \in (0, 1)$, we use v_p, w_p to denote the discrete random variables satisfying $\mathbb{P}(v_p = 0) = \mathbb{P}(w_p = -1) = p$ and $\mathbb{P}(v_p = 1) = \mathbb{P}(w_p = 1) = 1 - p$.

(I) For the first lower bound in (2.1), we construct some bump function $g \in W^{s,p}(\Omega)$ satisfying $\text{supp}(g) \subseteq \Omega_1$ and $I_g^q = \int_{\Omega_1} g(x)dx = \Theta(m^{q(-s + \frac{d}{p}) - d})$. Now let's take some sufficiently small constant $\epsilon \in (0, 1)$ and pick μ_0, μ_1 to be discrete measures supported on the two finite sets $\Big\{ v_{\frac{1+\epsilon}{2}} g \Big\}$ and $\Big\{ v_{\frac{1-\epsilon}{2}} g \Big\}$. On the one hand, the difference between the q-th moments under μ_0 and μ_1 can be

lower bounded by $\Theta(n^{q(\frac{1}{p}-\frac{s}{d})-1})$ with constant probability. On the other hand, $KL(\mathbb{P}_0\|\mathbb{P}_1)$ can be upper bounded by the KL divergence between $v_{\frac{1+\epsilon}{2}}$ and $v_{\frac{1-\epsilon}{2}}$, which is of constant magnitude.

(II) For the second lower bound in (2.1), we set $M > 0$ to be some sufficiently large constant and $\kappa = \Theta(\frac{1}{\sqrt{n}})$. For any $1 \le j \le m^d$, we construct bump functions $f_j \in W^{s,p}(\Omega)$ satisfying $\text{supp}(f_j) \subseteq \Omega_j$ and $I_{f_j}^k = \int_{\Omega_j} f_j(x)dx = \Theta(m^{-ks-d})$ for any $1 \le j \le m^d$ and $1 \le k \le s$. Now let's pick μ_0, μ_1 to be discrete measures supported on the two finite sets $\left\{M+\sum_{j=1}^{m^d} w_j^{(0)} f_j\right\}$ and $\left\{M+\sum_{j=1}^{m^d} w_j^{(1)} f_j\right\}$, where $\{w_j^{(0)}\}_{j=1}^{m^d}$ and $\{w_j^{(1)}\}_{j=1}^{m^d}$ are independent and identical copies of $w_{\frac{1+\kappa}{2}}$ and $w_{\frac{1-\kappa}{2}}$ respectively. On the one hand, applying Hoeffding's inequality yields that the q-th moments under μ_0 and μ_1 differ by $\Theta(n^{-\frac{s}{d}-\frac{1}{2}})$ with constant probability. On the other hand, note that $KL(\mathbb{P}_0\|\mathbb{P}_1)$ can be bounded by the KL divergence between two multivariate discrete distributions $(w_{j_1}^{(0)}, \cdots, w_{j_n}^{(0)})$ and $(w_{j_1}^{(1)}, \cdots, w_{j_n}^{(1)})$, where $\{w_{j_i}^{(0)}\}_{i=1}^n$ and $\{w_{j_i}^{(1)}\}_{i=1}^n$ are independent and identical copies of $w_{\frac{1+\kappa}{2}}$ and $w_{\frac{1-\kappa}{2}}$ respectively. Hence, $KL(\mathbb{P}_0\|\mathbb{P}_1)$ is of constant magnitude.

Combining the two cases above gives us the minimax lower bound in (2.1). We defer a complete proof of Theorem 2.1 to Appendix B.2.

3 Minimax Optimal Estimators for Moment Estimation

This section is devoted to constructing minimax optimal estimators of the q-th moment. We show that under the sufficient smoothness assumption, a regression-adjusted control variate is essential for building minimax optimal estimators. However, when the given function is not sufficiently smooth, we demonstrate that a truncated version of the Monte Carlo algorithm is minimax optimal, and control variates cannot give any improvement.

3.1 Sufficient Smoothness Regime: Non-parametric Regression-Adjusted Control Variate

This subsection is devoted to building a minimax optimal estimator of the q-th moment under the assumption that $\frac{s}{d} > \frac{1}{p} - \frac{1}{2q}$, which guarantees that functions in the space $W^{s,p}$ are sufficiently smooth. From the Sobolev Embedding theorem, we know that the sufficient smoothness assumption implies $W^{s,p}(\Omega) \subset L^{p^*}(\Omega) \subset L^{2q}(\Omega)$, where $\frac{1}{p^*} = \frac{1}{p} - \frac{s}{d}$. Given any function $f \in W^{s,p}(\Omega)$ along with n uniformly sampled quadrature points $\{x_i\}_{i=1}^n$ and corresponding observations $\{y_i = f(x_i)\}_{i=1}^n$ of f, the key idea behind the construction of our estimator $\hat{H}_C^q$ is to build a nonparametric estimation $\hat{f}$ of f based on a sub-dataset and use $\hat{f}$ as a control variate for Monte Carlo simulation. Consequently, it takes three steps to compute the numerical estimation of I_f^q for any estimator $\hat{H}_C^q : \Omega^n \times \mathbb{R}^n \to \mathbb{R}$. The first step is to divide the observed data into two subsets $\mathcal{S}_1 := \{(x_i, y_i)\}_{i=1}^{\frac{n}{2}}, \mathcal{S}_2 := \{(x_i, y_i)\}_{i=\frac{n}{2}+1}^n$ of equal size and use a machine learning algorithm to compute a nonparametric estimation $\hat{f}_{1:\frac{n}{2}}$ of f based on $\mathcal{S}_1$. Without loss of generality, we may assume that the number of data points is even. Secondly, we treat $\hat{f}_{1:\frac{n}{2}}$ as a control variate and compute the q-th moment $I_{\hat{f}}^q$. Using the other dataset $\mathcal{S}_2$, we may obtain a Monte Carlo estimate of $I_f^q - I_{\hat{f}_{1:\frac{n}{2}}}^q$ as follows: $I_f^q - I_{\hat{f}_{1:\frac{n}{2}}}^q \approx \frac{2}{n}\sum_{i=\frac{n}{2}+1}^n \left(y_i^q - \hat{f}_{1:\frac{n}{2}}^q(x_i)\right)$. Finally, combining the estimation of the q-th moment $I_{\hat{f}_{1:\frac{n}{2}}}^q = \int_\Omega \hat{f}_{1:\frac{n}{2}}^q(x)dx$ with the estimation of $I_f^q - I_{\hat{f}_{1:\frac{n}{2}}}^q$ gives us the numerical estimation returned by $\hat{H}_C^q$:

$$\hat{H}_C^q\Big(\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n\Big) := \int_\Omega \hat{f}_{1:\frac{n}{2}}^q(x)dx + \frac{2}{n}\sum_{i=\frac{n}{2}+1}^n \Big(y_i^q - \hat{f}_{1:\frac{n}{2}}^q(x_i)\Big). \tag{3.1}$$

We assume that our function estimation $\hat{f}$ is obtained from an $\frac{n}{2}$-oracle $K_{\frac{n}{2}} : \Omega^{\frac{n}{2}} \times \mathbb{R}^{\frac{n}{2}} \to W^{s,p}(\Omega)$ satisfying Assumption 3.1. For example, there are lines of research [49, 50, 59, 60, 61] considering how the moving least squares method [66, 67] can achieve the convergence rate in (3.2).

Assumption 3.1 (Optimal Function Estimator as an Oracle) *Given any function $f \in W^{s,p}(\Omega)$ and $n \in \mathbb{N}$, let $\{x_i\}_{i=1}^n$ be n data points sampled independently and identically from the uniform distribution on Ω. Assume that for $s > \frac{2dq-dp}{2pq}$, there exists an oracle $K_n : \Omega^n \times \mathbb{R}^n \to W^{s,p}(\Omega)$ that estimates f based on the n points $\{x_i\}_{i=1}^n$ along with the n observed function values $\{f(x_i)\}_{i=1}^n$ and satisfies the following bound for any r satisfying $\frac{1}{r} \in \left(\max\{\frac{d-sp}{pd}, 0\}, \max\{\frac{1}{p}, \mathbb{1}\{s > \frac{d}{p}\}\}\right]$:*

$$\left(\mathbb{E}_{\{x_i\}_{i=1}^n}\left[\|K_n(\{x_i\}_{i=1}^n, \{f(x_i)\}_{i=1}^n) - f\|_{L^r(\Omega)}^r\right]\right)^{\frac{1}{r}} \lesssim n^{-\frac{s}{d}+(\frac{1}{p}-\frac{1}{r})_+}. \tag{3.2}$$

A construction of the desired oracle and a complete proof of the upper bound above (up to logarithm factors) is deferred to Appendix E. Based on the oracle above, we can obtain the following upper bound that matches the information-theoretic lower bound in Theorem 2.1.

Theorem 3.1 (Upper Bound on Moment Estimation with Sufficient Smoothness) *Assume that $p > 2$, $q < p < 2q$ and $s > \frac{2dq-dp}{2pq}$. Let $\{x_i\}_{i=1}^n$ be n quadrature points independently and identically sampled from the uniform distribution on Ω and $\{y_i := f(x_i)\}_{i=1}^n$ be the corresponding n observations of $f \in W^{s,p}(\Omega)$. Then the estimator $\hat{H}_C^q$ constructed in (3.1) above satisfies*

$$\mathbb{E}_{\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n}\left[\left|\hat{H}_C^q\Big(\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n\Big) - I_f^q\right|\right] \lesssim n^{\max\{-q(\frac{s}{d}-\frac{1}{p})-1, -\frac{s}{d}-\frac{1}{2}\}}, \tag{3.3}$$

Proof Sketch Given a non-parametric estimator $\hat{f}$ of the function f, we may bound the variance of the Monte Carlo process by $(f^q - \hat{f}^q)^2$ and further upper bound it by the sum of the following two terms:

$$|f^q - \hat{f}^q|^2 \lesssim \underbrace{|f^{q-1}(f-\hat{f})|^2}_{\text{semi-parametric influnce}} + \underbrace{|(f-\hat{f})^q|^2}_{\text{estimation error propagation}}. \tag{3.4}$$

The first term above represents the semi-parametric influence part of the problem, as qf^{q-1} is the influence function for the estimation of the q-th moment f^q. The second term characterizes how function estimation affects functional estimation. If we consider the special case of estimating the mean instead of a general q-th moment, *i.e,* $q = 1$, the semi-parametric influence term will disappear. Consequently, the convergence rate won't transit from $n^{-\frac{1}{2}-\frac{s}{d}}$ to $n^{-q(\frac{s}{d}-\frac{1}{p})-1}$ in the special case.

Although the algorithm remains unchanged in the sufficient smooth regime, we need to consider three separate cases to obtain an upper bound on the integral of the semi-parametric influence term $|f^{q-1}(f-\hat{f})|^2$ in (3.4). An illustration of the three cases is given in Figure 2.

From Hölder's inequality, we know that $\int_\Omega f^{2q-2}(x)(f(x)-\hat{f}(x))^2dx$ can be upper bounded by $\|f^{2q-2}\|_{L^{r'}(\Omega)}\|(f-\hat{f})^2\|_{L^{r_*}(\Omega)}$, where $\|\cdot\|_{L^{r'}(\Omega)}$ and $\|\|_{L^{r_*}(\Omega)}$ are dual norms. Therefore, the main difficulty here is to embed the function f in different spaces via the Sobolev Embedding Theorem under different assumptions on the smoothness parameter s. When the function is smooth enough, *i.e.* $s > \frac{d}{p}$, we embed the function f in $L^\infty(\Omega)$ and evaluate the estimation error $f - \hat{f}$ under the L^2 norm. Then our assumption on the oracle (3.2) gives us an upper bound of magnitude $n^{-\frac{2s}{d}}$ on $\|f-\hat{f}\|_{L^2(\Omega)}^2$, which helps us further upper bound the semi-parametric influence part $\int_\Omega f^{2q-2}(x)(f(x)-\hat{f}(x))^2dx$ by $n^{-\frac{2s}{d}}$ up to constants. When $\frac{d(2q-p)}{p(2q-2)} < s < \frac{d}{p}$, we embed the function f in $L^{\frac{2pq-2p}{p-2}}(\Omega) \subseteq L^{\frac{pd}{d-sp}}(\Omega)$ and evaluate the estimation error $f - \hat{f}$ under the L^p norm. Applying our assumption on the oracle (3.2) again implies that the semi-parametric influence part $\int_\Omega f^{2q-2}(x)(f(x)-\hat{f}(x))^2dx$ can be upper bounded by $n^{-\frac{2s}{d}}$ up to constants. When $\frac{d(2q-p)}{2pq} < s < \frac{d(2q-p)}{p(2q-2)}$, we embed the function f in L^{p^*} and evaluate the error of the oracle in $L^{\frac{2p^*}{p^*+2-2q}}$, where $\frac{1}{p^*} = \frac{1}{p} - \frac{s}{d}$. Similarly, we can use (3.2) to upper bound the semi-parametric influence part $\int_{x\in\Omega} f^{2q-2}(x)(f(x)-\hat{f}(x))^2dx$ by $n^{2q(\frac{1}{p}-\frac{s}{d})-1}$.

The upper bound on the propagated estimation error $\int_{x\in\Omega}(f(x)-\hat{f}(x))^{2q}dx$ in (3.4) can be derived by evaluating the error of the oracle under the L^{2q} norm. *i.e,* by picking $r = 2q$ in (3.2) above, which yields an upper bound of magnitude $n^{2q(\frac{1}{p}-\frac{s}{d})-1}$.

The obtained upper bounds on the semi-parametric influence part and the propagated estimation error above provide us with a clear view of the upper bound on the variance of $f^q - \hat{f}^q$, which is the random variable we aim to simulate via Monte-Carlo in the second stage. Using the standard Monte-Carlo algorithm to simulate the expectation of $f^q - \hat{f}^q$ then gives us an extra $n^{-\frac{1}{2}}$ factor for the convergence rate, which helps us attain the final upper bounds given in (3.3). A complete proof of Theorem 3.1 is given in Appendix C.1.

3.2 Beyond the Sufficient Smoothness Regime: Truncated Monte Carlo

In this subsection, we study the case when the sufficient smoothness assumption breaks, *i.e.* $\frac{s}{d} < \frac{1}{p} - \frac{1}{2q}$. According to the Sobolev Embedding theorem, we have that W_s^p is embedded in $L^{\frac{dp}{d-sp}}$. Since $\frac{1}{p} - \frac{s}{d} > \frac{1}{2q}$ implies $\frac{dp}{d-sp} < 2q$, the underlying function f is not guaranteed to have bounded L^{2q} norm, which indicates the existence of rare and extreme events. Consequently, the Monte Carlo estimate of f's q-th moment must have infinite variance, which makes it hard to simulate. Here we present a truncated version of the Monte Carlo algorithm that can achieve the minimax optimal convergence rate. For any fixed parameter $M > 0$, our estimator is designed as follows:

$$\hat{H}_M^q\Big(\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n\Big) := \frac{1}{n}\sum_{i=1}^n \max\Big\{\min\{y_i, M\}, -M\Big\}^q. \tag{3.5}$$

In Theorem 3.2, we provide the convergence rate of the estimator (3.5) by choosing the truncation parameter M in an optimal way.

Theorem 3.2 (Upper Bound on Moment Estimation without Sufficient Smoothness) *Assuming that $p > 2$, $q < p < 2q$ and $s < \frac{2dq-dp}{2pq}$, we pick $M = \Theta(n^{\frac{1}{p}-\frac{s}{d}})$. Let $\{x_i\}_{i=1}^n$ be n quadrature points independently and identically sampled from the uniform distribution on Ω and $\{y_i := f(x_i)\}_{i=1}^n$ be the corresponding n observations of $f \in W^{s,p}(\Omega)$. Then we have that the estimator $\hat{H}_M^q$ constructed in (3.5) above satisfies*

$$\mathbb{E}_{\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n}\left[\left|\hat{H}_M^q\Big(\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n\Big) - I_f^q\right|\right] \lesssim n^{-q(\frac{s}{d}-\frac{1}{p})-1}. \tag{3.6}$$

Proof Sketch The error can be decomposed into bias and variance parts. The bias part is caused by the truncation in our algorithm, which is controlled by the parameter M and can be bounded by $\int_{\{x:|f(x)|>M\}} |f|^q dx$. According to the Sobolev Embedding Theorem, $W^{s,p}(\Omega)$ can be embedded in the space L^{p^*}, where $\frac{1}{p^*} = \frac{1}{p} - \frac{s}{d}$. As $|f(x)| > M$ implies $|f(x)|^q < M^{q-p^*}|f(x)|^{p^*}$, the bias can be upper bounded by M^{q-p^*}. Similarly, the variance is controlled by M and can be upper bounded by $M^{q-\frac{p^*}{2}}$. Combining the bias and variance bound, we can bound the final error as $M^{q-p^*} + \frac{M^{q-\frac{p^*}{2}}}{\sqrt{n}}$. By selecting $M = \Theta(n^{\frac{1}{p^*}}) = \Theta(n^{\frac{1}{p}-\frac{s}{d}})$, we obtain the final convergence rate $n^{-q(\frac{s}{d}-\frac{1}{p})-1}$. A complete proof of Theorem 3.2 is given in Appendix C.2.

Remark 3.1 *[64] has shown that the convergence rate of the optimal non-parametric regression-based estimation is $n^{-\frac{s}{d}+\frac{1}{p}-\frac{1}{q}}$, which is slower than the convergence rate of the truncated Monte Carlo estimator that we show above.*

4 Adapting to the Noise Level: a Case Study for Linear Functional

In this section, we study how the regression-adjusted control variate adapts to different noise levels. Here we consider the linear functional, *i.e.* estimating a function's definite integral via low-noise observations at random points.

Problem Setup We consider estimating $I_f = \int_\Omega f(x)dx$, the integral of f over Ω, for a fixed function $f \in C^s(\Omega)$ with uniformly sampled quadrature points $\{x_i\}_{i=1}^n \subset \Omega$. On each quadrature point x_i $(i = 1, \cdots, n)$, we have a noisy observation $y_i := f(x_i) + \epsilon_i$. Here the ϵ_i's are independently and identically distributed Gaussian noises sampled from $\mathcal{N}(0, n^{-2\gamma})$, where $\gamma \in [0, \infty]$.

4.1 Information-Theoretic Lower Bound on Mean Estimation

In this subsection, we present a minimax lower bound (Theorem 4.1) for all estimators $\hat{H} : \Omega^n \times \mathbb{R}^n \to \mathbb{R}$ of the integral I_f of a function $f \in C^s(\Omega)$ when one can only access noisy observations.

Theorem 4.1 (Lower Bound for Integral Estimation) *Let $\mathcal{H}_n^f$ denote the class of all the estimators that use n quadrature points $\{x_i\}_{i=1}^n$ and noisy observations $\{y_i = f(x_i) + \epsilon_i\}_{i=1}^n$ to estimate the integral of f, where $\{x_i\}_{i=1}^n$ and $\{\epsilon_i\}_{i=1}^n$ are independently and identically sampled from the uniform distribution on Ω and the normal distribution $\mathcal{N}(0, n^{-2\gamma})$ respectively. Assuming that $\gamma \in [0, \infty]$ and $s > 0$, we have*

$$\inf_{\hat{H} \in \mathcal{H}_n^f} \sup_{f \in C^s(\Omega)} \mathbb{E}_{\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n} \left[\left| \hat{H}\Big(\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n \Big) - I_f \right| \right] \gtrsim n^{\max\{-\frac{1}{2}-\gamma, -\frac{1}{2}-\frac{s}{d}\}}. \quad (4.1)$$

Remark 4.1 *Functional estimation is a well-studied problem in the literature of nonparametric statistics. However, current information-theoretic lower bounds for functional estimation [51, 52, 53, 56, 57, 58, 65, 68] assume a constant level of noise on the observed function values. One essential idea for proving these lower bounds is to leverage the existence of the observational noise, which enables us to upper bound the amount of information required to distinguish between two reduced hypotheses. In contrast, we provide a minimax lower bound that is applicable for noises at any level by constructing two priors with overlapping support and assigning distinct probabilities to the corresponding Bernoulli random variables, which separates the two hypotheses. A comprehensive proof of Theorem 4.1 is given in Appendix D.2.*

4.2 Optimal Nonparametric Regression-Adjusted Quadrature Rule

In the discussion below, we use the nearest-neighbor method as an example. For any $k \in \{1, 2, \cdots, \frac{n}{2}\}$, the k-nearest neighbor estimator $\hat{f}_{k\text{-NN}}$ of f is given by $\hat{f}_{k\text{-NN}}(z) := \frac{1}{k} \sum_{j=1}^k y_{i_j^{(z)}}$, where $\{x_{i_j^{(z)}}\}_{j=1}^{\frac{n}{2}}$ is a permutation of the quadrature points $\{x_i\}_{i=1}^{\frac{n}{2}}$ such that $\|x_{i_1^{(z)}} - z\| \leq \|x_{i_2^{(z)}} - z\| \leq \cdots \leq \|x_{i_{\frac{n}{2}}^{(z)}} - z\|$ holds for any $z \in \Omega$. Moreover, we use $\mathcal{T}_{k,z} := \{x_{i_j^{(z)}}\}_{j=1}^k$ to denote the collection of the k nearest neighbors of z among $\{x_i\}_{i=1}^{\frac{n}{2}}$ for any $z \in \Omega$. For any $1 \leq i \leq \frac{n}{2}$, we take $D_i \subset \Omega$ to be the region formed by all the points whose k nearest neighbors contain x_i, *i.e,* $D_i := \Big\{ z \in \Omega : x_i \in \mathcal{T}_{k,z} \Big\}$. Our estimator $\hat{H}_{k\text{-NN}}$ can be formally represented as

$$\hat{H}_{k\text{-NN}}\Big(\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n \Big) = \underbrace{\sum_{i=1}^{\frac{n}{2}} \frac{V(D_i)}{k} y_i}_{\int_\Omega \hat{f}_{k\text{-NN}}(x)dx} + \underbrace{\frac{2}{n} \sum_{i=\frac{n}{2}+1}^{n} y_i - \frac{2}{n} \sum_{i=\frac{n}{2}+1}^{n} \Big(\frac{1}{k} \sum_{j=1}^{\frac{n}{2}} \mathbb{1}\{x_i \in D_j\} y_j \Big)}_{\frac{2}{n} \sum_{i=\frac{n}{2}+1}^{n} \Big(y_i - \hat{f}_{k\text{-NN}}(x_i) \Big)}.$$

In the following theorem, we present an upper bound on the expected risk of the estimator $\hat{H}_{k\text{-NN}}$:

Theorem 4.2 (Matching Upper Bound for Integral Estimation) *Let $\{x_i\}_{i=1}^n$ be n quadrature points independently and identically sampled from the uniform distribution on Ω and $\{y_i := f(x_i) + \epsilon_i\}_{i=1}^n$ be the corresponding n noisy observations of $f \in C^s(\Omega)$, where $\{\epsilon_i\}_{i=1}^n$ are independently and identically sampled from the normal distribution $\mathcal{N}(0, n^{-2\gamma})$. Assuming that $\gamma \in [0, \infty]$ and $s \in (0, 1)$, we have that there exists $k \in \mathbb{N}$ such that the estimator $\hat{H}_{k\text{-NN}}$ constructed above satisfies*

$$\mathbb{E}_{\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n} \left[\left| \hat{H}_{k\text{-NN}}\Big(\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n \Big) - I_f \right| \right] \lesssim n^{\max\{-\frac{1}{2}-\gamma, -\frac{1}{2}-\frac{s}{d}\}}. \quad (4.2)$$

Remark 4.2 *Our upper bound in Theorem 4.2 matches our minimax lower bound in Theorem 4.1, which indicates that the regression-adjusted quadrature rule associated with the nearest neighbor estimator is minimax optimal. When the noise level is high ($\gamma < \frac{s}{d}$), the control variate helps*

to improve the rate from $n^{-\frac{1}{2}}$ (the Monte Carlo rate) to $n^{-\frac{1}{2}-\gamma}$ via eliminating **all** *the effects of simulating the smooth function. When the noise level is low ($\gamma > \frac{s}{d}$), we show that our estimator $\hat{H}_{k\text{-}NN}$ can achieve the optimal rate of quadrature rules [46]. We defer a complete proof of Theorem 4.2 to Appendix D.3.*

5 Discussion and Conclusion

In this paper, we have investigated whether a non-parametric regression-adjusted control variate can improve the rate of estimating functionals and its minimax optimality. Using the Sobolev Embedding Theorem, we discover that the existence of infinite variance rare and extreme events will change the answer to this question. We show that when infinite variance rare and extreme events are present, using a non-parametric machine learning algorithm as a control variate does not help to improve the convergence rate, and truncated Monte Carlo is minimax optimal. When the variance of the simulation problem is finite, using a regression-adjusted control variate via an optimal non-parametric estimator is minimax optimal.

The assumptions we made in this paper, such as boundedness of the domain Ω and constraints on the parameters p, q, might be too restrictive for some application scenarios. We left relaxations of these assumptions as future work. One other potential direction is to investigate how to combine importance sampling with regression-adjusted control variates. Also, the study of how regression-adjusted control variates adapt to the noise level for non-linear functionals [62, 63] may be of interest. Moreover, another intriguing project is to analyze how the data distribution's information [3, 7] can be used to achieve both better computational trackability and convergence rate [8].

Acknowledgments and Disclosure of Funding

Jose Blanchet is supported in part by the Air Force Office of Scientific Research (AFOSR) under award number FA9550-20-1-0397 and the National Science Foundation (NSF) under award number DMS-1915967. Haoxuan Chen is supported by the T. S. Lo Graduate Fellowship Fund. Yiping Lu is supported by the Stanford Interdisciplinary Graduate Fellowship (SIGF). Lexing Ying is supported by the National Science Foundation (NSF) under award number DMS-2011699 and DMS-2208163.

References

[1] Søren Asmussen and Peter W Glynn. *Stochastic simulation: algorithms and analysis*, volume 57. Springer, 2007.

[2] Russell Davidson and James G MacKinnon. Regression-based methods for using control variates in monte carlo experiments. *Journal of Econometrics*, 54(1-3):203–222, 1992.

[3] Chris Oates and Mark Girolami. Control functionals for quasi-monte carlo integration. In *Artificial Intelligence and Statistics*, pages 56–65. PMLR, 2016.

[4] Fred J Hickernell, Christiane Lemieux, and Art B Owen. Control variates for quasi-monte carlo. *Statistical Science*, 20(1):1 – 31, 2005.

[5] Roland Assaraf and Michel Caffarel. Zero-variance principle for monte carlo algorithms. *Physical review letters*, 83(23):4682, 1999.

[6] Antonietta Mira, Reza Solgi, and Daniele Imparato. Zero variance markov chain monte carlo for bayesian estimators. *Statistics and Computing*, 23:653–662, 2013.

[7] Chris J Oates, Mark Girolami, and Nicolas Chopin. Control functionals for monte carlo integration. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 695–718, 2017.

[8] Chris J Oates, Jon Cockayne, François-Xavier Briol, and Mark Girolami. Convergence rates for a class of estimators based on stein's method. *Bernoulli*, 25(2):1141 – 1159, 2019.

[9] Leah F South, CJ Oates, A Mira, and C Drovandi. Regularised zero-variance control variates. *arXiv preprint arXiv:1811.05073*, 2018.

[10] David Holzmüller and Francis Bach. Convergence rates for non-log-concave sampling and log-partition estimation. *arXiv preprint arXiv:2303.03237*, 2023.

[11] Robert A Adams and John JF Fournier. *Sobolev spaces*. Elsevier, 2003.

[12] Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. Prediction-powered inference. *arXiv preprint arXiv:2301.09633*, 2023.

[13] Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

[14] Raphael A Meyer, Cameron Musco, Christopher Musco, and David P Woodruff. Hutch++: Optimal stochastic trace estimation. In *Symposium on Simplicity in Algorithms (SOSA)*, pages 142–155. SIAM, 2021.

[15] Lin Lin. Randomized estimation of spectral densities of large matrices made accurate. *Numerische Mathematik*, 136:183–213, 2017.

[16] Aleksandros Sobczyk and Mathieu Luisier. Approximate euclidean lengths and distances beyond johnson-lindenstrauss. *arXiv preprint arXiv:2205.12307*, 2022.

[17] Hanzhong Liu and Yuehan Yang. Regression-adjusted average treatment effect estimates in stratified randomized experiments. *Biometrika*, 107(4):935–948, 2020.

[18] Thomas Müller, Fabrice Rousselle, Alexander Keller, and Jan Novák. Neural control variates. *ACM Transactions on Graphics (TOG)*, 39(6):1–19, 2020.

[19] Denis Belomestny, Artur Goldman, Alexey Naumov, and Sergey Samsonov. Theoretical guarantees for neural control variates in mcmc. *arXiv preprint arXiv:2304.01111*, 2023.

[20] Jiaxin Shi, Yuhao Zhou, Jessica Hwang, Michalis Titsias, and Lester Mackey. Gradient estimation with discrete stein operators. *Advances in Neural Information Processing Systems*, 35:25829–25841, 2022.

[21] Hao Liu, Yihao Feng, Yi Mao, Dengyong Zhou, Jian Peng, and Qiang Liu. Action-depedent control variates for policy optimization via stein's identity. *arXiv preprint arXiv:1710.11198*, 2017.

[22] Jonathan W Siegel and Jinchao Xu. High-order approximation rates for shallow neural networks with cosine and relu^k activation functions. *Applied and Computational Harmonic Analysis*, 58:1–26, 2022.

[23] Anthony O'Hagan. Bayes–hermite quadrature. *Journal of statistical planning and inference*, 29(3):245–260, 1991.

[24] Motonobu Kanagawa, Bharath K Sriperumbudur, and Kenji Fukumizu. Convergence guarantees for kernel-based quadrature rules in misspecified settings. *Advances in Neural Information Processing Systems*, 29, 2016.

[25] Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751, 2017.

[26] Toni Karvonen and Simo Sarkka. Fully symmetric kernel quadrature. *SIAM Journal on Scientific Computing*, 40(2):A697–A720, 2018.

[27] Motonobu Kanagawa and Philipp Hennig. Convergence guarantees for adaptive bayesian quadrature methods. *Advances in Neural Information Processing Systems*, 32, 2019.

[28] Ayoub Belhadji, Rémi Bardenet, and Pierre Chainais. Kernel quadrature with dpps. *Advances in Neural Information Processing Systems*, 32, 2019.

[29] Ayoub Belhadji. An analysis of ermakov-zolotukhin quadrature using kernels. *Advances in Neural Information Processing Systems*, 34:27278–27289, 2021.

[30] Rémi Bardenet and Adrien Hardy. Monte carlo with determinantal point processes. *Annals of Applied Probability*, 2020.

[31] Guillaume Gautier, Rémi Bardenet, and Michal Valko. On two ways to use determinantal point processes for monte carlo integration. *Advances in Neural Information Processing Systems*, 32, 2019.

[32] Satoshi Hayakawa, Harald Oberhauser, and Terry Lyons. Positively weighted kernel quadrature via subsampling. *arXiv preprint arXiv:2107.09597*, 2021.

[33] Satoshi Hayakawa, Harald Oberhauser, and Terry Lyons. Sampling-based nyström approximation and kernel quadrature. *arXiv preprint arXiv:2301.09517*, 2023.

[34] Yutian Chen, Max Welling, and Alex Smola. Super-samples from kernel herding. *arXiv preprint arXiv:1203.3472*, 2012.

[35] Simon Lacoste-Julien, Fredrik Lindsten, and Francis Bach. Sequential kernel herding: Frank-wolfe optimization for particle filtering. In *Artificial Intelligence and Statistics*, pages 544–552. PMLR, 2015.

[36] Ferenc Huszár and David Duvenaud. Optimally-weighted herding is bayesian quadrature. *arXiv preprint arXiv:1204.1664*, 2012.

[37] Wilson Ye Chen, Lester Mackey, Jackson Gorham, François-Xavier Briol, and Chris Oates. Stein points. In *International Conference on Machine Learning*, pages 844–853. PMLR, 2018.

[38] Raaz Dwivedi and Lester Mackey. Kernel thinning. *arXiv preprint arXiv:2105.05842*, 2021.

[39] Raaz Dwivedi and Lester Mackey. Generalized kernel thinning. *arXiv preprint arXiv:2110.01593*, 2021.

[40] Erich Novak. *Deterministic and stochastic error bounds in numerical analysis*, volume 1349. Springer, 2006.

[41] Joseph F Traub, GW Wasilkowski, H Wozniakowski, and Erich Novak. Information-based complexity. *SIAM Review*, 36(3):514–514, 1994.

[42] E Novak and H Wozniakowski. Tractability of multivariate problems, volume i: Linear information, european math. *Soc., Zürich*, 2(3), 2008.

[43] Erich Novak and Henryk Woźniakowski. *Tractability of Multivariate Problems: Standard Information for Functionals*, volume 2. European Mathematical Society, 2008.

[44] Nikolai Sergeevich Bakhvalov. On the approximate calculation of multiple integrals. *Journal of Complexity*, 31(4):502–516, 2015.

[45] Aicke Hinrichs, Erich Novak, Mario Ullrich, and H Woźniakowski. The curse of dimensionality for numerical integration of smooth functions. *Mathematics of Computation*, 83(290):2853–2863, 2014.

[46] Erich Novak. Some results on the complexity of numerical integration. *Monte Carlo and Quasi-Monte Carlo Methods: MCQMC, Leuven, Belgium, April 2014*, pages 161–183, 2016.

[47] Aicke Hinrichs, David Krieg, Erich Novak, Joscha Prochno, and Mario Ullrich. On the power of random information. *Multivariate Algorithms and information-based complexity*, 27:43–64, 2020.

[48] Aicke Hinrichs, David Krieg, Erich Novak, and Jan Vybíral. Lower bounds for integration and recovery in l_2. *Journal of Complexity*, 72:101662, 2022.

[49] David Krieg and Mathias Sonnleitner. Random points are optimal for the approximation of sobolev functions. *arXiv preprint arXiv:2009.11275*, 2020.

[50] David Krieg, Erich Novak, and Mathias Sonnleitner. Recovery of sobolev functions restricted to iid sampling. *Mathematics of Computation*, 91(338):2715–2738, 2022.

[51] Lucien Birgé and Pascal Massart. Estimation of integral functionals of a density. *The Annals of Statistics*, 23(1):11–29, 1995.

[52] David L Donoho and Michael Nussbaum. Minimax quadratic estimation of a quadratic functional. *Journal of Complexity*, 6(3):290–323, 1990.

[53] David L Donoho. One-sided inference about functionals of a density. *The Annals of Statistics*, pages 1390–1420, 1988.

[54] David L Donoho and Richard C Liu. Geometrizing rates of convergence, ii. *The Annals of Statistics*, pages 633–667, 1991.

[55] David L Donoho and Richard C Liu. Geometrizing rates of convergence, iii. *The Annals of Statistics*, pages 668–701, 1991.

[56] James Robins, Lingling Li, Eric Tchetgen, Aad van der Vaart, et al. Higher order influence functions and minimax estimation of nonlinear functionals. *Probability and statistics: essays in honor of David A. Freedman*, 2:335–421, 2008.

[57] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2015.

[58] Akshay Krishnamurthy, Kirthevasan Kandasamy, Barnabas Poczos, and Larry Wasserman. Nonparametric estimation of renyi divergence and friends. In *International Conference on Machine Learning*, pages 919–927. PMLR, 2014.

[59] Peter Mathé. Random approximation of sobolev embeddings. *Journal of Complexity*, 7(3):261–281, 1991.

[60] Stefan Heinrich. Randomized approximation of sobolev embeddings, ii. *Journal of Complexity*, 25(5):455–472, 2009.

[61] Stefan Heinrich. Randomized approximation of sobolev embeddings, iii. *Journal of Complexity*, 25(5):473–507, 2009.

[62] Yanjun Han, Jiantao Jiao, and Rajarshi Mukherjee. On estimation of l_r-norms in gaussian white noise models. *Probability Theory and Related Fields*, 177(3-4):1243–1294, 2020.

[63] Oleg Lepski, Arkady Nemirovski, and Vladimir Spokoiny. On estimation of the l_r norm of a regression function. *Probability theory and related fields*, 113:221–253, 1999.

[64] Stefan Heinrich. On the complexity of computing the l_q norm. *Journal of Complexity*, 49:1–26, 2018.

[65] Alexandre B Tsybakov. Introduction to nonparametric estimation, 2009. *URL https://doi.org/10.1007/b13794. Revised and extended from the*, 9(10), 2004.

[66] Holger Wendland. Local polynomial reproduction and moving least squares approximation. *IMA Journal of Numerical Analysis*, 21(1):285–300, 2001.

[67] Holger Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

[68] Yanjun Han, Jiantao Jiao, Tsachy Weissman, and Yihong Wu. Optimal rates of entropy estimation over lipschitz balls. *The Annals of Statistics*, 48(6):3228 – 3250, 2020.

[69] Gérard Biau and Luc Devroye. *Lectures on the nearest neighbor method*, volume 246. Springer, 2015.

[70] Aleksandr Reznikov and Edward B Saff. The covering radius of randomly distributed points on a manifold. *International Mathematics Research Notices*, 2016(19):6065–6094, 2016.

[71] Rafał Latała. Estimation of moments of sums of independent real random variables. *The Annals of Probability*, 25(3):1502–1513, 1997.

[72] Iosif Pinelis. Optimum bounds on moments of sums of independent random vectors. *Siberian Adv. Math*, 5(3):141–150, 1995.

Appendix

The appendix is organized as follows:

- In Appendix A, we list some notations and standard lemmas used in our proofs.
- Appendix B contains a comprehensive proof of the information-theoretic lower bound on the estimation of q-th moments, which is established in Theorem 2.1.
- In Appendix C, we provide a detailed proof of Theorem 3.1 and 3.2, which gives us the minimax optimal upper bound on estimating q-th moments.
- Appendix D consists of our proof for the information-theoretic lower bounds and minimax optimal upper bounds on integral estimation and function estimation, which are listed in Theorem 4.1 and 4.2.
- Appendix E provides our construction of the desired function estimator in Assumption 3.1 along with a proof of its convergence rate.

A Preliminaries and Basic Tools

A.1 Preliminaries

This subsection is devoted to presenting some basic notations used in our proofs. For any fixed convex function $f : \mathbb{R}^+ \to \mathbb{R}$ satisfying $f(1) = 0$, we use $D_f(\cdot\|\cdot)$ to denote the corresponding f-divergence, *i.e,* $D_f(P\|Q) = \int_{\mathcal{Y}} f\Big(\frac{dP}{dQ}\Big)dQ$ for any two probability distributions P and Q over some fixed space $\mathcal{Y}$. In particular, when $f(x) = \frac{1}{2}|x-1|$, $D_f(\cdot\|\cdot)$ is the total variation (TV) distance $TV(\cdot\|\cdot)$. When $f(x) = x\log x$, $D_f(\cdot\|\cdot)$ coincides with the Kullback–Leibler (KL) divergence $KL(\cdot\|\cdot)$. Moreover, for any $a \in \mathbb{R}$, we use $\delta_a(\cdot)$ to denote the Dirac delta distribution at point a, *i.e,* $\int_{-\infty}^{\infty} f(x)\delta_a(x) = f(a)$ for any function $f : \mathbb{R} \to \mathbb{R}$.

A.2 Basic Lemmas

In this subsection, we list some basic lemmas that serve as essential tools in our proofs.

Lemma 1 (Sobolev Embedding Theorem [11]) *For some fixed dimension $d \in \mathbb{N}$, we have that (I) For any $s, t \in \mathbb{N}_0$ and $p, q \in \mathbb{R}$ satisfying $s > t$, $p < d$ and $1 \le p < q \le \infty$, we have $W^{s,p}(\mathbb{R}^d) \subseteq W^{t,q}(\mathbb{R}^d)$ when the relation $\frac{1}{p} - \frac{s}{d} = \frac{1}{q} - \frac{t}{d}$ holds. In the special case when $t = 0$, we have $W^{s,p}(\mathbb{R}^d) \subseteq L^q(\mathbb{R}^d)$ for any $s \in \mathbb{N}$ and $p, q \in \mathbb{R}$ satisfying $1 \le p < q \le \infty$ and $\frac{1}{p} - \frac{s}{d} \le \frac{1}{q}$. (II) For any $\alpha \in (0, 1)$, let $\beta = \frac{d}{1-\alpha} \in (d, \infty]$. Then we have $C^1(\mathbb{R}^d) \cap W^{1,\beta}(\mathbb{R}^d) \subseteq C^\alpha(\mathbb{R}^d)$.*

Lemma 2 (Hölder's Inequality) *For any fixed domain Ω and $p, q \in [1, \infty]$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$, we have that $\|fg\|_{L^1(\Omega)} \le \|f\|_{L^p(\Omega)}\|g\|_{L^q(\Omega)}$ holds for any $f \in L^p(\Omega), g \in L^q(\Omega)$.*

Lemma 3 (Hoeffding's Inequality) *Let $X_1, X_2, \cdots, X_n$ be independent random variables satisfying $X_i \in [a_i, b_i]$ for any $1 \le i \le n$. Then for any $\epsilon > 0$, the sum $S_n := \sum_{i=1}^n X_i$ of these n random variables satisfies the following inequality:*

$$\begin{aligned} \mathbb{P}(S_n \ge \mathbb{E}[S_n] + t) &\le \exp\Big(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big) \\ \mathbb{P}(S_n \le \mathbb{E}[S_n] - t) &\le \exp\Big(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big) \end{aligned} \tag{A.1}$$

Lemma 4 (Data Processing Inequality) *Given some Markov Chain $X \to Z$, where X and Z are two random variables the measurable spaces $(\mathcal{X}, \mu)$ and $(\mathcal{Z}, \nu)$ respectively. Let K be the transition kernel of the Markov Chain above,* i.e, *for any $x \in \mathcal{X}$, the probability distribution of Z is given by $K(\cdot, x)$ when conditioned on $X = x$. For any two fixed two distributions P, Q over $\mathcal{X}$ with probability density functions p, q, we use $K_P(\cdot)$ and $K_Q(\cdot)$ to denote the corresponding marginal distributions respectively,* i.e, *$K_P(\cdot) := \int_{\mathcal{X}} K(\cdot, x)p(x)d\mu(x)$ and $K_Q(\cdot) = \int_{\mathcal{X}} K(\cdot, x)q(x)d\mu(x)$. Then we have $D_f(K_P||K_Q) \le D_f(P||Q)$ holds for any f-divergence $D_f(\cdot\|\cdot)$.*

B Proof of Lower Bounds in Section 2

B.1 A Key Lemma for Building Minimax Optimal Lower Bounds

In this subsection, we firstly present the method of two fuzzy hypotheses, which turns out to be the most essential tool for establishing all the minimax optimal lower bounds in our paper, before giving our complete proof of Theorem 2.1.

Lemma 5 (Method of Two Fuzzy Hypotheses: Theorem 2.15 (i), [65]) *Let $F : \Theta \to \mathbb{R}$ be some continuous functional defined on the measurable space $(\Theta, \mathcal{U})$ and taking values in $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, where $\mathcal{B}(\mathbb{R})$ denotes the Borel σ-algebra on $\mathbb{R}$. Suppose that each parameter $\theta \in \Theta$ is associated with a distribution $\mathbb{P}_\theta$, which together form a collection $\{\mathbb{P}_\theta : \theta \in \Theta\}$ of distributions.*

For any fixed $\theta \in \Theta$, assume that our observation $\mathbf{X}$ is distributed as $\mathbb{P}_\theta$. Let $\hat{F}$ be an arbitrary estimator of $F(\theta)$ based on $\mathbf{X}$. Let μ_0, μ_1 be two prior measures on Θ. Assume that there exist constants $c \in \mathbb{R}, \Delta \in (0, \infty)$ and $\beta_0, \beta_1 \in [0, 1)$, such that:

$$\begin{aligned} \mu_0(\theta \in \Theta : F(\theta) \leq c - \Delta) \geq 1 - \beta_0, \\ \mu_1(\theta \in \Theta : F(\theta) \geq c + \Delta) \geq 1 - \beta_1. \end{aligned} \tag{B.1}$$

For $j \in \{0, 1\}$, we use $\mathbb{P}_j(\cdot) := \int \mathbb{P}_\theta(\cdot)\mu_j(d\theta)$ to denote the marginal distribution $\mathbb{P}_j$ associated with the prior distribution μ_j. Then we have the following lower bound:

$$\inf_{\hat{F}} \sup_{\theta \in \Theta} \mathbb{P}_\theta(|\hat{F} - F(\theta)| \geq \Delta) \geq \frac{1 - TV(\mathbb{P}_0||\mathbb{P}_1) - \beta_0 - \beta_1}{2}. \tag{B.2}$$

B.2 Proof of Theorem 2.1 (Information-Theoretic Lower Bound on Moment Estimation)

In this subsection, we give a detailed proof of the two minimax lower bounds established in Theorem 2.1 above via the method of two fuzzy hypotheses (Lemma 5). We start off by introducing some preliminary tools used in our proof. Consider the function K_0 defined as follows:

$$K_0(x) := \prod_{i=1}^{d} \exp\Big(-\frac{1}{1 - x_i^2}\Big) \mathbb{1}(|x_i| \leq 1), \forall\, x = (x_1, x_2, \cdots, x_d) \in \mathbb{R}^d. \tag{B.3}$$

Moreover, we pick some function K satisfying

$$K(x) := K_0(2x),\ \forall\, x \in \mathbb{R}^d, \tag{B.4}$$

From our construction of K and K_0 above, we have that K_0 is in $C^\infty(\mathbb{R}^d)$ and compactly supported on $[-\frac{1}{2}, \frac{1}{2}]^d$. Furthermore, we set $m = (200n)^{\frac{1}{d}}$ and divide the domain Ω into m^d small cubes $\Omega_1, \Omega_2, \cdots, \Omega_{m^d}$, each of which has side length m^{-1}. For any $1 \leq j \leq m^d$, we use c_j to denote the center of the cube Ω_j. Similar to the proof sketch of Theorem 2.1, below we again use w_p to denote the discrete random variable satisfying $\mathbb{P}(w_p = -1) = p$ and $\mathbb{P}(w_p = 1) = 1 - p$ for any $p \in (0, 1)$. Furthermore, we use $\vec{x} := (x_1, x_2, \cdots, x_n)$ and $\vec{y} := (y_1, y_2, \cdots, y_n)$ to denote the two n-dimensional vectors formed by the quadrature points and observed function values, After introducing all preliminaries above, let's present the essential parts of our proof. Given that our lower bound in Theorem 2.1 consists of two terms, our proof is also divided into two parts:

(Case I) For the first lower bound in (2.1), let's consider two functions g_0 and g_1 defined as follows:

$$\begin{aligned} g_0(x) &\equiv 0\ (\forall\, x \in \Omega), \\ g_1(x) &= \begin{cases} m^{-s+\frac{d}{p}} K(m(x - c_1))\ (x \in \Omega_1), \\ 0\ (\text{otherwise}). \end{cases} \end{aligned} \tag{B.5}$$

Clearly we have $g_0 \in W^{s,p}(\Omega)$ and $I_{g_0}^q = 0$. Now let's verify that $g_1 \in W^{s,p}(\Omega)$ for any m. Note that the following bound holds for any $t \in \mathbb{N}_0^d$ satisfying $|t| \leq s$:

$$\begin{aligned} \|D^t g_1\|_{L^p(\Omega)} &= \left(\int_{\Omega_1} \Big| m^{-s+\frac{d}{p}} m^{|t|} (D^t K)(m(x - c_1)) \Big|^p dx \right)^{\frac{1}{p}} \\ &= m^{|t|-s+\frac{d}{p}} \left(\int_{[-\frac{1}{2}, \frac{1}{2}]^d} \Big| (D^t K)(y) \Big|^p \frac{1}{m^d} dy \right)^{\frac{1}{p}} = m^{|t|-s} \|D^t K\|_{L^p([-\frac{1}{2}, \frac{1}{2}]^d)} \lesssim 1. \end{aligned}$$

This implies $g_1 \in W^{s,p}(\Omega)$ for any m, as desired. Moreover, computing the q-th moment of g_1 yields

$$\begin{aligned} I_{g_1}^q &= \int_\Omega g_1^q(x)dx = \int_{\Omega_1} (m^{-s+\frac{d}{p}} K(m(x-c_1)))^q dx \\ &= m^{-q(s-\frac{d}{p})} \int_{[-\frac{1}{2},\frac{1}{2}]^d} (K(y))^q \frac{1}{m^d} dy = m^{-q(s-\frac{d}{p})-d} \|K\|^q_{L^q([-\frac{1}{2},\frac{1}{2}]^d)}. \end{aligned} \tag{B.6}$$

Now let us take $\epsilon = \frac{1}{2}$ and pick two discrete measures μ_0, μ_1 supported on the finite set $\{g_0, g_1\} \subset W^{s,p}(\Omega)$ as below:

$$\begin{aligned} \mu_0(\{g_0\}) = \frac{1+\epsilon}{2}, \mu_0(\{g_1\}) = \frac{1-\epsilon}{2}, \\ \mu_1(\{g_0\}) = \frac{1-\epsilon}{2}, \mu_1(\{g_1\}) = \frac{1+\epsilon}{2}. \end{aligned} \tag{B.7}$$

On the one hand, by taking $c = \Delta = \frac{1}{2} I_{g_1}^q$ and $\beta_0 = \beta_1 = \frac{1-\epsilon}{2}$, we may use (B.7) to deduce that

$$\begin{aligned} \mu_0(f \in W^{s,p}(\Omega) : I_f^q \le c - \Delta) = \mu_0(I_f^q \le 0) \ge \frac{1+\epsilon}{2} = 1 - \beta_0, \\ \mu_1(f \in W^{s,p}(\Omega) : I_f^q \ge c + \Delta) = \mu_1(I_f^q \ge I_{g_1}^q) \ge \frac{1+\epsilon}{2} = 1 - \beta_1. \end{aligned} \tag{B.8}$$

Hence, we have that (B.1) holds true. On the other hand, recall that the quadrature points $\{x_1, \cdots, x_n\}$ are identical and independent samples from the uniform distribution on Ω, which enables us to write the marginal distributions in an explicit form as follows:

$$\begin{aligned} \mathbb{P}_0(\vec{x}, \vec{y}) = \Big(\frac{1+\epsilon}{2} \prod_{i:x_i \in \Omega_1} \delta_0(y_i) + \frac{1-\epsilon}{2} \prod_{i:x_i \in \Omega_1} \delta_{g_1(x_i)}(y_i)\Big) \cdot \prod_{j=2}^{m^d} \Big(\prod_{i:x_i \in \Omega_j} \delta_0(y_i)\Big), \\ \mathbb{P}_1(\vec{x}, \vec{y}) = \Big(\frac{1-\epsilon}{2} \prod_{i:x_i \in \Omega_1} \delta_0(y_i) + \frac{1+\epsilon}{2} \prod_{i:x_i \in \Omega_1} \delta_{g_1(x_i)}(y_i)\Big) \cdot \prod_{j=2}^{m^d} \Big(\prod_{i:x_i \in \Omega_j} \delta_0(y_i)\Big). \end{aligned} \tag{B.9}$$

In particular, we have $\mathbb{P}_0 = \mathbb{P}_1$ when the set $\{i : x_i \in \Omega_1\}$ is empty. Combing this fact with (B.9) above allows us to compute the KL divergence between $\mathbb{P}_0$ and $\mathbb{P}_1$ as below

$$\begin{aligned} KL(\mathbb{P}_0 || \mathbb{P}_1) &= \int_\Omega \cdots \int_\Omega \Big(\int_{-\infty}^\infty \cdots \int_{-\infty}^\infty \log\Big(\frac{\mathbb{P}_0(\vec{x}, \vec{y})}{\mathbb{P}_1(\vec{x}, \vec{y})}\Big) \mathbb{P}_0(\vec{x}, \vec{y}) dy_1 \cdots dy_n \Big) dx_1 \cdots dx_n \\ &= \int_\Omega \cdots \int_\Omega \Bigg(\int_{-\infty}^\infty \cdots \int_{-\infty}^\infty \log\Big(\frac{\frac{1+\epsilon}{2} \prod_{i:x_i \in \Omega_1} \delta_0(y_i) + \frac{1-\epsilon}{2} \prod_{i:x_i \in \Omega_1} \delta_{g_1(x_i)}(y_i)}{\frac{1-\epsilon}{2} \prod_{i:x_i \in \Omega_1} \delta_0(y_i) + \frac{1+\epsilon}{2} \prod_{i:x_i \in \Omega_1} \delta_{g_1(x_i)}(y_i)} \Big) \\ &\cdot \Big(\frac{1+\epsilon}{2} \prod_{i:x_i \in \Omega_1} \delta_0(y_i) + \frac{1-\epsilon}{2} \prod_{i:x_i \in \Omega_1} \delta_{g_1(x_i)}(y_i)\Big) \cdot \Big(\prod_{j=2}^{m^d} \prod_{i:x_i \in \Omega_j} \delta_0(y_i)\Big) \prod_{i=1}^n dy_i \Bigg) \prod_{i=1}^n dx_i \\ &= \int_\Omega \cdots \int_\Omega \Bigg(\int_{-\infty}^\infty \cdots \int_{-\infty}^\infty \log\Big(\frac{\frac{1+\epsilon}{2} \prod_{i:x_i \in \Omega_1} \delta_0(y_i) + \frac{1-\epsilon}{2} \prod_{i:x_i \in \Omega_1} \delta_{g_1(x_i)}(y_i)}{\frac{1-\epsilon}{2} \prod_{i:x_i \in \Omega_1} \delta_0(y_i) + \frac{1+\epsilon}{2} \prod_{i:x_i \in \Omega_1} \delta_{g_1(x_i)}(y_i)} \Big) \\ &\cdot \Big(\frac{1+\epsilon}{2} \prod_{i:x_i \in \Omega_1} \delta_0(y_i) + \frac{1-\epsilon}{2} \prod_{i:x_i \in \Omega_1} \delta_{g_1(x_i)}(y_i)\Big) \prod_{i:x_i \in \Omega_1} dy_i \Bigg) \prod_{i=1}^n dx_i \\ &= \Big(\log\Big(\frac{1+\epsilon}{1-\epsilon}\Big)\frac{1+\epsilon}{2} + \log\Big(\frac{1-\epsilon}{1+\epsilon}\Big)\frac{1-\epsilon}{2}\Big)\mathbb{P}\Big(\{i : x_i \in \Omega_1\} \neq \varnothing\Big) \\ &= \epsilon \log\Big(\frac{1+\epsilon}{1-\epsilon}\Big)\mathbb{P}\Big(\{i : x_i \in \Omega_1\} \neq \varnothing\Big). \end{aligned} \tag{B.10}$$

Moreover, since the probability that $\{i : x_i \in \Omega_1\} = \varnothing$ equals to $(\frac{m^d-1}{m^d})^n = (\frac{m^d-1}{m^d})^{\frac{m^d}{200}}$, we have

$$\mathbb{P}\Big(\{i : x_i \in \Omega_1\} \neq \varnothing\Big) = 1 - \Big(1 - \frac{1}{m^d}\Big)^{\frac{m^d}{200}} \le 1 - \left(\frac{1}{e}\Big(1 - \frac{1}{m^d}\Big)\right)^{\frac{1}{200}} \le 1 - (2e)^{-\frac{1}{200}}. \tag{B.11}$$

Now we may combine (B.10), (B.11) and Pinkser's inequality to upper bound the TV distance between $\mathbb{P}_0$ and $\mathbb{P}_1$ as below:

$$TV(\mathbb{P}_0\|\mathbb{P}_1) \le \sqrt{\frac{1}{2}KL(\mathbb{P}_0\|\mathbb{P}_1)} \le \sqrt{\frac{1-(2e)^{-\frac{1}{200}}}{2}\epsilon\log\Big(\frac{1+\epsilon}{1-\epsilon}\Big)} \le \sqrt{\frac{3}{100}}\epsilon = \frac{\sqrt{3}}{10}\epsilon. \quad \text{(B.12)}$$

Finally, by substituting (B.6), (B.12), $\Delta = \frac{1}{2}I^q_{g_1}$ and $\beta_0 = \beta_1 = \frac{1-\epsilon}{2} = \frac{1}{4}$ into (B.2) and applying Markov's inequality, we obtain the final lower bound

$$\begin{aligned}
&\inf_{\hat{H}^q\in\mathcal{H}_n^{f,q}}\sup_{f\in W^{s,p}(\Omega)}\mathbb{E}_{\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n}\left[\Big|\hat{H}^q\Big(\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n\Big)-I_f^q\Big|\right]\\
&\ge \Delta\inf_{\hat{H}^q\in\mathcal{H}_n^{f,q}}\sup_{f\in W^{s,p}(\Omega)}\mathbb{P}_{\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n}\left[\Big|\hat{H}^q\Big(\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n\Big)-I_f^q\Big|\ge\Delta\right] \qquad \text{(B.13)}\\
&\ge \frac{1}{2}I^q_{g_1}\frac{1-TV(\mathbb{P}_0||\mathbb{P}_1)-\beta_0-\beta_1}{2} \ge \frac{1}{4}\Big(1-\frac{\sqrt{3}}{10}\Big)\epsilon I^q_{g_1}\\
&= \frac{1}{8}\Big(1-\frac{\sqrt{3}}{10}\Big)(200n)^{-\frac{q}{d}(s-\frac{d}{p})-1}\|K\|^q_{L^q([-\frac{1}{2},\frac{1}{2}]^d)} \gtrsim n^{-q(\frac{s}{d}-\frac{1}{p})-1},
\end{aligned}$$

which is exactly the first term in the RHS of (2.1).

(Case II) Now let us proceed to prove the second lower bound in (2.1). For any $1\le j\le m^d$, consider first some function f_j defined as follows

$$f_j(x) = \begin{cases} m^{-s}K(m(x-c_j))\ (x\in\Omega_j),\\ 0\ (\text{otherwise}),\end{cases} \quad \text{(B.14)}$$

which satisfies $\text{supp}(f_j)\subseteq\Omega_j$, $f_j\in C^\infty(\Omega)$ and $f_j(x)\ge 0$ $(\forall\ x\in\Omega)$. We further pick two constants α, M satisfying $\alpha := \|K\|_{L^\infty([-\frac{1}{2},\frac{1}{2}]^d)}$ and $M = 3\alpha$. Now consider the following finite set of 2^{m^d} functions:

$$\mathcal{S} := \Big\{M + \sum_{j=1}^{m^d}\eta_j f_j : \eta_j\in\{\pm1\},\ \forall\ 1\le j\le m^d\Big\}. \quad \text{(B.15)}$$

We will proceed to verify that any element in $\mathcal{S}$ must be in $W^{s,p}(\Omega)$ for any m. Note that for any $\eta_j\in\{\pm1\}$ $(1\le j\le m^d)$ and any $t\in\mathbb{N}_0^d$ satisfying $|t|\le s$, we have

$$\begin{aligned}
\Big\|D^t\Big(M+\sum_{j=1}^{m^d}\eta_j f_j\Big)\Big\|^p_{L^p(\Omega)} &\le \left(M+\Big\|\sum_{j=1}^{m^d}\eta_j(D^t f_j)\Big\|_{L^p(\Omega)}\right)^p\\
&\le 2^p\left(M^p+\Big\|\sum_{j=1}^{m^d}\eta_j(D^t f_j)\Big\|^p_{L^p(\Omega)}\right) \lesssim M^p+\sum_{j=1}^{m^d}\|D^t f_j\|^p_{L^p(\Omega_j)}\\
&= M^p+\sum_{j=1}^{m^d}\int_{\Omega_j}\Big|m^{-s+|t|}(D^tK)(m(x-c_j))\Big|^p dx\\
&= M^p+m^{(|t|-s)p}\sum_{j=1}^{m^d}\int_{[-\frac{1}{2},\frac{1}{2}]^d}\Big|(D^tK)(y)\Big|^p\frac{1}{m^d}dy\\
&\le M^p+\|D^tK\|^p_{L^p([-\frac{1}{2},\frac{1}{2}]^d)} \lesssim 1.
\end{aligned}$$

This gives us that $\mathcal{S}\subset W^{s,p}(\Omega)$ for any m, as desired. Now let's pick $\kappa = \frac{1}{3}\sqrt{\frac{2}{3n}}$ and take $\{w_j^{(0)}\}_{j=1}^{m^d}$ and $\{w_j^{(1)}\}_{j=1}^{m^d}$ to be independent and identical copies of $w_{\frac{1+\kappa}{2}}$ and $w_{\frac{1-\kappa}{2}}$ respectively. Then we define μ_0,μ_1 to be two discrete measures supported on the finite set $\mathcal{S}$ such that the following condition holds for any $\eta_j\in\{\pm1\}$ $(1\le j\le m^d)$:

$$\mu_k\Big(\Big\{M+\sum_{j=1}^{m^d}\eta_j f_j\Big\}\Big) = \prod_{j=1}^{m^d}\mathbb{P}(w_j^{(k)}=\eta_j),\ k\in\{0,1\}. \quad \text{(B.16)}$$

In order to determine the separation distance Δ between the two priors μ_0 and μ_1, we need to define two quantities $A := \int_{\Omega_j}(M+f_j(x))^q dx$ and $B := \int_{\Omega_j}(M-f_j(x))^q dx$, which both remain the same for any $1 \le j \le m^d$. Now consider deriving a lower bound on the quantity $\Delta' := A - B > 0$. Note that for any fixed $j \in \{1, 2, \cdots, m^d\}$, we have $M > 2\alpha \ge 2m^{-s}\|K\|_{L^\infty([-\frac{1}{2},\frac{1}{2}]^d)} = 2\|f_j\|_{L^\infty(\Omega_j)}$, which implies $M + y > \frac{1}{2}M > 0$ for any $y \in [-\|f_j\|_{L^\infty(\Omega_j)}, \|f_j\|_{L^\infty(\Omega_j)}]$. This helps us obtain the following lower bound on Δ':

$$\begin{aligned}
\Delta' &= \int_{\Omega_j}(M+f_j(x))^q dx - \int_{\Omega_j}(M-f_j(x))^q dx = \int_{\Omega_j}\Big(\int_{-f_j(x)}^{f_j(x)} q(M+y)^{q-1}dy\Big)dx \\
&\ge \int_{\Omega_j}\Big(\int_{-f_j(x)}^{f_j(x)} q(\frac{1}{2}M)^{q-1}dy\Big)dx = \frac{q}{2^{q-1}}M^{q-1}\int_{\Omega_j}\Big(2f_j(x)\Big)dx \gtrsim \int_{\Omega_j} f_j(x)dx \\
&= \int_{\Omega_j} m^{-s}K(m(x-c_j))dx = m^{-s}\int_{[-\frac{1}{2},\frac{1}{2}]^d} K(y)\frac{1}{m^d}dy = m^{-s-d}\|K\|_{L^1([-\frac{1}{2},\frac{1}{2}]^d)}.
\end{aligned} \tag{B.17}$$

Moreover, let us pick $\lambda = \frac{1}{2}$ and apply Hoeffding's Inequality (Lemma 3) to the bounded random variables $\{w_j^{(0)}\}_{j=1}^{m^d}$ and $\{w_j^{(1)}\}_{j=1}^{m^d}$ to deduce that

$$\begin{aligned}
\mathbb{P}\Big(\sum_{j=1}^{m^d} w_j^{(0)} \ge -(1-\lambda)m^d\kappa\Big) &\le \exp\Big(-\frac{2(\lambda m^d\kappa)^2}{4m^d}\Big) = \exp\Big(-\frac{1}{2}\lambda^2\kappa^2 m^d\Big), \\
\mathbb{P}\Big(\sum_{j=1}^{m^d} w_j^{(1)} \le (1-\lambda)m^d\kappa\Big) &\le \exp\Big(-\frac{2(\lambda m^d\kappa)^2}{4m^d}\Big) = \exp\Big(-\frac{1}{2}\lambda^2\kappa^2 m^d\Big).
\end{aligned} \tag{B.18}$$

By taking $c := \frac{m^d}{2}(A+B), \Delta := (1-\lambda)\kappa m^d(A-B) = (1-\lambda)\kappa m^d\Delta'$ and $\beta_0 = \beta_1 = \exp\Big(-\frac{1}{2}\lambda^2\kappa^2 m^d\Big)$, we may combine (B.17) and (B.18) justified above to get that

$$\begin{aligned}
&\mu_0(f \in W^{s,p}(\Omega) : I_f^q \le c - \Delta) = \mathbb{P}\Big(\sum_{j=1}^{m^d} I^q_{M+w_j^{(0)}f_j} \le \frac{1-(1-\lambda)\kappa}{2}m^d A + \frac{1+(1-\lambda)\kappa}{2}m^d B\Big) \\
&\ge \mathbb{P}\Big(\sum_{j=1}^{m^d} w_j^{(0)} \le -(1-\lambda)m^d\kappa\Big) = 1 - \mathbb{P}\Big(\sum_{j=1}^{m^d} w_j^{(0)} \ge -(1-\lambda)m^d\kappa\Big) \\
&\ge 1 - \exp\Big(-\frac{1}{2}\lambda^2\kappa^2 m^d\Big) = 1 - \beta_0, \\
&\mu_1(f \in W^{s,p}(\Omega) : I_f^q \ge c + \Delta) = \mathbb{P}\Big(\sum_{j=1}^{m^d} I^q_{M+w_j^{(1)}f_j} \ge \frac{1+(1-\lambda)\kappa}{2}m^d A + \frac{1-(1-\lambda)\kappa}{2}m^d B\Big) \\
&\ge \mathbb{P}\Big(\sum_{j=1}^{m^d} w_j^{(1)} \ge (1-\lambda)m^d\kappa\Big) = 1 - \mathbb{P}\Big(\sum_{j=1}^{m^d} w_j^{(0)} \le (1-\lambda)m^d\kappa\Big) \\
&\ge 1 - \exp\Big(-\frac{1}{2}\lambda^2\kappa^2 m^d\Big) = 1 - \beta_1,
\end{aligned} \tag{B.19}$$

which indicates that (B.1) holds true. Now let's consider bounding the KL divergence between the two marginal distributions $\mathbb{P}_0, \mathbb{P}_1$ associated with μ_0, μ_1, respectively. Using the fact that $\{x_1, \cdots, x_n\}$ are identical and independent samples from the uniform distribution on Ω again allows us to write the marginal distributions in an explicit form as follows:

$$\begin{aligned}
\mathbb{P}_0(\vec{x}, \vec{y}) &= \prod_{j=1}^{m^d}\Big(\frac{1+\kappa}{2}\prod_{i:x_i\in\Omega_j}\delta_{M-f_j(x_i)}(y_i) + \frac{1-\kappa}{2}\prod_{i:x_i\in\Omega_j}\delta_{M+f_j(x_i)}(y_i)\Big), \\
\mathbb{P}_1(\vec{x}, \vec{y}) &= \prod_{j=1}^{m^d}\Big(\frac{1-\kappa}{2}\prod_{i:x_i\in\Omega_j}\delta_{M-f_j(x_i)}(y_i) + \frac{1+\kappa}{2}\prod_{i:x_i\in\Omega_j}\delta_{M+f_j(x_i)}(y_i)\Big).
\end{aligned} \tag{B.20}$$

Furthermore, for any n quadrature points $\{x_i\}_{i=1}^n$, we use $\mathcal{J}_n$ to denote the set of all indices j satisfying that Ω_j contains at least one of the points in $\{x_i\}_{i=1}^n$, *i.e,*

$$\mathcal{J}_n := \mathcal{J}_n(x_1, \cdots, x_n) = \Big\{ j : 1 \le j \le m^d \text{ and } \Omega_j \cap \{x_1, \cdots, x_n\} \neq \varnothing \Big\} \tag{B.21}$$

Given that $m^d = 200n > n$, we have $|\mathcal{J}_n| \le n$ for any n quadrature points $\{x_i\}_{i=1}^n$. Using this upper bound on $|\mathcal{J}_n|$ allows us to bound the KL divergence between $\mathbb{P}_0$ and $\mathbb{P}_1$ in the following way:

$$\begin{aligned}
&KL(\mathbb{P}_0||\mathbb{P}_1) = \int_\Omega \cdots \int_\Omega \Big(\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \log\Big(\frac{\mathbb{P}_0(\vec{x},\vec{y})}{\mathbb{P}_1(\vec{x},\vec{y})}\Big)\mathbb{P}_0(\vec{x},\vec{y}) dy_1 \cdots dy_n \Big) dx_1 \cdots dx_n \\
&= \int_\Omega \cdots \int_\Omega \Bigg(\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \log\Big(\prod_{j=1}^{m^d} \frac{\frac{1+\kappa}{2}\prod_{i:x_i\in\Omega_j}\delta_{M-f_j(x_i)}(y_i) + \frac{1-\kappa}{2}\prod_{i:x_i\in\Omega_j}\delta_{M+f_j(x_i)}(y_i)}{\frac{1-\kappa}{2}\prod_{i:x_i\in\Omega_j}\delta_{M-f_j(x_i)}(y_i) + \frac{1+\kappa}{2}\prod_{i:x_i\in\Omega_j}\delta_{M+f_j(x_i)}(y_i)} \Big) \\
&\cdot \prod_{j=1}^{m^d} \Big(\frac{1+\kappa}{2} \prod_{i:x_i\in\Omega_j} \delta_{M-f_j(x_i)}(y_i) + \frac{1-\kappa}{2} \prod_{i:x_i\in\Omega_j} \delta_{M+f_j(x_i)}(y_i) \Big) \prod_{i=1}^n dy_i \Bigg) \prod_{i=1}^n dx_i \\
&= \int_\Omega \cdots \int_\Omega \Bigg(\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \log\Big(\prod_{j\in\mathcal{J}_n} \frac{\frac{1+\kappa}{2}\prod_{i:x_i\in\Omega_j}\delta_{M-f_j(x_i)}(y_i) + \frac{1-\kappa}{2}\prod_{i:x_i\in\Omega_j}\delta_{M+f_j(x_i)}(y_i)}{\frac{1-\kappa}{2}\prod_{i:x_i\in\Omega_j}\delta_{M-f_j(x_i)}(y_i) + \frac{1+\kappa}{2}\prod_{i:x_i\in\Omega_j}\delta_{M+f_j(x_i)}(y_i)} \Big) \\
&\cdot \prod_{j\in\mathcal{J}_n} \Big(\frac{1+\kappa}{2} \prod_{i:x_i\in\Omega_j} \delta_{M-f_j(x_i)}(y_i) + \frac{1-\kappa}{2} \prod_{i:x_i\in\Omega_j} \delta_{M+f_j(x_i)}(y_i) \Big) \prod_{i=1}^n dy_i \Bigg) \prod_{i=1}^n dx_i \\
&= \int_\Omega \cdots \int_\Omega \Bigg(\sum_{j\in\mathcal{J}_n} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \log\Big(\frac{\frac{1+\kappa}{2}\prod_{i:x_i\in\Omega_j}\delta_{M-f_j(x_i)}(y_i) + \frac{1-\kappa}{2}\prod_{i:x_i\in\Omega_j}\delta_{M+f_j(x_i)}(y_i)}{\frac{1-\kappa}{2}\prod_{i:x_i\in\Omega_j}\delta_{M-f_j(x_i)}(y_i) + \frac{1+\kappa}{2}\prod_{i:x_i\in\Omega_j}\delta_{M+f_j(x_i)}(y_i)} \Big) \\
&\cdot \Big(\frac{1+\kappa}{2} \prod_{i:x_i\in\Omega_j} \delta_{M-f_j(x_i)}(y_i) + \frac{1-\kappa}{2} \prod_{i:x_i\in\Omega_j} \delta_{M+f_j(x_i)}(y_i) \Big) \prod_{i:x_i\in\Omega_j} dy_i \Bigg) \prod_{i=1}^n dx_i \\
&= \int_\Omega \cdots \int_\Omega |\mathcal{J}_n| \Big(\log\Big(\frac{1+\kappa}{1-\kappa}\Big)\frac{1+\kappa}{2} + \log\Big(\frac{1-\kappa}{1+\kappa}\Big)\frac{1-\kappa}{2} \Big) \prod_{i=1}^n dx_i \le n\kappa \log\Big(\frac{1+\kappa}{1-\kappa}\Big).
\end{aligned} \tag{B.22}$$

Now we may combine (B.22) and Pinkser's inequality to upper bound the TV distance between $\mathbb{P}_0$ and $\mathbb{P}_1$ as below:

$$TV(\mathbb{P}_0\|\mathbb{P}_1) \le \sqrt{\frac{1}{2}KL(\mathbb{P}_0\|\mathbb{P}_1)} \le \sqrt{\frac{n\kappa}{2}\log\Big(\frac{1+\kappa}{1-\kappa}\Big)} \le \sqrt{\frac{3n}{2}}\kappa = \frac{1}{3}. \tag{B.23}$$

Finally, by substituting (B.17), (B.23), $\Delta = (1-\lambda)\kappa m^d \Delta'$ and $\beta_0 = \beta_1 = \exp\Big(-\frac{1}{2}\lambda^2\kappa^2 m^d\Big) = \exp(-\frac{50}{27}) < \frac{1}{6}$ into (B.2) and applying Markov's inequality, we obtain the final lower bound

$$\begin{aligned}
&\inf_{\hat{H}^q \in \mathcal{H}_n^{f,q}} \sup_{f\in W^{s,p}(\Omega)} \mathbb{E}_{\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n} \Bigg[\Big| \hat{H}^q\Big(\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n\Big) - I_f^q \Big| \Bigg] \\
&\ge \Delta \inf_{\hat{H}^q \in \mathcal{H}_n^{f,q}} \sup_{f\in W^{s,p}(\Omega)} \mathbb{P}_{\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n} \Bigg[\Big| \hat{H}^q\Big(\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n\Big) - I_f^q \Big| \ge \Delta \Bigg] \\
&\ge (1-\lambda)\kappa m^d \Delta' \frac{1 - TV(\mathbb{P}_0||\mathbb{P}_1) - \beta_0 - \beta_1}{2} \ge \frac{1}{2}\frac{\sqrt{2}}{3\sqrt{3n}} \cdot (200n) \cdot \frac{\Delta'}{6} \\
&\gtrsim \sqrt{n}\Delta' \gtrsim \sqrt{n}(200n)^{-\frac{s+d}{d}} \|K\|_{L^1([-\frac{1}{2},\frac{1}{2}]^d)} \gtrsim n^{-\frac{s}{d}-\frac{1}{2}},
\end{aligned} \tag{B.24}$$

which is exactly the second term in the RHS of (2.1). Combining the two lower bounds proved in (B.13) and (B.24) concludes our proof of Theorem 2.1.

C Proof of Upper Bounds in Section 3

C.1 Proof of Theorem 3.1 (Regression-Adjusted Control Variate)

In this subsection, we present a detailed proof of Theorem 3.1. With the first half of the quadrature points $\{x_i\}_{i=1}^{\frac{n}{2}}$ and observed function values $\{y_i\}_{i=1}^{\frac{n}{2}}$ as inputs, we pick the regression adjusted control variate $\hat{f}_{1:\frac{n}{2}}$ to be the estimator returned by the oracle $K_{\frac{n}{2}}$ specified in Assumption 3.1. Moreover, we use the following expression to denote the variance of the function $\hat{f}^q_{1:\frac{n}{2}}(x) - f^q(x)$ with respect to the uniform distribution on Ω:

$$\mathrm{Var}(\hat{f}^q_{1:\frac{n}{2}} - f^q) := \int_\Omega (f^q(x) - \hat{f}^q_{1:\frac{n}{2}}(x))^2 dx - \Big(\int_\Omega (f^q(x) - \hat{f}^q_{1:\frac{n}{2}}(x))dx\Big)^2. \tag{C.1}$$

By plugging in the expression of $\hat{H}^q_C, I^q_f$ and using the fact that $\{x_i\}_{i=1}^n$ are identical and independent copies of the uniform random variable over Ω, we have

$$\begin{aligned}
&\mathbb{E}_{\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n}\left[\Big|\hat{H}^q_C\Big(\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n\Big) - I^q_f\Big|^2\right] \\
&= \mathbb{E}_{\{x_i\}_{i=1}^n}\left[\Big|\int_\Omega \hat{f}^q_{1:\frac{n}{2}}(x)dx + \frac{2}{n}\sum_{i=\frac{n}{2}+1}^{n}\Big(f^q(x_i) - \hat{f}^q_{1:\frac{n}{2}}(x_i)\Big) - \int_\Omega f^q(x)dx\Big|^2\right] \\
&= \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\mathbb{E}_{\{x_i\}_{i=\frac{n}{2}+1}^n}\left[\Big|\frac{1}{\frac{n}{2}}\sum_{i=\frac{n}{2}+1}^{n}\Big(f^q(x_i) - \hat{f}^q_{1:\frac{n}{2}}(x_i) - \int_\Omega (f^q(x) - \hat{f}^q_{1:\frac{n}{2}}(x))dx\Big)\Big|^2\right]\right] \\
&= \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\frac{4}{n^2}\sum_{i=\frac{n}{2}+1}^{n}\mathbb{E}_{x_i}\left[\Big|\Big(f^q(x_i) - \hat{f}^q_{1:\frac{n}{2}}(x_i) - \int_\Omega (f^q(x) - \hat{f}^q_{1:\frac{n}{2}}(x))dx\Big)\Big|^2\right]\right] \\
&= \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\Big[\frac{4}{n^2}\sum_{i=\frac{n}{2}+1}^{n}\mathrm{Var}(\hat{f}^q_{1:\frac{n}{2}} - f^q)\Big] = \frac{2}{n}\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\Big[\mathrm{Var}(\hat{f}^q_{1:\frac{n}{2}} - f^q)\Big].
\end{aligned} \tag{C.2}$$

From the identity above, we know that it suffices to upper bound the term $\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\Big[\mathrm{Var}(\hat{f}^q_{1:\frac{n}{2}} - f^q)\Big]$. Let $g_{1:\frac{n}{2}} := \hat{f}_{1:\frac{n}{2}} - f$ denote the difference between the estimator $\hat{f}_{1:\frac{n}{2}}$ and underlying function f. Then we may further upper bound the expression $\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\Big[\mathrm{Var}(\hat{f}^q_{1:\frac{n}{2}} - f^q)\Big]$ as follows:

$$\begin{aligned}
&\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\Big[\mathrm{Var}(\hat{f}^q_{1:\frac{n}{2}} - f^q)\Big] = \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\Big[\int_\Omega (f^q(x) - \hat{f}^q_{1:\frac{n}{2}}(x))^2 dx - \Big(\int_\Omega (f^q(x) - \hat{f}^q_{1:\frac{n}{2}}(x))dx\Big)^2\Big] \\
&\le \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\Big[\int_\Omega \Big(f^q(x) - \hat{f}^q_{1:\frac{n}{2}}(x)\Big)^2 dx\Big] = \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\int_\Omega \Big((f(x) + g_{1:\frac{n}{2}}(x))^q - f^q(x)\Big)^2 dx\right] \\
&= \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\int_\Omega \Big(\int_0^{g_{1:\frac{n}{2}}(x)} q(f(x)+y)^{q-1}dy\Big)^2 dx\right] \\
&\le \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\int_\Omega \Big|\int_0^{g_{1:\frac{n}{2}}(x)} 1dy\Big|\Big|\int_0^{g_{1:\frac{n}{2}}(x)} q^2(|f(x)+y|^2)^{q-1}dy\Big|dx\right] \\
&\lesssim \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\int_\Omega |g_{1:\frac{n}{2}}(x)|\cdot|g_{1:\frac{n}{2}}(x)|\max\Big\{|f^{2q-2}(x)|, |g^{2q-2}_{1:\frac{n}{2}}(x)|\Big\}dx\right] \\
&\lesssim \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\int_\Omega |g^{2q}_{1:\frac{n}{2}}(x)|dx\right] + \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\int_\Omega |g^2_{1:\frac{n}{2}}(x)f^{2q-2}(x)|dx\right].
\end{aligned} \tag{C.3}$$

Now let's proceed to bound from above the two expected integrals in the last line of (C.3). For the first expected integral, since $s > \frac{2dq-dp}{2pq} \Rightarrow \frac{1}{2q} > \frac{d-sp}{pd}$, we may apply (3.2) in Assumption 3.1 to

deduce that

$$\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\int_\Omega |g_{1:\frac{n}{2}}^{2q}(x)|dx\right] = \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\|\hat{f}_{1:\frac{n}{2}} - f\|_{L^{2q}(\Omega)}^{2q}\right] \\ \lesssim ((\frac{n}{2})^{-\frac{s}{d}+(\frac{1}{p}-\frac{1}{2q})_+})^{2q} \lesssim n^{2q(-\frac{s}{d}+\frac{1}{p}-\frac{1}{2q})} = n^{2q(\frac{1}{p}-\frac{s}{d})-1}, \tag{C.4}$$

where the last equality above follows from the given assumption that $p < 2q$. Now let's proceed to bound from above the second expected integral in (C.3). Here we define $p^* = (\max\{\frac{1}{p} - \frac{s}{d}, 0\})^{-1}$, *i.e,* $p^* = \frac{pd}{d-sp}$ when $s < \frac{d}{p}$ and $p^* = \infty$ otherwise. From Sobolev Embedding Theorem (Lemma 1), we have that $W^{s,p}(\Omega) \subseteq L^{p^*}(\Omega)$. Based on the value of the smoothness parameter s, we have three separate cases as below:

(Case I) When $s \in (\frac{d}{p}, \infty)$, we have $p^* = \infty$ and $f \in W^{s,p}(\Omega) \subset L^\infty(\Omega)$. Since $\hat{f}_{1:\frac{n}{2}}$ and f are both in the Sobolev space $W^{s,p}(\Omega) \subseteq L^\infty(\Omega)$, we may further deduce that $g_{1:\frac{n}{2}} = \hat{f}_{1:\frac{n}{2}} - f \in W^{s,p}(\Omega) \subseteq L^\infty(\Omega) \subseteq L^2(\Omega)$. By picking $r = 2$ in (3.2) of Assumption 3.1, we may use the facts that $p > 2$ and $f \in L^\infty(\Omega)$ to deduce that

$$\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\int_\Omega |g_{1:\frac{n}{2}}^2(x) f^{2q-2}(x)|dx\right] \lesssim \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\int_\Omega |g_{1:\frac{n}{2}}^2(x)|dx\right] \\ = \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\|\hat{f}_{1:\frac{n}{2}} - f\|_{L^2(\Omega)}^2\right] \lesssim \left(n^{-\frac{s}{d}+(\frac{1}{p}-\frac{1}{2})_+}\right)^2 = n^{-\frac{2s}{d}}, \tag{C.5}$$

which is our final upper bound on the second expected integral in (C.3) under the assumption that $s \in (\frac{d}{p}, \infty)$.

(Case II) When $s \in (\frac{d(2q-p)}{p(2q-2)}, \frac{d}{p})$, we have $p^* = \frac{pd}{d-sp} > \frac{pd}{d-p\frac{d(2q-p)}{p(2q-2)}} = \frac{p(2q-2)}{p-2}$, which implies $f \in W^{s,p}(\Omega) \subseteq L^{p^*}(\Omega) \subseteq L^{\frac{p(2q-2)}{p-2}}(\Omega) \subseteq L^p(\Omega)$. Given that $\frac{p}{p-2} > 1$, we can further deduce that $f^{2q-2} \in L^{\frac{p}{p-2}}(\Omega)$. Moreover, since $\hat{f}_{1:\frac{n}{2}} \in W^{s,p}(\Omega) \subseteq L^p(\Omega)$, we have that $g_{1:\frac{n}{2}} = \hat{f}_{1:\frac{n}{2}} - f \in L^p(\Omega)$. Given that $p > 2$, we can further deduce that $g_{1:\frac{n}{2}}^2 \in L^{\frac{p}{2}}(\Omega)$. Then we may apply Hölder's inequality (Lemma 2) to $g_{1:\frac{n}{2}}^2 \in L^{\frac{p}{2}}(\Omega)$ and $f^{2q-2} \in L^{\frac{p}{p-2}}(\Omega)$ to obtain that

$$\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\int_\Omega |g_{1:\frac{n}{2}}^2(x) f^{2q-2}(x)|dx\right] = \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\|g_{1:\frac{n}{2}}^2 f^{2q-2}\|_{L^1(\Omega)}\right] \\ \leq \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\left\|g_{1:\frac{n}{2}}^2\right\|_{L^{\frac{p}{2}}(\Omega)}\left\|f^{2q-2}\right\|_{L^{\frac{p}{p-2}}(\Omega)}\right] \leq \left\|f\right\|_{L^{\frac{p(2q-2)}{p-2}}(\Omega)}^{2q-2}\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\left\|g_{1:\frac{n}{2}}\right\|_{L^p(\Omega)}^2\right]. \tag{C.6}$$

Note that the function $h(t) = t^{\frac{2}{p}}$ is concave and $\frac{1}{p} \in (\frac{d-sp}{pd}, 1]$ when $p > 2$. Hence, applying Jensen's inequality and picking $r = p$ in (3.2) of Assumption 3.1 further allows us to upper bound the last term in (C.6) as follows:

$$\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\|g_{1:\frac{n}{2}}\|_{L^p(\Omega)}^2\right] = \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\left(\|g_{1:\frac{n}{2}}\|_{L^p(\Omega)}^p\right)^{\frac{2}{p}}\right] \\ \leq \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\|g_{1:\frac{n}{2}}\|_{L^p(\Omega)}^p\right]^{\frac{2}{p}} = \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\|\hat{f}_{1:\frac{n}{2}} - f\|_{L^p(\Omega)}^p\right]^{\frac{2}{p}} \\ \lesssim \left((\frac{n}{2})^{-\frac{s}{d}+(\frac{1}{p}-\frac{1}{p})_+}\right)^2 \lesssim n^{-\frac{2s}{d}}. \tag{C.7}$$

Substituting (C.7) into (C.6) then gives us the final upper bound on the second expected integral in (C.3) under the assumption that $s \in (\frac{d(2q-p)}{p(2q-2)}, \infty)$:

$$\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\int_\Omega |g_{1:\frac{n}{2}}^2(x) f^{2q-2}(x)|dx\right] \lesssim n^{-\frac{2s}{d}}. \tag{C.8}$$

(Case III) When $s \in (\frac{d(2q-p)}{2pq}, \frac{d(2q-p)}{p(2q-2)})$, we have that $s < \frac{d}{p}$, which indicates that $p^* = \frac{pd}{d-sp}$ satisfies $2q < p^* < \frac{p(2q-2)}{p-2}$. Given that $p^* > 2q > 2q-2$ and $f \in W^{s,p}(\Omega) \subseteq L^{p^*}(\Omega)$, we can deduce

that $f^{2q-2} \in L^{\frac{p^*}{2q-2}}(\Omega)$. Furthermore, note that $p^* > 2q$ implies $\frac{2p^*}{p^*+2-2q} < p^*$ and $p^* < \frac{p(2q-2)}{p-2}$ implies $\frac{2p^*}{p^*+2-2q} > p$. Since $\hat{f}_{1:\frac{n}{2}}$ and f are both in the Sobolev space $W^{s,p}(\Omega) \subseteq L^{p^*}(\Omega)$, we may further deduce that $g_{1:\frac{n}{2}} = \hat{f}_{1:\frac{n}{2}} - f \in W^{s,p}(\Omega) \subseteq L^{p^*}(\Omega) \subseteq L^{\frac{2p^*}{p^*+2-2q}}(\Omega)$. Given that $q \geq 1 \Rightarrow \frac{p^*}{p^*+2-2q} \geq 1$, we have $g^2_{1:\frac{n}{2}} \in L^{\frac{p^*}{p^*+2-2q}}(\Omega)$. Then we may apply Hölder's inequality (Lemma 2) to $g^2_{1:\frac{n}{2}} \in L^{\frac{p^*}{p^*+2-2q}}(\Omega)$ and $f^{2q-2} \in L^{\frac{p^*}{2q-2}}(\Omega)$, which yields the following upper bound:

$$\begin{aligned}
\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\int_\Omega |g^2_{1:\frac{n}{2}}(x) f^{2q-2}(x)|dx\right] &= \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\|g^2_{1:\frac{n}{2}} f^{2q-2}\|_{L^1(\Omega)}\right] \\
&\leq \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\left\|g^2_{1:\frac{n}{2}}\right\|_{L^{\frac{p^*}{p^*+2-2q}}(\Omega)} \left\|f^{2q-2}\right\|_{L^{\frac{p^*}{2q-2}}(\Omega)}\right] \quad \text{(C.9)} \\
&\leq \left\|f\right\|^{2q-2}_{L^{p^*}(\Omega)} \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\left\|g_{1:\frac{n}{2}}\right\|^2_{L^{\frac{2p^*}{p^*+2-2q}}(\Omega)}\right].
\end{aligned}$$

Note that the function $\omega(t) = t^{\frac{p^*+2-2q}{p^*}}$ is concave since $q \geq 1$. Moreover, using the given assumption $s \in (\frac{d(2q-p)}{2pq}, \frac{d(2q-p)}{p(2q-2)})$ we get that $\frac{pd}{d-sp} > 2q$, which further yields

$$\frac{p^*+2-2q}{2p^*} = \frac{\frac{pd}{d-sp}+2-2q}{2\frac{pd}{d-sp}} > \frac{2}{2\frac{pd}{d-sp}} = \frac{d-sp}{pd},$$

i.e, $\frac{(p^*+2-2q)}{2p^*} \in (\frac{d-sp}{pd}, 1]$. Hence, we may apply Jensen's inequality and (3.2) in Assumption 3.1 to upper-bound the last term in (C.9) as follows:

$$\begin{aligned}
\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\left\|g_{1:\frac{n}{2}}\right\|^2_{L^{\frac{2p^*}{p^*+2-2q}}(\Omega)}\right] &= \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\left(\left\|g_{1:\frac{n}{2}}\right\|^{\frac{2p^*}{p^*+2-2q}}_{L^{\frac{2p^*}{p^*+2-2q}}(\Omega)}\right)^{\frac{p^*+2-2q}{p^*}}\right] \\
&\leq \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\left\|g_{1:\frac{n}{2}}\right\|^{\frac{2p^*}{p^*+2-2q}}_{L^{\frac{2p^*}{p^*+2-2q}}(\Omega)}\right]^{\frac{(p^*+2-2q)}{p^*}} \\
&= \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\left\|\hat{f}_{1:\frac{n}{2}} - f\right\|^{\frac{2p^*}{p^*+2-2q}}_{L^{\frac{2p^*}{p^*+2-2q}}(\Omega)}\right]^{\frac{(p^*+2-2q)}{p^*}} \quad \text{(C.10)} \\
&\lesssim \left((\frac{n}{2})^{-\frac{s}{d}+(\frac{1}{p}-\frac{p^*+2-2q}{2p^*})_+}\right)^2 \\
&\lesssim n^{-\frac{2s}{d}+2(\frac{1}{p}-\frac{p^*+2-2q}{2p^*})_+}.
\end{aligned}$$

In order to simplify the last expression in (C.10), let's recall the fact that $p^* \in (2q, \frac{p(2q-2)}{p-2})$ proved above. This gives us that $p^*(p-2) < p(2q-2) \Rightarrow 2p^* > p(p^*+2-2q)$, *i.e,* $\frac{1}{p} > \frac{p^*+2-2q}{2p^*}$. Then we may simplify the power term in the last expression of (C.10) as follows:

$$-\frac{2s}{d} + 2\Big(\frac{1}{p} - \frac{p^*+2-2q}{2p^*}\Big)_+ = -\frac{2s}{d} + \frac{2}{p} - \Big(1 + \frac{2}{p^*} - \frac{2q}{p^*}\Big) = \frac{2q}{p^*} - 1 = 2q\Big(\frac{1}{p} - \frac{s}{d}\Big) - 1.$$

Now let's substitute (C.10) into (C.9), which gives us the final upper bound on the second expected integral in (C.3) under the assumption that $s \in (\frac{d(2q-p)}{2pq}, \frac{d(2q-p)}{p(2q-2)})$:

$$\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\int_\Omega |g^2_{1:\frac{n}{2}}(x) f^{2q-2}(x)|dx\right] \lesssim n^{2q(\frac{1}{p}-\frac{s}{d})-1}. \quad \text{(C.11)}$$

Combining the upper bounds derived in (C.4), (C.5), (C.8) and (C.11) finally allows us to upper bound the expected variance $\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\Big[\text{Var}(\hat{f}^q_{1:\frac{n}{2}} - f^q)\Big]$ as below:

$$\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\Big[\text{Var}(\hat{f}^q_{1:\frac{n}{2}} - f^q)\Big] \lesssim n^{2q(\frac{1}{p}-\frac{s}{d})-1} + \max\{n^{-\frac{2s}{d}}, n^{2q(\frac{1}{p}-\frac{s}{d})-1}\}. \quad \text{(C.12)}$$

Finally, substituting (C.12) into C.2) derived at the beginning gives us the final upper bound:

$$\begin{aligned}
&\mathbb{E}_{\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n}\left[\left|\hat{H}_C^q\Big(\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n\Big)-I_f^q\right|\right]\\
&\leq\sqrt{\mathbb{E}_{\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n}\left[\left|\hat{H}_C^q\Big(\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n\Big)-I_f^q\right|^2\right]}=\sqrt{\frac{2}{n}\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\Big[\mathrm{Var}(\hat{f}^q_{1:\frac{n}{2}}-f^q)\Big]}\\
&\lesssim n^{-\frac{1}{2}}\sqrt{n^{2q(\frac{1}{p}-\frac{s}{d})-1}+\max\{n^{-\frac{2s}{d}},n^{2q(\frac{1}{p}-\frac{s}{d})-1}\}}\lesssim\max\{n^{-\frac{s}{d}-\frac{1}{2}},n^{-q(\frac{s}{d}-\frac{1}{p})-1}\}.
\end{aligned}\tag{C.13}$$

This concludes our proof of Theorem 3.1.

C.2 Proof of Theorem 3.2 (Truncated Monte Carlo)

In this subsection, we provide a complete proof of Theorem 3.2. For any fixed parameter $M>0$, we may divide Ω into the following two regions:

$$\Omega_M^+:=\{x\in\Omega:|f(x)|\geq M\},\ \Omega_M^-:=\{x\in\Omega:|f(x)|<M\},\tag{C.14}$$

where $\Omega_M^+\cap\Omega_M^-=\varnothing$ and $\Omega_M^+\cup\Omega_M^-=\Omega$. Let $f_M(x):=\max\Big\{\min\{f(x),M\},-M\Big\}\ (\forall\ x\in\Omega)$ denote a truncated version of the given function f, where M is the threshold. Also, we use the following expression to denote the expectation of the q-th power of the truncated function f_M with respect to the uniform distribution on Ω:

$$\mathbb{E}(f_M^q(x))=\int_\Omega\max\Big\{\min\{f(x),M\},-M\Big\}^q dx=\int_{\Omega_M^+}M^q dx+\int_{\Omega_M^-}f(x)^q dx,\tag{C.15}$$

where the last identity in (C.15) above follows from our definition of the two regions defined in (C.14). In a similar way, we can define the variance of the function f_M^q as below:

$$\begin{aligned}
&\mathrm{Var}(f_M^q(x))=\mathbb{E}(f_M^{2q}(x))-\mathbb{E}(f_M^q(x))^2\\
&=\int_\Omega\max\Big\{\min\{f(x),M\},-M\Big\}^{2q}dx-\Big(\int_\Omega\max\Big\{\min\{f(x),M\},-M\Big\}^q dx\Big)^2.
\end{aligned}\tag{C.16}$$

Furthermore, as $\{x_i\}_{i=1}^n$ are identical and independent samples of the uniform distribution on Ω, we have that for any $1\leq i\leq n$, the following identity holds

$$\begin{aligned}
&\mathbb{E}_{\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n}\Big[\hat{H}_M^q\Big(\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n\Big)\Big]\\
&=\mathbb{E}_{\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n}\Big[\frac{1}{n}\sum_{i=1}^n\max\Big\{\min\{y_i,M\},-M\Big\}^q\Big]\\
&=\mathbb{E}_{x_i}\Big[\max\Big\{\min\{f(x_i),M\},-M\Big\}^q\Big]=\mathbb{E}_{x_i}[f_M^q(x_i)]=\mathbb{E}(f_M^q(x)).
\end{aligned}\tag{C.17}$$

Now we may use (C.17) and the bias-variance decomposition to derive an upper bound on the squared expected risk of the estimator $\hat{H}_M^q$ as follows:

$$\begin{aligned}
&\mathbb{E}_{\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n}\left[\left|\hat{H}_M^q\Big(\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n\Big)-I_f^q\right|^2\right]\\
&=\mathbb{E}_{\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n}\left[\left|\hat{H}_M^q\Big(\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n\Big)-\mathbb{E}(f_M^q(x))+\mathbb{E}(f_M^q(x))-I_f^q\right|^2\right]\\
&\leq 2\mathbb{E}_{\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n}\left[\left|\frac{1}{n}\sum_{i=1}^n\max\Big\{\min\{y_i,M\},-M\Big\}^q-\mathbb{E}(f_M^q(x))\right|^2\right]\\
&+2\mathbb{E}_{\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n}\left[\left|\mathbb{E}(f_M^q(x))-I_f^q\right|^2\right]
\end{aligned}\tag{C.18}$$

where the first and the second term in the last line of (C.18) above denotes the variance and the bias part, respectively. Again, we define $p^* = (\max\{\frac{1}{p} - \frac{s}{d}, 0\})^{-1}$, *i.e,* $p^* = \frac{pd}{d-sp}$ when $s < \frac{d}{p}$ and $p^* = \infty$ otherwise. Under the assumption that $s < \frac{2dq-dp}{2pq} < \frac{d}{p}$, we have $p^* = \frac{pd}{d-sp} \in (p, 2q)$. Moreover, from Sobolev Embedding Theorem (Lemma 1), we have that $f \in W^{s,p}(\Omega) \subseteq L^{p^*}(\Omega)$.

On the one hand, since $p < 2q$, we can deduce that $|f(x)|^{2q} \leq M^{2q-p^*}|f(x)|^{p^*}$ for any $x \in \Omega_M^-$ and $M^{2q} \leq M^{2q-p^*}|f(x)|^{p^*}$ for any $x \in \Omega_M^+$, which helps us upper bound the variance part as below:

$$
\begin{aligned}
&\mathbb{E}_{\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n}\left[\left|\frac{1}{n}\sum_{i=1}^n \max\Big\{\min\{y_i, M\}, -M\Big\}^q - \mathbb{E}(f_M^q(x))\right|^2\right] \\
&= \mathbb{E}_{\{x_i\}_{i=1}^n}\left[\left|\frac{1}{n}\sum_{i=1}^n \Big(f_M^q(x_i) - \mathbb{E}_{x_i}\Big[f_M^q(x_i)\Big]\Big)\right|^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}_{x_i}\Big[\Big(f_M^q(x_i) - \mathbb{E}_{x_i}\Big[f_M^q(x_i)\Big]\Big)^2\Big] = \frac{1}{n}\mathrm{Var}(f_M^q(x)) \\
&\leq \frac{1}{n}\mathbb{E}(f_M^{2q}(x)) = \frac{1}{n}\Big(\int_{\Omega_M^+} M^{2q}dx + \int_{\Omega_M^-} f(x)^{2q}dx\Big) \\
&\leq \frac{1}{n}\Big(\int_{\Omega_M^+} M^{2q-p^*}|f(x)|^{p^*}dx + \int_{\Omega_M^-} M^{2q-p^*}|f(x)|^{p^*}dx\Big) \\
&= \frac{1}{n}\int_{\Omega} M^{2q-p^*}|f(x)|^{p^*}dx \lesssim \frac{M^{2q-p^*}}{n},
\end{aligned}
\tag{C.19}
$$

where the last step of (C.19) above follows from the fact that $f \in W^{s,p}(\Omega) \subseteq L^{p^*}(\Omega)$.

On the other hand, using the fact that $p^* > p > q \Rightarrow |f(x)|^q \leq M^{q-p^*}|f(x)|^{p^*}$ for any $x \in \Omega_M^+$, we may upper-bound the bias part as follows:

$$
\begin{aligned}
&\mathbb{E}_{\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n}\left[\left|\mathbb{E}(f_M^q(x)) - I_f^q\right|^2\right] \\
&= \left|\int_{\Omega_M^+} M^q dx + \int_{\Omega_M^-} f(x)^q dx - \int_{\Omega_M^+} f^q(x)dx - \int_{\Omega_M^-} f(x)^q dx\right|^2 \\
&= \left|\int_{\Omega_M^+} \Big(M^q - f^q(x)\Big)dx\right|^2 \leq \left|\int_{\Omega_M^+} \Big|M^q - f^q(x)\Big|dx\right|^2 \leq \left|\int_{\Omega_M^+} \Big(M^q + |f(x)|^q\Big)dx\right|^2 \\
&\leq \left|2\int_{\Omega_M^+} |f(x)|^q dx\right|^2 \lesssim \left|\int_{\Omega_M^+} M^{q-p^*}|f(x)|^{p^*}dx\right|^2 \leq M^{2q-2p^*}\left|\int_{\Omega} |f(x)|^{p^*}dx\right|^2 \\
&\lesssim M^{2q-2p^*},
\end{aligned}
\tag{C.20}
$$

where the last step above again follows from the fact that $f \in W^{s,p}(\Omega) \subseteq L^{p^*}(\Omega)$. By substituting (C.19) and (C.20) into (C.18), we obtain that

$$
\begin{aligned}
&\mathbb{E}_{\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n}\left[\left|\hat{H}_M^q\Big(\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n\Big) - I_f^q\right|\right] \\
&\leq \sqrt{\mathbb{E}_{\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n}\left[\left|\hat{H}_M^q\Big(\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n\Big) - I_f^q\right|^2\right]} \lesssim \sqrt{\frac{M^{2q-p^*}}{n} + M^{2q-2p^*}}.
\end{aligned}
\tag{C.21}
$$

By balancing the variance part $\frac{M^{2q-p^*}}{n}$ and the bias part M^{2q-2p^*} above, we may get the optimal choice of M as follows: $\frac{M^{2q-p^*}}{n} = M^{2q-2p^*} \Rightarrow M = \Theta(n^{\frac{1}{p^*}})$. Plugging in the optimal choice of

M gives us the final upper bound:

$$\mathbb{E}_{\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n}\left[\left|\hat{H}_M^q\Big(\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n\Big)-I_f^q\right|\right]\lesssim\sqrt{n^{\frac{2q-2p^*}{p^*}}}=n^{\frac{q}{p^*}-1}=n^{-q(\frac{s}{d}-\frac{1}{p})-1}, \tag{C.22}$$

which finishes our proof of Theorem 3.2.

D Proof of Minimax Lower and Upper Bounds in Section 4

This section is organized as follows. The first subsection consists of one important lemma used in our proof. In the second subsection, we provide complete proof for the minimax optimal lower bound on the estimation of integrals under any level of noise. In the third subsection, a complete proof for the upper bound on the estimation of integrals is given.

D.1 A Key Lemma for Establishing the Upper Bound on Integral Estimation

Lemma 6 (Bound on the Expected k-Nearest Neighbor Distance: Theorem 2.4, [69]) *Assume that $x_1, x_2, \cdots, x_n$ are independent and identical samples from the uniform distribution on the domain $\Omega=[0,1]^d$. For any $k\in\{1,2,\cdots,n\}$ and $z\in\Omega$, we use $x_{i_k^{(z)}}$ to denote the k-th nearest neighbor of z among $\{x_i\}_{i=1}^n$. When z is also uniformly distributed over the domain Ω, we have the following upper bound on the expected distance between z and $x_{i_k^{(z)}}$:*

$$\mathbb{E}_{z,\{x_i\}_{i=1}^n}\left[\|z-x_{i_k^{(z)}}\|^2\right]\lesssim(\frac{k}{n})^{\frac{2}{d}}. \tag{D.1}$$

D.2 Proof of Theorem 4.1 (Lower Bound on Integral Estimation)

Here we present a comprehensive proof of the two lower bounds given in Theorem 4.1 above by applying the method of two fuzzy hypotheses (Lemma 5). Below we again use $\vec{x}:=(x_1,x_2,\cdots,x_n)$ and $\vec{y}:=(y_1,y_2,\cdots,y_n)$ to denote the two n-dimensional vectors formed by the quadrature points and observed function values. Since our lower bound in Theorem 4.1 consists of two terms, we need to prove the two bounds in the following two separate cases:

(Case I) For the first lower bound in (4.1), let's consider two constant functions g_0 and g_1 defined as follows:

$$g_0(x)\equiv 0\ (\forall\, x\in\Omega),\ g_1(x)\equiv n^{-\gamma-\frac{1}{2}}\ (\forall\, x\in\Omega) \tag{D.2}$$

Clearly we have $g_0, g_1\in C^s(\Omega)$. Then let's take μ_k to be a Dirac delta measure supported on the set $\{g_j\}$, *i.e,* $\mu_k(\{g_k\})=1$, for $k\in\{0,1\}$. By picking $c=\Delta=\frac{1}{2}I_{g_1}=\frac{1}{2}n^{-\gamma-\frac{1}{2}}$ and $\beta_0=\beta_1=0$, we then obtain that

$$\begin{aligned}&\mu_0(f\in W^{s,p}(\Omega):I_f\le c-\Delta)=\mu_0(I_f\le 0)=1=1-\beta_0,\\&\mu_1(f\in W^{s,p}(\Omega):I_f\ge c+\Delta)=\mu_1(I_f\ge I_{g_1})=1=1-\beta_1,\end{aligned} \tag{D.3}$$

which indicates that (B.1) holds true. Now let's consider bounding the KL divergence between the two marginal distributions $\mathbb{P}_0,\mathbb{P}_1$ associated with μ_0,μ_1, respectively. Given that the quadrature points $\{x_i\}_{i=1}^n$ and the observational noises $\{\epsilon_i\}_{i=1}^n$ are independent and identical samples from the uniform distribution on Ω and the normal distribution $\mathcal{N}(0,n^{-2\gamma})$, we can write the marginal distributions in an explicit form as follows:

$$\mathbb{P}_0(\vec{x},\vec{y})=\prod_{i=1}^n\Big(\frac{1}{\sqrt{2\pi}n^{-\gamma}}e^{-\frac{1}{2n^{-2\gamma}}y_i^2}\Big),\ \mathbb{P}_1(\vec{x},\vec{y})=\prod_{i=1}^n\Big(\frac{1}{\sqrt{2\pi}n^{-\gamma}}e^{-\frac{1}{2n^{-2\gamma}}(y_i-n^{-\gamma-\frac{1}{2}})^2}\Big). \tag{D.4}$$

From (D.4) we can see that $\mathbb{P}_0$ and $\mathbb{P}_1$ are two n-dimensional normal distributions having the same covariance matrix but different mean vectors. Computing the KL divergence between them and applying Pinsker's inequality then give us that

$$TV(\mathbb{P}_0\|\mathbb{P}_1)\le\sqrt{\frac{1}{2}KL(\mathbb{P}_0\|\mathbb{P}_1)}=\sqrt{\frac{n(n^{-\gamma-\frac{1}{2}})^2}{4n^{-2\gamma}}}=\frac{1}{2}. \tag{D.5}$$

Substituting (D.5), $\Delta = \frac{1}{2} I_{g_1} = \frac{1}{2} n^{-\gamma-\frac{1}{2}}$ and $\beta_0 = \beta_1 = 0$ into (B.2) and applying Markov's inequality yield the final lower bound

$$\begin{aligned}
&\inf_{\hat{H} \in \mathcal{H}_n^f} \sup_{f \in C^s(\Omega)} \mathbb{E}_{\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n} \left[\left| \hat{H}\Big(\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n \Big) - I_f \right| \right] \\
&\geq \Delta \inf_{\hat{H} \in \mathcal{H}_n^f} \sup_{f \in C^s(\Omega)} \mathbb{P}_{\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n} \left[\left| \hat{H}\Big(\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n \Big) - I_f \right| \geq \Delta \right] \\
&\geq \frac{1}{2} I_{g_1} \frac{1 - TV(\mathbb{P}_0 || \mathbb{P}_1) - \beta_0 - \beta_1}{2} \geq \frac{1}{8} n^{-\gamma-\frac{1}{2}} \gtrsim n^{-\gamma-\frac{1}{2}},
\end{aligned} \tag{D.6}$$

which is exactly the first term in the RHS of (4.1).

(Case II) For the second lower bound in (4.1), our proof is similar to the proof of the second lower bound in Theorem 2.1 presented in Appendix B.2 above. Again, we select $m = (200n)^{\frac{1}{d}}$ and divide the domain Ω into m^d small cubes $\Omega_1, \Omega_2, \cdots, \Omega_{m^d}$, each of which has side length m^{-1}. For any $1 \leq j \leq m^d$, we use c_j to denote center of the cube Ω_j. Then let's consider the same bump function K defined in (B.3) and (B.4) above, which satisfies $\text{supp}(K) \subseteq [-\frac{1}{2}, \frac{1}{2}]^d$ and $K \in C^\infty([-\frac{1}{2}, \frac{1}{2}]^d)$. In an analogous way, for any $1 \leq j \leq m^d$, we associate each cube Ω_j with a bump function f_j defined as follows:

$$f_j(x) = \begin{cases} m^{-s} K(m(x - c_j)) \ (x \in \Omega_j), \\ 0 \ (\text{otherwise}), \end{cases} \tag{D.7}$$

where $\text{supp}(f_j) \subseteq \Omega_j$, $f_j \in C^\infty(\Omega)$ and $f_j(x) \geq 0$ $(\forall\ x \in \Omega)$. Then let's consider the following finite set of 2^{m^d} functions:

$$\mathcal{S} := \Big\{ \sum_{j=1}^{m^d} \eta_j f_j : \eta_j \in \{\pm 1\},\ \forall\, 1 \leq j \leq m^d \Big\}. \tag{D.8}$$

We will first verify that $\mathcal{S} \subseteq C^s(\Omega)$. Fix any element $f_* = \sum_{j=1}^{m^d} \eta_j f_j \in \mathcal{S}$. On the one hand, from our construction of the f_j's given in (D.7) above, we have

$$\begin{aligned}
\max_{|t| \leq \lfloor s \rfloor} \| D^t f_* \|_{L^\infty(\Omega)} &= \max_{|t| \leq \lfloor s \rfloor} m^{-s+|t|} \| D^t K \|_{L^\infty([-\frac{1}{2}, \frac{1}{2}]^d)} \\
&\leq \max_{|t| \leq \lfloor s \rfloor} \| D^t K \|_{L^\infty([-\frac{1}{2}, \frac{1}{2}]^d)}.
\end{aligned} \tag{D.9}$$

On the other hand, for any $1 \leq i \neq j \leq m^d$, we consider the function $\psi_i f_i + \psi_j f_j$, where the scalars $\psi_j, \psi_j \in \{0, \pm 1\}$. Now let's may pick $\beta := \frac{d}{1 - \{s\}}$, where $\{s\} = s - \lfloor s \rfloor \in (0, 1)$ denotes the fractional part of s. Given that $f_j \in C^\infty(\Omega)$, we may upper bound the Sobolev norm $\| \cdot \|_{W^{1,\beta}}$ of the function $D^t(\psi_i f_i + \psi_j f_j)$ for any $t \in \mathbb{N}_0^d$ satisfying $|t| = \lfloor s \rfloor$ as follows:

$$\begin{aligned}
&\Big\| D^t(\psi_i f_i + \psi_j f_j) \Big\|_{W^{1,\beta}(\Omega)}^\beta = |\psi_i|^\beta \Big\| D^t f_i \Big\|_{W^{1,\beta}(\Omega_i)}^\beta + |\psi_j|^\beta \Big\| D^t f_j \Big\|_{W^{1,\beta}(\Omega_j)}^\beta \\
&\leq \Big\| D^t f_i \Big\|_{L^\beta(\Omega_i)}^\beta + \sum_{r=1}^d \Big\| \frac{\partial}{\partial x_r} D^t f_i \Big\|_{L^\beta(\Omega_i)}^\beta + \Big\| D^{\lfloor s \rfloor} f_j \Big\|_{L^\beta(\Omega_j)}^\beta + \sum_{r=1}^d \Big\| \frac{\partial}{\partial x_r} D^t f_j \Big\|_{L^\beta(\Omega_j)}^\beta \\
&= \sum_{l \in \{i,j\}} \int_{\Omega_l} \Big(m^{-s+|t|} D^t K(m(x - c_l)) \Big)^\beta dx \\
&+ \sum_{l \in \{i,j\}} \sum_{r=1}^d \int_{\Omega_l} \Big(m^{-s+|t|+1} \frac{\partial}{\partial x_r} D^t K(m(x - c_l)) \Big)^\beta dx.
\end{aligned} \tag{D.10}$$

From our choice of β and assumption on the bump function K, we may further upper bound the Sobolev norm $\left\|D^t(\psi_i f_i + \psi_j f_j)\right\|_{W^{1,\beta}(\Omega)}$ as below:

$$\begin{aligned}
&\left\|D^t(\psi_i f_i + \psi_j f_j)\right\|^\beta_{W^{1,\beta}(\Omega)} \leq \sum_{l\in\{i,j\}} m^{-\beta\{s\}} \int_{[-\frac{1}{2},\frac{1}{2}]^d} \left(D^t K(y)\right)^\beta \frac{1}{m^d} dy \\
&+ \sum_{l\in\{i,j\}} dm^{\beta(1-\{s\})} \sup_{|t'|\leq\lfloor s\rfloor+1} \left(\int_{[-\frac{1}{2},\frac{1}{2}]^d} \left(D^{t'} K(y)\right)^\beta \frac{1}{m^d} dy \right) \\
&\leq 2m^{-\beta\{s\}-d} \left\|D^t K\right\|^\beta_{L^\beta([-\frac{1}{2},\frac{1}{2}]^d)} + 2dm^{\beta(1-\{s\})-d} \cdot \sup_{|t'|\leq\lfloor s\rfloor+1} \left\|D^{t'} K\right\|^\beta_{L^\beta([-\frac{1}{2},\frac{1}{2}]^d)} \\
&\lesssim \left\|D^t K\right\|^\beta_{L^\beta([-\frac{1}{2},\frac{1}{2}]^d)} + \sup_{|t'|\leq\lfloor s\rfloor+1} \left\|D^{t'} K\right\|^\beta_{L^\beta([-\frac{1}{2},\frac{1}{2}]^d)},
\end{aligned} \tag{D.11}$$

where the last inequality above follows from our choice of β. From (D.11) and the second part of the Sobolev Embedding Theorem (Lemma 1), we can deduce that $D^t(\psi_i f_i + \psi_j f_j) \in C^1(\Omega) \cap W^{1,\frac{d}{1-\{s\}}}(\Omega) \subseteq C^{\{s\}}(\Omega)$ and the following inequality holds:

$$\begin{aligned}
\left\|D^t(\psi_i f_i + \psi_j f_j)\right\|_{C^{\{s\}}(\Omega)} &\lesssim \left\|D^t(\psi_i f_i + \psi_j f_j)\right\|_{W^{1,\beta}(\Omega)} \\
&\lesssim \left(\sup_{|t'|=\lfloor s\rfloor} \left\|D^{t'} K\right\|^\beta_{L^\beta([-\frac{1}{2},\frac{1}{2}]^d)} + \sup_{|t'|=\lfloor s\rfloor+1} \left\|D^{t'} K\right\|^\beta_{L^\beta([-\frac{1}{2},\frac{1}{2}]^d)} \right)^{\frac{1}{\beta}},
\end{aligned} \tag{D.12}$$

Furthermore, combining (D.12) with our construction of the f_j's given in (D.7) above gives us that

$$\begin{aligned}
&\max_{|t|=\lfloor s\rfloor} \sup_{x,y\in\Omega, x\neq y} \frac{|D^t f_*(x) - D^t f_*(y)|}{\|x-y\|^{s-\lfloor s\rfloor}} \\
&\leq \max_{\substack{1\leq i\neq j\leq k \\ \psi_i,\psi_j\in\{0,\pm 1\}}} \max_{|t|=\lfloor s\rfloor} \sup_{x\neq y\in\Omega} \frac{|D^t(\psi_i f_i + \psi_j f_j)(x) - D^t(\psi_i f_i + \psi_j f_j)(x)|}{\|x-y\|^{\{s\}}} \\
&\leq \max_{\substack{1\leq i\neq j\leq k, |t|=\lfloor s\rfloor \\ \psi_i,\psi_j\in\{0,\pm 1\}}} \left\|D^t(\psi_i f_i + \psi_j f_j)\right\|_{C^{\{s\}}(\Omega)} \\
&\lesssim \left(\sup_{|t'|=\lfloor s\rfloor} \left\|D^{t'} K\right\|^\beta_{L^\beta([-\frac{1}{2},\frac{1}{2}]^d)} + \sup_{|t'|=\lfloor s\rfloor+1} \left\|D^{t'} K\right\|^\beta_{L^\beta([-\frac{1}{2},\frac{1}{2}]^d)} \right)^{\frac{1}{\beta}}
\end{aligned} \tag{D.13}$$

Finally, adding the two inequalities (D.9) and (D.13) gives us that for any $f_* \in \mathcal{S}$, we have

$$\begin{aligned}
\|f_*\|_{C^s(\Omega)} &= \max_{|t|\leq\lfloor s\rfloor} \|D^t f_*\|_{L^\infty(\Omega)} + \max_{|t|=\lfloor s\rfloor} \sup_{x,y\in\Omega,x\neq y} \frac{|D^t f_*(x) - D^t f_*(y)|}{\|x-y\|^{s-\lfloor s\rfloor}} \\
&\lesssim \max_{|t|\leq\lfloor s\rfloor} \|D^t K\|_{L^\infty([-\frac{1}{2},\frac{1}{2}]^d)} \\
&+ \left(\sup_{|t'|=\lfloor s\rfloor} \left\|D^{t'} K\right\|^\beta_{L^\beta([-\frac{1}{2},\frac{1}{2}]^d)} + \sup_{|t'|=\lfloor s\rfloor+1} \left\|D^{t'} K\right\|^\beta_{L^\beta([-\frac{1}{2},\frac{1}{2}]^d)} \right)^{\frac{1}{\beta}} \lesssim 1.
\end{aligned} \tag{D.14}$$

From the arbitrariness of f_*, we can then deduce that $\mathcal{S} \subseteq C^s(\Omega)$, as desired. For any $p \in (0,1)$, below we again use w_p to denote the discrete random variable satisfying $\mathbb{P}(w_p = -1) = p$ and $\mathbb{P}(w_p = 1) = 1-p$. Now let's pick $\kappa = \frac{1}{3}\sqrt{\frac{2}{3n}}$ and take $\{w_j^{(0)}\}_{j=1}^{m^d}$ and $\{w_j^{(1)}\}_{j=1}^{m^d}$ to be independent and identical copies of $w_{\frac{1+\kappa}{2}}$ and $w_{\frac{1-\kappa}{2}}$ respectively. Then we define μ_0, μ_1 to be two discrete measures supported on the finite set $\mathcal{S}$ such that the following condition holds for any $\eta_j \in \{\pm 1\}$ $(1 \leq j \leq m^d)$:

$$\mu_k\Big(\Big\{\sum_{j=1}^{m^d} \eta_j f_j\Big\}\Big) = \prod_{j=1}^{m^d} \mathbb{P}(w_j^{(k)} = \eta_j),\ k \in \{0,1\}. \tag{D.15}$$

Then we proceed to determine the separation distance Δ between the two priors μ_0 and μ_1. Similar to what we did in the proof of Theorem 2.1, we need to first define the following quantity $C := \int_{\Omega_j} f_j(x)dx$, which remains the same for any $1 \le j \le m^d$. Moreover, applying (D.7) helps us evaluate the quantity C directly as follows

$$\begin{aligned} C &= \int_{\Omega_j} f_j(x)dx = \int_{\Omega_j} m^{-s}K(m(x-c_j))dx \\ &= m^{-s}\int_{[-\frac{1}{2},\frac{1}{2}]^d} K(y)\frac{1}{m^d}dy = m^{-s-d}\|K\|_{L^1([-\frac{1}{2},\frac{1}{2}]^d)}. \end{aligned} \tag{D.16}$$

Moreover, by picking $\lambda = \frac{1}{2}$, we may apply Hoeffding's Inequality (Lemma 3) to the bounded random variables $\{w_j^{(0)}\}_{j=1}^{m^d}$ and $\{w_j^{(1)}\}_{j=1}^{m^d}$ to deduce that

$$\begin{aligned} \mathbb{P}\Big(\sum_{j=1}^{m^d} w_j^{(0)} \ge -(1-\lambda)m^d\kappa\Big) &\le \exp\Big(-\frac{2(\lambda m^d\kappa)^2}{4m^d}\Big) = \exp\Big(-\frac{1}{2}\lambda^2\kappa^2 m^d\Big), \\ \mathbb{P}\Big(\sum_{j=1}^{m^d} w_j^{(1)} \le (1-\lambda)m^d\kappa\Big) &\le \exp\Big(-\frac{2(\lambda m^d\kappa)^2}{4m^d}\Big) = \exp\Big(-\frac{1}{2}\lambda^2\kappa^2 m^d\Big). \end{aligned} \tag{D.17}$$

By taking $c := 0, \Delta := (1-\lambda)\kappa m^d C$ and $\beta_0 = \beta_1 = \exp\Big(-\frac{1}{2}\lambda^2\kappa^2 m^d\Big)$, we may use (D.17) justified above to get that

$$\begin{aligned} &\mu_0(f \in C^s(\Omega) : I_f \le c-\Delta) = \mathbb{P}\Big(\sum_{j=1}^{m^d} I_{w_j^{(0)} f_j} \le \frac{1-(1-\lambda)\kappa}{2}m^d C - \frac{1+(1-\lambda)\kappa}{2}m^d C\Big) \\ &\ge \mathbb{P}\Big(\sum_{j=1}^{m^d} w_j^{(0)} \le -(1-\lambda)m^d\kappa\Big) = 1 - \mathbb{P}\Big(\sum_{j=1}^{m^d} w_j^{(0)} \ge -(1-\lambda)m^d\kappa\Big) \\ &\ge 1 - \exp\Big(-\frac{1}{2}\lambda^2\kappa^2 m^d\Big) = 1-\beta_0, \\ &\mu_1(f \in C^s(\Omega) : I_f \ge c+\Delta) = \mathbb{P}\Big(\sum_{j=1}^{m^d} I_{w_j^{(1)} f_j} \ge \frac{1+(1-\lambda)\kappa}{2}m^d C - \frac{1-(1-\lambda)\kappa}{2}m^d C\Big) \\ &\ge \mathbb{P}\Big(\sum_{j=1}^{m^d} w_j^{(1)} \ge (1-\lambda)m^d\kappa\Big) = 1 - \mathbb{P}\Big(\sum_{j=1}^{m^d} w_j^{(0)} \le (1-\lambda)m^d\kappa\Big) \\ &\ge 1 - \exp\Big(-\frac{1}{2}\lambda^2\kappa^2 m^d\Big) = 1-\beta_1, \end{aligned} \tag{D.18}$$

which indicates that (B.1) holds true. Now let's consider bounding the KL divergence between the two marginal distributions $\mathbb{P}_0, \mathbb{P}_1$ associated with μ_0, μ_1, respectively. Applying the fact that $\{x_1, \cdots, x_n\}$ and $\{\epsilon_1, \cdots, \epsilon_n\}$ are identical and independent samples from the uniform distribution on Ω and the normal distribution $\mathcal{N}(0, n^{-2\gamma})$ allows us to write the marginal distributions in an explicit form as follows:

$$\begin{aligned} \mathbb{P}_0(\vec{x},\vec{y}) &= \prod_{j=1}^{m^d}\Big(\frac{1-\kappa}{2}\prod_{i:x_i\in\Omega_j}\frac{1}{\sqrt{2\pi}n^{-\gamma}}e^{-\frac{(y_i-f_j(x_i))^2}{2n^{-2\gamma}}} + \frac{1+\kappa}{2}\prod_{i:x_i\in\Omega_j}\frac{1}{\sqrt{2\pi}n^{-\gamma}}e^{-\frac{(y_i+f_j(x_i))^2}{2n^{-2\gamma}}}\Big), \\ \mathbb{P}_1(\vec{x},\vec{y}) &= \prod_{j=1}^{m^d}\Big(\frac{1+\kappa}{2}\prod_{i:x_i\in\Omega_j}\frac{1}{\sqrt{2\pi}n^{-\gamma}}e^{-\frac{(y_i-f_j(x_i))^2}{2n^{-2\gamma}}} + \frac{1-\kappa}{2}\prod_{i:x_i\in\Omega_j}\frac{1}{\sqrt{2\pi}n^{-\gamma}}e^{-\frac{(y_i+f_j(x_i))^2}{2n^{-2\gamma}}}\Big). \end{aligned} \tag{D.19}$$

Furthermore, for any n fixed quadrature points $\vec{x} = (x_1, x_2, \cdots, x_n)$, we use $\mathbb{P}_k(\cdot \mid \vec{x})$ to denote the marginal distribution of the observed function values $\vec{y} = (y_1, y_2, \cdots, y_n)$ conditioned on $\vec{x}$ for $k \in \{0, 1\}$. Since $\{x_i\}_{i=1}^n$ are identically and independently sampled from the uniform distribution

on Ω, we have that the two probability densities $\mathbb{P}_k(\vec{x}, \vec{y})$ and $\mathbb{P}_k(\vec{y} \mid \vec{x})$ have the same mathematical expression for any $k \in \{0, 1\}$. Then we may further rewrite the KL divergence between the two marginal distributions $\mathbb{P}_0, \mathbb{P}_1$ as follows:

$$\begin{aligned} KL(\mathbb{P}_0||\mathbb{P}_1) &= \int_\Omega \cdots \int_\Omega \Big(\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \log\Big(\frac{\mathbb{P}_0(\vec{x}, \vec{y})}{\mathbb{P}_1(\vec{x}, \vec{y})}\Big) \mathbb{P}_0(\vec{x}, \vec{y}) dy_1 \cdots dy_n \Big) dx_1 \cdots dx_n \\ &= \int_\Omega \cdots \int_\Omega \Big(\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \log\Big(\frac{\mathbb{P}_0(\vec{y} \mid \vec{x})}{\mathbb{P}_1(\vec{y} \mid \vec{x})}\Big) \mathbb{P}_0(\vec{y} \mid \vec{x}) dy_1 \cdots dy_n \Big) dx_1 \cdots dx_n \\ &= \int_\Omega \cdots \int_\Omega \left(KL\Big(\mathbb{P}_0(\cdot \mid \vec{x}) \| \mathbb{P}_1(\cdot \mid \vec{x})\Big) \right) dx_1 \cdots dx_n. \end{aligned} \tag{D.20}$$

It now remains to upper bound the KL divergence between the two conditional distributions $\mathbb{P}_0(\cdot \mid \vec{x})$ and $\mathbb{P}_1(\cdot \mid \vec{x})$ for any fixed $\vec{x} = (x_1, \cdots, x_n)$. In order to derive such an upper bound, we need to introduce the following notations first. For any n quadrature points $\{x_i\}_{i=1}^n$, we use $\mathcal{J}_n$ to denote the set of all indices j satisfying that Ω_j contains at least one of the points in $\{x_i\}_{i=1}^n$, *i.e,*

$$\mathcal{J}_n := \mathcal{J}_n(\vec{x}) = \Big\{ j : 1 \le j \le m^d \text{ and } \Omega_j \cap \{x_1, \cdots, x_n\} \neq \varnothing \Big\}. \tag{D.21}$$

Moreover, we use $\vec{\omega}_{\mathcal{J}_n}^{(k)}$ to denote $|\mathcal{J}_n|$-dimensional vector formed by the random variables $\{\omega_j^{(k)} : j \in \mathcal{J}_n\}$ and $p_{\mathcal{J}_n}^{(k)}(\cdot)$ to denote the probability density function of $\vec{\omega}_{\mathcal{J}_n}^{(k)}$, where $k \in \{0, 1\}$. From our assumption on the distribution of the weights $\{w_j^{(0)}\}_{j=1}^n$ and $\{w_j^{(1)}\}_{j=1}^n$, we have that for any $\vec{\omega}_{\mathcal{J}_n} \in \{\pm 1\}^{|\mathcal{J}_n|}$,

$$\begin{aligned} p_{\mathcal{J}_n}^{(0)}(\vec{\omega}_{\mathcal{J}_n}) &= \prod_{j \in \mathcal{J}_n} \Big(\frac{1+\kappa}{2}\Big)^{\frac{1}{2}(1-\omega_j)} \Big(\frac{1-\kappa}{2}\Big)^{\frac{1}{2}(1+\omega_j)} \\ p_{\mathcal{J}_n}^{(1)}(\vec{\omega}_{\mathcal{J}_n}) &= \prod_{j \in \mathcal{J}_n} \Big(\frac{1+\kappa}{2}\Big)^{\frac{1}{2}(1+\omega_j)} \Big(\frac{1-\kappa}{2}\Big)^{\frac{1}{2}(1-\omega_j)} \end{aligned} \tag{D.22}$$

Furthermore, for any fixed quadrature points $\vec{x} = (x_1, \cdots, x_n)$ and weights $\vec{\omega}_{\mathcal{J}_n} := \{\omega_j : j \in \mathcal{J}_n\} \subseteq \{\pm 1\}^{|\mathcal{J}_n|}$, we may define the transition kernel $G(\vec{x}, \vec{\omega}_{\mathcal{J}_n})$ as below

$$G(\vec{x}, \vec{\omega}_{\mathcal{J}_n}) := \prod_{j \in \mathcal{J}_n} \Big(\prod_{i : x_i \in \Omega_j} \frac{1}{\sqrt{2\pi} n^{-\gamma}} e^{-\frac{(y_i + \omega_j f_j(x_i))^2}{2n^{-2\gamma}}} \Big) \tag{D.23}$$

Combining the expressions in (D.19),(D.22) and (D.23) allows us to rewrite the two conditional distributions $\mathbb{P}_k(\cdot \mid \vec{x})$ as below:

$$\mathbb{P}_k(\vec{y} \,|\vec{x}) = \mathbb{P}_k(\vec{x}, \vec{y}) = \int_{\{\pm 1\}^{|\mathcal{J}_n|}} G(\vec{x}, \vec{\omega}_{\mathcal{J}_n}) p_{\mathcal{J}_n}^{(k)}(\vec{\omega}_{\mathcal{J}_n}) d\vec{\omega}_{\mathcal{J}_n} \tag{D.24}$$

where $k \in \{0, 1\}$. Applying the data processing inequality (Lemma 4) to (D.24) above then enables us to derive the following upper bound on $KL\Big(\mathbb{P}_0(\cdot \mid \vec{x}) \| \mathbb{P}_1(\cdot \mid \vec{x})\Big)$ for any n fixed quadrature points $\vec{x} = (x_1, \cdots, x_n)$:

$$\begin{aligned} KL\Big(\mathbb{P}_0(\cdot \mid \vec{x}) \| \mathbb{P}_1(\cdot \mid \vec{x})\Big) &\le KL\Big(p_{\mathcal{J}_n}^{(0)} \,\|\, p_{\mathcal{J}_n}^{(1)}\Big) \\ &= |\mathcal{J}_n| \Big(\log\Big(\frac{1+\kappa}{1-\kappa}\Big) \frac{1+\kappa}{2} + \log\Big(\frac{1-\kappa}{1+\kappa}\Big) \frac{1-\kappa}{2} \Big) \\ &\le n\kappa \log\Big(\frac{1+\kappa}{1-\kappa}\Big) \end{aligned} \tag{D.25}$$

where the equality in (D.25) above follows from the fact that $\{w_j^{(0)}\}_{j=1}^{m^d}$ and $\{w_j^{(1)}\}_{j=1}^{m^d}$ are independent and identical copies of $w_{\frac{1+\kappa}{2}}$ and $w_{\frac{1-\kappa}{2}}$ respectively. The last inequality of (D.25) above, however, is deduced from the fact that $m^d = 200n > n$, which implies $|\mathcal{J}_n| \le n$ for any n quadrature points $\{x_i\}_{i=1}^n$. Substituting (D.25) into (D.20) and applying Pinkser's inequality yields the final upper bound on the TV distance between $\mathbb{P}_0$ and $\mathbb{P}_1$:

$$
\begin{aligned}
TV(\mathbb{P}_0\|\mathbb{P}_1) &\le \sqrt{\frac{1}{2}KL(\mathbb{P}_0\|\mathbb{P}_1)} \le \sqrt{\int_\Omega \cdots \int_\Omega \frac{n\kappa}{2}\log\Big(\frac{1+\kappa}{1-\kappa}\Big)dx_1\cdots dx_n} \\
&= \sqrt{\frac{n\kappa}{2}\log\Big(\frac{1+\kappa}{1-\kappa}\Big)} \le \sqrt{\frac{3n}{2}}\kappa = \frac{1}{3}.
\end{aligned} \tag{D.26}
$$

Finally, by substituting (D.16), (D.26), $\Delta = (1-\lambda)\kappa m^d C$ and $\beta_0 = \beta_1 = \exp\Big(-\frac{1}{2}\lambda^2\kappa^2 m^d\Big) = \exp(-\frac{50}{27}) < \frac{1}{6}$ into (B.2) and applying Markov's inequality, we obtain the final lower bound

$$
\begin{aligned}
&\inf_{\hat{H}\in\mathcal{H}_n^f}\sup_{f\in C^s(\Omega)} \mathbb{E}_{\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n}\left[\left|\hat{H}\Big(\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n\Big) - I_f\right|\right] \\
&\ge \Delta \inf_{\hat{H}\in\mathcal{H}_n^f}\sup_{f\in C^s(\Omega)} \mathbb{P}_{\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n}\left[\left|\hat{H}\Big(\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n\Big) - I_f\right| \ge \Delta\right] \\
&\ge (1-\lambda)\kappa m^d C\frac{1 - TV(\mathbb{P}_0||\mathbb{P}_1) - \beta_0 - \beta_1}{2} \ge \frac{1}{2}\frac{\sqrt{2}}{3\sqrt{3n}}\cdot(200n)\cdot\frac{C}{6} \\
&\gtrsim \sqrt{n}C \gtrsim \sqrt{n}(200n)^{-\frac{s+d}{d}}\|K\|_{L^1([-\frac{1}{2},\frac{1}{2}]^d)} \gtrsim n^{-\frac{s}{d}-\frac{1}{2}},
\end{aligned} \tag{D.27}
$$

which is exactly the second term in the RHS of (4.1). Combining the two lower bounds proved in (D.6) and (D.27) concludes our proof of Theorem 4.1

D.3 Proof of Theorem 4.2 (Upper Bound on Integral Estimation)

Before proving the upper bound on integral estimation, we need to derive an upper bound on the expected error of the k-nearest neighbor estimator $\hat{f}_{k\text{-NN}}$, which is built based on the first half of the given dataset $\{(x_i, y_i)\}_{i=1}^n$, with respect to the L^2 norm. From our construction of $\hat{f}_{k\text{-NN}}$ given in Section 4.2, we have that for any fixed $\frac{n}{2}$ quadrature points $\{x_i\}_{i=1}^{\frac{n}{2}}$, $z \in \Omega$ and $k \in \{1, 2, \cdots, \frac{n}{2}\}$, the expected value of $\hat{f}_{k\text{-NN}}(z)$ with respect to the observational noises $\{\epsilon_i\}_{i=1}^{\frac{n}{2}}$ is given by

$$
\mathbb{E}_{\{\epsilon_i\}_{i=1}^{\frac{n}{2}}}\Big[\hat{f}_{k\text{-NN}}(z)\Big] = \frac{1}{k}\sum_{j=1}^k \mathbb{E}_{\{\epsilon_i\}_{i=1}^{\frac{n}{2}}}\Big[f(x_{i_j^{(z)}}) + \epsilon_{i_j^{(z)}}\Big] = \frac{1}{k}\sum_{j=1}^k f(x_{i_j^{(z)}}), \tag{D.28}
$$

where $\{x_{i_j^{(z)}}\}_{j=1}^k$ above are the k nearest neighbors of z among $\{x_i\}_{i=1}^{\frac{n}{2}}$. Now let's consider using the bias-variance decomposition to upper bound the error $\|\hat{f}_{k\text{-NN}}(z) - f(z)\|_{L^2(\Omega)}^2$. Based on the expected value computed in (D.28) above, we may decompose the function $\hat{f}_{k\text{-NN}} - f$ as a sum of the bias part and the variance part as follows:

$$
B(z) := \mathbb{E}_{\{\epsilon_i\}_{i=1}^{\frac{n}{2}}}\Big[\hat{f}_{k\text{-NN}}(z)\Big] - f(z) = \frac{1}{k}\sum_{j=1}^k f(x_{i_j^{(z)}}) - f(z) = \frac{1}{k}\sum_{j=1}^k \Big(f(x_{i_j^{(z)}}) - f(z)\Big), \tag{D.29}
$$

$$
V(z) := \hat{f}_{k\text{-NN}}(z) - \mathbb{E}_{\{\epsilon_i\}_{i=1}^{\frac{n}{2}}}\Big[\hat{f}_{k\text{-NN}}(z)\Big] = \hat{f}_{k\text{-NN}}(z) - \frac{1}{k}\sum_{j=1}^k f(x_{i_j^{(z)}}) = \frac{1}{k}\sum_{j=1}^k \epsilon_{i_j^{(z)}}, \tag{D.30}
$$

where the function B corresponds to the bias part and the function V corresponds to the variance part. Using the decomposition $\hat{f}_{k\text{-NN}} - f = B + V$ allows us to upper bound the expected error of

$\hat{f}_{k\text{-NN}}$ with respect to the L^2 norm as below:

$$\begin{aligned}
&\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}},\{y_i\}_{i=1}^{\frac{n}{2}}}\Big[\|\hat{f}_{k\text{-NN}}-f\|_{L^2(\Omega)}^2\Big]=\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}},\{y_i\}_{i=1}^{\frac{n}{2}}}\Big[\|B+V\|_{L^2(\Omega)}^2\Big]\\
&\leq \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}},\{y_i\}_{i=1}^{\frac{n}{2}}}\left[\Big(\|B\|_{L^2(\Omega)}+\|V\|_{L^2(\Omega)}\Big)^2\right]\\
&\leq \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}},\{y_i\}_{i=1}^{\frac{n}{2}}}\Big[2\|B\|_{L^2(\Omega)}^2+2\|V\|_{L^2(\Omega)}^2\Big]\\
&\lesssim \mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}},\{y_i\}_{i=1}^{\frac{n}{2}}}\Big[\|V\|_{L^2(\Omega)}^2\Big]+\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}},\{y_i\}_{i=1}^{\frac{n}{2}}}\Big[\|B\|_{L^2(\Omega)}^2\Big]\\
&=\mathbb{E}_{z,\{x_i\}_{i=1}^{\frac{n}{2}},\{\epsilon_i\}_{i=1}^{\frac{n}{2}}}\Big[|V(z)|^2\Big]+\mathbb{E}_{z,\{x_i\}_{i=1}^{\frac{n}{2}},\{\epsilon_i\}_{i=1}^{\frac{n}{2}}}\Big[|B(z)|^2\Big],
\end{aligned}\tag{D.31}$$

where z above is uniformly distributed over the domain Ω and independent of x_i for any $1\leq i\leq \frac{n}{2}$. On the one hand, using the expression of the variance part V derived in (D.30) above and the fact that $\{\epsilon_i\}_{i=1}^{\frac{n}{2}}$ are independent and identical distributed noises, we may compute the first term in (D.31) above as follows:

$$\begin{aligned}
\mathbb{E}_{z,\{x_i\}_{i=1}^{\frac{n}{2}},\{\epsilon_i\}_{i=1}^{\frac{n}{2}}}\Big[|V(z)|^2\Big]&=\mathbb{E}_z\left[\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}},\{\epsilon_i\}_{i=1}^{\frac{n}{2}}}\left[\Big|\frac{1}{k}\sum_{j=1}^{k}\epsilon_{i_j^{(z)}}\Big|^2\right]\right]\\
&=\mathbb{E}_z\left[\frac{1}{k^2}\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}},\{\epsilon_i\}_{i=1}^{\frac{n}{2}}}\left[\sum_{j=1}^{k}\epsilon_{i_j^{(z)}}^2\right]\right]\\
&=\mathbb{E}_z\Big[\frac{n^{-2\gamma}k}{k^2}\Big]=\frac{n^{-2\gamma}}{k}.
\end{aligned}\tag{D.32}$$

On the other hand, since $s\in(0,1)$ and the given function f is s-Hölder smooth, we have that the inequality $|f(x)-f(y)|\lesssim\|x-y\|^s$ holds true for any $x,y\in\Omega$. Combining this inequality with the expression of the bias part B derived in (D.30) above helps us upper bound the second term in (D.31) as below:

$$\begin{aligned}
\mathbb{E}_{z,\{x_i\}_{i=1}^{\frac{n}{2}},\{\epsilon_i\}_{i=1}^{\frac{n}{2}}}\Big[|B(z)|^2\Big]&=\mathbb{E}_z\left[\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}},\{\epsilon_i\}_{i=1}^{\frac{n}{2}}}\left[\Big|\frac{1}{k}\sum_{j=1}^{k}\Big(f(x_{i_j^{(z)}})-f(z)\Big)\Big|^2\right]\right]\\
&\leq\frac{1}{k}\mathbb{E}_z\left[\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\sum_{j=1}^{k}\Big|f(x_{i_j^{(z)}})-f(z)\Big|^2\right]\right]\\
&\lesssim\frac{1}{k}\mathbb{E}_z\left[\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\sum_{j=1}^{k}\Big|x_{i_j^{(z)}}-z\Big|^{2s}\right]\right]\\
&\leq\mathbb{E}_{z,\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\Big|x_{i_k^{(z)}}-z\Big|^{2s}\right]\\
&\leq\left(\mathbb{E}_{z,\{x_i\}_{i=1}^{\frac{n}{2}}}\left[\Big|x_{i_k^{(z)}}-z\Big|^2\right]\right)^s\lesssim\Big(\frac{k}{n}\Big)^{\frac{2s}{d}}.
\end{aligned}\tag{D.33}$$

The second least inequality follows from the fact that $\omega(t):=t^s$ is a concave function when $s\in(0,1)$, while the last inequality is obtained by plugging in (D.1) given in Lemma 6. Substituting (D.32) and (D.33) into (D.31) then yields that for any $k\in\{1,2,\cdots,\frac{n}{2}\}$, the expected error of $\hat{f}_{k\text{-NN}}$ with respect to the L^2 norm can be upper bounded as follows:

$$\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}},\{y_i\}_{i=1}^{\frac{n}{2}}}\Big[\|\hat{f}_{k\text{-NN}}-f\|_{L^2(\Omega)}^2\Big]\lesssim\frac{n^{-2\gamma}}{k}+\Big(\frac{k}{n}\Big)^{\frac{2s}{d}}.\tag{D.34}$$

Furthermore, from our construction of the integral estimator $\hat{H}_{k\text{-NN}}$ given in Section 4.2, we may upper bound the expectation of the estimator $\hat{H}_{k\text{-NN}}$'s squared error via the expected error of $\hat{f}_{k\text{-NN}}$

with respect to the L^2 norm as below:

$$
\begin{aligned}
&\mathbb{E}_{\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n}\left[\left|\hat{H}_{k\text{-NN}}\Big(\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n\Big)-I_f\right|^2\right]\\
&=\mathbb{E}_{\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n}\left[\left|\int_\Omega \hat{f}_{k\text{-NN}}(x)dx+\frac{2}{n}\sum_{i=\frac{n}{2}+1}^{n}\Big(y_i-\hat{f}_{k\text{-NN}}(x_i)\Big)-\int_\Omega f(x)dx\right|^2\right]\\
&\lesssim \mathbb{E}_{\substack{\{x_i\}_{i=1}^{\frac{n}{2}},\\ \{y_i\}_{i=1}^{\frac{n}{2}}}}\left[\mathbb{E}_{\substack{\{x_i\}_{i=\frac{n}{2}+1}^{n},\\ \{y_i\}_{i=\frac{n}{2}+1}^{n}}}\left[\left|\frac{1}{\frac{n}{2}}\sum_{i=\frac{n}{2}+1}^{n}\Big(f(x_i)-\hat{f}_{k\text{-NN}}(x_i)-\int_\Omega(f(x)-\hat{f}_{k\text{-NN}}(x))dx\Big)\right|^2\right]\right]\\
&+\mathbb{E}_{\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n}\left[\left|\frac{2}{n}\sum_{i=\frac{n}{2}+1}^{n}\epsilon_i\right|^2\right]\\
&=\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}},\{y_i\}_{i=1}^{\frac{n}{2}}}\left[\frac{4}{n^2}\sum_{i=\frac{n}{2}+1}^{n}\mathbb{E}_{x_i}\left[\left|\Big(f(x_i)-\hat{f}_{k\text{-NN}}(x_i)-\int_\Omega(f(x)-\hat{f}_{k\text{-NN}}(x))dx\Big)\right|^2\right]\right]\\
&+\frac{4}{n^2}\sum_{i=\frac{n}{2}+1}^{n}\mathbb{E}_{x_i,y_i}\left[\epsilon_i^2\right]\lesssim\frac{1}{n}\left(\mathbb{E}_{\{x_i\}_{i=1}^{\frac{n}{2}},\{y_i\}_{i=1}^{\frac{n}{2}}}\left[\|\hat{f}_{k\text{-NN}}-f\|_{L^2(\Omega)}^2\right]+n^{-2\gamma}\right)\\
&\lesssim\frac{1}{n}\Big(\frac{n^{-2\gamma}}{k}+\Big(\frac{k}{n}\Big)^{\frac{2s}{d}}\Big)+n^{-2\gamma-1}.
\end{aligned}
\tag{D.35}
$$

Based on the magnitude of the noises, we have the following two cases for the final upper bound: When $\gamma\in[0,\frac{s}{d})$, the optimal k is determined by balancing the two terms $\frac{n^{-2\gamma}}{k}$ and $\left(\frac{k}{n}\right)^{\frac{2s}{d}}$ in (D.35), which yields $\frac{n^{-2\gamma}}{k}=\left(\frac{k}{n}\right)^{\frac{2s}{d}}\Rightarrow k=\Theta(n^{\frac{2(s-\gamma d)}{d+2s}})$. The corresponding upper bound is given by

$$
\begin{aligned}
\frac{1}{n}\Big(\frac{n^{-2\gamma}}{k}+\Big(\frac{k}{n}\Big)^{\frac{2s}{d}}\Big)+n^{-2\gamma-1}&\lesssim\frac{1}{n}n^{-2\gamma-\frac{2(s-\gamma d)}{d+2s}}+n^{-1-2\gamma}=n^{-\frac{2s(1+2\gamma)}{2s+d}-1}+n^{-2\gamma-1}\\
&\lesssim\max\{n^{-\frac{2s(1+2\gamma)}{2s+d}-1},n^{-2\gamma-1}\}=n^{-2\gamma-1}.
\end{aligned}
\tag{D.36}
$$

When $\gamma\in[\frac{s}{d},\infty]$, we note that $k\in\{1,2,\cdots,\frac{n}{2}\}$ must be of at least constant level. Therefore, the optimal k is determined by balancing the two terms $\frac{n^{-2\gamma-1}}{k}$ and $n^{-2\gamma-1}$, which yields that $k=\Theta(1)$ is of constant level. The corresponding upper bound is given by

$$
\begin{aligned}
\frac{1}{n}\Big(\frac{n^{-2\gamma}}{k}+\Big(\frac{k}{n}\Big)^{\frac{2s}{d}}\Big)+n^{-2\gamma-1}&\lesssim n^{-\frac{2s}{d}-1}+n^{-2\gamma-1}\\
&\lesssim\max\{n^{-\frac{2s}{d}-1},n^{-2\gamma-1}\}=n^{-\frac{2s}{d}-1}.
\end{aligned}
\tag{D.37}
$$

Finally, substituting (D.36) and (D.37) into (D.35) gives us the final upper bound:

$$
\begin{aligned}
&\mathbb{E}_{\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n}\left[\left|\hat{H}_{k\text{-NN}}\Big(\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n\Big)-I_f\right|\right]\\
&\le\sqrt{\mathbb{E}_{\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n}\left[\left|\hat{H}_{k\text{-NN}}\Big(\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n\Big)-I_f\right|^2\right]}\\
&\lesssim\sqrt{\frac{1}{n}\Big(\frac{n^{-2\gamma}}{k}+\Big(\frac{k}{n}\Big)^{\frac{2s}{d}}\Big)+n^{-2\gamma-1}}\lesssim\sqrt{\max\{n^{-\frac{2s}{d}-1},n^{-2\gamma-1}\}}\\
&=n^{\max\{-\frac{1}{2}-\gamma,-\frac{1}{2}-\frac{s}{d}\}},
\end{aligned}
\tag{D.38}
$$

which concludes our proof of Theorem 4.2.

E Construction of the Oracle in Assumption 3.1

For the cases when $s > \frac{d}{p}$, the optimal function estimator for any $r \in [1, \infty)$ is already constructed in a series of earlier work [49, 50], whose convergence rate has aleardy been proved to be $(n^{-\frac{s}{d}+(\frac{1}{p}-\frac{1}{r})_+})^r$. Hence, here we only need to focus on the cases when $s \in (\frac{2dq-dp}{2pq}, \frac{d}{p})$. Our goal is construct a desired oracle such that for any r satisfying $\frac{1}{r} \in (\frac{d-sp}{pd}, \frac{1}{p}]$, the convergence rate of function estimation is given by $(n^{-\frac{s}{d}+\frac{1}{p}-\frac{1}{r}})^r$ up to logarithm factors. We need to introduce a few notations and key lemmas in the following two subsections beforehand.

E.1 Preliminaries and Notations

We introduce a few notations used in our proofs in this subsection. For any compact region $R \subset \mathbb{R}^d$, the diameter R is defined as $\text{diam}(R) := \sup_{x,y\in R} \|x - y\|$. Moreover, for any two compact regions $R_1, R_2 \subset \mathbb{R}^d$, the distance $\text{dist}(R_1, R_2)$ between them is defined as $\text{dist}(R_1, R_2) := \|c_{R_1} - c_{R_2}\|_\infty$, where $\|\cdot\|_\infty$ denotes the l_∞ norm in $\mathbb{R}^d$ and c_{R_1}, c_{R_2} are the centroids of R_1, R_2 respectively. For any collection of n data points $P := \{x_1, x_2, \cdots, x_n\}$, we use $\rho(P, \Omega)$ to denote the covering radius of P in $\Omega = [0, 1]^d$, which is defined as follows:

$$\rho(P, \Omega) := \sup_{y\in\Omega} \inf_{1\le i\le n} \|y - x_i\| \tag{E.1}$$

E.2 Key Lemmas

In this subsection, we first list corollaries of some well-known theorems as lemmas used in our proofs.

Lemma 7 (Bramble-Hilbert Lemma) *For any $s \in \mathbb{N}$ and $u \in W^{s,p}(\Omega)$, there exists some polynomial π with $\deg(\pi) \le s - 1$, such that for any $k \in \mathbb{N}$ and $0 \le k \le s$, the following inequality holds:*

$$\|u - \pi\|_{W^{k,p}(\Omega)} \lesssim \textit{diam}(\Omega)^{s-k} \|u\|_{W^{s,p}(\Omega)} \tag{E.2}$$

Lemma 8 (Gagliardo–Nirenberg interpolation inequality) *Fix $s \in \mathbb{N}, p \in [1, \infty)$ and $r \in [1, \infty)$ such that $\frac{1}{r} \in (\frac{d-sp}{pd}, \frac{1}{p}]$. Let $\theta = \frac{d}{s}(\frac{1}{p} - \frac{1}{r}) \in (0, 1)$ such that the relation $\frac{1}{r} = \theta(\frac{1}{p} - \frac{s}{d}) + (1 - \theta)\frac{1}{p}$, holds. Then we have the following inequality:*

$$\|u\|_{L^r(\Omega)} \lesssim \|u\|^\theta_{W^{s,p}(\Omega)} \|u\|^{1-\theta}_{L^p(\Omega)} \tag{E.3}$$

Now let's proceed to list some other results developed in earlier works [66, 67, 70] as lemmas used for constructing the oracle here.

Lemma 9 (Bound on the covering radius (Theorem 2.1 in [70])) *Given $P := \{x_1, \cdots, x_n\}$ sampled independently and identically from the uniform distribution on $\Omega = [0, 1]^d$, we have that there exist constants $c_1, c_2 > 0$ and $\alpha_0 > 0$, which are all independent of n, such that the following inequality holds for any $\alpha > \alpha_0$:*

$$\mathbb{P}\left(\rho(P, \Omega) \ge c_1\Big(\frac{\alpha \log n}{n}\Big)^{\frac{1}{d}}\right) \lesssim n^{1-c_2\alpha} \tag{E.4}$$

Lemma 10 (Properties of the moving least squares estimator (Theorem 4.7 in [67])) *For any given collection of n points $P = \{x_1, \cdots, x_n\}$ with covering radius $\rho(P, \Omega)$, there exist constants a_1, a_2 independent of n and continuous functions $u_{x_i} : \Omega \to \mathbb{R}$ $(1 \le i \le n)$, such that*

- $\pi(y) = \sum_{i=1}^n \pi(x_i)u_{x_i}(y)$ *for any $y \in \Omega$ and any polynomial π with $\deg(\pi) \le s - 1$*
- $\sum_{i=1}^n |u_{x_i}(y)| \le a_1$ *for any $y \in \Omega$*
- $u_x(y) = 0$ *for any $y \in \Omega$ and $x \in P$ with $\|x - y\| \ge a_2\rho(P, \Omega)$*

Based on the functions u_{x_i} $(1 \leq i \leq n)$ given in Lemma 10, one may define a function estimator $K_n = K_n(\{x_i\}_{i=1}^n, \{f(x_i)\}_{i=1}^n)$ of f as $K_n(\{x_i\}_{i=1}^n, \{f(x_i)\}_{i=1}^n) := \sum_{i=1}^n f(x_i)u_{x_i}$. Such an estimator is obtained via the moving least square approximation, which was first proposed in [66]. One may refer to [67] for more detail about it.

In addition, we also need upper bounds on the moments of any binomial random variable, which is given as the lemma below.

Lemma 11 (Bound on the moment of binomial random variable ([71, 72])) *Let $Z \sim Bin(m, p)$ be a binomial random variable binomial distribution with parameters m and p. Then for any $k \in \mathbb{N}$, the k-th moment $\mathbb{E}[Z^k]$ of Z can be upper bounded as below:*

$$\mathbb{E}[Z^k] \leq \Big(c' \frac{k}{\log(1 + \frac{k}{mp})}\Big)^k \tag{E.5}$$

where $c' > 1$ above is some universal constant.

E.3 Construction of the oracle and proof of its convergence rate

Finally, we will explain how the desired oracle is constructed and present a complete proof of its convergence rate in this subsection. We remark that our proof is similar to the one presented in section 2.1 of [49].

Let $c_1, c_2, \alpha_0, a_1, a_2$ be the positive constants specified in Lemma 9 and Lemma 10 above. We pick $\alpha > 0$ to be sufficiently large such that $\alpha > \max\{\alpha_0, \frac{1}{c_2}(2 + \frac{sr}{d} - \frac{r}{p})\}$, which implies that $1 - c_2\alpha < -\frac{sr}{d} + \frac{r}{p} - 1$. Now we pick $k = \min\Big\{\frac{\sqrt{d}}{c_1}(\frac{n}{\alpha \log n})^{\frac{1}{d}}, (\frac{n}{2\log n})^{\frac{1}{d}}\Big\}$ and divide $\Omega = [0,1]^d$ into small cubes $C_1, C_2, \cdots, C_{k^d}$, each of which has side length $k^{-1} \geq 2c_1(\frac{\alpha \log n}{n})^{\frac{1}{d}}$. Furthermore, since the observed data points $\mathcal{X}_n := \{x_i\}_{i=1}^n$ are i.i.d samples from the uniform distribution on Ω, we may define A to be the following event:

$$A := \Big\{\mathcal{X}_n : |\mathcal{X}_n \cap C_l| \geq 1, \ \forall\, 1 \leq l \leq k^d\Big\} \tag{E.6}$$

For any $1 \leq l \leq k^d$, we use B_l to denote the event $\{\mathcal{X}_n : |\mathcal{X}_n \cap C_l| \geq 1\}$. Then we have $A = \cap_{l=1}^{k^d} B_l$ by definition. Based on A defined above, we may describe our choice of the oracle $K_n = K_n(\{x_i\}_{i=1}^n, \{f(x_i)\}_{i=1}^n)$ as follows: When A is false, we simply pick $K_n(\{x_i\}_{i=1}^n, \{f(x_i)\}_{i=1}^n)$ to be the zero function. When A is true, we pick $K_n(\{x_i\}_{i=1}^n, \{f(x_i)\}_{i=1}^n) := \sum_{i=1}^n f(x_i)u_{x_i}$ to be the moving least square estimator specified in Lemma 10 above.

Before proving that the estimator given by the oracle above satisfies the desired upper bound, let's firstly derive lower bound the probability $\mathbb{P}(A)$ at first. We may apply the assumption $k^d \leq \frac{n}{2\log n}$ above and union bound to deduce that

$$\begin{aligned}
&\mathbb{P}(A) = \mathbb{P}(\cap_{l=1}^{k^d} B_l) = 1 - \mathbb{P}(\cup_{l=1}^{k^d} B_l^c) \geq 1 - \sum_{l=1}^{k^d} \mathbb{P}(B_l^c) = 1 - k^d(1 - \frac{1}{k^d})^n \\
&= 1 - k^d\Big((1 - \frac{1}{k^d})^{k^d}\Big)^{\frac{n}{k^d}} \geq 1 - k^d e^{-\frac{n}{k^d}} \geq 1 - \frac{n}{2\log n} e^{-2\log n} = 1 - \frac{1}{2n\log n} \geq \frac{1}{2}
\end{aligned} \tag{E.7}$$

Moreover, we also need to derive an upper bound on the probability $\mathbb{P}(A^c)$ via Lemma 9. We use E to denote the event that $\Big\{\mathcal{X}_n : \rho(\mathcal{X}_n, \Omega) > \frac{\sqrt{d}}{k}\Big\}$. Below we will firstly justify that E implies A^c, which indicates that $\mathbb{P}(A^c) \leq \mathbb{P}(E)$. For the sake of contradiction, assume that A and E both hold true. Then for any point $y \in \Omega$, there must exist some $x_{i_y} \in \mathcal{X}_n$ and some $l_y \in \{1, 2, \cdots, k^d\}$, such that $\{y, x_{i_y}\} \in C_{l_y}$. This implies that $\inf_{1 \leq i \leq n} \|y - x_i\| \leq \|y - x_{i_y}\| \leq \text{diam}(C_{l_y}) \leq \frac{\sqrt{d}}{k}$. Taking supremum with respect to all y then implies $\rho(\mathcal{X}_n, \Omega) \leq \frac{\sqrt{d}}{k}$, which contradicts the definition of E. Therefore we must have $\mathbb{P}(A^c) \leq \mathbb{P}(E)$. Applying our assumption $k \leq \frac{\sqrt{d}}{c_1}(\frac{n}{\alpha \log n})^{\frac{1}{d}}$ and Lemma 9 above further implies that

$$\mathbb{P}(A^c) \leq \mathbb{P}(E) = \mathbb{P}\Big(\rho(\mathcal{X}_n, \Omega) > \frac{\sqrt{d}}{k}\Big) \leq \mathbb{P}\bigg(\rho(\mathcal{X}_n, \Omega) \geq c_1\Big(\frac{\alpha \log n}{n}\Big)^{\frac{1}{d}}\bigg) \lesssim n^{1-c_2\alpha} \tag{E.8}$$

Using the law of total expectation and the two upper bounds derived in E.7 and E.8 above, we may obtain the following upper bound on the error of the function estimator K_n as follows:

$$\begin{aligned}
&\mathbb{E}_{\{x_i\}_{i=1}^n}\Big[\|K_n(\{x_i\}_{i=1}^n, \{f(x_i)\}_{i=1}^n) - f\|_{L^r(\Omega)}^r\Big] \\
&= \mathbb{E}_{\{x_i\}_{i=1}^n}\left[\Big\|\sum_{i=1}^n f(x_i)u_{x_i} - f\Big\|_{L^r(\Omega)}^r \Big| A\right]\mathbb{P}(A) + \mathbb{E}_{\{x_i\}_{i=1}^n}\Big[\|0 - f\|_{L^r(\Omega)}^r \Big| A^c\Big]\mathbb{P}(A^c) \\
&\lesssim \mathbb{E}_{\{x_i\}_{i=1}^n}\left[\Big\|\sum_{i=1}^n f(x_i)u_{x_i} - f\Big\|_{L^r(\Omega)}^r \Big| A\right]\mathbb{P}(A) + \|f\|_{L^r(\Omega)}^r n^{1-c_2\alpha} \\
&\lesssim \mathbb{E}_{\{x_i\}_{i=1}^n}\left[\Big\|\sum_{i=1}^n f(x_i)u_{x_i} - f\Big\|_{L^r(\Omega)}^r \Big| A\right] + n^{-\frac{sr}{d}+\frac{r}{p}-1} \\
&\leq \frac{\mathbb{E}_{\{x_i\}_{i=1}^n}\Big[\|\sum_{i=1}^n f(x_i)u_{x_i} - f\|_{L^r(\Omega)}^r\Big]}{\mathbb{P}(A)} + n^{-\frac{sr}{d}+\frac{r}{p}-1} \\
&\leq 2\mathbb{E}_{\{x_i\}_{i=1}^n}\left[\Big\|\sum_{i=1}^n f(x_i)u_{x_i} - f\Big\|_{L^r(\Omega)}^r\right] + n^{-\frac{sr}{d}+\frac{r}{p}-1} \\
&\lesssim \mathbb{E}_{\{x_i\}_{i=1}^n}\left[\Big\|\sum_{i=1}^n f(x_i)u_{x_i} - f\Big\|_{L^r(\Omega)}^r\right] + n^{-\frac{sr}{d}+\frac{r}{p}-1}
\end{aligned} \tag{E.9}$$

where the second inequality above follows from our assumption $\alpha > \max\{\alpha_0, \frac{1}{c_2}(2 + \frac{sr}{d} - \frac{r}{p})\}$ specified earlier. From the last expression above, we can see that now it suffices to show that the first term $\mathbb{E}_{\{x_i\}_{i=1}^n}\Big[\|\sum_{i=1}^n f(x_i)u_{x_i} - f\|_{L^r(\Omega)}^r\Big]$ in the last expression above is no larger than $n^{-\frac{sr}{d}+\frac{r}{p}-1}$ up to constants and logarithm factors.

Moreover, for any $1 \leq l \leq k^d$, we define $\mathcal{N}_l$ to be a collection of "neighbors" of the cube C_l as below:

$$\mathcal{N}_l := \Big\{C_j : \text{dist}(C_j, C_l) \leq \max\{a_2, 1\}\frac{\sqrt{d}}{c_1}\big(\frac{n}{\alpha \log n}\big)^{\frac{1}{d}}\Big\} \tag{E.10}$$

Correspondingly, the union of all the cubes in $\mathcal{N}_l$ is defined as $M_l := \cup_{C_j \in \mathcal{N}_l} C_j$. Since the distance between any two cubes is measured by the l_∞ distance between their centroids, we may deduce that each M_l remains to be a cube for any $1 \leq l \leq k^d$. Also, the number of cubes in each $\mathcal{N}_l$ can be upper bounded by some constant C_* independent of n, which implies that the diameter of each M_l satisfies $\text{diam}(M_l) \lesssim k^{-s}$. Then for any $1 \leq l \leq k^d$, we may apply Lemma 7 above to deduce that there exists some polynomial π_l such that $\deg(\pi_l) \leq s - 1$ and the following two inequalities hold:

$$\begin{aligned}
\|f - \pi_l\|_{W^{s,p}(M_l)} &\lesssim \|f\|_{W^{s,p}(M_l)} \\
\|f - \pi_l\|_{L^p(M_l)} = \|f - \pi_l\|_{W^{0,p}(M_l)} &\lesssim \text{diam}(M_l)^s \|f\|_{W^{s,p}(M_l)} \lesssim k^{-s}\|f\|_{W^{s,p}(M_l)}
\end{aligned} \tag{E.11}$$

Combining the two inequalities above with Gagliardo–Nirenberg interpolation inequality E.3 listed in Lemma 8 further implies that for any $1 \leq l \leq k^d$,

$$\|f - \pi_l\|_{L^r(M_l)}^r \lesssim \Big(\|f - \pi_l\|_{W^{s,p}(M_l)}^\theta \|f - \pi_l\|_{L^p(M_l)}^{1-\theta}\Big)^r \lesssim k^{-sr(1-\theta)}\|f\|_{W^{s,p}(M_l)}^r \tag{E.12}$$

Using the first and third property of the moving least square estimator listed in Lemma 10 above, we may deduce that the following inequality holds for any $1 \leq l \leq k^d$ and any $y \in C_l$:

$$
\begin{aligned}
\Big|f(y)-\sum_{i=1}^{n} f(x_i)u_{x_i}(y)\Big|^r &= \Big|f(y)-\pi_l(y)-\Big(\sum_{i=1}^{n} f(x_i)u_{x_i}(y)-\sum_{i=1}^{n}\pi_l(x_i)u_{x_i}(y)\Big)\Big|^r \\
&= \Big|f(y)-\pi_l(y)-\sum_{i:x_i\in M_l}(f(x_i)-\pi_l(x_i))u_{x_i}(y)\Big|^r \\
&\leq \left(\Big|(f-\pi_l)(y)\Big|+\sum_{i:x_i\in M_l}\Big|(f-\pi_l)(x_i)u_{x_i}(y)\Big|\right)^r \\
&\leq \Big(1+|\mathcal{X}_n\cap M_l|\Big)^{r-1}\left(\Big|(f-\pi_l)(y)\Big|^r+\sum_{i:x_i\in M_l}\Big|(f-\pi_l)(x_i)u_{x_i}(y)\Big|^r\right) \\
&\lesssim \max\Big\{1,|\mathcal{X}_n\cap M_l|^{r-1}\Big\}\cdot\max\{1,a_1^r\}\cdot\left(\Big|(f-\pi_l)(y)\Big|^r+\sum_{i:x_i\in M_l}\Big|(f-\pi_l)(x_i)\Big|^r\right) \\
&\lesssim \max\Big\{1,|\mathcal{X}_n\cap M_l|^{r-1}\Big\}\left(\Big|(f-\pi_l)(y)\Big|^r+\sum_{i:x_i\in M_l}\Big|(f-\pi_l)(x_i)\Big|^r\right)
\end{aligned}
\tag{E.13}
$$

By rewriting $\|\sum_{i=1}^n f(x_i)u_{x_i}-f\|_{L^r(\Omega)}^r$ as an integral and plugging in the inequality derived above, we then obtain that

$$
\begin{aligned}
&\mathbb{E}_{\{x_i\}_{i=1}^n}\left[\Big\|\sum_{i=1}^{n} f(x_i)u_{x_i}-f\Big\|_{L^r(\Omega)}^r\right]=\mathbb{E}_{\{x_i\}_{i=1}^n}\left[\sum_{l=1}^{k^d}\int_{C_l}\Big|\sum_{i=1}^{n} f(x_i)u_{x_i}(y)-f(y)\Big|^r dy\right] \\
&\lesssim \mathbb{E}_{\{x_i\}_{i=1}^n}\left[\sum_{l=1}^{k^d}\int_{C_l}\max\Big\{1,|\mathcal{X}_n\cap M_l|^{r-1}\Big\}\Big(\Big|(f-\pi_l)(y)\Big|^r+\sum_{i:x_i\in M_l}\Big|(f-\pi_l)(x_i)\Big|^r\Big)dy\right] \\
&\lesssim \sum_{l=1}^{k^d}\left(\int_{C_l}\Big|(f-\pi_l)(y)\Big|^r dy\right)\mathbb{E}_{\{x_i\}_{i=1}^n}\Big[\max\Big\{1,|\mathcal{X}_n\cap M_l|^{r-1}\Big\}\Big] \\
&+\sum_{l=1}^{k^d}\mathbb{E}_{\{x_i\}_{i=1}^n}\left[\max\Big\{1,|\mathcal{X}_n\cap M_l|^{r-1}\Big\}\Big(\sum_{i:x_i\in M_l}\Big|(f-\pi_l)(x_i)\Big|^r\Big)k^{-d}\right] \\
&\leq \sum_{l=1}^{k^d}\|f-\pi_l\|_{L^r(M_l)}^r\mathbb{E}_{\{x_i\}_{i=1}^n}\Big[1+|\mathcal{X}_n\cap M_l|^{r-1}\Big] \\
&+k^{-d}\sum_{l=1}^{k^d}\mathbb{E}_{\{x_i\}_{i=1}^n}\left[\max\Big\{1,|\mathcal{X}_n\cap M_l|^{r-1}\Big\}\Big(\sum_{i:x_i\in M_l}\Big|(f-\pi_l)(x_i)\Big|^r\Big)\right]
\end{aligned}
\tag{E.14}
$$

Before deriving upper bounds on the final expression in E.14, let's consider bounding the two expectations $\mathbb{E}_{\{x_i\}_{i=1}^n}\Big[|\mathcal{X}_n\cap M_l|^{r-1}\Big]$ and $\mathbb{E}_{\{x_i\}_{i=1}^n}\Big[|\mathcal{X}_n\cap M_l|^{r}\Big]$ $(1\leq l\leq k^d)$ at first. Given that the datapoints $\mathcal{X}_n=\{x_i\}_{i=1}^n$ are i.i.d samples from the uniform distribution on Ω, we have that each $|\mathcal{X}_n\cap M_l|\sim \text{Bin}(k^d,p_l)$ is a binomial random variable, where $|\mathcal{N}_l|\leq c_*\Rightarrow p_l\leq\frac{c_*}{k^d}$ for any $1\leq l\leq k^d$. Applying Lemma 11 above then yields that for any $1\leq l\leq k^d$:

$$
\begin{aligned}
\mathbb{E}_{\{x_i\}_{i=1}^n}\Big[|\mathcal{X}_n\cap M_l|^{r-1}\Big]&\leq\Big(c'\frac{r-1}{\log(1+\frac{r-1}{k^dp_l})}\Big)^{r-1}\leq\Big(c'\frac{r-1}{\log(1+\frac{r-1}{c_*})}\Big)^{r-1}, \\
\mathbb{E}_{\{x_i\}_{i=1}^n}\Big[|\mathcal{X}_n\cap M_l|^{r}\Big]&\leq\Big(c'\frac{r}{\log(1+\frac{r}{k^dp_l})}\Big)^{r}\leq\Big(c'\frac{r}{\log(1+\frac{r}{c_*})}\Big)^{r}
\end{aligned}
\tag{E.15}
$$

where the two upper bounds in E.15 are all constants independent of the sample size n.

Secondly, we need to derive an upper bound on the summation $\sum_{l=1}^{k^d} \|f-\pi_l\|^r_{L^r(M_l)}$. Recall our definition of the cubes M_l ($1 \le l \le k^d$), we can deduce that each small cube C_i is contained within at most constantly many big cubes M_l. That is to say, there exists some constant c'_* independent of sample size n, such that $|l : C_i \subset M_l| \le c'_*$ for any $1 \le i \le k^d$. By combining this fact with inequality E.12 proved above, we can deduce that

$$\begin{aligned}
&\sum_{l=1}^{k^d} \|f-\pi_l\|^r_{L^r(M_l)} \lesssim \sum_{l=1}^{k^d} k^{-sr(1-\theta)} \|f\|^r_{W^{s,p}(M_l)} = k^{-sr(1-\theta)} \sum_{l=1}^{k^d} \sum_{i:C_i \in M_l} \|f\|^r_{W^{s,p}(C_i)} \\
&= k^{-sr(1-\theta)} \sum_{i=1}^{k^d} \sum_{l:C_i \in M_l} \|f\|^r_{W^{s,p}(C_i)} \le k^{-sr(1-\theta)} \sum_{i=1}^{k^d} c'_* \|f\|^r_{W^{s,p}(C_i)} \\
&\lesssim k^{-sr(1-\theta)} \sum_{i=1}^{k^d} \|f\|^r_{W^{s,p}(C_i)} \le k^{-sr(1-\theta)} \Big(\sum_{i=1}^{k^d} \|f\|^p_{W^{s,p}(C_i)} \Big)^{\frac{r}{p}} = k^{-sr(1-\theta)} \|f\|^r_{W^{s,p}(\Omega)}
\end{aligned} \tag{E.16}$$

Furthermore, to derive the final upper bound, let's simplify the second term in E.14 and try to upper bound it. Using the law of total expectation again, we have that

$$\begin{aligned}
&k^{-d} \sum_{l=1}^{k^d} \mathbb{E}_{\{x_i\}_{i=1}^n} \Bigg[\max\Big\{1, |\mathcal{X}_n \cap M_l|^{r-1}\Big\} \Big(\sum_{i:x_i \in M_l} \Big|(f-\pi_l)(x_i)\Big|^r \Big) \Bigg] \\
&= k^{-d} \sum_{l=1}^{k^d} \sum_{l_1,\cdots,l_n=1}^{k^d} \mathbb{E}_{\{x_i\}_{i=1}^n} \Bigg[\max\Big\{1, |\mathcal{X}_n \cap M_l|^{r-1}\Big\} \cdot \\
&\Big(\sum_{i:x_i \in M_l} \Big|(f-\pi_l)(x_i)\Big|^r \Big) \Bigg| x_1 \in C_{l_1}, \cdots, x_n \in C_{l_n} \Bigg] \cdot \mathbb{P}(x_1 \in C_{l_1}, \cdots, x_n \in C_{l_n}) \\
&= k^{-d} \sum_{l=1}^{k^d} \sum_{l_1,\cdots,l_n=1}^{k^d} \max\Big\{1, |\mathcal{X}_n \cap M_l|^{r-1}\Big\} \Big(\sum_{i:x_i \in M_l} \int_{C_{l_i}} \Big|(f-\pi_l)(x_i)\Big|^r dx_i \Big) \\
&\cdot \mathbb{P}(x_1 \in C_{l_1}, \cdots, x_n \in C_{l_n}) \\
&= k^{-d} \sum_{l=1}^{k^d} \sum_{l_1,\cdots,l_n=1}^{k^d} \max\Big\{1, |\mathcal{X}_n \cap M_l|^{r-1}\Big\} \Big(\sum_{i:x_i \in M_l} \|f-\pi_l\|^r_{L^r(C_{l_i})} \Big) \\
&\cdot \mathbb{P}(x_1 \in C_{l_1}, \cdots, x_n \in C_{l_n}) \\
&\le k^{-d} \sum_{l=1}^{k^d} \sum_{l_1,\cdots,l_n=1}^{k^d} \max\Big\{1, |\mathcal{X}_n \cap M_l|^{r-1}\Big\} \Big(|\mathcal{X}_n \cap M_l| \sum_{C_j \in N_l} \|f-\pi_l\|^r_{L^r(C_j)} \Big) \\
&\cdot \mathbb{P}(x_1 \in C_{l_1}, \cdots, x_n \in C_{l_n}) \\
&\le k^{-d} \sum_{l=1}^{k^d} \Bigg(\sum_{l_1,\cdots,l_n=1}^{k^d} \max\Big\{1, |\mathcal{X}_n \cap M_l|^{r}\Big\} \mathbb{P}(x_1 \in C_{l_1}, \cdots, x_n \in C_{l_n}) \Bigg) \|f-\pi_l\|^r_{L^r(M_l)} \\
&= k^{-d} \sum_{l=1}^{k^d} \mathbb{E}_{\{x_i\}_{i=1}^n} \Big[\max\Big\{1, |\mathcal{X}_n \cap M_l|^{r}\Big\} \Big] \cdot \|f-\pi_l\|^r_{L^r(M_l)} \\
&\le k^{-d} \sum_{l=1}^{k^d} \mathbb{E}_{\{x_i\}_{i=1}^n} \Big[1 + |\mathcal{X}_n \cap M_l|^{r} \Big] \cdot \|f-\pi_l\|^r_{L^r(M_l)}
\end{aligned} \tag{E.17}$$

Finally, by substituting the bounds derived in E.15, E.16, E.17 into E.14, we may further plug in our choice of k and ignore the logarithm terms to obtain the desired upper bound:

$$\begin{aligned}\mathbb{E}_{\{x_i\}_{i=1}^n}\left[\Big\|\sum_{i=1}^n f(x_i)u_{x_i} - f\Big\|_{L^r(\Omega)}^r\right] &\lesssim (1+k^{-d})\sum_{l=1}^{k^d}\|f-\pi_l\|_{L^r(M_l)}^r \\ &\lesssim \sum_{l=1}^{k^d}\|f-\pi_l\|_{L^r(M_l)}^r \lesssim k^{-sr(1-\theta)}\|f\|_{W^{s,p}(\Omega)}^r \\ &\lesssim \left((\frac{n}{\log n})^{\frac{1}{d}}\right)^{-sr(1-\frac{d}{s}(\frac{1}{p}-\frac{1}{r}))} \lesssim n^{-\frac{sr}{d}+\frac{r}{p}-1}\end{aligned} \tag{E.18}$$

where we ignore the logarithm term in the last step above. This concludes our proof of the convergence rate of the oracle we specified above. Substituting E.18 into E.9 completes our proof of Assumption 3.1.