SPEEDING UP MONTE CARLO INTEGRATION: CONTROL NEIGHBORS FOR OPTIMAL CONVERGENCE

Rémi Leluc, François Portier, Johan Segers, Aigerim Zhuman

REPRINT | 2025 / 02

UCLouvain

ISBA | LIDAM Louvain Institute of Data Analysis and Modeling in economics and statistics

ISBA

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/isba/publication

arXiv:2305.06151v3 [math.NA] 4 Apr 2024

Speeding up Monte Carlo Integration: Control Neighbors for Optimal Convergence

Rémi Leluc
CMAP, École Polytechnique,
Institut Polytechnique de Paris, France
remi.leluc@gmail.com

François Portier
CREST
ENSAI, France
francois.portier@gmail.com

Aigerim Zhuman
LIDAM, ISBA
UCLouvain, Belgium
aigerim.zhuman@uclouvain.be

Johan Segers
LIDAM, ISBA
UCLouvain, Belgium
johan.segers@uclouvain.be

Abstract

A novel linear integration rule called *control neighbors* is proposed in which nearest neighbor estimates act as control variates to speed up the convergence rate of the Monte Carlo procedure on metric spaces. The main result is the $\mathcal{O}(n^{-1/2}n^{-s/d})$ convergence rate – where n stands for the number of evaluations of the integrand and d for the dimension of the domain – of this estimate for Hölder functions with regularity $s \in (0, 1]$, a rate which, in some sense, is optimal. Several numerical experiments validate the complexity bound and highlight the good performance of the proposed estimator.

1 Introduction

Consider the classical numerical integration problem of approximating the value of an integral $\mu(\varphi) = \int \varphi \, \mathrm{d}\mu$ where μ is a probability measure on a metric space (M, ρ) and the integrand φ is a real-valued function on the support of μ. Suppose that random draws from the measure μ are available and calls to the function φ are possible. The standard Monte Carlo estimate consists in averaging $\varphi(X_i)$ over $i = 1, \dots, n$, where the particles X_i are drawn independently from μ. For square-integrable integrands, the Monte Carlo estimate has convergence rate $\mathcal{O}(n^{-1/2})$ as $n \to \infty$, whatever the dimension of the domain. In some applications, calls to the integrand may be expensive (Sacks et al., 1989; Doucet et al., 2001) and one may only have access to a small number of evaluations $\varphi(X_i)$, such as in Bayesian inference on complex models (Higdon et al., 2015; Toscano-Palmerin and Frazier, 2022). The $n^{-1/2}$-rate of the standard Monte Carlo estimate then becomes too slow and leads to highly variable estimates.

As detailed in Novak (2016), the complexity of integration algorithms may be analyzed through the convergence rate of the error. Any randomized procedure based on n particles yields an estimate $\hat{\mu}_n(\varphi)$ of the integral $\mu(\varphi)$ and the root mean-square error of the procedure is $\mathbb{E}[|\hat{\mu}_n(\varphi) - \mu(\varphi)|^2]^{1/2}$. For the specific problem of integration with respect to the uniform measure over the unit cube $[0, 1]^d$, the complexity rate of randomized methods for Lipschitz integrands is known to be $\mathcal{O}(n^{-1/2}n^{-1/d})$ (Bakhvalov, 2015; Novak, 2016). Furthermore, when the first s derivatives of the integrand are bounded, the convergence rate becomes $\mathcal{O}(n^{-1/2}n^{-s/d})$. These complexity rates are informative as they show that for smooth integrands, the Monte Carlo estimate is suboptimal and leaves room for improvement by relying on the regularity of the integrand. Several approaches are already known for

Accepted to Bernoulli (2024).

improving upon the Monte Carlo benchmark. They can be classified according to their convergence rates while keeping in mind the lower bound $\mathcal{O}(n^{-1/2}n^{-s/d})$.

The control variate method (Rubinstein, 1981; Newton, 1994; Caflisch, 1998; Evans and Swartz, 2000; Glynn and Szechtman, 2002; Glasserman, 2004) is a powerful technique that allows to reduce the variance of the Monte Carlo estimate by approximating the integrand with a function whose integral is known. The Monte Carlo rate can be improved by combining control variates to model φ (Oates et al., 2017; Portier and Segers, 2019; Leluc et al., 2021; South et al., 2022). In Portier and Segers (2019), when μ is the uniform distribution on $[-1,1]^d$ and when using m orthogonal polynomials as control variates, the convergence rate is $\mathcal{O}(n^{-1/2}m^{-s/d})$, with s is the regularity of φ. However, m is required to be of smaller order than $n^{1/3}$, so that the optimal rate cannot be achieved. By relying on control functions constructed in a reproducing kernel Hilbert space, Oates et al. (2019) obtained the rate $\mathcal{O}(n^{-1/2-(a\wedge b)/d+\varepsilon})$ for a specific class of integrands where a is related to the smoothness of the target density, b is related to the smoothness of the integrand φ and $\varepsilon > 0$ hides logarithmic factors and can be arbitrary small. This is done at the expense of computational complexity in $\mathcal{O}(n^3)$ due to matrix inversion when solving the underlying kernel ridge regression problem.

Another reliable technique to improve the rate of convergence of standard Monte Carlo is stratification. The sample space is partitioned and particles are sampled from each piece of the partition separately. It has allowed to improve the convergence rate of Monte Carlo estimates (Haber, 1966, 1967) and to derive a general framework called stochastic quadrature rules (Haber, 1969). Recently, Haber's work has been extended to take advantage of higher smoothness in the integrand (Chopin and Gerber, 2022). To the best of our knowledge, the works of Haber (1966) and Chopin and Gerber (2022) are the first ones to provide a feasible method achieving the best rate of convergence for smooth integrands; an algorithm with optimal rate of convergence but difficult to implement is proposed in Krieg and Novak (2017). Some related ideas for building optimal algorithms were already introduced in the seminal work of Bakhvalov (2015); see also Novak (2016) for a short review and Novak (1988) for more details. A drawback of the methods proposed in Haber (1966) and Chopin and Gerber (2022) is that they are only valid for integration over the unit cube and involve a geometric number ($n = \ell^d$) of evaluations of the integrand φ, which may be somewhat restrictive in case the integrand is hard to compute, as in complex Bayesian models.

Interestingly, the stratification method in Chopin and Gerber (2022) relies on a piecewise constant control function with a low bias compared to a traditional regression estimate. Such type of estimator is shown to attain the optimal rate of convergence in case of noiseless evaluations of the function to approximate (Kohler and Krzyżak, 2013; Hinrichs et al., 2020) and this explains the success of their approach to reach to optimal rate of convergence for the integration problem. The same idea of using a low-bias estimate is the starting point of this paper too and is relevant because the integrand φ is accessible without noise. Low-bias estimates have also been employed successfully in adaptive rejection sampling (Achddou et al., 2019), allowing to reach the optimal rate. Alternatively, stochastic quadrature rules based on determinantal point processes have been proposed in Bardenet and Hardy (2020) and Coeurjolly et al. (2021). Such methods allow to reduce the root mean squared error to $\mathcal{O}(n^{-1/2}n^{-1/(2d)})$ when the integrand φ has a continuous derivative. This interesting acceleration still remains slower than the optimal lower bound.

In this paper, a new Monte Carlo method called *control neighbors* is introduced. By using 1-nearest neighbor estimates as control variates, it produces an estimate $\hat{\mu}_n(\varphi)$ of the integral $\mu(\varphi)$ for a probability measure μ on a metric space (M,ρ) such that the measure of a ball of radius $r > 0$ is of the order r^d as $r \to 0$, uniformly over the space. This novel estimate is shown to achieve the convergence rate $\mathcal{O}(n^{-1/2}n^{-s/d})$ for Hölder integrands of regularity $s \in (0,1]$. The main properties of the control neighbors estimate are as follows:

(a) It can be obtained under the same framework as standard Monte Carlo, i.e., as soon as one can draw random particles from μ and evaluate the integrand φ. In contrast to the classical method of control variates, the existence of control variates with known integrals is not required.

(b) The method takes the form of a linear integration rule $\sum_{i=1}^n w_{i,n}\varphi(X_i)$ with weights $w_{i,n}$ not depending on the integrand φ but only on the particles $X_1,\dots,X_n$. This property is computationally beneficial when several integrals are to be computed with the same measure μ.

(c) For Lipschitz integrands φ, when $s = 1$, the convergence rate is the optimal one $\mathcal{O}(n^{-1/2}n^{-1/d})$ (Novak, 2016). Other recent approaches for general μ (e.g. Oates et al., 2017; Portier and Segers, 2019) do not achieve this rate. Such a rate is achieved by the methods in Haber (1966) and in Chopin and Gerber (2022) provided μ is the uniform distribution on a d-dimensional cube and the number of evaluations of the integrand is geometric, $n = \ell^d$ for some integer $\ell \geq 1$.

(d) The approach is *post-hoc* in the sense that it can be run after sampling the particles and independently from the sampling mechanism. Although the theory in our paper is restricted to independent random variables, the method can be implemented for other sampling designs including MCMC or adaptive importance sampling.

(e) The method not only applies to integration with respect to smooth measures on compact subsets of Euclidean space but also with respect to the normalized volume measure on compact Riemannian manifolds such as the unit sphere or the orthogonal group.

A similar approach to the one in this paper (and developed independently) has been proposed in Blanchet et al. (2023) as an example of an algorithm that achieves the complexity bounds for the integration problem; see their Section 4 and in particular their Theorem 4.2. Their study is focused on the derivation of complexity rates and on the existence of methods to achieve these rates. They consider the case where the integrand is evaluated with noise, and this leads to different choices of the number of neighbors (as explained in the proof of their Theorem 4.2). When the noise vanishes, their approach is similar to our 1-nearest neighbor control variate estimate. One major difference remains that, for the proof of their theorem, they need to split the sample into two independent parts, leading to a loss of efficiency in practice. This latter drawback is avoided in this study with the help of a careful analysis of the leave-one-out algorithm.

The outline of the paper is as follows. The mathematical foundations of nearest neighbor estimates are gathered in Section 2 with a formal introduction of two different nearest neighbors estimates. The theoretical properties of the control neighbor estimates are stated in Section 3. Finally, Section 4 reports on several numerical experiments along with some practical remarks on the implementation of the proposed estimates, and Section 5 concludes. The supplement contains proofs, auxiliary results and additional experiments.

2 From nearest neighors to control neighbors

This section presents the mathematical framework of nearest neighbor estimates with reminders on Voronoi cells and central quantities for the analysis, namely the degree of a point and the average cell volume. Next, two *control neighbor* estimates are introduced and some basic properties are stated.

2.1 Nearest neighbors

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables taking values in a metric space (M, ρ), drawn from a distribution μ without atoms. Let $x \in M$ denote a generic point.

Definition 1 (Nearest neighbors and distances)**.** *The nearest neighbor $\hat{N}_n(x)$ of x among $X_1, \ldots, X_n$ and the associated distance $\hat{\tau}_n(x)$ are*

$$\hat{N}_n(x) \in \operatorname*{arg\,min}_{Y \in \{X_1, \ldots, X_n\}} \rho(x, Y), \qquad \hat{\tau}_n(x) = \rho\left(\hat{N}_n(x), x\right).$$

When the $\arg\min$ *is not unique, $\hat{N}_n(x)$ is defined as the one point among the* $\arg\min$ *having the smallest index. More generally, for $k \in \{1, \ldots, n\}$, let $\hat{N}_{n,k}(x)$ denote the k-nearest neighbor of x and $\hat{\tau}_{n,k}(x) = \rho\left(\hat{N}_{n,k}(x), x\right)$ the associated distance, breaking ties by the lexicographic order.*

The sample $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ defines a natural (random) partition of the integration domain when considering the associated Voronoi cells. Any such cell is associated to a given sample point, say X_i, and contains all the points x of which the nearest neighbor is X_i.

Definition 2 (Voronoi cells and volumes)**.** *The Voronoi cells of $X_1, \ldots, X_n$ are*

$$\forall i = 1, \ldots, n, \qquad S_{n,i} = \left\{x \in M : \hat{N}_n(x) = X_i\right\},$$

with Voronoi volumes $V_{n,i} = \mu(S_{n,i})$.

The 1-NN estimate of φ is defined as $\hat{\varphi}_n(x) = \varphi\left(\hat{N}_n(x)\right)$ for all $x \in \mathbb{R}^d$ and is piece-wise constant on the Voronoi cells, i.e., $\hat{\varphi}_n(x) = \sum_{i=1}^n \varphi(X_i)\, \mathbb{1}_{S_{n,i}}(x)$.

The leave-one-out rule is a general technique to introduce independence between the prediction and the evaluation points. It is used as a cross-validation strategy in order to tune hyper-parameters of statistical procedures (Stone, 1974; Craven and Wahba, 1978). The leave-one-out version of $\hat{\varphi}_n$ without X_i is denoted by $\hat{\varphi}_n^{(i)}$ and is obtained in the exact same way as $\hat{\varphi}_n$ except that a slightly different sample—in which the i-th observation has been removed—is used.

Definition 3 (Leave-one-out neighbors, Voronoi cells and volumes)**.** *Let* $i \in \{1, \dots, n\}$ *and* $\mathcal{X}_n^{(i)} = \{X_1, \dots, X_n\} \setminus \{X_i\}$. *The leave-one-out neighbor of* $x \in M$ *is*

$$\hat{N}_n^{(i)}(x) \in \operatorname*{arg\,min}_{Y \in \mathcal{X}_n^{(i)}} \rho(x, Y).$$

When the above $\arg\min$ *is not unique,* $\hat{N}_n^{(i)}(x)$ *is defined as the one point among the* $\arg\min$ *having the smallest index. The leave-one-out Voronoi cell* $S_{n,j}^{(i)}$ *denotes the* j*-th Voronoi cell in* $\mathcal{X}_n^{(i)}$*, i.e.,*

$$\forall j \in \{1, \dots, n\} \setminus \{i\}, \qquad S_{n,j}^{(i)} = \left\{x \in M : \hat{N}_n^{(i)}(x) = X_j\right\}.$$

The leave-one-out Voronoi volume is defined as $V_{n,j}^{(i)} = \mu\big(S_{n,j}^{(i)}\big)$.

The leave-one-out 1-NN predictor used in Section 2.2 to define the proposed integral estimate is $\hat{\varphi}_n^{(i)}(x) = \varphi\big(\hat{N}_n^{(i)}(x)\big)$. A key property is that $\hat{\varphi}_n^{(i)}$ and $\hat{\varphi}_n$ coincide on $S_{n,j}$ for $j \neq i$. On the cell $S_{n,i}$, when the function φ is Hölder with regularity $s \in (0, 1]$, the supremum distance between $\hat{\varphi}_n^{(i)}$ and $\hat{\varphi}_n$ is of the same order as the nearest neighbor distance to the power s. In terms of the $L^1(\mu)$-norm, their difference is even smaller if the cell $S_{n,i}$ has a small μ-volume. Relevant for our numerical integration problem is that the average of the integrals $\mu\big(\hat{\varphi}_n^{(i)}\big)$ is close to $\mu(\hat{\varphi}_n)$.

Lemma 1. *In terms of* $\bar{\varphi}_n(x) = \sum_{i=1}^n \hat{\varphi}_n^{(i)}(x)\, \mathbb{1}_{S_{n,i}}(x)$*, we have* $\sum_{i=1}^n \big(\mu\big(\hat{\varphi}_n^{(i)}\big) - \mu(\hat{\varphi}_n)\big) = \mu(\bar{\varphi}_n - \hat{\varphi}_n)$.

Enumerating how often a point X_j is the nearest neighbor of points X_i for $i \neq j$ reflects how much X_j is surrounded within the sample. Another quantification of the isolation of a point X_j is obtained by summing the Voronoi volumes $V_{n,j}^{(i)}$ over $i \neq j$. These two notions are formally stated in the next definition.

Definition 4 (Degree and cumulative volume)**.** *For all* $j = 1, \dots, n$*, the degree* $\hat{d}_{n,j}$ *represents the number of times* X_j *is a nearest neighbor of a point* X_i *for* $i \neq j$*. The associated* j*-th cumulative Voronoi volume is denoted by* $\hat{c}_{n,j}$*, that is,*

$$\hat{d}_{n,j} = \sum_{i: i \neq j} \mathbb{1}_{S_{n,j}^{(i)}}(X_i) \qquad \textit{and} \qquad \hat{c}_{n,j} = \sum_{i: i \neq j} V_{n,j}^{(i)}.$$

Interestingly, the degree of a point and its cumulative Voronoi volume have the same expectation: $\mathbb{E}[\hat{d}_{n,j}] = \mathbb{E}[\hat{c}_{n,j}] = 1$.

The two quantities $\hat{d}_{n,j}$ and $\hat{c}_{n,j}$ will be useful in Proposition 1 below to express the control neighbors estimate as a linear integration rule. For now, we note that weighted sums of $\varphi(X_j)$ using $\hat{d}_{n,j}$ and $\hat{c}_{n,j}$ as weights are related to the leave-one-out estimate.

Lemma 2. *It holds that*

$$\sum_{i=1}^n \varphi(X_i)\, \hat{d}_{n,i} = \sum_{i=1}^n \hat{\varphi}_n^{(i)}(X_i) \qquad \textit{and} \qquad \sum_{i=1}^n \varphi(X_i)\, \hat{c}_{n,i} = \sum_{i=1}^n \mu\big(\hat{\varphi}_n^{(i)}\big).$$

2.2 Control neighbors

With the help of previous notation, we now introduce the two *control neighbor* estimates

$$\hat{\mu}_n^{(\mathrm{NN})}(\varphi) = \frac{1}{n}\sum_{i=1}^{n}\left[\varphi(X_i) - \left\{\hat{\varphi}_n^{(i)}(X_i) - \mu(\hat{\varphi}_n)\right\}\right] \tag{1}$$

$$\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi) = \frac{1}{n}\sum_{i=1}^{n}\left[\varphi(X_i) - \left\{\hat{\varphi}_n^{(i)}(X_i) - \mu(\hat{\varphi}_n^{(i)})\right\}\right]. \tag{2}$$

The first one, $\hat{\mu}_n^{(\mathrm{NN})}(\varphi)$, is the output in the nearest neighbor algorithm in Section 4.1. The second one, $\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi)$, is a slight modification of $\hat{\mu}_n^{(\mathrm{NN})}(\varphi)$ that is an unbiased estimate of $\mu(\varphi)$. Indeed, a simple conditioning argument implies

$$\mathbb{E}\left[\hat{\varphi}_n^{(i)}(X_i) - \mu\big(\hat{\varphi}_n^{(i)}\big)\right] = \mathbb{E}\left[\mathbb{E}\left[\hat{\varphi}_n^{(i)}(X_i) - \mu\big(\hat{\varphi}_n^{(i)}\big)\,\middle|\,\mathcal{X}_n^{(i)}\right]\right] = 0,$$

which is sufficient to get the zero-bias property

$$\mathbb{E}\left[\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi)\right] = \mu(\varphi).$$

The estimate $\hat{\mu}_n^{(\mathrm{NN})}$ is not unbiased, but under the conditions of Theorem 1 below, its bias is of smaller order than its root mean squared error; see Remark 4 below. Moreover, since $\hat{\varphi}_n^{(i)}$ is similar to the 1-NN estimate $\hat{\varphi}_n$ of φ based on the full sample $X_1, \ldots, X_n$, their integrals should be close. This intuition is confirmed in Proposition 3 in the Supplementary material. Consequently, the estimate (2) will play an important role in the theory. However, computing the terms $\mu\big(\hat{\varphi}_n^{(i)}\big)$ for $i = 1, \ldots, n$ requires the evaluation of n additional integrals. In practice, the working estimate is (1) as it involves fewer computations. Further, it is worthwhile to note that our two control neighbor estimates are *not* based on a single approximation of the integrand φ but rather on n different approximations $\hat{\varphi}_n^{(i)}$.

The two control neighbor estimates in (1) and (2) can be expressed as linear integration rules of the form $\sum_{i=1}^{n} w_{i,n}\,\varphi(X_i)$ with weights $w_{i,n}$ not depending on the integrand φ. The weights involve the degrees $\hat{d}_{n,i}$ and the (cumulative) volumes $V_{n,i}$ and $\hat{c}_{n,i}$ in Definitions 2 and 4.

Proposition 1 (Quadrature rules)**.** *The estimates $\hat{\mu}_n^{(\mathrm{NN})}(\varphi)$ and $\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi)$ can be expressed as linear estimates of the form*

$$\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi) = \sum_{i=1}^{n} w_{i,n}^{(\mathrm{NN-loo})}\varphi(X_i) \quad \textit{and} \quad \hat{\mu}_n^{(\mathrm{NN})}(\varphi) = \sum_{i=1}^{n} w_{i,n}^{(\mathrm{NN})}\varphi(X_i)$$

where $w_{i,n}^{(\mathrm{NN-loo})} = (1 + \hat{c}_{n,i} - \hat{d}_{n,i})/n$ and $w_{i,n}^{(\mathrm{NN})} = (1 + nV_{n,i} - \hat{d}_{n,i})/n$.

The weights of the two estimates satisfy $\sum_{i=1}^{n} w_{i,n} = 1$, meaning that the integration rules are exact for constant functions. In the light of Proposition 1, the proposed estimate $\hat{\mu}_n^{(\mathrm{NN})}(\varphi)$ consists in a simple modification of $\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi)$ by replacing $\hat{c}_{n,i}$, which involves $n-1$ Voronoi volumes, by $nV_{n,i}$. For Hölder integrands of order $s \in (0, 1]$, the difference between both estimates is of the order $n^{-1/2-s/d}$, in view of Eq. (3) below and Proposition 3 in the Supplementary material.

3 Main results

This section gathers some preliminary considerations on distributions on metric spaces and a bound on moments of nearest-neighbor distances before stating the main theoretical properties of the control neighbor estimates (1) and (2). Under suitable conditions on the measure μ and the integrand φ, their root mean squared errors are shown to reach the rate $\mathbb{E}[|\hat{\mu}_n(\varphi) - \mu(\varphi)|^2]^{1/2} \lesssim n^{-1/2}n^{-s/d}$, where two positive sequences a_n and b_n satisfy $a_n \lesssim b_n$ provided $a_n \leq Cb_n$ for some constant $C > 0$. Further, their finite-sample performance is studied through a concentration inequality providing a high-probability bound on the error $|\hat{\mu}_n(\varphi) - \mu(\varphi)|$.

3.1 Distributions on metric spaces

Let μ be a Borel probability measure on a bounded metric space (M, ρ) with diameter $\mathrm{diam}(M)$. To control the nearest neighbor distances, we need a handle on the probabilities of small balls. Let $B(x, r) = \{y \in M : \rho(x, y) \leq r\}$ denote the ball with center $x \in M$ and radius $r > 0$. Intuitively, if μ is an absolutely continuous measure on a compact subset of $\mathbb{R}^d$, then $\mu(B(x, r))$ is of the order r^d. The same is true if μ is a smooth measure on a d-dimensional manifold. These considerations motivate the following assumption, which implicitly defines a kind of intrinsic dimension of M as measured by μ.

(A1) There exist $d > 0$ and $0 < C_0 \leq C_1 < \infty$ such that

$$\forall x \in M,\ \forall r \in (0, \mathrm{diam}(M)], \qquad C_0 r^d \leq \mu(B(x, r)) \leq C_1 r^d.$$

Remark 1. In (A1), it is sufficient to require the stated inequalities for all $r \in (0, r_0]$ for some $r_0 > 0$. Upon changing C_0 and C_1, the inequalities then hold for all $r \in (r_0, \mathrm{diam}(M)]$ too. Indeed, for such r, we then have the upper bound $\mu(B(x, r)) \leq 1 \leq r_0^{-d} \cdot r^d$ and the lower bound $\mu(B(x, r)) \geq \mu(B(x, r_0)) \geq C_0 r_0^d \geq C_0 \{r_0 / \mathrm{diam}(M)\}^d \cdot r^d$.

Remark 2. In theory, the exponent d in (A1) could be non-integer. As an example, suppose μ is the normalized Hausdorff measure on a fractal-like set, like the Cantor set ($d = \log 2 / \log 3$) or the Sierpinski triangle ($d = \log 3 / \log 2$), see Stein and Shakarchi (2005). For the integration problem we study, this does not seem of much practical interest.

A sufficient hypothesis for (A1) is that μ is supported by a bounded subset of $\mathbb{R}^d$ with positive d-dimensional Lebesgue measure and is absolutely continuous with a well-behaved density:

(A1') The measure μ is supported by a compact subset M of $\mathbb{R}^d$ and admits a density f with respect to the d-dimensional Lebesgue measure λ_d. There exist constants $b, U, c, r_0 \in (0, \infty)$ with $b \leq U$ such that

- $\forall x \in M, \quad b \leq f(x) \leq U$;
- $\forall r \in (0, r_0], \forall x \in M, \quad \lambda_d(M \cap B(x, r)) \geq c\, \lambda_d(B(x, r))$.

Assumption (A1') is related to the *strong density assumption* of Audibert and Tsybakov (2007). It ensures that the density f of the measure μ has a sufficiently regular support and that it is bounded away from zero and infinity. In view of Remark 1, it is easy to see that (A1') implies (A1): since $\mu(B(x, r)) = \int_{M \cap B(x,r)} f(y)\, \mathrm{d}y$, we have, for $x \in M$ and $r \in (0, r_0]$, the bounds

$$\mu(B(x, r)) \leq U\, \lambda_d(B(x, r)) \leq U V_d \cdot r^d, \qquad \text{and}$$
$$\mu(B(x, r)) \geq b\, \lambda_d(M \cap B(x, r)) \geq bc\, \lambda(B(x, r)) \geq bc V_d \cdot r^d,$$

where $V_d = \lambda_d(B(0, 1))$ is the volume of the unit ball in $\mathbb{R}^d$.

Another case of interest is when μ is supported by a subset of a closed d-dimensional Riemannian manifold $(\mathcal{M}_g, g)$. Let ρ_g be the geodesic distance on $\mathcal{M}_g$ induced by the metric g and let $B_g(x, r) = \{y \in \mathcal{M}_g : \rho_g(x, y) \leq r\}$ denote the geodesic ball with center $x \in \mathcal{M}_g$ and radius $r > 0$.

(A1") The measure μ admits a density f having support $M = \{x \in \mathcal{M}_g : f(x) > 0\}$ with respect to the Riemannian volume measure ν_g on $\mathcal{M}_g$. The sectional curvatures of $\mathcal{M}_g$ are not greater than some $\delta > 0$ and there exist constants $b, U, c, r_0 \in (0, \infty)$ with $b \leq U$ such that

- $\forall x \in M, \quad b \leq f(x) \leq U$;
- $\forall 0 < r \leq r_0, \forall x \in M, \quad \nu_g(\mathcal{M}_g \cap B_g(x, r)) \geq c\, \nu_g(B_g(x, r))$.

Since $\mu(B_g(x, r)) = \int_{M \cap B_g(x,r)} f(y)\, \nu_g(\mathrm{d}y)$, the argument that (A1") implies (A1) is the same as the one of the sufficiency of (A1') above, provided we can relate the volume of a geodesic ball to its radius. This is the content of the next result.

Theorem (Corollary 2.2 in Kokarev (2021))**.** *Let $(\mathcal{M}_g, g)$ be a closed d-dimensional Riemannian manifold whose sectional curvatures are not greater than δ, where $\delta > 0$. Let $\mathrm{inj}(\mathcal{M}_g)$ denote the injectivity radius of $\mathcal{M}_g$. Then, for any point $x \in \mathcal{M}_g$ and any positive radius $r \leq \mathrm{rad}(g)$, the volume of the geodesic ball $B_g(x, r)$ satisfies the inequalities*

$$2^{1-d} V_d r^d \leq \nu_g(B_g(x, r)) \leq 2^{d-1} \frac{\nu_g(\mathcal{M}_g)}{\mathrm{rad}(g)^d} r^d,$$

where V_d is the volume of a unit ball in the d-dimensional Euclidean space, $\mathrm{rad}(g)$ *stands for* $\min\{\mathrm{inj}(\mathcal{M}_g), \pi/(2\sqrt{\delta})\}$.

In Section 4, we perform numerical experiments featuring manifolds such as the unit sphere $\mathbb{S}^{m-1}$ in $\mathbb{R}^m$ and the orthogonal group. The cited theorem is valid for $\mathbb{S}^{m-1}$ since the sectional curvature of a $(m-1)$-sphere of radius r is $1/r^2$ (Lee, 2019). The orthogonal group is a special case of the Stiefel manifold, which has a positive bounded sectional curvature (Zimmermann and Stoye, 2024). Other examples are the complex projective space CP^d and the quaternion projective space HP^d, both having sectional curvatures bounded between 1 and 4 (Ichida, 2000).

3.2 Moments of nearest-neighbor distance

Consider drawing independent random samples from a distribution on a metric space.

(A2) $X, X_1, X_2, \ldots$ are independent and identically distributed random variables in a bounded metric space (M, ρ) with common distribution μ.

For Hölder-continuous functions, the quality of the approximation of the integrand φ by its nearest-neighbor estimate $\hat{\varphi}_n$ depends on the nearest-neighbor distance $\hat{\tau}_n$ (see Section 2.1). The next lemma provides control on the k-nearest neighbor distance $\hat{\tau}_{n,k}(x)$ and is inspired by the k-NN literature (Biau and Devroye, 2015). Let Γ denote the Euler gamma function.

Lemma 3 (Bounding moments of nearest neighbor distances)**.** *Under (A1) and (A2), we have, for any $q > 0$ and for integer $k \in \{1, \ldots, n\}$,*

$$\mathbb{E}\left[\hat{\tau}_{n,k}(x)^q\right] \leq (C_0 n)^{-q/d} \frac{\Gamma(k + q/d)}{\Gamma(k)}.$$

Remark 3 (Lower bounds)**.** In Lemma 3, we only need the lower bound in (A1), not the upper bound. This uniform lower bound in Condition (A1) plays an important role in our analysis as it allows a uniform control on the Voronoi cell diameters. Some refinement might be possible given the recent progress in k-NN regression for covariates with unbounded support (Kohler et al., 2006) and k-NN classification using some tail assumption on the covariates (Gadat et al., 2016). Extending the present analysis to such general measures is left for further research.

3.3 Root mean squared error bounds

The main result of this section provides a finite-sample bound on the root mean squared error of the two control neighbors estimates in (1) and (2). The bound depends on the regularity of the integrand.

(A3) The function $\varphi : M \to \mathbb{R}$ is Hölder continuous, i.e., there exist $s \in (0, 1]$ and $L > 0$ such that

$$\forall x, y \in M, \qquad |\varphi(x) - \varphi(y)| \leq L\rho(x, y)^s.$$

The leave-one-out version $\hat{\mu}_n^{(\mathrm{NN-loo})}$ involves n additional integrals and is computationally cumbersome. The proposed estimate $\hat{\mu}_n^{(\mathrm{NN})}$ requires only a single additional integral while remaining close to the leave-one-out estimate, since their difference is

$$\hat{\mu}_n^{(\mathrm{NN})}(\varphi) - \hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi) = \mu(\hat{\varphi}_n) - \frac{1}{n}\sum_{i=1}^{n} \mu(\hat{\varphi}_n^{(i)}). \tag{3}$$

Using this property and Lemma 1, we obtain that the root mean squared distance between the leave-one-out version and the proposed estimate is of the order $\mathcal{O}(n^{-1/2-s/d})$ as $n \to \infty$; see Proposition 3 in the Supplementary material for a precise statement. Therefore, the two estimates share the same convergence rate.

In the next theorem, we assume (A1), for which each of (A1') and (A1") is a sufficient condition.

Theorem 1 (Root mean squared error bounds)**.** *Under (A1), (A2) and (A3), if $n \geq 4$, then*

$$\mathbb{E}\left[\left|\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi) - \mu(\varphi)\right|^2\right]^{1/2} \leq C_{\mathrm{NN-loo}} n^{-1/2} n^{-s/d},$$
$$\mathbb{E}\left[\left|\hat{\mu}_n^{(\mathrm{NN})}(\varphi) - \mu(\varphi)\right|^2\right]^{1/2} \leq C_{\mathrm{NN}} n^{-1/2} n^{-s/d},$$

where

$$C_{\mathrm{NN-loo}} = L C_0^{-s/d} (C_1/C_0)^{1/2} 2^{s/d} \sqrt{10\Gamma(2s/d+2)},$$
$$C_{\mathrm{NN}} = L C_0^{-s/d} \sqrt{\Gamma(2s/d+2)} \left(2\sqrt{s/d+1} + 2^{s/d}(C_1/C_0)^{1/2}\sqrt{10}\right).$$

For Lipschitz functions ($s = 1$), the rates obtained in Theorem 1 match the complexity rate stated in Novak (2016), see Section 1. The results in the aforementioned paper are concerned about a slightly more precise context as they assert that no random integration rule (see the article for more details) can reach a higher accuracy—measured in terms of mean-square error—than $\mathcal{O}(n^{-1-2/d})$ when the integration measure is the uniform distribution on the unit cube and the integrand is Lipschitz. Theorem 1 states that the optimal rate is in fact achieved for more general spaces and measures.

The control neighbors estimates involve an approximation of φ by piece-wise constant functions $\hat{\varphi}_n^{(i)}$ and therefore cannot achieve an acceleration of the convergence rate of more than $n^{-1/d}$. It is an interesting question whether a higher-order nearest neighbor approximation can achieve a greater accuracy for integrands that are smoother than just Lipschitz. The simulation experiments in Section 4 will give some feeling on the actual improvement of the control neighbors estimate over the basic Monte Carlo method at finite sample sizes.

Remark 4. As mentioned in Section 2.2, the control neighbor estimate $\hat{\mu}_n^{(\mathrm{NN})}$ is in general not unbiased, in contrast to the leave-one-out version $\hat{\mu}_n^{(\mathrm{NN-loo})}$. Still, by the first statement of Proposition 3 in the Supplement, the bias of $\hat{\mu}_n^{(\mathrm{NN})}$ is of the order $n^{-1-s/d}$ and thus constitutes an asymptotically negligible part of the estimator's root mean squared error, which decreases with rate $n^{-1/2-s/d}$.

3.4 Concentration inequalities

In order to obtain a finite-sample performance guarantee of the proposed estimates, we apply an extension of McDiarmid's concentration inequality for functions with bounded differences on a high probability set A. The inequality is stated in Theorem 3 in Appendix A of the supplementary material and is itself a minor extension of an inequality due to Combes (2015).

Because of our reliance on theory for nearest neighbor processes in Portier (2023), our bound is established only for measures supported by a compact subset M of Euclidean space, which is why we assume the more specific setting of (A1') rather than (A1).

Assumption (A3) implies that φ is uniformly bounded on M, since M has a finite diameter. Write $C_\varphi = \frac{1}{2}\{\sup_{x\in M}\varphi(x) - \inf_{x\in M}\varphi(x)\}$. The two control neighbor estimates satisfy the following concentration inequalities.

Theorem 2 (Concentration inequalities)**.** *Under (A1'), (A2) and (A3), we have, for any $\varepsilon \in (0,1)$, with probability at least $1-\varepsilon$,*

$$\left|\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi) - \mu(\varphi)\right| \leq K_1 \Delta_{n,\varepsilon} + \frac{6C_\varphi}{n^{1/2+s/d}} \qquad \textit{and}$$
$$\left|\hat{\mu}_n^{(\mathrm{NN})}(\varphi) - \mu(\varphi)\right| \leq K_2 \Delta_{n,\varepsilon} + \frac{6C_\varphi}{n^{1/2+s/d}} + \frac{(s/d+2)L(V_d bc)^{-s/d}}{n^{1+s/d}},$$

where

$$\Delta_{n,\varepsilon} = \frac{\sqrt{\log(2/\sqrt{\varepsilon})} + \frac{1}{n^{s/d}}}{n^{1/2+s/d}} \cdot \begin{cases} \log(24n/\varepsilon)^{1+s/d} & \textit{for } \varepsilon \leq \dfrac{2}{n^{1/2+s/d}}, \\ (3\log(3n))^{1+s/d} & \textit{for } \varepsilon > \dfrac{2}{n^{1/2+s/d}}. \end{cases}$$

The values of $K_1, K_2 > 0$ depend on (b, c, d, L, s, U) and are provided explicitly in the proof.

For small but fixed $\varepsilon \in (0,1)$, the concentration inequalities state that the probability that the error $|\hat{\mu}_n(\varphi) - \mu(\varphi)|$ is larger than a constant multiple of $\sqrt{\log(1/\varepsilon)}(\log n)^{1+s/d} n^{-1/2} n^{-s/d}$ is less than ε. The additional logarithmic factor in Theorem 2 in comparison to Theorem 1 comes from the need to exert uniform control on the largest nearest neighbor distance via Lemma 5 in Appendix A in the supplement, which in turn originates from a concentration inequality in Portier (2023).

3.5 Optimal L_p-approximation with nearest neighbors

The upper bounds on nearest neighbor distances stated in Lemma 3 have been key to examine the control neighbor estimators $\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi)$ and $\hat{\mu}_n^{(\mathrm{NN})}(\varphi)$ as claimed in Theorems 1 and 2. Another application of these upper bounds is the L_p-approximation of the nearest neighbor function estimator $\hat{\varphi}_{n,k}(x) = k^{-1} \sum_{i=1}^{k} \varphi\big(\hat{N}_{n,i}(x)\big)$ of $\varphi(x)$. Note that this type of estimator is often used in the setting of noiseless regression where one has access to $\varphi(X_i)$ without noise. A precise statement is given in the next proposition, the proof of which is given in the supplement.

Proposition 2 (L_p-approximation with nearest neighbors)**.** *Under (A1), (A2), and (A3), we have, for any $k \in \{1, \dots, n\}$ and $p \geq 1$,*

$$\forall x \in M, \qquad \mathbb{E}\left[|\hat{\varphi}_{n,k}(x) - \varphi(x)|^p\right]^{1/p} \leq L(C_0 n)^{-s/d} \left(\frac{\Gamma(k + sp/d)}{\Gamma(k)}\right)^{1/p}.$$

Since $\Gamma(x+\alpha) \sim x^\alpha \Gamma(x)$ as $x \to \infty$ for fixed $\alpha \in \mathbb{R}$, the upper bound is of the order $(k/n)^{s/d}$ as $k = k_n \to \infty$. We emphasize that similar results as in Proposition 2 can be found among the noiseless regression literature. However, to the best of our knowledge, most of the results seem to have been obtained for Euclidean covariates, i.e., when $M = \mathbb{R}^d$. In Chapter 15 in Biau and Devroye (2015), the weak convergence of $n^{1/d}(\hat{\varphi}_{n,k}(x) - \varphi(x))$ is established. In Theorem 1 from Kohler and Krzyżak (2013), the same rate of convergence for the integrated error is obtained. Some other results are given in Hinrichs et al. (2020); Krieg et al. (2022) where the authors investigate whether the distribution of the points can deteriorate the L_p-error or not. They show that the optimal rate of convergence can be obtained using either uniformly distributed points or deterministic points from the uniform grid. Our Proposition 2 confirms, not only for the Euclidean space but for general metric spaces, the following principle, that when there are enough points everywhere, as implied by (A1), the optimal rates of convergence for the L_p-error can be achieved.

Even though the analysis of the control neighbor estimate cannot be deduced from Proposition 2 because of the probabilistic dependence that is introduced in the algorithm (see the proofs of Theorems 1 and 2), the previous statement can be used for a better understanding of the control neighbor estimator as it shows that the rate obtained in Theorem 1 can be decomposed as the usual Monte Carlo error, $n^{-1/2}$, times the L_p-approximation error of the control variates estimating the integrand, $n^{-s/d}$. This is in line with the results from Portier and Segers (2018) where the same type of decomposition can be found for another control variates based estimator.

4 Numerical experiments

In Section 4.1, we start by presenting the *control neighbors* algorithm[1] along with some practical remarks related to its implementation. We present in Section 4.2 some examples involving integration problems on different spaces. Finally, Section 4.3 deals with the application of Monte Carlo estimates for optimal transport when computing the Sliced-Wasserstein distance. Additional experiments on a finance application with Monte Carlo exotic option pricing are available in Appendix E.

4.1 Practical implementation

The procedure for computing the control neighbor estimate $\hat{\mu}_n^{(\mathrm{NN})}(\varphi)$ is presented in Algorithm 1. Required is a collection of points $X_1, \dots, X_n$ generated from a distribution μ. The estimate is based on the evaluations $\varphi(X_i)$ of the integrand and the evaluations $\hat{\varphi}_n^{(i)}(X_i)$ of the leave-one-out 1-nearest neighbors (1-NN) estimates of $\varphi(X_i)$. Here, $\hat{\varphi}_n^{(i)}$ denotes the 1-NN estimate of φ constructed from the sample $\mathcal{X}_n^{(i)} = \{X_1, \dots, X_n\} \setminus \{X_i\}$ without the i-th particle, for $i \in \{1, \dots, n\}$; see Section 2.1 for precise definitions. Also required is the integral $\mu(\hat{\varphi}_n)$ of the 1-NN estimate $\hat{\varphi}_n$ based on the whole sample. Practical remarks regarding the computation of all these quantities are given next.

[1] For reproducibility, the code is available at Control Neighbors.

Algorithm 1 Control Neighbors for Monte Carlo integration

Require: integrand φ, probability measure μ, number of particles n.

1: Generate an independent random sample $X_1, \dots, X_n$ from μ
2: Compute evaluations $\varphi(X_1), \dots, \varphi(X_n)$
3: Compute nearest neighbor evaluations $\hat{\varphi}_n^{(1)}(X_1), \dots, \hat{\varphi}_n^{(n)}(X_n)$
4: Compute the integral of the nearest neighbor estimate $\mu(\hat{\varphi}_n)$
5: Return $\frac{1}{n}\sum_{i=1}^n [\varphi(X_i) - \{\hat{\varphi}_n^{(i)}(X_i) - \mu(\hat{\varphi}_n)\}]$

Nearest neighbors and tree search. The naive neighbor search implementation involves the computation of distances between all pairs of points in the training samples and may be computationally prohibitive. To address such practical inefficiencies, a variety of tree-based data structures have been invented to reduce the cost of a nearest neighbors search. The KD-Tree (Bentley, 1975) is a binary tree structure which recursively partitions the space along the coordinate axes, dividing it into nested rectangular regions into which data points are filed. The construction of such a tree requires $\mathcal{O}(dn \log n)$ operations (Friedman et al., 1977). Once constructed, the query of a nearest neighbor in a KD-Tree can be done in $\mathcal{O}(\log n)$ operations. However, in high dimension, the query cost increases and the structure of Ball-Tree (Omohundro, 1989) is favoured. Where KD trees partition data points along the Cartesian axes, Ball trees do so in a series of nested hyper-spheres, making tree construction more costly than for a KD tree, but resulting in an efficient data structure even in high dimensions. Many software libraries contain KD-tree and Ball-Tree implementations with efficient compression and parallelization (Pedregosa et al., 2011; Johnson et al., 2019).

Evaluation of $\hat{\varphi}_n^{(i)}(X_i)$. When the computing time is measured through the evaluation of the integrand, the n additional evaluations $\hat{\varphi}_n^{(i)}(X_i)$ are not computationally difficult as no additional calls to φ are necessary. The leave-one-out nearest neighbor evaluation can be easily obtained as follows: first fit a KD-Tree on the particles $X_1, \dots, X_n$, then query the 2-nearest neighbor of each X_i to produce the vector of leave-one-out nearest neighbors.

Evaluation of $\mu(\hat{\varphi}_n)$ and Voronoi volumes. The quantity $\mu(\hat{\varphi}_n)$ is the sum of the evaluations $\varphi(X_i)$ weighted by the values of the Voronoi volumes associated to the sample points X_i (see Definition 2 in the next section). The Voronoi volumes may be hard to compute but can always be approximated. The integral $\mu(\hat{\varphi}_n)$ of the nearest neighbor estimate may be replaced by a Monte Carlo approximation based on N particles, such as $\tilde{\mu}_N(\hat{\varphi}_n) = N^{-1}\sum_{i=1}^N \hat{\varphi}_n(\tilde{X}_i)$, where the variables $\tilde{X}_i$ are drawn independently from μ. No additional evaluations of φ are required. Conditionally on the first sample $X_1, \dots, X_n$, the error of this additional Monte Carlo approximation is $\mathcal{O}(N^{-1/2})$, meaning that large values of the form $N = n^{1+2/d}$ permit to preserve the $\mathcal{O}(n^{-1/2}n^{-1/d})$ rate of the control neighbors estimate for Lipschitz functions.

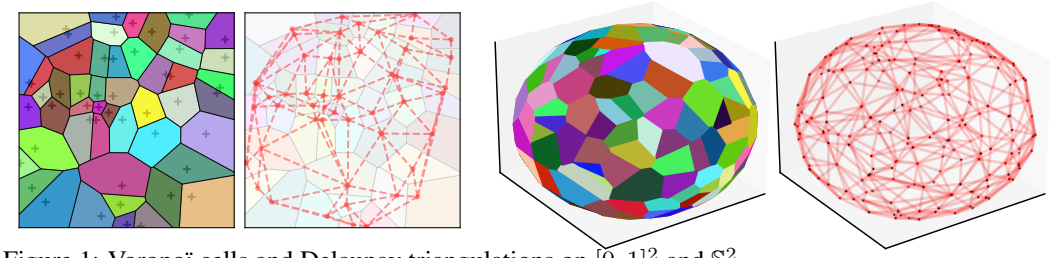


Figure 1: Voronoï cells and Delaunay triangulations on $[0,1]^2$ and $\mathbb{S}^2$.
(Made with Python package Matplotlib (Hunter, 2007))

In case the measure μ is the uniform measure on $[0,1]^d$, one may be able to explicitly compute the Voronoi volumes. Starting from the pioneering work of Richards (1974) in the context of protein structures, there have been advances to perform efficient Voronoi volume computations using Delaunay triangulations and taking advantage of graphic hardware (Hoff III et al., 1999). Computations for Voronoi tessellations in dimensions $d = 2$ and $d = 3$ are implemented in the Voro++ software (Rycroft, 2009). However, this type of algorithm might become inefficient when d is large.

Computing time. When using the standard KD-tree approach, the computation time of our algorithm can be estimated in the light of the previous remarks. The different computation times are: $\mathcal{O}(nd \log n)$ for building the KD-tree, $\mathcal{O}(n \log n)$ for the evaluations $\{\hat{\varphi}_n^{(i)}(X_i)\}_{i=1}^n$ and finally, $\mathcal{O}(N \log n)$ operations for the estimation of $\mu(\hat{\varphi}_n)$. Choosing N as recommended before, the overall complexity is $\mathcal{O}\big((nd + n^{1+2/d}) \log n\big)$ operations. Note that when the Voronoi volumes are available, the computation time reduces to $\mathcal{O}(nd \log n)$.

Extensions. A natural variant of the proposed method is obtained by replacing the 1-NN estimate $\hat{\varphi}_n$ in Eq. (1) by a k-NN estimate $\hat{\varphi}_n^{(k)}$ which averages the evaluations of the k nearest neighbors of a given point. The estimate is then defined by $\hat{\varphi}_n^{(k)}(x) = k^{-1} \sum_{j=1}^k \varphi(\hat{N}_{n,j}(x))$ where $\hat{N}_{n,j}(x)$ is the j-nearest neighbor of x. This involves both the tuning of the hyperparameter $k \geq 1$ and some extra computation due to the associated nearest neighbors search. In regression or classification, high values of k can reduce the variance of the estimate by averaging the model noise at the cost of added computations. In contrast, the control neighbors estimate $(k = 1)$ is free of these additional costs and takes advantage of the noiseless evaluations (Biau and Devroye, 2015, Chapter 15) of the integrand.

4.2 Integration on various spaces: $[0,1]^d, \mathbb{R}^d, O_m(\mathbb{R})$ and $\mathbb{S}^{q-1}$

The aim of this section is to empirically validate the $\mathcal{O}(n^{-1/2} n^{-1/d})$ convergence rate of the control neighbors estimate in a wide variety of integration problems ranging from the unit cube $[0,1]^d$, the group of orthogonal matrices $O_m(\mathbb{R})$, which is an $m(m-1)/2$ dimensional manifold embedded in $\mathbb{R}^{m \times m}$, and the unit sphere $\mathbb{S}^{q-1}$ in $\mathbb{R}^q$. In the different settings, the sample size evolves from $n = 10^1$ to $n = 10^4$. The figures report the evolution of the root mean squared error $n \mapsto \mathbb{E}[|\hat{\mu}_n^{(\mathrm{NN})}(\varphi) - \mu(\varphi)|^2]^{1/2}$, where the expectation is computed over 100 independent replications. In all experiments, MC represents the naive Monte Carlo estimate and CVNN returns the value of $\hat{\mu}_n^{(\mathrm{NN})}(\varphi)$ for which the integral $\mu(\varphi_n)$ is replaced by a Monte Carlo estimate that uses n^2 particles, with n the number of evaluations of φ.

Integration on $[0,1]^d$ and $\mathbb{R}^d$. Consider first the integration problem $\mu(\varphi) = \int \varphi \, \mathrm{d}\mu$ where the measure μ is either the uniform distribution over the unit cube $[0,1]^d$ or the multivariate Gaussian measure on $\mathbb{R}^d$. The latter case is not covered by our theory, since the support is unbounded. The goal is to compute $\int \varphi_1(x) \, \mathbb{1}_{[0,1]^d}(x) \, \mathrm{d}x$ and $\int \varphi_2(x) \, \psi(x) \, \mathrm{d}x$ with

$$\varphi_1(x_1, \ldots, x_d) = \sin\left(\pi \left(\tfrac{2}{d} \textstyle\sum_{i=1}^d x_i - 1\right)\right) \quad \text{and} \quad \varphi_2(x_1, \ldots, x_d) = \sin\left(\tfrac{\pi}{d} \textstyle\sum_{i=1}^d x_i\right) \tag{4}$$

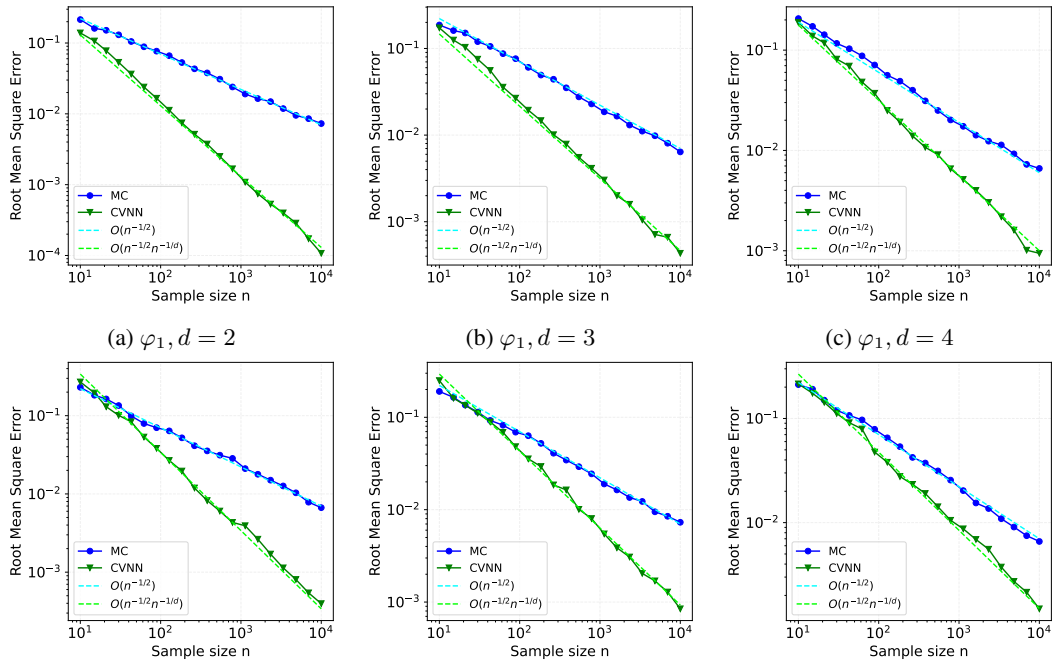
and with $\psi(\cdot)$ the probability density function of the multivariate Gaussian distribution $\mathcal{N}(0, I_d)$. Figure 2 displays the evolution of the root mean squared error for the two integrals in dimensions $d \in \{2, 3, 4\}$. The different error curves confirm the optimal convergence rate $\mathcal{O}(n^{-1/2} n^{-1/d})$ for the control neighbors estimate and illustrate the error reduction with respect to the Monte Carlo method.

Integration on the orthogonal group $O_m(\mathbb{R})$. Consider the group of real orthogonal matrices $O_m(\mathbb{R}) = \big\{X \in \mathrm{GL}_m(\mathbb{R}) : X^\top X = X X^\top = \mathrm{I}_m\big\}$. As a toy example, we compute moments of the trace of a random orthogonal matrix, i.e., integrands of the form $O_m(\mathbb{R}) \to \mathbb{R} : X \mapsto \mathrm{tr}(X)^k$, when X is drawn uniformly on $O_m(\mathbb{R})$. In fact, exact expressions are known; see for instance Diaconis and Evans (2001) and Pastur and Vasilchuk (2004) for background and results. The goal is thus to compute the integrals

$$a_k = \int_{O_m(\mathbb{R})} \mathrm{tr}(X)^k \, \mathrm{d}\mu(X), \tag{5}$$

with $\mathrm{d}\mu(X)$ the unit Haar measure on $O_m(\mathbb{R})$. Assumption (A1) is satisfied with $d = m(m-1)/2$. In practice, random orthogonal matrices $X_1, \ldots, X_n \in O_m(\mathbb{R})$ are generated using the function `ortho_group` of the Python package `scipy` (Virtanen et al., 2020). This function returns random orthogonal matrices drawn from the Haar distribution using a careful QR decomposition[2] as in Mezzadri (2007). The nearest neighbors are computed using the norm associated to the Frobenius

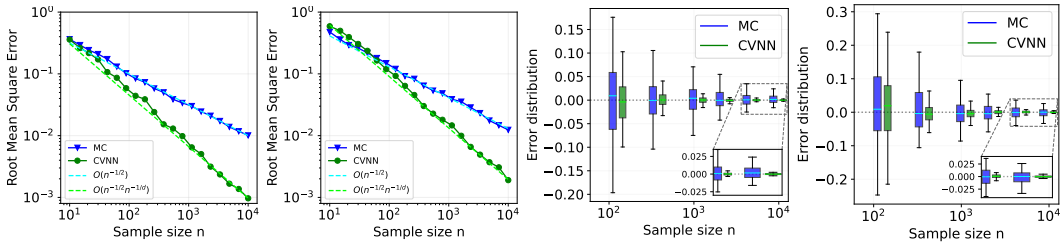
[2] QR decomposition refers to the factorization of a matrix X into a product $X = QR$ of an orthonormal matrix Q and an upper triangular matrix R.



(a) $\varphi_1, d = 2$ (b) $\varphi_1, d = 3$ (c) $\varphi_1, d = 4$

(d) $\varphi_2, d = 2$ (e) $\varphi_2, d = 3$ (f) $\varphi_2, d = 4$

Figure 2: Root mean squared errors obtained over 100 replications for functions φ_1 (top) and φ_2 (bottom) in Eq. (4) in dimension $d \in \{2, 3, 4\}$ (left to right).



(a) RSME for a_1 (b) RMSE for a_2 (c) Boxplot for a_1 (d) Boxplot for a_2

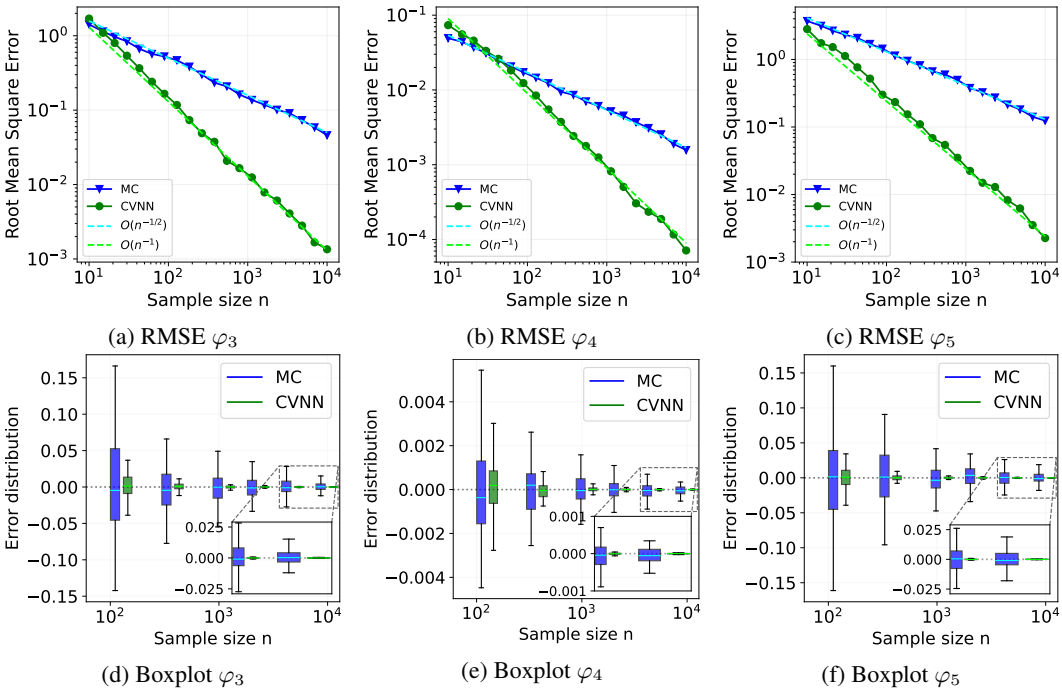
Figure 3: Root mean squared errors (left) and boxplots of the errors (right) obtained over 100 replications for integrands $X \mapsto \mathrm{tr}(X)^k$ for $k \in \{1, 2\}$ in Eq. (5) with respect to the unit Haar measure on the orthogonal group $O_3(\mathbb{R})$.

inner product $\langle X, Y \rangle = \mathrm{tr}(X^\top Y)$ for $X, Y \in O_m(\mathbb{R})$. Figure 3 reports the evolution of the root mean squared error and boxplots of the integration error over the group $O_3(\mathbb{R})$ which corresponds to dimension $d = 3$. Both integrands are continuously differentiable and thus Lipschitz, since $O_m(\mathbb{R})$ is compact. Once again, the experiments empirically validate the convergence rates of the Monte Carlo methods and reveal the variance reduction obtained with the control neighbors estimate.

Integration on the sphere $\mathbb{S}^2$. The unit sphere $\mathbb{S}^2$ in $\mathbb{R}^3$ is a compact Riemannian manifold of dimension $d = 2$ with positive curvature. The distances on the sphere are computed using the equations of the great circles. Consider the integral $\int_{\mathbb{S}^2} \varphi \, \mathrm{d}\Omega$ with Ω the uniform distribution on $\mathbb{S}^2$ and integrands

$$\varphi_3(x, y, z) = \cos(x + y + z), \quad \varphi_4(x, y, z) = \cos(x)\cos(y)\cos(z), \quad \varphi_5(x, y, z) = \exp(x - y), \tag{6}$$

so that $\mu(\varphi_3) = \mu(\varphi_4) = (4\pi/\sqrt{3})\sin(\sqrt{3})$ and $\mu(\varphi_5) = \pi\sqrt{8}\sinh(\sqrt{2})$. Figure 4 reports the evolution of the root mean squared error and the boxplots of the errors over the sphere $\mathbb{S}^2$. The error curves empirically validate the $\mathcal{O}(n^{-1})$ convergence rate for the control neighbors estimate while the boxplots highlight the variance reduction of the proposed method.



(a) RMSE φ_3 (b) RMSE φ_4 (c) RMSE φ_5

(d) Boxplot φ_3 (e) Boxplot φ_4 (f) Boxplot φ_5

Figure 4: Root mean squared errors (top) and boxplots of the errors (bottom) obtained over 100 replications for functions φ_3 (left), φ_4 (middle) and φ_5 (right) in Eq. (6) when integrating over $\mathbb{S}^2$.

4.3 Monte Carlo Integration for Optimal Transport

Optimal transport. Optimal transport (OT) is a mathematical framework for measuring the distance between probability distributions. It has proven to be particularly useful in machine learning applications, where it can be used for tasks such domain adaptation (Courty et al., 2017) and image generation (Gulrajani et al., 2017; Genevay et al., 2018). The recent development of efficient algorithms for solving optimal transport problems, e.g., the Sliced-Wasserstein (SW) distance (Rabin et al., 2012) and Sinkhorn distance (Cuturi, 2013), has made it a practical tool for large-scale machine learning applications.

OT distances. For $\mathsf{X} \subseteq \mathbb{R}^q$, let $\mathcal{P}(\mathsf{X})$ denote the set of probability measures supported on X and let $p \in [1, \infty)$. The Wasserstein distance of order p between $P, Q \in \mathcal{P}(\mathsf{X})$ is

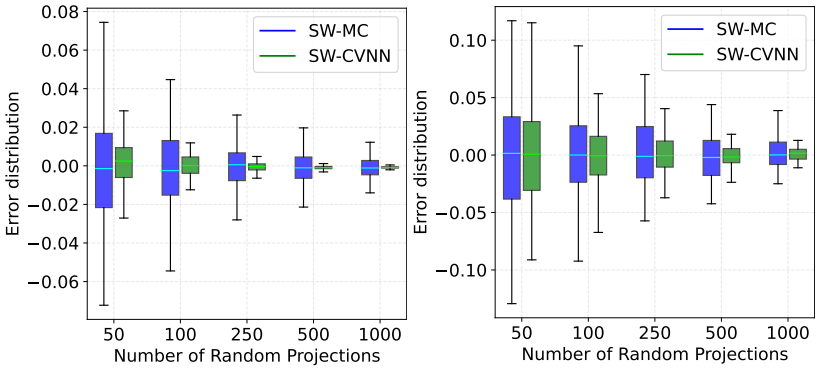
$$W_p^p(P, Q) = \inf_{\pi \in \Pi(P,Q)} \int_{\mathsf{X}\times\mathsf{X}} \|x - y\|^p \, \mathrm{d}\pi(x, y),$$

where $\Pi(P, Q) \subset \mathcal{P}(\mathsf{X} \times \mathsf{X})$ denotes the set of couplings for (P, Q), i.e., probability measures whose marginals with respect to the first and second variables are P and Q respectively. While the Wasserstein distance enjoys attractive theoretical properties (Villani, 2009, Chapter 6), it suffers from a high computational cost. When computing $W_p(P_m, Q_m)$ for discrete distributions P_m and Q_m supported on m points, the worst-case computational complexity scales as $\mathcal{O}(m^3 \log m)$ (Peyré et al., 2019). To overcome this issue, the Sliced-Wasserstein distance takes advantage of the fast computation of the Wasserstein distance between univariate distributions $P, Q \in \mathcal{P}(\mathbb{R})$. Indeed, for $P_m = (1/m)\sum_{i=1}^m \delta_{x_i}$ and $Q_m = (1/m)\sum_{i=1}^m \delta_{y_i}$ with $x_i, y_i \in \mathbb{R}$, the W_p-distance involves sorting the atoms $x_{(1)} \leq \ldots \leq x_{(m)}$ and $y_{(1)} \leq \ldots \leq y_{(m)}$, yielding

$$W_p^p(P_m, Q_m) = \frac{1}{m} \sum_{i=1}^m |x_{(i)} - y_{(i)}|^p,$$

leading to a complexity of $\mathcal{O}(m \log m)$ operations induced by the sorting step.

Recall that $\mathbb{S}^{q-1} = \{\theta \in \mathbb{R}^q : \|\theta\| = 1\}$ is the unit sphere in $\mathbb{R}^q$ and for $\theta \in \mathbb{S}^{q-1}$ let $\theta^\star : \mathbb{R}^q \to \mathbb{R}$ denote the linear map $\theta^\star(x) = \langle \theta, x \rangle$ for $x \in \mathbb{R}^q$. Let μ denote the uniform distribution on $\mathbb{S}^{q-1}$,



(a) $SW_2,\ q = 3$ (b) $SW_2,\ q = 6$

Figure 5: Boxplots of Sliced-Wasserstein estimates SW-MC and SW-CVNN for Gaussian distributions on $\mathbb{R}^q$ with $q \in \{3; 6\}$. The boxplots are obtained over 100 replications.

that is, the normalized volume measure. The Sliced-Wasserstein (SW) distance (Rabin et al., 2012; Bonneel et al., 2015; Kolouri et al., 2019) of order p based on μ is defined for $P, Q \in \mathcal{P}(\mathsf{X})$ as

$$SW_p^p(P, Q, \rho) = \mathbb{E}_\theta[W_p^p(\theta^\star_\# P, \theta^\star_\# Q)] = \int_{\mathbb{S}^{q-1}} W_p^p(\theta^\star_\# P, \theta^\star_\# Q)\, \mathrm{d}\mu(\theta), \tag{7}$$

where $f_\# \xi$ is the *push-forward* measure of $\xi \in \mathcal{P}(\mathbb{R}^q)$ by a measurable function f on $\mathbb{R}^q$. For $\theta \in \mathbb{S}^{q-1}$, the measures $\theta^\star_\# P$ and $\theta^\star_\# Q$ are the distributions of the projections $\langle \theta, X \rangle$ and $\langle \theta, Y \rangle$ of random vectors X and Y with distributions P and Q, respectively. The integrand is Lipschitz in view of Theorem 2.4 in Han (2023).

In practice, the random directions $\theta_1, \ldots, \theta_n$ are sampled independently from the uniform distribution on the unit sphere $\mathbb{S}^{q-1}$ and the SW-distance of Eq. (7) is approximated using a standard Monte Carlo estimate with n random projections as

$$\widehat{SW}_p^p(P_m, Q_m) = \frac{1}{n} \sum_{i=1}^{n} W_p^p\left((\theta_i)^\star_\# P_m, (\theta_i)^\star_\# Q_m\right).$$

The use of more accurate numerical integration methods to compute the SW-distance is an active area of current research; see, e.g., Nguyen and Ho (2024) and Leluc et al. (2024). Below, we illustrate the benefits of using control neighbors, while at the same time we anticipate that more specialized methods may do even better for this particular integration problem.

Multivariate Gaussians. The goal is to compare the variance of the standard Monte Carlo estimate (SW-MC) with the proposed control neighbors estimate (SW-CVNN) when computing the Sliced-Wasserstein distance between two Gaussian distributions. More precisely, we want to compute $SW_2(P, Q)$, where $P = \mathcal{N}_q(m_X, \sigma_X^2 \mathrm{I}_q)$ and $Q = \mathcal{N}_q(m_Y, \sigma_Y^2 \mathrm{I}_q)$ with $m_X, m_Y \sim \mathcal{N}(0, \mathrm{I}_q)$ and $\sigma_X = 2$ and $\sigma_Y = 5$. We consider the corresponding empirical distributions P_m and Q_m based on $m = 2\,000$ samples and compute the Monte Carlo estimates of the SW_2 distance using a number of projections $n \in \{50; 100; 250; 500; 1000\}$ in dimension $q \in \{3; 6\}$. In the notation of the general theory, we have $d = q - 1 \in \{2; 5\}$ since we are integrating over the $(q-1)$-dimensional sphere in $\mathbb{R}^q$. Figure 5 shows the error distribution of the different Monte Carlo estimates (SW-MC and SW-CVNN) of the SW_2 distance where the exact value (see Nadjahi et al., 2021, Appendix S3.1) is given by

$$SW_2^2\left(\mathcal{N}_q(m_X, \sigma_X^2 \mathrm{I}_q), \mathcal{N}_q(m_Y, \sigma_Y^2 \mathrm{I}_q)\right) = q^{-1} \|m_X - m_Y\|_2^2 + (\sigma_X - \sigma_Y)^2.$$

The different boxplots highlight the good performance of the control neighbors estimate in terms of variance reduction.

5 Discussion

We have explored the use of nearest neighbors in the construction of control variates for variance reduction in Monte Carlo integration. We have shown that for Hölder integrands of regularity

$s \in (0, 1]$ on bounded metric spaces of dimension d as measured by a sufficiently regular probability distribution, a faster rate of convergence, namely $\mathcal{O}(n^{-1/2}n^{-s/d})$ as $n \to \infty$, is possible through the construction of a control variate via leave-one-out neighbors. Theoretical guarantees are given both in terms of bounds on the root mean squared error and as concentration inequalities (requiring an additional logarithmic factor). In numerical experiments, the method enjoyed a notable error reduction with respect to Monte Carlo integration.

A drawback is that our method is not able to leverage higher-order smoothness of the integrand besides the Lipschitz property: this comes from the fact that the control variate is piecewise constant on the Voronoi cells induced by the Monte Carlo sample. As a consequence, the accuracy gain is limited on high-dimensional domains. Further, integration of the control function requires an additional Monte Carlo sample, of larger size than the original one, although without needing additional evaluations of the integrand.

Acknowledgments

Aigerim Zhuman gratefully acknowledges a research grant from the *National Bank of Belgium* and of the Research Council of the UCLouvain.

References

Juliette Achddou, Joseph Lam-Weil, Alexandra Carpentier, and Gilles Blanchard. A minimax near-optimal algorithm for adaptive rejection sampling. In *Algorithmic Learning Theory*, pages 94–126. PMLR, 2019.

J-Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Statist.*, 35(2): 608–633, 2007.

Nikolai Sergeevich Bakhvalov. On the approximate calculation of multiple integrals. *J. Complex.*, 31 (4):502–516, 2015.

Rémi Bardenet and Adrien Hardy. Monte carlo with determinantal point processes. *Ann. Appl. Probab.*, 30(1):368–417, 2020.

Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *CACM*, 18 (9):509–517, 1975.

Gérard Biau and Luc Devroye. *Lectures on the Nearest Neighbor Method*. Springer International Publishing, 2015.

Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of political economy*, 81(3):637–654, 1973.

Jose Blanchet, Haoxuan Chen, Yiping Lu, and Lexing Ying. When can regression-adjusted control variates help? In *Adv. Neural Inf. Process. Syst.*, volume 36, pages 36566–36578, 2023.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein barycenters of measures. *J. Math. Imaging Vision*, 51:22–45, 2015.

Russel E Caflisch. Monte Carlo and quasi-Monte Carlo methods. *Acta Numer.*, 7:1–49, 1998.

Nicolas Chopin and Mathieu Gerber. Higher-order stochastic integration through cubic stratification. *arXiv preprint arXiv:2210.01554*, 2022.

Jean-François Coeurjolly, Adrien Mazoyer, and Pierre-Olivier Amblard. Monte Carlo integration of non-differentiable functions on $[0, 1]^\iota$, $\iota = 1, \ldots, d$, using a single determinantal point pattern defined on $[0, 1]^d$. *Electron. J. Statist.*, 15(2):6228–6280, 2021.

Richard Combes. An extension of McDiarmid's inequality. *arXiv preprint arXiv:1511.05240*, 2015.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Adv. Neural Inf. Process. Syst.*, 30:3733–3742, 2017.

Peter Craven and Grace Wahba. Smoothing noisy data with spline functions. *Numer. Math.*, 31(4): 377–403, 1978.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Adv. Neural Inf. Process. Syst.*, volume 26, pages 2292–2300, 2013.

Persi Diaconis and Steven N. Evans. Linear functionals of eigenvalues of random matrices. *Trans. Amer. Math. Soc.*, 353(7):2615–2633, 2001.

Arnaud Doucet, Nando De Freitas, Neil James Gordon, et al. *Sequential Monte Carlo Methods in Practice*. Number 2. Springer, New York, 2001.

Michael Evans and Tim Swartz. *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford Statistical Science Series. Oxford University Press, Oxford, 2000.

Jerome H Friedman, Jon Louis Bentley, and Raphael Ari Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Math. Softw.*, 3(3):209–226, 1977.

Sébastien Gadat, Thierry Klein, and Clément Marteau. Classification in general finite dimensional spaces with the k-nearest neighbor rule. *Ann. Statist.*, 44(3):982–1009, 2016.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In *Int. Conf. Artif. Intell. Stat.*, pages 1608–1617. PMLR, 2018.

Paul Glasserman. *Monte Carlo Methods in Financial Engineering*. Springer, New York, 2004.

Peter W Glynn and Roberto Szechtman. Some new perspectives on the method of control variates. In *Monte Carlo and Quasi-Monte Carlo Methods 2000: Proceedings of a Conference held at Hong Kong Baptist University, Hong Kong SAR, China, November 27–December 1, 2000*, pages 27–49. Springer, 2002.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. In *Adv. Neural Inf. Process. Syst.*, volume 30, page 5769–5779, 2017.

Seymour Haber. A modified Monte-Carlo quadrature. *Math. Comp.*, 20(95):361–368, 1966.

Seymour Haber. A modified Monte-Carlo quadrature. II. *Math. Comp.*, 21(99):388–397, 1967.

Seymour Haber. Stochastic quadrature formulas. *Math. Comp.*, 23(108):751–764, 1969.

Ruiyu Han. Sliced Wasserstein distance between probability measures on Hilbert spaces, 2023. arXiv preprint arXiv:2307.05802.

Steven L Heston. A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The review of financial studies*, 6(2):327–343, 1993.

Dave Higdon, Jordan D McDonnell, Nicolas Schunck, Jason Sarich, and Stefan M Wild. A Bayesian approach for parameter estimation and prediction using a computationally intensive model. *J. Phys. G Nucl. Part. Phys.*, 42(3):034009, 2015.

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.

Kenneth E Hoff III, John Keyser, Ming Lin, Dinesh Manocha, and Tim Culver. Fast computation of generalized Voronoi diagrams using graphics hardware. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 277–286, 1999.

John D Hunter. Matplotlib: A 2d graphics environment. *Comput. Sci. Eng.*, 9(03):90–95, 2007.

Ryosuke Ichida. The injectivity radius and the fundamental group of compact homogeneous riemannian manifolds of positive curvature. *Tsukuba J. Math.*, 24(1):139–156, 2000.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Trans. Big Data*, 7(3):535–547, 2019.

GA Kabatyanskiı and VI Levenshteın. Bounds for packings on a sphere and in space. *Problems of Information Transmission*, 95:148–158, 1974.

Michael Kohler and Adam Krzyżak. Optimal global rates of convergence for interpolation problems with random design. *Statist. & Probabil. Lett.*, 83(8):1871–1879, 2013.

Michael Kohler, Adam Krzyżak, and Harro Walk. Rates of convergence for partitioning and nearest neighbor regression estimates with unbounded data. *J. Multivariate Anal.*, 97(2):311–323, 2006.

Gerasim Kokarev. Berger's inequality in the presence of upper sectional curvature bound. *Int. Math. Res. Not.*, 2022(15):11262–11303, 2021.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced Wasserstein distances. *Adv. Neural Inf. Process. Syst.*, 32:261–272, 2019.

David Krieg and Erich Novak. A universal algorithm for multivariate integration. *Found. Comput. Math.*, 17(4):895–916, 2017.

David Krieg, Erich Novak, and Mathias Sonnleitner. Recovery of Sobolev functions restricted to iid sampling. *Math. Comp.*, 91(338):2715–2738, 2022.

J.M. Lee. *Introduction to Riemannian Manifolds*. Graduate Texts in Mathematics. Springer International Publishing, 2019.

R. Leluc, F. Portier, and J. Segers. Control variate selection for Monte Carlo integration. *Stat. Comput.*, 31:50, 07 2021.

Rémi Leluc, Aymeric Dieuleveut, François Portier, Johan Segers, and Aigerim Zhuman. Sliced-wasserstein estimation with spherical harmonics as control variates. *arXiv preprint arXiv:2402.01493*, 2024.

Robert C Merton. Theory of rational option pricing. *The Bell Journal of economics and management science*, pages 141–183, 1973.

Francesco Mezzadri. How to generate random matrices from the classical compact groups. *Not. Am. Math. Soc.*, 54(5):592–604, 2007.

Kimia Nadjahi, Alain Durmus, Pierre E Jacob, Roland Badeau, and Umut Simsekli. Fast approximation of the sliced-wasserstein distance using concentration of random projections. *Adv. Neural Inf. Process. Syst.*, 34:12411–12424, 2021.

Nigel J Newton. Variance reduction for simulated diffusions. *SIAM J. Appl. Math.*, 54(6):1780–1805, 1994.

Khai Nguyen and Nhat Ho. Sliced Wasserstein estimation with control variates, 2024. arXiv preprint arXiv:2307.05802.

Erich Novak. *Deterministic and Stochastic Error Bounds in Numerical Analysis*. Springer, Berlin, Heidelberg, 1988.

Erich Novak. Some results on the complexity of numerical integration. In *Monte Carlo and Quasi-Monte Carlo Methods*, pages 161–183. Springer, 2016.

Chris J Oates, Mark Girolami, and Nicolas Chopin. Control functionals for Monte Carlo integration. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(3):695–718, 2017.

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.

Stephen M Omohundro. *Five balltree construction algorithms*. International Computer Science Institute Berkeley, 1989.

L. Pastur and V. Vasilchuk. On the moments of traces of matrices of classical groups. *Comm. Math. Phys.*, 252(1-3):149–166, 2004.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Found. Trends Mach. Learn.*, 11(5-6):355–607, 2019.

F. Portier and J. Segers. Monte Carlo integration with a growing number of control variates. *J. Appl. Probab.*, 56(4):1168–1186, 2019.

François Portier and Johan Segers. On the weak convergence of the empirical conditional copula under a simplifying assumption. *J. Multivariate Anal.*, 166:160–181, 2018.

François Portier. Nearest neighbor process: weak convergence and non-asymptotic bound. *arXiv preprint arXiv:2110.15083*, 2023.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision: Third International Conference, SSVM 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers 3*, pages 435–446. Springer, 2012.

Frederic M Richards. The interpretation of protein structures: total volume, group volume distributions and packing density. *J. Mol. Biol.*, 82(1):1–14, 1974.

Reuven Y. Rubinstein. *Simulation and the Monte Carlo method*. John Wiley & Sons, Inc., New York, 1981.

Chris Rycroft. Voro++: A three-dimensional Voronoi cell library in C++. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2009.

Jerome Sacks, William J Welch, Toby J Mitchell, and Henry P Wynn. Design and analysis of computer experiments. *Statist. Sci.*, 4(4):409–423, 1989.

LF South, CJ Oates, A Mira, and C Drovandi. Regularized zero-variance control variates. *Bayesian Anal.*, 1(1):1–24, 2022.

Elias M. Stein and Rami Shakarchi. *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*. Princeton University Press, Princeton, 2005.

Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 36(2):111–133, 1974.

Saul Toscano-Palmerin and Peter I Frazier. Bayesian optimization with expensive integrands. *SIAM J. Optim.*, 32(2):417–444, 2022.

Cédric Villani. *Optimal Transport: Old and New*. Springer, Berlin, Heidelberg, 2009.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*, 17(3):261–272, 2020.

Kenneth Zeger and Allen Gersho. Number of nearest neighbors in a Euclidean code. *IEEE Transactions on Information Theory*, 40(5):1647–1649, 1994.

Ralf Zimmermann and Jakob Stoye. High curvature means low-rank: On the sectional curvature of grassmann and stiefel manifolds and the underlying matrix trace inequalities. *arXiv preprint arXiv:2403.01879*, 2024.

APPENDIX:
SPEEDING UP MONTE CARLO INTEGRATION: CONTROL NEIGHBORS FOR OPTIMAL CONVERGENCE

Appendix A contains some auxiliary results. Appendix B gathers the proofs of all the lemmas, while Appendix C is concerned with the proofs of the different propositions. Appendix D comprises the proofs of the theorems and Appendix E presents additional numerical experiments for option pricing.

A Auxiliary results **19**

B Proofs of Lemmas **20**

B.1 Proof of Lemma 1 20

B.2 Proof of Lemma 2 20

B.3 Proof of Lemma 3 21

B.4 Proof of Lemma 7 22

C Proofs of Propositions **22**

C.1 Proof of Proposition 1 22

C.2 Proof of Proposition 2 22

C.3 Proof of Proposition 3 23

D Proofs of Theorems **24**

D.1 Proof of Theorem 1 for $\hat{\mu}_n^{(\mathrm{NN-loo})}$ 24

D.2 Proof of Theorem 1 for $\hat{\mu}_n^{(\mathrm{NN})}$ 27

D.3 Proof of Theorem 2 for $\hat{\mu}_n^{(\mathrm{NN-loo})}$ 27

D.4 Proof of Theorem 2 for $\hat{\mu}_n^{(\mathrm{NN})}$ 31

D.5 Proof of Theorem 3 33

E Additional Experiments **33**

A Auxiliary results

The following theorem is a corollary to Theorem 2.1 in Combes (2015).

Theorem 3 (Extension of McDiarmid's inequality)**.** *Let $X_1, \dots, X_n$ independent random variables with X_ℓ taking values in some measurable space Ω_ℓ. Define $X = (X_1, \dots, X_n) \in \prod_{\ell=1}^n \Omega_\ell =: \Omega$. Let $\phi : \Omega \to \mathbb{R}$ be a measurable function and assume there exists an event $A \subseteq \Omega$ and constants $c_1, \dots, c_n \geq 0$ such that, for any $\ell \in \{1, \dots, n\}$ and for any $x \in A$ and $x' \in A$ that differ only in the ℓ-th coordinate, we have*

$$|\phi(x) - \phi(x')| \leq c_\ell. \tag{8}$$

Then, for any $t \geq 0$, we have, writing $m = \mathbb{E}[\phi \mid A]$, $p = 1 - \mathbb{P}(A)$ and $\bar{c} = \sum_{\ell=1}^n c_\ell$,

$$\mathbb{P}\left(|\phi(X) - m| \geq t\right) \leq p + 2\exp\left(-\frac{2\max(0, t - p\bar{c})^2}{\sum_{\ell=1}^n c_\ell^2}\right).$$

Lemma 4 (Kabatyanskiı and Levenshteın (1974); Zeger and Gersho (1994))**.** *Given a point set $P \subset \mathbb{R}^d$ and $x \in P$, the maximum number of points in P that can have x as nearest neighbor is bounded by the d-dimensional kissing number ψ_d, where $\psi_d \leq 2^{0.401d(1+\mathrm{o}(1))}$ as $d \to \infty$.*

Proposition 3. *Under Assumptions (A1), (A2) and (A3) in the paper, if $n \geq 4$, we have*

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n \mu(\hat{\varphi}_n^{(i)}) - \mu(\hat{\varphi}_n)\right|\right] \leq (s/d + 2)\, L C_0^{-s/d}\, \Gamma(s/d + 1)\, n^{-1-s/d}$$

and

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\mu(\hat{\varphi}_n^{(i)})-\mu(\hat{\varphi}_n)\right|^2\right]\leq 4\,(s/d+1)\,L^2C_0^{-2s/d}\,\Gamma(2s/d+1)\,n^{-1-2s/d}.$$

Lemma 5 (Portier (2023))**.** *Under Assumptions (A1') and (A2) in the paper, for all $n\geq 1$, all $\delta\in(0,1)$ and all $k\in\{1,\dots,n\}$ such that $8d\log(12n/\delta)\leq k\leq r_0^d nbcV_d/2$, it holds, with probability at least $1-\delta$, that*

$$\sup_{x\in\mathcal{X}}\hat{\tau}_{n,k}(x)\leq\left(\frac{2k}{nbcV_d}\right)^{1/d}=:\bar{\tau}_{n,k}.$$

Lemma 6 (Combes (2015))**.** *Let $\phi:\Omega\to\mathbb{R}$ be a bounded, measurable function and let $A\subseteq\Omega$ be an event with $\mathbb{P}(A)>0$. Writing $m=\mathbb{E}[\phi\mid A]$, $p=1-\mathbb{P}(A)$ and $\|\phi\|_\infty=\sup_{x\in\Omega}|\phi(x)|$, we have*

$$|\mathbb{E}[\phi]-m|\leq 2p\,\|\phi\|_\infty\,.$$

Lemma 7. *For $x,a>0$, we have $\Gamma(x+1+a)>x^a\Gamma(x+1)$.*

B Proofs of Lemmas

B.1 Proof of Lemma 1

Given any collection $X_1,\dots,X_n$ of n distinct points, if $j\neq i$, then $\hat{\varphi}_n^{(i)}$ and $\hat{\varphi}_n$ are the same on $S_{n,j}$. It holds that

$$\hat{\varphi}_n^{(i)}(x)-\hat{\varphi}_n(x)=\left\{\hat{\varphi}_n^{(i)}(x)-\hat{\varphi}_n(x)\right\}\mathbb{1}_{S_{n,i}}(x).$$

Now using that $\bar{\varphi}_n$ and $\hat{\varphi}_n^{(i)}$ are the same on $S_{n,i}$, it follows that

$$\hat{\varphi}_n^{(i)}(x)-\hat{\varphi}_n(x)=\left\{\bar{\varphi}_n(x)-\hat{\varphi}_n(x)\right\}\mathbb{1}_{S_{n,i}}(x).$$

Taking the sum and using $\sum_{i=1}^n\mathbb{1}_{S_{n,i}}(x)=1$ gives

$$\sum_{i=1}^{n}\left\{\hat{\varphi}_n^{(i)}(x)-\hat{\varphi}_n(x)\right\}=\bar{\varphi}_n(x)-\hat{\varphi}_n(x),$$

and the result follows by integrating with respect to μ.

B.2 Proof of Lemma 2

Because the Voronoi cells define a partition of M, we have for any $x\in M$,

$$\hat{\varphi}_n^{(i)}(x)=\sum_{j:j\neq i}\varphi(X_j)\,\mathbb{1}_{S_{n,j}^{(i)}}(x)$$

and in particular

$$\hat{\varphi}_n^{(i)}(X_i)=\sum_{j:j\neq i}\varphi(X_j)\,\mathbb{1}_{S_{n,j}^{(i)}}(X_i),$$

from which we deduce

$$\sum_{i=1}^{n}\hat{\varphi}_n^{(i)}(X_i)=\sum_{j=1}^{n}\varphi(X_j)\sum_{i:i\neq j}\mathbb{1}_{S_{n,j}^{(i)}}(X_i)=\sum_{j=1}^{n}\varphi(X_j)\,\hat{d}_{n,j}.$$

Further, we have

$$\mu(\hat{\varphi}_n)=\sum_{i=1}^{n}\mu\left(\varphi(X_i)\,\mathbb{1}_{S_{n,i}}\right)=\sum_{i=1}^{n}\varphi(X_i)\,\mu(S_{n,i})=\sum_{i=1}^{n}\varphi(X_i)\,V_{n,i}$$

and

$$\sum_{i=1}^{n}\mu\big(\hat{\varphi}_n^{(i)}\big)=\sum_{i=1}^{n}\sum_{j:j\neq i}\mu\left(\varphi(X_j)\,\mathbb{1}_{S_{n,j}^{(i)}}\right)=\sum_{j=1}^{n}\varphi(X_j)\sum_{i:i\neq j}V_{n,j}^{(i)}=\sum_{j=1}^{n}\varphi(X_j)\,\hat{c}_{n,j}.$$

B.3 Proof of Lemma 3

For $x \in M$, $r \in (0, \operatorname{diam}(M)]$ and integer $k \in \{1, \dots, n\}$, we have

$$\begin{aligned}
\mathbb{P}\left[\hat{\tau}_{n,k}(x) > r\right] &= \mathbb{P}\left[\sum_{i=1}^{n} \mathbb{1}_{\{\rho(X_i,x)\le r\}} \le k-1\right] \\
&= \mathbb{P}\left[\operatorname{Bin}(n, \mu(B(x,r)) \le k-1\right] \\
&\le \mathbb{P}\left[\operatorname{Bin}(n, C_0 r^d) \le k-1\right] \\
&= \sum_{j=0}^{k-1} \binom{n}{j} (C_0 r^d)^j (1 - C_0 r^d)^{n-j}.
\end{aligned}$$

The equality remains valid for $r > 0$ such that $C_0 r^d \le 1$, since if $r > \operatorname{diam}(M)$, the probability is zero. We find

$$\begin{aligned}
\mathbb{E}\left[\hat{\tau}_{n,k}(x)^q\right] &= \int_0^\infty \mathbb{P}\left[\hat{\tau}_{n,k}(x) > t^{1/q}\right] \mathrm{d}t \\
&\le \sum_{j=0}^{k-1} \binom{n}{j} \int_0^{C_0^{-q/d}} \left(C_0 t^{d/q}\right)^j \left(1 - C_0 t^{d/q}\right)^{n-j} \mathrm{d}t \\
&= \sum_{j=0}^{k-1} \binom{n}{j} \int_0^1 u^j (1-u)^{n-j} \cdot \frac{q}{d} C_0^{-q/d} u^{q/d-1} \,\mathrm{d}u \\
&= \frac{q}{d} C_0^{-q/d} \sum_{j=0}^{k-1} \binom{n}{j} \frac{\Gamma(j+q/d)\Gamma(n-j+1)}{\Gamma(n+q/d+1)},
\end{aligned}$$

since the integral can be expressed in terms of the Beta function, which can in turn be developed in terms of the Gamma function. Writing the binomial coefficient in terms of the Gamma function too and simplifying yields

$$\mathbb{E}\left[\hat{\tau}_{n,k}(x)^q\right] \le \frac{q}{d} C_0^{-q/d} \frac{\Gamma(n+1)}{\Gamma(n+q/d+1)} \sum_{j=0}^{k-1} \frac{\Gamma(j+q/d)}{\Gamma(j+1)} \le \frac{q}{d} \left(C_0 n\right)^{-q/d} \sum_{j=0}^{k-1} \frac{\Gamma(j+q/d)}{\Gamma(j+1)},$$

where we used Lemma 7 in the second inequality. By induction, one can obtain

$$\frac{q}{d} \sum_{j=0}^{k-1} \frac{\Gamma(j+q/d)}{\Gamma(j+1)} = \frac{\Gamma(k+q/d)}{\Gamma(k)}. \tag{9}$$

For $k = 1$, (9) reduces to

$$\frac{(q/d)\Gamma(q/d)}{\Gamma(1)} = \frac{\Gamma(q/d+1)}{\Gamma(1)}.$$

Assume that (9) holds for $k-1$, i.e.,

$$\frac{q}{d} \sum_{j=0}^{k-2} \frac{\Gamma(j+q/d)}{\Gamma(j+1)} = \frac{\Gamma(k-1+q/d)}{\Gamma(k-1)}. \tag{10}$$

Then, by (10) and $\Gamma(k) = (k-1)\Gamma(k-1)$, we have

$$\begin{aligned}
\frac{q}{d} \sum_{j=0}^{k-1} \frac{\Gamma(j+q/d)}{\Gamma(j+1)} &= \frac{q}{d} \sum_{j=0}^{k-2} \frac{\Gamma(j+q/d)}{\Gamma(j+1)} + \frac{q}{d} \frac{\Gamma(k-1+q/d)}{\Gamma(k)} \\
&= \frac{\Gamma(k-1+q/d)}{\Gamma(k-1)} + \frac{q}{d} \frac{\Gamma(k-1+q/d)}{\Gamma(k)} \\
&= \frac{(k-1+q/d)\Gamma(k-1+q/d)}{\Gamma(k)} \\
&= \frac{\Gamma(k+q/d)}{\Gamma(k)}.
\end{aligned}$$

Therefore,

$$\mathbb{E}\left[\hat{\tau}_{n,k}(x)^q\right] \leq (C_0 n)^{-q/d} \frac{\Gamma(k+q/d)}{\Gamma(k)}.$$

B.4 Proof of Lemma 7

The Gamma function is strictly logarithmically convex. Writing

$$x+1 = (1-\lambda)x + \lambda(x+1+a) \text{ with } \lambda = \frac{1}{1+a}$$

yields

$$\Gamma(x+1) < \Gamma(x)^{1-\lambda}\,\Gamma(x+1+a)^{\lambda}.$$

Since $\Gamma(x) = x^{-1}\Gamma(x+1)$, it follows that

$\Gamma(x+1+a) > \left\{\Gamma(x+1)\Gamma(x)^{\lambda-1}\right\}^{1/\lambda} = \left\{\Gamma(x+1)^{1+\lambda-1}x^{1-\lambda}\right\}^{1/\lambda} = \Gamma(x+1)x^{1/\lambda-1} = \Gamma(x+1)x^a$, as required.

C Proofs of Propositions

C.1 Proof of Proposition 1

Using Lemma 2, we find

$$\begin{aligned}
\hat{\mu}_n^{(\mathrm{NN})}(\varphi) &= \frac{1}{n}\sum_{i=1}^{n}[\varphi(X_i) - \{\hat{\varphi}_n^{(i)}(X_i) - \mu(\hat{\varphi}_n)\}] \\
&= \frac{1}{n}\sum_{i=1}^{n}\varphi(X_i) - \frac{1}{n}\sum_{j=1}^{n}\varphi(X_j)\,\hat{d}_{n,j} + \sum_{i=1}^{n}\varphi(X_i)\,V_{n,i} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(1 - \hat{d}_{n,i} + nV_{n,i}\right)\varphi(X_i)
\end{aligned}$$

and, similarly,

$$\begin{aligned}
\hat{\mu}_n^{(\mathrm{NN-loo})} &= \frac{1}{n}\sum_{i=1}^{n}[\varphi(X_i) - \{\hat{\varphi}_n^{(i)}(X_i) - \mu(\hat{\varphi}_n^{(i)})\}] \\
&= \frac{1}{n}\sum_{i=1}^{n}\varphi(X_i) - \frac{1}{n}\sum_{j=1}^{n}\varphi(X_j)\hat{d}_{n,j} + \frac{1}{n}\sum_{j=1}^{n}\varphi(X_j)\hat{c}_{n,j} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(1 - \hat{d}_{n,i} + \hat{c}_{n,i}\right)\varphi(X_i),
\end{aligned}$$

as required.

C.2 Proof of Proposition 2

Let $N_{n,k}(x)$ denote the set of indices $i \in \{1,\ldots,n\}$ such that X_i is among the k nearest neighbors of $x \in M$. Using the Hölder property and the definition of $\hat{\tau}_{n,k}$ we have, for each $i \in N_{n,k}(x)$,

$$|\varphi(X_i) - \varphi(x)| \leq L\rho(X_i, x)^s \leq L\hat{\tau}_{n,k}(x)^s.$$

It follows that

$$|\hat{\varphi}_{n,k}(x) - \varphi(x)| \leq \frac{1}{k}\sum_{i\in N_{n,k}(x)}|\varphi(X_i) - \varphi(x)| \leq L\hat{\tau}_{n,k}(x)^s.$$

Next we apply Lemma 3 with $q = sp$ to obtain

$$\begin{aligned}
\mathbb{E}[|\hat{\varphi}_{n,k}(x) - \varphi(x)|^p] &\leq L^p\,\mathbb{E}[\hat{\tau}_{n,k}(x)^{sp}] \\
&\leq L^p (C_0 n)^{-sp/d}\frac{\Gamma(k+sp/d)}{\Gamma(k)}.
\end{aligned}$$

Raising both sides to the power $1/p$ yields the stated inequality.

C.3 Proof of Proposition 3

Using the fact that $\hat{\varphi}_n^{(i)}$ and $\hat{\varphi}_n$ coincide outside $S_{n,i}$ and that $\hat{\varphi}_n(x) = \varphi(X_i)$ for $x \in S_{n,i}$, we have

$$\begin{aligned}\mu(\hat{\varphi}_n^{(i)}) - \mu(\hat{\varphi}_n) &= \mu(\hat{\varphi}_n^{(i)} - \hat{\varphi}_n) \\ &= \mu\big((\hat{\varphi}_n^{(i)} - \hat{\varphi}_n)\, \mathbb{1}_{S_{n,i}}\big) \\ &= \mu\big(\{\hat{\varphi}_n^{(i)} - \varphi(X_i)\}\, \mathbb{1}_{S_{n,i}}\big).\end{aligned}$$

Write $R_n = \left|\frac{1}{n}\sum_{i=1}^n \{\mu(\hat{\varphi}_n^{(i)}) - \mu(\hat{\varphi}_n)\}\right|$. Then

$$\begin{aligned}R_n &\le \frac{1}{n}\sum_{i=1}^n \left|\mu(\hat{\varphi}_n^{(i)}) - \mu(\hat{\varphi}_n)\right| \\ &= \frac{1}{n}\sum_{i=1}^n \left|\mu\big(\{\hat{\varphi}_n^{(i)} - \varphi(X_i)\}\, \mathbb{1}_{S_{n,i}}\big)\right| \\ &\le \frac{1}{n}\sum_{i=1}^n \int_{S_{n,i}} \left|\hat{\varphi}_n^{(i)}(x) - \varphi(X_i)\right| \,\mathrm{d}\mu(x) \\ &= \frac{1}{n}\sum_{i=1}^n \int_{S_{n,i}} \left|\varphi\big(\hat{N}_n^{(i)}(x)\big) - \varphi(X_i)\right| \,\mathrm{d}\mu(x).\end{aligned}$$

Further, since φ is Hölder continuous with exponent $s \in (0, 1]$, we have, by the triangle inequality and the sub-additivity of the function $x \mapsto x^s$,

$$\begin{aligned}R_n &\le \frac{L}{n}\sum_{i=1}^n \int_{S_{n,i}} \rho\big(\hat{N}_n^{(i)}(x), X_i\big)^s \,\mathrm{d}\mu(x) \\ &\le \frac{L}{n}\sum_{i=1}^n \int_{S_{n,i}} \Big\{\rho\big(\hat{N}_n^{(i)}(x), x\big) + \rho(x, X_i)\Big\}^s \,\mathrm{d}\mu(x) \\ &\le \frac{L}{n}\sum_{i=1}^n \int_{S_{n,i}} \Big\{\rho\big(\hat{N}_n^{(i)}(x), x\big)^s + \rho(x, X_i)^s\Big\} \,\mathrm{d}\mu(x) \\ &= \frac{L}{n}\sum_{i=1}^n \int_{S_{n,i}} \Big\{\hat{\tau}_n^{(i)}(x)^s + \hat{\tau}_n(x)^s\Big\} \,\mathrm{d}\mu(x).\end{aligned}$$

Concerning R_n^2, one can use Jensen's inequality to obtain

$$\begin{aligned}R_n^2 &= \left|\frac{1}{n}\sum_{i=1}^n \Big\{\mu(\hat{\varphi}_n^{(i)}) - \mu(\hat{\varphi}_n)\Big\}\right|^2 \\ &\le \frac{1}{n}\sum_{i=1}^n \left|\mu(\hat{\varphi}_n^{(i)}) - \mu(\hat{\varphi}_n)\right|^2 \\ &\le \frac{L^2}{n}\sum_{i=1}^n \int_{S_{n,i}} \Big\{\rho\big(\hat{N}_n^{(i)}(x), x\big)^s + \rho(x, X_i)^s\Big\}^2 \,\mathrm{d}\mu(x) \\ &\le \frac{2L^2}{n}\sum_{i=1}^n \int_{S_{n,i}} \Big\{\rho\big(\hat{N}_n^{(i)}(x), x\big)^{2s} + \rho(x, X_i)^{2s}\Big\} \,\mathrm{d}\mu(x).\end{aligned}$$

For $x \in S_{n,i}$, the nearest neighbor in $\{X_1, \ldots, X_n\}$ is X_i, and thus

$$\hat{\tau}_n^{(i)}(x) = \hat{\tau}_{n,2}(x)$$

is the distance to the second nearest neighbor in $\{X_1, \ldots, X_n\}$. We get

$$\begin{aligned}R_n &\le \frac{L}{n}\sum_{i=1}^n \int_{S_{n,i}} \{\hat{\tau}_{n,2}(x)^s + \hat{\tau}_n(x)^s\} \,\mathrm{d}\mu(x) \\ &= \frac{L}{n}\int_M \{\hat{\tau}_{n,2}(x)^s + \hat{\tau}_n(x)^s\} \,\mathrm{d}\mu(x)\end{aligned}$$

and, similarly,

$$
\begin{aligned}
R_n^2 &\leq \frac{2L^2}{n} \sum_{i=1}^{n} \int_{S_{n,i}} \left\{ \hat{\tau}_{n,2}(x)^{2s} + \hat{\tau}_n(x)^{2s} \right\} \mathrm{d}\mu(x) \\
&= \frac{2L^2}{n} \int_M \left\{ \hat{\tau}_{n,2}(x)^{2s} + \hat{\tau}_n(x)^{2s} \right\} \mathrm{d}\mu(x).
\end{aligned}
$$

Consequently, by Lemma 3,

$$
\begin{aligned}
\mathbb{E}[R_n] &\leq \frac{L}{n} \int_M \mathbb{E}\left[\hat{\tau}_{n,2}(x)^s + \hat{\tau}_n(x)^s\right] \mathrm{d}\mu(x) \\
&\leq \frac{L}{n} \left(\sup_{x \in M} \mathbb{E}\left[\hat{\tau}_{n,2}(x)^s\right] + \sup_{x \in M} \mathbb{E}\left[\hat{\tau}_n(x)^s\right] \right) \\
&\leq \frac{L}{n} \left\{ (C_0 n)^{-s/d} \Gamma(s/d+2) + (C_0 n)^{-s/d} \Gamma(s/d+1) \right\} \\
&= (s/d+2)\, L C_0^{-s/d} \Gamma(s/d+1)\, n^{-1-s/d}
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}[R_n^2] &\leq \frac{2L^2}{n} \int_M \mathbb{E}\left[\hat{\tau}_{n,2}(x)^{2s} + \hat{\tau}_n(x)^{2s}\right] \mathrm{d}\mu(x) \\
&\leq \frac{2L^2}{n} \left(\sup_{x \in M} \mathbb{E}\left[\hat{\tau}_{n,2}(x)^{2s}\right] + \sup_{x \in M} \mathbb{E}\left[\hat{\tau}_n(x)^{2s}\right] \right) \\
&\leq \frac{2L^2}{n} \left\{ (C_0 n)^{-2s/d} \Gamma(2s/d+2) + (C_0 n)^{-2s/d} \Gamma(2s/d+1) \right\} \\
&= 4\,(s/d+1)\, L^2 C_0^{-2s/d} \Gamma(2s/d+1)\, n^{-1-2s/d}.
\end{aligned}
$$

D Proofs of Theorems

D.1 Proof of Theorem 1 for $\hat{\mu}_n^{(\mathrm{NN-loo})}$

Let $Y_{n,i} = \hat{\varphi}_n^{(i)}(X_i) - \mu(\hat{\varphi}_n^{(i)})$ and write

$$
\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi) - \mu(\varphi) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - Y_{n,i})
$$

with $Y_i = \varphi(X_i) - \mu(\varphi)$. Then write

$$
\begin{aligned}
n^2\, \mathbb{E}\left[\left\{ \hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi) - \mu(\varphi) \right\}^2 \right] &= \sum_{i=1}^{n} \mathbb{E}\left[(Y_i - Y_{n,i})^2\right] + \sum_{i \neq j} \mathbb{E}\left[(Y_i - Y_{n,i})(Y_j - Y_{n,j})\right] \\
&= n\, \mathbb{E}\left[(Y_1 - Y_{n,1})^2\right] + n(n-1)\, \mathbb{E}\left[(Y_1 - Y_{n,1})(Y_2 - Y_{n,2})\right].
\end{aligned} \tag{11}
$$

We decompose $Y_{n,1}$ into two terms, one of which does not depend on X_2. We also use the fact that the Voronoi partition made with $(n-1)$ points is more detailed than the one constructed with $(n-2)$ points, i.e., $S_{n-1,i}^{(1)} \subset S_{n-2,i}^{(1,2)}$ for $i = 3, \ldots, n$. Define the map $\mathcal{N}^{(1,2)} : M \to M$ such that $\mathcal{N}^{(1,2)}(x)$ is the nearest neighbor to x among the sample $\{X_3, X_4, \ldots, X_n\}$. Using the identity

$\mathcal{N}^{(1,2)}(x) = X_i$ whenever $x \in S^{(1)}_{n-1,i}$ for $i \geq 3$, we write

$$\begin{aligned}
\hat{\varphi}_n^{(1,2)}(x) &= \varphi\big(\mathcal{N}^{(1,2)}(x)\big) \\
&= \varphi\big(\mathcal{N}^{(1,2)}(x)\big) \left(\sum_{i=2}^{n} \mathbb{1}_{S^{(1)}_{n-1,i}}(x) \right) \\
&= \varphi\big(\mathcal{N}^{(1,2)}(x)\big)\, \mathbb{1}_{S^{(1)}_{n-1,2}}(x) + \sum_{i=3}^{n} \varphi(X_i)\, \mathbb{1}_{S^{(1)}_{n-1,i}}(x) \\
&= \Big\{ \varphi\big(\mathcal{N}^{(1,2)}(x)\big) - \varphi(X_2) \Big\}\, \mathbb{1}_{S^{(1)}_{n-1,2}}(x) + \sum_{i=2}^{n} \varphi(X_i)\, \mathbb{1}_{S^{(1)}_{n-1,i}}(x).
\end{aligned}$$

It follows that

$$\hat{\varphi}_n^{(1)}(x) = \hat{L}^{(1)}(x) + \hat{\varphi}_n^{(1,2)}(x)$$

with $\hat{L}^{(1)}(x) = \big\{ \varphi(X_2) - \varphi\big(\mathcal{N}^{(1,2)}(x)\big) \big\}\, \mathbb{1}_{S^{(1)}_{n-1,2}}(x)$. Therefore, since $Y_{n,1} = \hat{\varphi}_n^{(1)}(X_1) - \mu(\hat{\varphi}_n^{(1)})$, we get

$$Y_1 - Y_{n,1} = Y_1 - \Big\{ \hat{L}^{(1)}(X_1) - \mu(\hat{L}^{(1)}) \Big\} - \Big\{ \hat{\varphi}_n^{(1,2)}(X_1) - \mu(\hat{\varphi}_n^{(1,2)}) \Big\}.$$

Write

$$\begin{aligned}
A_1 &= Y_1 = \varphi(X_1) - \mu(\varphi), \\
A_2 &= Y_2 = \varphi(X_2) - \mu(\varphi), \\
B_1 &= \hat{L}^{(1)}(X_1) - \mu(\hat{L}^{(1)}), \\
B_2 &= \hat{L}^{(2)}(X_2) - \mu(\hat{L}^{(2)}), \\
C_1 &= \hat{\varphi}_n^{(1,2)}(X_1) - \mu(\hat{\varphi}_n^{(1,2)}), \\
C_2 &= \hat{\varphi}_n^{(1,2)}(X_2) - \mu(\hat{\varphi}_n^{(1,2)}),
\end{aligned}$$

where $\hat{L}^{(2)}(x) = \big\{ \varphi(X_1) - \varphi\big(\mathcal{N}^{(1,2)}(x)\big) \big\}\, \mathbb{1}_{S^{(2)}_{n-1,1}}(x)$. Then

$$\begin{aligned}
\mathbb{E}\left[(Y_1 - Y_{n,1})\,(Y_2 - Y_{n,2}) \right] &= \mathbb{E}\left[(A_1 - B_1 - C_1)\,(A_2 - B_2 - C_2) \right] \\
&= \mathbb{E}[A_1 A_2] - \mathbb{E}[A_1 B_2] - \mathbb{E}[A_1 C_2] \\
&\quad - \mathbb{E}[B_1 A_2] + \mathbb{E}[B_1 B_2] + \mathbb{E}[B_1 C_2] \\
&\quad - \mathbb{E}[C_1 A_2] + \mathbb{E}[C_1 B_2] + \mathbb{E}[C_1 C_2].
\end{aligned}$$

Since A_1 and A_2 are independent, $\mathbb{E}[A_1 A_2] = 0$. This also applies to $\mathbb{E}[A_1 C_2]$ and $\mathbb{E}[C_1 A_2]$. Considering $\mathbb{E}[A_1 B_2]$ gives

$$\begin{aligned}
\mathbb{E}[A_1 B_2] &= \mathbb{E}\left[Y_1 \left\{ \hat{L}^{(2)}(X_2) - \mu(\hat{L}^{(2)}) \right\} \right] \\
&= \mathbb{E}\left[\mathbb{E}\left[Y_1 \left\{ \hat{L}^{(2)}(X_2) - \mu(\hat{L}^{(2)}) \right\} \mid X_1, X_3, \ldots, X_n \right] \right] \\
&= \mathbb{E}\left[Y_1\, \mathbb{E}\left[\hat{L}^{(2)}(X_2) - \mu(\hat{L}^{(2)}) \mid X_1, X_3, \ldots, X_n \right] \right] = 0.
\end{aligned}$$

Due to similar reasoning, $\mathbb{E}[B_1 A_2] = 0$, $\mathbb{E}[B_1 C_2] = 0$ and $\mathbb{E}[C_1 B_2] = 0$. For $\mathbb{E}[C_1 C_2]$, we have

$$\begin{aligned}
\mathbb{E}[C_1 C_2] &= \mathbb{E}\left[\Big\{ \hat{\varphi}_n^{(1,2)}(X_1) - \mu\big(\hat{\varphi}_n^{(1,2)}\big) \Big\} \Big\{ \hat{\varphi}_n^{(1,2)}(X_2) - \mu\big(\hat{\varphi}_n^{(1,2)}\big) \Big\} \right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\Big\{ \hat{\varphi}_n^{(1,2)}(X_1) - \mu\big(\hat{\varphi}_n^{(1,2)}\big) \Big\} \Big\{ \hat{\varphi}_n^{(1,2)}(X_2) - \mu\big(\hat{\varphi}_n^{(1,2)}\big) \Big\} \mid X_3, \ldots, X_n \right] \right] = 0.
\end{aligned}$$

Therefore, we get

$$\mathbb{E}\left[(Y_1 - Y_{n,1})\,(Y_2 - Y_{n,2}) \right] = \mathbb{E}[B_1 B_2] = \mathbb{E}\left[\Big\{ \hat{L}^{(1)}(X_1) - \mu(\hat{L}^{(1)}) \Big\} \Big\{ \hat{L}^{(2)}(X_2) - \mu(\hat{L}^{(2)}) \Big\} \right].$$

Write the four terms on the right-hand side as A, B, C, D, respectively. The Cauchy–Schwarz inequality gives $\|(A - B)(C - D)\|_1 \leq \|A - B\|_2\, \|C - D\|_2$, while the fact that B and D are

conditional expectations of A and C, respectively, leads to $\|(A-B)(C-D)\|_1 \leq \|A\|_2 \|C\|_2 = \|A\|_2^2$. As a result,

$$\mathbb{E}\left[(Y_1 - Y_{n,1})(Y_2 - Y_{n,2})\right] \leq \mathbb{E}\left[\hat{L}^{(1)}(X_1)^2\right].$$

Using the Hölder property, we obtain, by the triangle inequality and the sub-additivity of $x \mapsto x^s$,

$$\begin{aligned}
\left|\hat{L}^{(1)}(x)\right| &= \left|\varphi(X_2) - \varphi\big(\mathcal{N}^{(1,2)}(x)\big)\right| \mathbb{1}_{S^{(1)}_{n-1,2}}(x) \\
&= \left|\varphi\big(\mathcal{N}^{(1)}(x)\big) - \varphi\big(\mathcal{N}^{(1,2)}(x)\big)\right| \mathbb{1}_{S^{(1)}_{n-1,2}}(x) \\
&\leq L\, \rho\big(\mathcal{N}^{(1)}(x), \mathcal{N}^{(1,2)}(x)\big)^s \, \mathbb{1}_{S^{(1)}_{n-1,2}}(x) \\
&\leq 2L\, \rho\big(x, \mathcal{N}^{(1,2)}(x)\big)^s \, \mathbb{1}_{S^{(1)}_{n-1,2}}(x) \\
&\leq 2L\, \rho\big(x, \mathcal{N}^{(1,2)}(x)\big)^s \, \mathbb{1}_{B(x,\hat{\tau}^{(1)}(x))}(X_2) \\
&\leq 2L\, \rho\big(x, \mathcal{N}^{(1,2)}(x)\big)^s \, \mathbb{1}_{B(x,\hat{\tau}^{(1,2)}(x))}(X_2).
\end{aligned}$$

Hence

$$\mathbb{E}\left[\left|\hat{L}^{(1)}(x)\right|^2 \mid X_3, \ldots, X_n\right] \leq 4L^2 \rho\big(x, \mathcal{N}^{(1,2)}(x)\big)^{2s} \mu\left(B\big(x, \hat{\tau}^{(1,2)}(x)\big)\right).$$

Moreover,

$$\mu\left(B\big(x, \hat{\tau}^{(1,2)}(x)\big)\right) \leq C_1 \hat{\tau}^{(1,2)}(x)^d.$$

Using the equality $\hat{\tau}^{(1,2)}(x) = \rho\big(x, \mathcal{N}^{(1,2)}(x)\big)$, we obtain

$$\mathbb{E}\left[\left|\hat{L}^{(1)}(x)\right|^2 \mid X_3, \ldots, X_n\right] \leq 4C_1 L^2\, \rho\big(x, \mathcal{N}^{(1,2)}(x)\big)^{2s+d}.$$

Applying to the term

$$(Y_1 - Y_{n,1})^2 = \left[\varphi(X_1) - \hat{\varphi}_n^{(1)}(X_1) - \left\{\mu(\varphi) - \mu(\hat{\varphi}_n^{(1)})\right\}\right]^2$$

the same reasoning as above with $A = C = \varphi(X_1) - \hat{\varphi}_n^{(1)}(X_1)$ and $B = D = \mu(\varphi) - \mu(\hat{\varphi}_n^{(1)})$, we get

$$\begin{aligned}
\mathbb{E}\left[(Y_1 - Y_{n,1})^2\right] &\leq \mathbb{E}\left[\left\{\varphi(X_1) - \hat{\varphi}_n^{(1)}(X_1)\right\}^2\right] \\
&= \mathbb{E}\left[\left\{\varphi(X_1) - \varphi\big(\mathcal{N}^{(1)}(X_1)\big)\right\}^2\right] \\
&\leq L^2\, \mathbb{E}\left[\rho\big(X_1, \mathcal{N}^{(1)}(X_1)\big)^{2s}\right].
\end{aligned}$$

Combining all this with (11) gives

$$\begin{aligned}
&\mathbb{E}\left[\left|\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi) - \mu(\varphi)\right|^2\right] \\
&\leq L^2 n^{-1}\, \mathbb{E}\left[\rho\big(X_1, \mathcal{N}^{(1)}(X_1)\big)^{2s}\right] + 4C_1 L^2\, \mathbb{E}\left[\rho\big(X_1, \mathcal{N}^{(1,2)}(X_1)\big)^{2s+d}\right] \\
&= L^2 n^{-1}\, \mathbb{E}[\hat{\tau}_{n-1}(X_1)^{2s}] + 4C_1 L^2\, \mathbb{E}[\hat{\tau}_{n-2}(X_1)^{2s+d}],
\end{aligned}$$

where X_1 is understood to be independent of the nearest neighbor distance functions $\hat{\tau}_{n-1}$ (based on $X_2, \ldots, X_n$) and $\hat{\tau}_{n-2}$ (based on $X_3, \ldots, X_n$). Applying Lemma 3 to $\mathbb{E}[\hat{\tau}_{n-1}(X_1)^{2s} \mid X_1]$ and to $\mathbb{E}[\hat{\tau}_{n-2}(X_1)^{2s+d} \mid X_1]$, we get

$$\begin{aligned}
\mathbb{E}\left[\hat{\tau}_{n-1}(X_1)^{2s}\right] &\leq \{(n-1)C_0\}^{-2s/d}\, \Gamma(2s/d+1), \\
\mathbb{E}\left[\hat{\tau}_{n-2}(X_1)^{2s+d}\right] &\leq \{(n-2)C_0\}^{-2s/d-1}\, \Gamma(2s/d+2).
\end{aligned}$$

Therefore,

$$\mathbb{E}\left[\left|\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi)-\mu(\varphi)\right|^2\right] \leq L^2 n^{-1}\left\{(n-1)C_0\right\}^{-2s/d}\Gamma(2s/d+1) + 4C_1L^2\left\{(n-2)C_0\right\}^{-2s/d-1}\Gamma(2s/d+2)$$

and thus

$$n^{1+2s/d}\,\mathbb{E}\left[\left|\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi)-\mu(\varphi)\right|^2\right] \leq L^2 C_0^{-2s/d}\Gamma(2s/d+1)\left\{n/(n-1)\right\}^{2s/d} + 4C_1L^2C_0^{-2s/d-1}\Gamma(2s/d+2)\left\{n/(n-2)\right\}^{2s/d+1}.$$

Since $n/(n-2) \leq 2$ (from $n \geq 4$), the upper bound can be simplified to

$$5L^2C_1C_0^{-2s/d-1}\Gamma(2s/d+2)2^{2s/d+1}.$$

D.2 Proof of Theorem 1 for $\hat{\mu}_n^{(\mathrm{NN})}$

The proof follows from combining the previous part of Theorem 1 and the following inequality from Proposition 3,

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n \mu(\hat{\varphi}_n^{(i)})-\mu(\hat{\varphi}_n)\right|^2\right] \leq 4\,(s/d+1)\,L^2C_0^{-2s/d}\,\Gamma(2s/d+1)\,n^{-1-2s/d}.$$

By Minkowski's inequality, we have

$$\left(\mathbb{E}\left[\left|\hat{\mu}_n^{(\mathrm{NN})}(\varphi)-\mu(\varphi)\right|^2\right]\right)^{1/2} \leq \left(\mathbb{E}\left[\left|\hat{\mu}_n^{(\mathrm{NN})}(\varphi)-\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi)\right|^2\right]\right)^{1/2} + \left(\mathbb{E}\left[\left|\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi)-\mu(\varphi)\right|^2\right]\right)^{1/2}.$$

In view of (3), $\mathbb{E}\left[\left|\hat{\mu}_n^{(\mathrm{NN})}(\varphi)-\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi)\right|^2\right] = \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n \mu(\hat{\varphi}_n^{(i)})-\mu(\hat{\varphi}_n)\right|^2\right]$. Therefore,

$$\left(\mathbb{E}\left[\left|\hat{\mu}_n^{(\mathrm{NN})}(\varphi)-\mu(\varphi)\right|^2\right]\right)^{1/2} \leq \sqrt{L^2C_0^{-2s/d}\Gamma(2s/d+2)n^{-1-2s/d}}\left(2\sqrt{s/d+1}+2^{s/d}(C_1/C_0)^{1/2}\sqrt{10}\right).$$

D.3 Proof of Theorem 2 for $\hat{\mu}_n^{(\mathrm{NN-loo})}$

We will apply Theorem 3, showing the bounded difference property in two parts (Step 2). First, in Step 1, we construct a large-probability event A on which the bounded difference property will hold. In order to bound the gap between $\mathbb{E}[\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi) \mid A]$ and $\mu(\varphi)$, we rely in Step 3 on the identity $\mathbb{E}[\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi)] = \mu(\varphi)$ and on Lemma 6.

Step 1. Let $\lceil x \rceil$ denote the smallest integer upper bound to $x \in \mathbb{R}$. By Lemma 5, there exists an event A with probability $\mathbb{P}(A) \geq 1-\delta_n$ such that on A, we have, for $k = \lceil 8d\log(12n/\delta_n)\rceil$,

$$\sup_{x\in\mathcal{X}} \hat{\tau}_{n,k}(x) \leq \left(\frac{2k}{nbcV_d}\right)^{1/d} = \left(\frac{2\lceil 8d\log(12n/\delta_n)\rceil}{nbcV_d}\right)^{1/d} \leq \left(\frac{17d\log(12n/\delta_n)}{nbcV_d}\right)^{1/d} = \bar{\tau}_n.$$

Step 2. On the event A, we consider separately two terms of $\hat{\mu}_n^{(\mathrm{NN-loo})} = n^{-1}\sum_{i=1}^n\{\varphi(X_i) - \hat{\varphi}_n^{(i)}(X_i) + \mu(\hat{\varphi}_n^{(i)})\}$, namely $n^{-1}\sum_{i=1}^n\{\varphi(X_i)-\hat{\varphi}_n^{(i)}(X_i)\}$ and $n^{-1}\sum_{i=1}^n \mu(\hat{\varphi}_n^{(i)})$, in order to make the bounded differences property (8) for $\hat{\mu}_n^{(\mathrm{NN-loo})}$ satisfied. Further, we apply Theorem 3.

Step 2a. Fix $\ell \in \{1, \ldots, n\}$. In the original set of points $X_1, \ldots, X_n$, replace X_ℓ by $\tilde{X}_\ell$. Let $\hat{\tilde{N}}_n^{(i)}(x)$ be the nearest neighbor of x among $X_1, \ldots, \tilde{X}_\ell, \ldots, X_n$ without X_i and define $\hat{\tilde{\varphi}}_n^{(i)}(x) = \varphi(\hat{\tilde{N}}_n^{(i)}(x))$. Note that $\hat{\tilde{N}}_n^{(\ell)}(x) = \hat{N}_n^{(\ell)}(x)$ and hence $\hat{\tilde{\varphi}}_n^{(\ell)}(x) = \hat{\varphi}_n^{(\ell)}(x)$. Let

$$\phi_1 \equiv \phi_1(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} \{\varphi(X_i) - \hat{\varphi}_n^{(i)}(X_i)\}. \tag{12}$$

Since $\hat{\tilde{\varphi}}_n^{(\ell)}(\tilde{X}_\ell) = \hat{\varphi}_n^{(\ell)}(\tilde{X}_\ell)$, we have

$$\begin{aligned}
D_{\ell,1} &= \phi_1(X_1, \ldots, X_\ell, \ldots, X_n) - \phi_1(X_1, \ldots, \tilde{X}_\ell, \ldots, X_n) \\
&= \frac{1}{n} \sum_{\substack{i=1 \\ i \neq \ell}}^{n} \{\varphi(X_i) - \hat{\varphi}_n^{(i)}(X_i)\} + \frac{1}{n} \{\varphi(X_\ell) - \hat{\varphi}_n^{(\ell)}(X_\ell)\} \\
&\quad - \frac{1}{n} \sum_{\substack{i=1 \\ i \neq \ell}}^{n} \{\varphi(X_i) - \hat{\tilde{\varphi}}_n^{(i)}(X_i)\} - \frac{1}{n} \{\varphi(\tilde{X}_\ell) - \hat{\tilde{\varphi}}_n^{(\ell)}(\tilde{X}_\ell)\} \\
&= \frac{1}{n} \sum_{\substack{i=1 \\ i \neq \ell}}^{n} \{\hat{\tilde{\varphi}}_n^{(i)}(X_i) - \hat{\varphi}_n^{(i)}(X_i)\} + \frac{1}{n} \left[\{\varphi(X_\ell) - \varphi(\tilde{X}_\ell)\} - \{\hat{\varphi}_n^{(\ell)}(X_\ell) - \hat{\varphi}_n^{(\ell)}(\tilde{X}_\ell)\} \right].
\end{aligned}$$

Considering the first term of $D_{\ell,1}$, we have, in view of (A3),

$$\begin{aligned}
\frac{1}{n} \sum_{\substack{i=1 \\ i \neq \ell}}^{n} \left| \hat{\tilde{\varphi}}_n^{(i)}(X_i) - \hat{\varphi}_n^{(i)}(X_i) \right| &= \frac{1}{n} \sum_{\substack{i=1 \\ i \neq \ell}}^{n} \left| \varphi(\hat{\tilde{N}}_n^{(i)}(X_i)) - \varphi(\hat{N}_n^{(i)}(X_i)) \right| \\
&\leq \frac{L}{n} \sum_{\substack{i=1 \\ i \neq \ell}}^{n} \left\| \hat{\tilde{N}}_n^{(i)}(X_i) - \hat{N}_n^{(i)}(X_i) \right\|^s \\
&= \frac{L}{n} \sum_{\substack{i=1 \\ i \neq \ell}}^{n} \left\| \hat{\tilde{N}}_n^{(i)}(X_i) - \hat{N}_n^{(i)}(X_i) \right\|^s \mathbb{1}_{\{\hat{\tilde{N}}_n^{(i)}(X_i) = \tilde{X}_\ell \text{ or } \hat{N}_n^{(i)}(X_i) = X_\ell\}}.
\end{aligned}$$

By the triangle inequality and the sub-additivity of $x \mapsto x^s$, we get, in hopefully obvious notation,

$$\begin{aligned}
\left\| \hat{\tilde{N}}_n^{(i)}(X_i) - \hat{N}_n^{(i)}(X_i) \right\|^s &\leq \left\| X_i - \hat{\tilde{N}}_n^{(i)}(X_i) \right\|^s + \left\| X_i - \hat{N}_n^{(i)}(X_i) \right\|^s \\
&\leq \left\| X_i - \hat{\tilde{N}}_n^{(i,\ell)}(X_i) \right\|^s + \left\| X_i - \hat{N}_n^{(i,\ell)}(X_i) \right\|^s \\
&\leq 2 \left\| X_i - \hat{N}_n^{(i,\ell)}(X_i) \right\|^s \leq 2 \sup_{x \in \mathcal{X}} \hat{\tau}_{n,3}(x)^s,
\end{aligned} \tag{13}$$

where $\hat{\tau}_{n,3}(x)$ is the distance to the third nearest neighbor of x among $X_1, \ldots, X_n$. By Lemma 4,

$$\sum_{\substack{i=1 \\ i \neq \ell}}^{n} \mathbb{1}_{\{\hat{\tilde{N}}_n^{(i)}(X_i) = \tilde{X}_\ell \text{ or } \hat{N}_n^{(i)}(X_i) = X_\ell\}} \leq 2\psi_d. \tag{14}$$

Therefore, by (13) and (14),

$$\frac{1}{n} \sum_{\substack{i=1 \\ i \neq \ell}}^{n} \left| \hat{\tilde{\varphi}}_n^{(i)}(X_i) - \hat{\varphi}_n^{(i)}(X_i) \right| \leq \frac{4L\psi_d}{n} \sup_{x \in \mathcal{X}} \hat{\tau}_{n,3}(x)^s. \tag{15}$$

Considering the second term of $D_{\ell,1}$, we have

$$\begin{aligned}
&\frac{1}{n} \left| \{\varphi(X_\ell) - \varphi(\tilde{X}_\ell)\} - \{\hat{\varphi}_n^{(\ell)}(X_\ell) - \hat{\varphi}_n^{(\ell)}(\tilde{X}_\ell)\} \right| \\
&\leq \frac{L}{n} \left[\left\| X_\ell - \hat{N}_n^{(\ell)}(X_\ell) \right\|^s + \left\| \tilde{X}_\ell - \hat{N}_n^{(\ell)}(\tilde{X}_\ell) \right\|^s \right] \\
&\leq \frac{2L}{n} \sup_{x \in \mathcal{X}} \hat{\tau}_{n,3}(x)^s.
\end{aligned} \tag{16}$$

On the event A, by (15) and (16), we therefore have, since $k \geq 3$ for our choice of k in Step 1, the bound

$$|D_{\ell,1}| \leq \frac{2L}{n}(2\psi_d + 1)\bar{\tau}_n^s.$$

Step 2b. Let $\phi_2 \equiv \phi_2(X_1, \ldots, X_n) = n^{-1}\sum_{i=1}^n \mu(\hat{\varphi}_n^{(i)})$. Therefore,

$$\begin{aligned} D_{l,2} &= \phi_2(X_1, \ldots, X_\ell, \ldots, X_n) - \phi_2(X_1, \ldots, \tilde{X}_\ell, \ldots, X_n) \\ &= \frac{1}{n}\sum_{i=1}^n \{\mu(\hat{\varphi}_n^{(i)}) - \mu(\hat{\tilde{\varphi}}_n^{(i)})\} = \frac{1}{n}\sum_{i=1}^n \mu(\hat{\varphi}_n^{(i)} - \hat{\tilde{\varphi}}_n^{(i)}). \end{aligned}$$

Let $S_{n+1,j}$, for $j = 1, \ldots, n+1$, be the Voronoi cells induced by $\mathcal{X}_{n+1} = \{X_1, \ldots, X_n, \tilde{X}_\ell\}$: for $x \in S_{n+1,j}$ and $j = 1, \ldots, n+1$, the nearest neighbor of x among $\mathcal{X}_{n+1}$ is X_j, where $X_{n+1} := \tilde{X}_\ell$. Clearly,

$$\forall j \in \{1, \ldots, n+1\} \setminus \{i, \ell, n+1\},\ \forall x \in S_{n+1,j}, \qquad \hat{\tilde{\varphi}}_n^{(i)}(x) = \varphi(X_j) = \hat{\varphi}_n^{(i)}(x).$$

It follows that

$$D_{\ell,2} = \frac{1}{n}\sum_{i=1}^n \mu\left((\hat{\varphi}_n^{(i)} - \hat{\tilde{\varphi}}_n^{(i)})\,\mathbb{1}_{S_{n+1,i} \cup S_{n+1,\ell} \cup S_{n+1,n+1}}\right).$$

Moreover, by the triangle inequality, the sub-additivity of $x \mapsto x^s$ and (A3), we have, for $x \in S_{n+1,i} \cup S_{n+1,\ell} \cup S_{n+1,n+1}$,

$$\begin{aligned} \left|\hat{\tilde{\varphi}}_n^{(i)}(x) - \hat{\varphi}_n^{(i)}(x)\right| &\leq \left|\hat{\tilde{\varphi}}_n^{(i)}(x) - \varphi(x)\right| + \left|\hat{\varphi}_n^{(i)}(x) - g(x)\right| \\ &\leq L \cdot \left\|\hat{\tilde{N}}_n^{(i)}(x) - x\right\|^s + L \cdot \left\|\hat{N}_n^{(i)}(x) - x\right\|^s \\ &\leq 2L \sup_{x \in \mathcal{X}} \hat{\tau}_{n,3}(x)^s. \end{aligned}$$

Clearly, $\sum_{i=1}^n \mu(S_{n+1,i}) = \mu\left(\bigcup_{i=1}^n S_{n+1,i}\right) \leq 1$. On the event A, we thus obtain

$$\begin{aligned} |D_{\ell,2}| &\leq \frac{1}{n}\sum_{i=1}^n \mu\left(\left|\hat{\varphi}_n^{(i)} - \hat{\tilde{\varphi}}_n^{(i)}\right| \mathbb{1}_{S_{n+1,i} \cup S_{n+1,\ell} \cup S_{n+1,n+1}}\right) \\ &\leq 2L\bar{\tau}_n^s \cdot \frac{1}{n}\sum_{i=1}^n \mu(S_{n+1,i} \cup S_{n+1,\ell} \cup S_{n+1,n+1}) \\ &\leq 2L\bar{\tau}_n^s \cdot \left\{n^{-1} + \mu(S_{n+1,\ell}) + \mu(S_{n+1,n+1})\right\}. \end{aligned}$$

Recall that λ_d denotes the d-dimensional Lebesgue-measure and that the density, f, of μ satisfies $\sup_{x \in M} f(x) \leq U$. For any $j = 1, \ldots, n+1$, the μ-volume of a Voronoi cell satisfies

$$\mu(S_{n+1,j}) = \int_{S_{n+1,j}} f \,\mathrm{d}\lambda_d \leq U \cdot \lambda_d(S_{n+1,j} \cap M).$$

Letting $\hat{\tau}_{n+1,1}(x)$ denote the nearest neighbor of $x \in M$ in $\mathcal{X}_{n+1}$, we have

$$\forall x \in S_{n+1,j} \cap M, \qquad \|x - X_j\| = \hat{\tau}_{n+1,1}(x) \leq \sup_{x' \in M} \hat{\tau}_{n+1,1}(x').$$

On the event A, the latter is bounded by $\bar{\tau}_n$, which implies that $S_{n+1,j} \cap M$ is contained in a ball of radius $\bar{\tau}_n$ centered at X_j. We find

$$\max_{j=1,\ldots,n+1} \mu(S_{n+1,j}) \leq U \cdot \bar{\tau}_n^d V_d, \tag{17}$$

with V_d the volume of the unit ball in $\mathbb{R}^d$. We conclude that, on the event A,

$$|D_{\ell,2}| \leq 2L\bar{\tau}_n^s \cdot \left(n^{-1} + 2UV_d\bar{\tau}_n^d\right).$$

Step 2c. Let $\phi_3 = \phi_1 + \phi_2 = n^{-1}\sum_{i=1}^n \{\varphi(X_i) - \hat{\varphi}_n^{(i)}(X_i) + \mu(\hat{\varphi}_n^{(i)})\} = \hat{\mu}_n^{(\mathrm{NN-loo})}$. Then we get, with probability at least $1-\delta_n$,

$$\begin{aligned}
D_{\ell,3} &= \left|\phi_3(X_1,\ldots,X_\ell,\ldots,X_n) - \phi_3(X_1,\ldots,\tilde{X}_\ell,\ldots,X_n)\right| \le |D_{\ell,1}| + |D_{\ell,2}| \\
&\le 2Ln^{-1}(2\psi_d+1)\bar{\tau}_n^s + 2L\bar{\tau}_n^s \cdot \left(n^{-1} + 2UV_d\bar{\tau}_n^d\right) \\
&= 2Ln^{-1}(2\psi_d+2)\bar{\tau}_n^s + 4LUV_d\bar{\tau}_n^{s+d} \\
&= 4L(\psi_d+1)n^{-1}\left(\frac{17d\log(12n/\delta_n)}{nbcV_d}\right)^{s/d} + 4LUV_d\left(\frac{17d\log(12n/\delta_n)}{nbcV_d}\right)^{1+s/d} \\
&\le K_1\left(\frac{\log(12n/\delta_n)}{n}\right)^{1+s/d} = c_{\ell,3} \quad \text{with} \quad K_1 = 4L\left(\frac{17d}{bcV_d}\right)^{s/d}\left(\psi_d + 1 + \frac{17dU}{bc}\right).
\end{aligned}$$

We have

$$\begin{aligned}
\sum_{\ell=1}^n c_{\ell,3} &= nc_{\ell,3} = K_1 n^{-s/d}(\log(12n/\delta_n))^{1+s/d}, \\
\sum_{\ell=1}^n c_{\ell,3}^2 &= nc_{\ell,3}^2 = K_1^2 n^{-1-2s/d}(\log(12n/\delta_n))^{2+2s/d}.
\end{aligned}$$

We apply Theorem 3 to $\phi_3 = \phi_3(X_1,\ldots,X_n) = n^{-1}\sum_{i=1}^n\{\varphi(X_i) - \hat{\varphi}_n^{(i)}(X_i) + \mu(\hat{\varphi}_n^{(i)})\}$. For $t \ge \delta_n \sum_{\ell=1}^n c_{\ell,3} = \delta_n K_1 n^{-s/d}(\log(12n/\delta_n))^{1+s/d}$, we get

$$\mathbb{P}\left(|\phi_3 - \mathbb{E}(\phi_3 \mid A)| \ge t\right) \le \delta_n + 2\exp\left(-\frac{2\left\{t - \delta_n K_1 n^{-s/d}(\log(12n/\delta_n))^{1+s/d}\right\}^2}{K_1^2 n^{-1-2s/d}(\log(12n/\delta_n))^{2+2s/d}}\right). \tag{18}$$

Let $0 < \delta < 1-\delta_n$. Then

$$\mathbb{P}(|\phi_3 - \mathbb{E}(\phi_3 \mid A)| \ge t) \le \delta + \delta_n$$

provided t is such that the exponential function in (18) is bounded by $\delta/2$, which happens if

$$\begin{aligned}
t &\ge \sqrt{\frac{K_1^2\log(2/\delta)(\log(12n/\delta_n))^{2+2s/d}}{2n^{1+2s/d}}} + \frac{K_1\delta_n(\log(12n/\delta_n))^{1+s/d}}{n^{s/d}} \\
&= K_1\left(n^{-1/2}\sqrt{\log(2/\delta)/2} + \delta_n\right)\frac{(\log(12n/\delta_n))^{1+s/d}}{n^{s/d}}.
\end{aligned}$$

Hence, for $0 < \delta \le 1-\delta_n$, with probability at least $1-(\delta+\delta_n)$, we have

$$|\phi_3 - \mathbb{E}(\phi_3 \mid A)| < K_1\left(n^{-1/2}\sqrt{\log(2/\delta)/2} + \delta_n\right)\frac{(\log(12n/\delta_n))^{1+s/d}}{n^{s/d}}$$

Step 3. We have $\|\phi_3\|_\infty \le 3\|\varphi\|_\infty$. By Lemma 6,

$$|\mathbb{E}(\phi_3) - \mathbb{E}(\phi_3 \mid A)| \le 2\delta_n\|\phi_3\|_\infty \le 6\delta_n\|\varphi\|_\infty.$$

Thus, we obtain, with probability at least $1-(\delta+\delta_n)$, for $0 < \delta \le 1-\delta_n$,

$$\begin{aligned}
\left|\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi) - \mu(\varphi)\right| &= |\phi_3 - \mathbb{E}(\phi_3)| \\
&\le |\phi_3 - \mathbb{E}(\phi_3 \mid A)| + |\mathbb{E}(\phi_3 \mid A) - \mathbb{E}(\phi_3)| \\
&\le K_1\left(n^{-1/2}\sqrt{\log(2/\delta)/2} + \delta_n\right)\frac{(\log(12n/\delta_n))^{1+s/d}}{n^{s/d}} + 6\delta_n\|\varphi\|_\infty.
\end{aligned}$$

Step 4. Fix $\varepsilon > 0$ and choose $\delta_n = \min(\varepsilon/2, n^{-1/2-s/d})$, $\delta = \varepsilon - \delta_n$. Then $\delta = \varepsilon - \min(\varepsilon/2, n^{-1/2-s/d})$

$= \max(\varepsilon/2, \varepsilon - n^{-1/2-s/d})$. Considering the case when $\varepsilon \le \frac{2}{n^{1/2+s/d}}$, we get

$$\begin{aligned}
&\left|\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi) - \mu(\varphi)\right| \\
&\le K_1 \left(\frac{\sqrt{\log(4/\varepsilon)/2}}{n^{1/2}} + \frac{\varepsilon}{2} \right) \frac{(\log(24n/\varepsilon))^{1+s/d}}{n^{s/d}} + 3\varepsilon \left\|\varphi\right\|_\infty \\
&\le K_1 \left(\frac{\sqrt{\log(4/\varepsilon)/2}}{n^{1/2}} + \frac{1}{n^{1/2+s/d}} \right) \frac{(\log(24n/\varepsilon))^{1+s/d}}{n^{s/d}} + \frac{6\left\|\varphi\right\|_\infty}{n^{1/2+s/d}} \\
&\le K_1 \left(\sqrt{\log\left(\frac{4}{\varepsilon}\right) \Big/ 2} + \frac{1}{n^{s/d}} \right) \frac{(\log(24n/\varepsilon))^{1+s/d}}{n^{1/2+s/d}} + \frac{6\left\|\varphi\right\|_\infty}{n^{1/2+s/d}}.
\end{aligned}$$

Otherwise, when $\varepsilon > \frac{2}{n^{1/2+s/d}}$, using that $\log(12n^{3/2+s/d}) \le \log((3n)^3)$, we have

$$\begin{aligned}
&\left|\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi) - \mu(\varphi)\right| \\
&\le K_1 \left(\sqrt{\log\left(\frac{2}{\varepsilon - n^{-1/2-s/d}}\right) \Big/ 2} + \frac{1}{n^{s/d}} \right) \frac{(\log(12n^{3/2+1/d}))^{1+s/d}}{n^{1/2+s/d}} + \frac{6\left\|\varphi\right\|_\infty}{n^{1/2+s/d}} \\
&\le K_1 \left(\sqrt{\log\left(\frac{4}{\varepsilon}\right) \Big/ 2} + \frac{1}{n^{s/d}} \right) \frac{(3\log(3n))^{1+s/d}}{n^{1/2+s/d}} + \frac{6\left\|\varphi\right\|_\infty}{n^{1/2+s/d}}.
\end{aligned}$$

Therefore, for any $\varepsilon \in (0,1)$, we have, with probability at least $1-\varepsilon$,

$$\begin{aligned}
&\left|\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi) - \mu(\varphi)\right| \le \frac{6\left\|\varphi\right\|_\infty}{n^{1/2+s/d}} \\
&+K_1 \left(\sqrt{\log\left(\frac{4}{\varepsilon}\right) \Big/ 2} + \frac{1}{n^{s/d}} \right) \times \begin{cases} \dfrac{(\log(24n/\varepsilon))^{1+s/d}}{n^{1/2+s/d}}, & \text{for } \varepsilon \le \frac{2}{n^{1/2+s/d}}, \\ \dfrac{(3\log(3n))^{1+s/d}}{n^{1/2+s/d}}, & \text{for } \varepsilon > \frac{2}{n^{1/2+s/d}}. \end{cases}
\end{aligned}$$

Since $\sum_{i=1}^n w_{n,i} = 1$, we can replace φ by $\varphi - C$ for any constant C and thus, choosing C optimally, replace $\|\varphi\|_\infty$ by half the diameter of the range of φ, that is, by $C_\varphi = \frac{1}{2}\{\sup_{x\in M}\varphi(x) - \inf_{x\in M}\varphi(x)\}$.

D.4 Proof of Theorem 2 for $\hat{\mu}_n^{(\mathrm{NN})}$

Consider the same event A and the same upper bound $\bar{\tau}_n$ as in Step 1 in the proof of Theorem 2 for $\hat{\mu}_n^{(\mathrm{NN-loo})}$. We write

$$\hat{\mu}_n^{(\mathrm{NN})}(\varphi) = \phi_5 \equiv \phi_5(X_1,\ldots,X_n) = \phi_1(X_1,\ldots,X_n) + \phi_4(X_1,\ldots,X_n)$$

with ϕ_1 as in Eq. (12) and with

$$\phi_4 \equiv \phi_4(X_1,\ldots,X_n) = \mu(\hat{\varphi}_n).$$

Step 1. Let $\hat{\tilde{N}}_n(x)$ be the nearest neigbor of x among $X_1,\ldots,\tilde{X}_\ell,\ldots,X_n$ and let $\hat{\tilde{\varphi}}_n(x) = g(\hat{\tilde{N}}_n(x))$. Then

$$\begin{aligned}
D_{\ell,4} &= \phi_4(X_1,\ldots,X_\ell,\ldots,X_n) - \phi_4(X_1,\ldots,\tilde{X}_\ell,\ldots,X_n) \\
&= \mu(\hat{\varphi}_n) - \mu(\hat{\tilde{\varphi}}_n) = \mu(\hat{\varphi}_n - \hat{\tilde{\varphi}}_n).
\end{aligned}$$

For $i = 1,\ldots,n+1$, define $S_{n+1,i}$ as in Step 2b in the proof of Theorem 2 for $\hat{\mu}_n^{(\mathrm{NN-loo})}$. It follows that

$$D_{\ell,4} = \mu\left((\hat{\varphi}_n - \hat{\tilde{\varphi}}_n)\, \mathbb{1}_{S_{n+1,\ell}\cup S_{n+1,n+1}} \right).$$

Further, by the triangle inequality, the sub-additivity of the function $x \mapsto x^s$ and the Hölder-s property, we have, for $x \in M \cap (S_{n+1,\ell} \cup S_{n+1,n+1})$,

$$\begin{aligned}\left|\hat{\varphi}_n(x) - \hat{\tilde{\varphi}}_n(x)\right| &\leq |\hat{\varphi}_n(x) - g(x)| + \left|\hat{\tilde{\varphi}}_n - g(x)\right| \\ &\leq L \cdot \left\|\hat{\tilde{N}}_n(x) - x\right\|^s + L \cdot \left\|\hat{N}_n(x) - x\right\|^s \\ &\leq 2L \sup_{x' \in M} \hat{\tau}_{n,2}(x')^s.\end{aligned}$$

On the event A, we thus obtain

$$|D_{\ell,4}| \leq 2L\bar{\tau}_n^s \cdot \mu(S_{n+1,\ell} \cup S_{n+1,n+1}) \leq 2L\bar{\tau}_n^s \cdot \{\mu(S_{n+1,\ell}) + \mu(S_{n+1,n+1})\}.$$

By Eq. (17) in Step 2b in the proof of Theorem 2 for $\hat{\mu}_n^{(\mathrm{NN-loo})}$, we have, on the event A,

$$\max_{j=1,\ldots,n+1} \mu(S_{n+1,j}) \leq U \cdot \bar{\tau}_n^d V_d,$$

with V_d the volume of the unit ball in $\mathbb{R}^d$. Therefore, we conclude that, on the event A, we have

$$|D_{\ell,4}| \leq 4LUV_d\bar{\tau}_n^{s+d}. \tag{19}$$

Step 2. By (19) and by the result of Step 2a in the proof of Theorem 2 for $\hat{\mu}_n^{(\mathrm{NN-loo})}$, we have

$$\begin{aligned}D_{\ell,5} &= \left|\phi_5(X_1,\ldots,X_\ell,\ldots,X_n) - \phi_5(X_1,\ldots,\tilde{X}_\ell,\ldots,X_5)\right| \\ &\leq |D_{\ell,1}| + |D_{\ell,4}| \\ &\leq 2Ln^{-1}(2\psi_d+1)\bar{\tau}_n^s + 4LUV_d\bar{\tau}_n^{s+d} \\ &= 2Ln^{-1}(2\psi_d+1)\left(\frac{17d\log(12n/\delta_n)}{nbcV_d}\right)^{s/d} + 4LUV_d\left(\frac{17d\log(12n/\delta_n)}{nbcV_d}\right)^{1+s/d} \\ &\leq K_2\left(\frac{\log(12n/\delta_n)}{n}\right)^{1+s/d} = c_{\ell,5} \quad \text{with} \quad K_2 = 2L\left(\frac{17d}{bcV_d}\right)^{s/d}\left(2\psi_d + 1 + \frac{34dU}{bc}\right).\end{aligned}$$

We have

$$\begin{aligned}\sum_{\ell=1}^n c_{\ell,5} &= nc_{\ell,5} = K_2 n^{-s/d}(\log(12n/\delta_n))^{1+s/d}, \\ \sum_{\ell=1}^n c_{\ell,5}^2 &= nc_{\ell,5}^2 = K_2^2 n^{-1-2s/d}(\log(12n/\delta_n))^{2+2s/d}.\end{aligned}$$

Thanks to Theorem 3, it follows that, for $t \geq \delta_n \sum_{\ell=1}^n c_{\ell,5} = \delta_n K_2 n^{-s/d}(\log(12n/\delta_n))^{1+s/d}$,

$$\mathbb{P}\left(|\phi_5 - \mathbb{E}(\phi_5 \mid A)| \geq t\right) \leq \delta_n + 2\exp\left(-\frac{2\left\{t - \delta_n K_2 n^{-s/d}(\log(12n/\delta_n))^{1+s/d}\right\}^2}{K_2^2 n^{-1-2s/d}(\log(12n/\delta_n))^{2+2s/d}}\right).$$

Let $0 < \delta \leq 1 - \delta_n$. Then

$$\mathbb{P}\left(|\phi_5 - \mathbb{E}(\phi_5 \mid A)| \geq t\right) \leq \delta + \delta_n$$

provided

$$t \geq K_2\left(n^{-1/2}\sqrt{\log(2/\delta)/2} + \delta_n\right)\frac{(\log(12n/\delta_n))^{1+s/d}}{n^{s/d}}.$$

Hence, for $0 < \delta < 1 - \delta_n$, with probability at least $1 - (\delta + \delta_n)$, we have

$$|\phi_5 - \mathbb{E}(\phi_5 \mid A)| < K_2\left(n^{-1/2}\sqrt{\log(2/\delta)/2} + \delta_n\right)\frac{(\log(12n/\delta_n))^{1+s/d}}{n^{s/d}}. \tag{20}$$

Step 3. Since $\|\phi_5\|_\infty \leq 3\|\varphi\|_\infty$, we have, by Lemma 6,

$$|\mathbb{E}(\phi_5 \mid A) - \mathbb{E}(\phi_5)| \leq 2\delta_n \|\phi_5\|_\infty \leq 6\delta_n \|\varphi\|_\infty.$$

Recall $\mathbb{E}[\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi)] = \mu(\varphi)$. On the event on which (20) holds, we have, in view of Proposition 3 with $C_0 = V_d bc$, for $n \geq 4$,

$$\begin{aligned}
&\left|\hat{\mu}_n^{(\mathrm{NN})}(\varphi) - \mu(\varphi)\right| \\
\leq& |\phi_5 - \mathbb{E}(\phi_5 \mid A)| + |\mathbb{E}(\phi_5 \mid A) - \mathbb{E}(\phi_5)| + \left|\mathbb{E}\left[\hat{\mu}_n^{(\mathrm{NN})}(\varphi)\right] - \mathbb{E}\left[\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi)\right]\right| \\
\leq& K_2\left(n^{-1/2}\sqrt{\log(2/\delta)/2} + \delta_n\right)\frac{(\log(12n/\delta_n))^{1+s/d}}{n^{s/d}} + 6\delta_n \|\varphi\|_\infty \\
&+ \frac{(s/d+2)L(V_d bc)^{-s/d}\Gamma(s/d+1)}{n^{1+s/d}} \\
\leq& K_2\left(n^{-1/2}\sqrt{\log(2/\delta)/2} + \delta_n\right)\frac{(\log(12n/\delta_n))^{1+s/d}}{n^{s/d}} + 6\delta_n \|\varphi\|_\infty \\
&+ \frac{(s/d+2)L(V_d bc)^{-s/d}}{n^{1+s/d}}.
\end{aligned}$$

Step 4. Similarly as in Step 4 of the proof of Theorem 2 for $\hat{\mu}_n^{(\mathrm{NN-loo})}$, fix $\varepsilon > 0$ and choose $\delta_n = \min(\varepsilon/2, n^{-1/2-1/d})$, $\delta = \varepsilon - \delta_n$. Then, for any $\varepsilon \in (0,1)$, we have, with probability at least $1-\varepsilon$,

$$\begin{aligned}
\left|\hat{\mu}_n^{(\mathrm{NN})}(\varphi) - \mu(\varphi)\right| \leq \frac{6\|\varphi\|_\infty}{n^{1/2+s/d}} + \frac{(s/d+2)L(V_d bc)^{-s/d}}{n^{1+s/d}} + K_2\left(\sqrt{\log\left(\frac{4}{\varepsilon}\right)\Big/2} + \frac{1}{n^{s/d}}\right) \\
\times \begin{cases} \dfrac{(\log(24n/\varepsilon))^{1+s/d}}{n^{1/2+s/d}}, & \text{for } \varepsilon \leq \frac{2}{n^{1/2+s/d}}, \\ \dfrac{(3\log(3n))^{1+s/d}}{n^{1/2+s/d}}, & \text{for } \varepsilon > \frac{2}{n^{1/2+s/d}}. \end{cases}
\end{aligned}$$

D.5 Proof of Theorem 3

We have

$$\begin{aligned}
\mathbb{P}\left(|\phi(X) - m| \geq t\right) &\leq \mathbb{P}\left(|\phi(X) - m| \geq t, X \in A\right) + \mathbb{P}(X \notin A) \\
&\leq \mathbb{P}\left(\phi(X) - m \geq t, X \in A\right) + \mathbb{P}\left(-\phi(X) + m \geq t, X \in A\right) + p.
\end{aligned}$$

From the proof of Theorem 2.1 in Combes (2015), we get

$$\mathbb{P}\left(\phi(X) - m \geq t, X \in A\right) \leq \exp\left(-\frac{2\max(0, t - p\bar{c})^2}{\sum_{\ell=1}^n c_\ell^2}\right).$$

By symmetry,

$$\mathbb{P}\left(-\phi(X) + m \geq t, X \in A\right) \leq \exp\left(-\frac{2\max(0, t - p\bar{c})^2}{\sum_{\ell=1}^n c_\ell^2}\right).$$

Thus, we obtain the stated result.

E Additional Experiments

Finance background. Options are financial derivatives based on the value of underlying securities. They give the buyer the right to buy (call option) or sell (put option) the underlying asset at a pre-determined price within a specific time frame. The price of an option may be expressed as the expectation, under the so-called risk-neutral measure, of the payoff discounted to the present value. Consider a contract of European type, which specifies a payoff $V(S_T)$, depending on the level of the underlying asset S_t at maturity $t = T$. The value V of the contract at time $t = 0$ conditional on an underlying value S_0 is

$$V(S_0) = \mathbb{E}_Q[e^{-rT}V(S_T)], \tag{21}$$

where $\mathbb{E}_Q$ denotes the expectation under the risk-neutral measure and r is the risk-free interest rate. Such a representation suggests a straightforward Monte Carlo based method for its calculation by

simulating random paths of the underlying asset, calculating each time the resulting payoff and taking the average of the result. This approach is particularly useful when dealing with exotic options, for which the above expectation often does not permit a closed-form expression.

The payoff of a European call option with strike price K is given by $V(S_T) = (S_T - K)_+$ and depends only on the level of the underlying asset S_t at maturity time $t = T$. In contrast, the payoff a of barrier option (Merton, 1973) depends on the whole path $(S_t)_{t\in[0,T]}$. The option becomes worthless or may be activated upon the crossing of a price point barrier denoted H. More precisely, Knock-Out (KO) options expire worthlessly when the underlying's spot price crosses the pre-specified barrier level whereas Knock-In (KI) options only come into existence if the pre-specified barrier level is crossed by the underlying asset's price. The payoffs of *up-in* (UI) and *up-out* (UO) barrier options with barrier price K are given by

$$V_{(\mathrm{UI})}(S) = (S_T - K)_+ \mathbb{1}\big\{\max_{t\in[0,T]} S_t \geq H\big\}, \quad V_{(\mathrm{UO})}(S) = (S_T - K)_+ \mathbb{1}\big\{\max_{t\in[0,T]} S_t < H\big\}. \tag{22}$$

Market Dynamics. The Black–Scholes model (Black and Scholes, 1973) is a mathematical model for pricing option contracts. It is based on geometric Brownian motion with constant drift and volatility so that the underlying stock S_t satisfies the following stochastic differential equation:

$$\mathrm{d}S_t = \mu S_t \,\mathrm{d}t + \sigma S_t \,\mathrm{d}W_t,$$

where μ represents the drift rate of growth of the underlying stock, σ is the volatility and W denotes a Wiener process. Although simple and widely used in practice, the Black–Scholes model has some limitations. In particular, it assumes constant values for the risk-free rate of return and volatility over the option duration. Neither of those necessarily remains constant in the real world. The Heston model (Heston, 1993) is a type of stochastic volatility model that can be used for pricing options on various securities. For the Heston model, the previous constant volatility σ is replaced by a stochastic volatility v_t which follows an Ornstein–Uhlenbeck process. The underlying stock S_t satisfies the following equations

$$\begin{cases} \mathrm{d}S_t = \mu S_t \,\mathrm{d}t + \sqrt{v_t} S_t \,\mathrm{d}W_t^S, \\ \mathrm{d}v_t = \kappa(\theta - v_t)\,\mathrm{d}t + \xi\sqrt{v_t}\,\mathrm{d}W_t^v, \qquad \mathrm{d}W_t^S\,\mathrm{d}W_t^v = \rho\,\mathrm{d}t. \end{cases}$$

with stochastic volatility v_t, drift term μ, long run average variance θ, rate of mean reversion κ and volatility of volatility ξ. Essentially the Heston model is a geometric Brownian motion with non-constant volatility, where the change in S has relationship ρ with the change in volatility.

Monte Carlo procedures. The application of standard Monte Carlo methods to option pricing takes the following form:

(1) Simulate a large number n of price paths for the underlying asset: $(S_{(1)}, \ldots, S_{(n)})$.

(2) For each path, compute the associated payoff of the option, e.g., as in Eq. (22): $(V_1, \ldots, V_n)$.

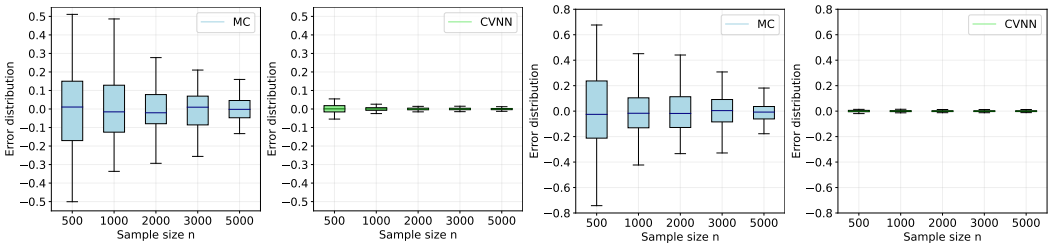
(3) Average the payoffs and discount them to present value: $\hat{V}_n = (e^{-rT}/n)\sum_{i=1}^n V_i$.

In practice, the price paths are simulated using an Euler scheme with a discretization of the time period $[0, T]$ comprised of m times $t_1 = 0 < t_2 < \ldots < t_m = T$. Each price path $S_{(i)}$ for $i = 1, \ldots, n$ is actually a vector $(S_{(i)}^{(1)}, \ldots, S_{(i)}^{(m)})$, so that the indicator function of the barrier options is computed on the discretized prices. Common values for m are the number of trading days per year which is $m = 252$ for $T = 1$ year.

Parameters. Several numerical experiments are performed for the pricing of European barrier call options "up-in" and "up-out". The number of sampled paths is $n \in \{500, 1\,000, 2\,000, 3\,000, 5\,000\}$ and the granularity of the grid is equal to $m = 240$. Two different mathematical models are considered when simulating the underlying asset price trajectories:

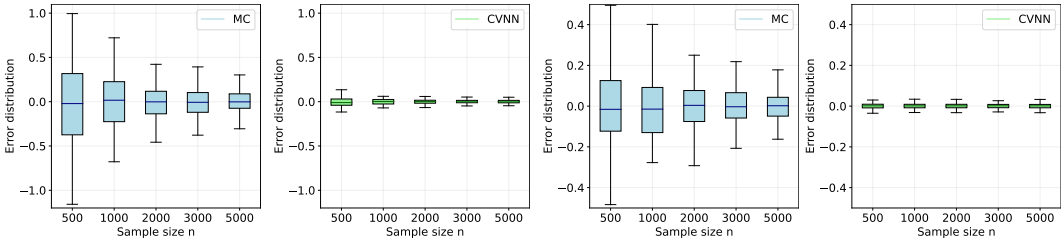
(1) the Black–Scholes model with constant volatility $\sigma = 0.30$;

(2) the Heston model with initial volatility $v_0 = 0.1$, long-run average variance $\theta = 0.02$, rate of mean reversion $\kappa = 4$, instantaneous correlation $\rho = 0.8$ and volatility of volatility $\xi = 0.9$.



(a) “Up-In”, MC (b) “Up-In”, CVNN (c) “Up-Out”, MC (d) “Up-Out”, CVNN

Figure 6: Barrier option pricing under Black–Scholes model with spot price $S_0 = 100$, strike $K = S_0$, maturity $T = 2$ months, risk-free rate $r = 0.1$, constant volatility $\sigma = 0.3$, barrier price $H = 130$. The boxplots are obtained over 100 replications.



(a) “Up-In”, MC (b) “Up-In”, CVNN (c) “Up-Out”, MC (d) “Up-Out”, CVNN

Figure 7: Boxplots of barrier option pricing with Heston Model with spot price $S_0 = 100$, strike $K = S_0$, barrier price $H = 130$, maturity $T = 2$ months, risk-free rate $r = 0.1$, initial volatility $v_0 = 0.1$, long-run average variance $\theta = 0.02$, rate of mean reversion $\kappa = 4$, instantaneous correlation $\rho = 0.8$ and volatility of volatility $\xi = 0.9$. The boxplots are obtained over 100 replications.

In both cases the fixed parameters are: spot price $S_0 = 100$, interest rate $r = 0.10$, maturity $T = 2$ months, strike price $K = S_0 = 100$ and barrier price $H = 130$.

Results. Figure 6 shows the error distribution of the different Monte Carlo estimates (naive MC and CVNN) for the pricing of Barrier call options “up-in” and “up-out” in the Black–Scholes model. The boxplots are computed over 100 independent replications and the true values of the options are approximated using the Python package QuantLib. Similarly, Figure 7 gathers the results for the Heston model. The variance is greatly reduced when using the control neighbors estimate compared to the standard Monte Carlo approach.