SPEEDING UP MONTE CARLO INTEGRATION: CONTROL NEIGHBORS FOR OPTIMAL CONVERGENCE

Rémi Leluc, François Portier, Aigerim Zhuman, Johan Segers

LIDAM Discussion Paper ISBA
2023 / 19

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.html

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. The main result is the $\mathcal{O}(n^{-1/2}n^{-1/d})$ 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 Lipschitz functions, 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 $\mathbb{R}^d$ and the integrand φ is a real-valued function defined 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. While easy to implement and fast to compute, a recognized drawback of the Monte Carlo estimate is its slow convergence rate $\mathcal{O}(n^{-1/2})$ as $n \to \infty$. 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 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 error of the procedure is defined as $\mathbb{E}[|\hat{\mu}_n(\varphi) - \mu(\varphi)|^2]^{1/2}$. For the specific problem of integration with respect to the uniform measure over the unit cube $[0,1]^d$ with $d \geq 1$, the complexity rate of randomized methods for Lipschitz integrands is known to be $\mathcal{O}(n^{-1/2}n^{-1/d})$ (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 advocate the use of randomized methods over deterministic integration rules: randomization yields an acceleration of the order $\mathcal{O}(n^{-1/2})$ compared to deterministic methods with complexity rates $\mathcal{O}(n^{-s/d})$. In addition, they show that for smooth integrands, the naive Monte Carlo estimate is suboptimal and that there is room for improvement by relying on the regularity of the integrand.

Preprint. Under review.

Several approaches are already known for 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. 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 using m control variates, the convergence rate is $\mathcal{O}(n^{-1/2}m^{-s/d})$ where s is the regularity of φ and the measure μ is arbitrary. However, it is required that m be of a smaller order than n, 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 $\mathcal{O}(n^{-1/2-\frac{a\wedge b}{d}+\varepsilon})$ 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 only ones achieving the best rate of convergence for smooth integrands.

The methods in Haber (1966) and in Chopin and Gerber (2022), even though they achieve the optimal convergence rate, are valid only for integration over the unit cube. In addition they involve a geometric number ($n = \ell^d$) of evaluations of the integrand φ, which is problematic in practice for applications with small computational budget, 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. This precise idea of using an estimate with small bias is the starting point of this paper. It is relevant to the considered framework because the function φ is accessible without noise. Low-bias estimates have also been used successfully in adaptive rejection sampling (Achddou et al., 2019), allowing to reach the optimal rate. The different properties of the mentioned Monte Carlo estimates are summarized in Table 1.

Monte Carlo method	Super-$\sqrt{n}$ convergence	Optimal rate	General μ
Vanilla Monte Carlo (Metropolis and Ulam, 1949)	×	×	✓
OLS-based Control Functionals (Oates et al., 2017; Portier and Segers, 2019)	✓	×	✓
Cubic Stratification (Haber, 1966; Chopin and Gerber, 2022)	✓	✓	×
Control neighbors (this paper)	✓	✓	✓

Figure 1: Convergence rates of randomized numerical integration methods

In this paper, a new Monte Carlo method called *control neighbors* is introduced. This method yields an estimate $\hat{\mu}_n(\varphi)$ of the integral $\mu(\varphi)$ for general probability measures μ (subject to regularity conditions) by using 1-nearest neighbor estimates as control variates. This novel estimate is shown to achieve the optimal convergence rate $\mathcal{O}(n^{-1/2}n^{-1/d})$ for Lipschitz integrands. To the best of our knowledge, achieving the optimal convergence rate for general probability measures makes this method the first of its kind. The most remarkable properties of the control neighbors estimate are:

(a) It can be obtained under the same framework as standard Monte Carlo, i.e., as soon as one can both *(i)* draw random particles from μ and *(ii)* evaluate the integrand φ. In contrast

to the classical control variates framework, 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 key property allows computational benefits when several integrals are to be computed with respect to the same measure μ.

(c) For Lipschitz integrands φ, 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.

(d) Since the weights $w_{n,i}$ are built using nearest neighbor estimates, practical tools are already available, including tree-based effective nearest neighbor search (Bentley, 1975) and efficient compression and parallelization (Pedregosa et al., 2011; Johnson et al., 2019).

(e) The approach is *post-hoc* in the sense that it can be run after sampling the particles and independently from the sampling mechanism. In particular, it can be implemented for other sampling designs including MCMC or adaptive importance sampling.

The outline of the paper is as follows. Section 2 presents a unified view of the control functionals framework and motivates the use of nearest neighbor estimates acting as control variates. Then, the mathematical foundations of nearest neighbor estimates are gathered in Section 3. The theoretical properties of the control neighbor estimates are stated in Section 4. Finally, Section 5 reports on several numerical experiments.

2 From control functionals to the method of control neighbors

2.1 General view of control functionals

The goal of this section is to introduce the framework and the main ideas of control functionals. By considering several examples, we present the key ingredients of the proposed approach of control neighbors.

Consider the classical numerical integration problem where, given a probability measure μ on $\mathbb{R}^d$ and a square-integrable function $\varphi \in L^2(\mu)$, the goal is to compute

$$\mu(\varphi) = \mathbb{E}_\mu[\varphi(X)] = \int_{\mathbb{R}^d} \varphi(x)\, \mu(\mathrm{d}x).$$

The standard Monte Carlo estimate approximates this value by using an independent random sample $X_1, \dots, X_n$ drawn from μ and takes the average

$$\hat{\mu}_n^{(\mathrm{MC})}(\varphi) = \frac{1}{n}\sum_{i=1}^{n} \varphi(X_i).$$

This unbiased estimate is consistent and, provided $\varphi(X_1)$ has finite variance, satisfies the central limit theorem. In particular, it converges to $\mu(\varphi)$ at the rate $\mathcal{O}(n^{-1/2})$, which may be prohibitive for complex statistical methods where the integrand φ is expensive to evaluate. While the use of control variates has been recognized as a useful variance reduction tool in many situations (Glasserman, 2004; Owen, 2013), it is only recently that control variates have been cast into a general functional approximation problem (Oates et al., 2017; Portier and Segers, 2019). This method of control variates or *control functionals* consists of two steps:

1. Build a surrogate function $\hat{\varphi}$ with known integral $\mu(\hat{\varphi})$.
2. Use the centred variables $\hat{\varphi}(X_i) - \mu(\hat{\varphi})$ to build the control variate Monte Carlo estimate

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

Whenever the function $\hat{\varphi}$ is constructed from another *surrogate* sample $\tilde{X}_1, \dots, \tilde{X}_N$ being either deterministic or independent from $X_1, \dots X_n$, the error analysis is simple and can be conducted using the mean squared error conditionally on $\tilde{X}_1, \dots, \tilde{X}_N$. This yields the following proposition in which the integrated mean squared error $\int \mathbb{E}[(\hat{\varphi} - \varphi)^2]\, \mathrm{d}\mu$ plays an important role.

Proposition 1. *Let $X_1, \dots, X_n$ be independent and identically distributed random variables with common distribution μ. If the estimate $\hat{\varphi}$ is independent of $X_1, \dots, X_n$, then*

$$\mathbb{E}\left[\left|\hat{\mu}_n^{(\mathrm{CV})}(\varphi) - \mu(\varphi)\right|^2\right] \leq n^{-1}\,\mathbb{E}\left[\int (\hat{\varphi} - \varphi)^2 \,\mathrm{d}\mu\right].$$

The success of the approach depends on the size of the (random) squared L^2-error $\int(\hat{\varphi} - \varphi)^2 \,\mathrm{d}\mu$. This promotes the use of the most accurate estimate $\hat{\varphi}$ of φ in $L^2(\mu)$. When the function φ is accessible without noise, the expected convergence rate of the L^2-error is then $n^{-1/d}$ (Kohler and Krzyżak, 2013) rather than $n^{-1/(d+2)}$ as in standard regression (Stone, 1982). Reaching the optimal convergence rate when estimating φ by control variates is the cornerstone to speed up the convergence rate of Monte Carlo integration.

Consider now the control variate approach described in Chopin and Gerber (2022) which is related to the stratification method of Haber (1966, Section 2.1). Suppose that the support of μ is $[0,1]^d$ and for integer $\ell \geq 1$, let $\{\tilde{X}_1, \dots, \tilde{X}_N\}$ be the $(1/\ell)$-equidistant grid of $[0,1]^d$ with $N = \ell^d$. Their control variate estimate is given by $\hat{\varphi}(x) = \sum_{i=1}^N \varphi(\tilde{X}_i)\,\mathbb{1}_{R_i}(x)$ where $(R_i)_{i=1,\dots,N}$ is the partition of $[0,1]^d$ made of the rectangles induced by the elements of the grid. Standard results give $\int(\hat{\varphi} - \varphi)^2 = \mathcal{O}(N^{-2/d})$ and, from Proposition 1, one obtains that the associated integration method has convergence rate $\mathcal{O}(n^{-1/2}N^{-1/d})$. Minimizing the previous upper bound under a fixed budget $(n + N)$ implies choosing N and n of a similar order. This leads to the convergence rate $\mathcal{O}(n^{-1/2}n^{-1/d})$.

Although this control variate method achieves the optimal convergence rate, the basic nature of the implied partitioning leads to some practical issues. First, the support of μ needs to be the unit cube for the equidistant grid to form a reasonable partitioning. Second, the number of evaluations needs to be of the form $N = \ell^d$, which could be restrictive in case d is large.

The method we propose, called *control neighbors*, is based on the following idea: use a nearest neighbor (NN) estimate for $\hat{\varphi}$ instead of a regular grid-based estimate. Given a surrogate sample $\tilde{X}_1, \dots, \tilde{X}_N$, let $\hat{\varphi}$ be the 1-NN estimate of φ, that is, $\hat{\varphi}(x) = \sum_{i=1}^N \varphi(\tilde{X}_i)\,\mathbb{1}_{S_{N,i}}(x)$, where $(S_{N,i})_{i=1,\dots,N}$ are the Voronoi cells associated to the sample, i.e., each cell $S_{N,i}$ contains all the points in $\mathbb{R}^d$ that are closer to $\tilde{X}_i$ than to any other point of the surrogate sample. Similarly to the rectangle approach described above, the resulting method employs a partitioning estimate of φ. However, the surrogate sample $\tilde{X}_1, \dots, \tilde{X}_N$ can now be any set of points in the support of μ. Hence, neither a strong assumption on the support of μ nor a restriction on the sample size are needed.

Finally, by following a "leave-one-out" strategy, the control variate estimate is built directly from the initial sample $X_1, \dots, X_n$. This allows to ultimately reduce the number of evaluations of φ from $n + N$ to n.

2.2 Control Neighbors estimates

For $i \in \{1, \dots, n\}$, let $\hat{\varphi}_n^{(i)}$ denote the 1-NN estimate of φ constructed on the sample $\mathcal{X}_n^{(i)} = \{X_1, \dots, X_n\} \setminus \{X_i\}$ without the i-th sample point; see Section 3 for precise definitions. We introduce the following *control neighbors* Monte Carlo estimate

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

in which the functions $x \mapsto \hat{\varphi}_n^{(i)}(x) - \mu(\hat{\varphi}_n^{(i)})$ act as control variates. A simple conditioning argument implies that

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

which is sufficient to get

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

Note that the n additional evaluations $\hat{\varphi}_n^{(i)}(X_i)$ are not computationally difficult as no additional evaluations of φ are necessary. However, computing the terms $\mu(\hat{\varphi}_n^{(i)})$ for $i = 1, \dots, n$ requires the evaluation of n additional integrals. Intuitively, since $\hat{\varphi}_n^{(i)}$ is similar to the 1-NN estimate $\hat{\varphi}_n$ of φ based on the full sample $X_1, \dots, X_n$, their integrals should be close. This is stated in the Appendix (Proposition 3) and one consequence is that

$$\frac{1}{n}\sum_{i=1}^{n} \mu(\hat{\varphi}_n^{(i)}) = \mu(\hat{\varphi}_n) + \mathcal{O}_{\mathbb{P}}(n^{-1/2}n^{-1/d}), \qquad n \to \infty.$$

Based on this remark, one may replace the n integrals $\mu(\hat{\varphi}_n^{(i)})$ by only a single integral $\mu(\hat{\varphi}_n)$. This gives the following *control neighbors* Monte Carlo estimate

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

Both estimates (1) and (2) may be written as linear integration rules with weights that do not depend on the integrand; see Section 4 below. In practice, the working estimate is the control neighbor estimate (2) as it involves less computations.

2.3 Control Neighbors implementation

We specify the algorithm for computing the control neighbor estimate (2). This 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 nearest neighbors estimates. Also required is the integral $\mu(\hat{\varphi}_n)$ of the 1-NN estimate $\hat{\varphi}_n$. Several practical remarks regarding the computation of all these quantities are given right after Algorithm 1.

Algorithm 1 Control Neighbors for Monte Carlo integration

Require: integrand φ, probability measure μ, number of samples 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 $\hat{\mu}_n^{(\mathrm{NN})}(\varphi) = \frac{1}{n}\sum_{i=1}^{n}[\varphi(X_i) - \{\hat{\varphi}_n^{(i)}(X_i) - \mu(\hat{\varphi}_n)\}]$

Remark 1 (Tree search)**.** The naive neighbor search implementation involves the brute-force 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 so the cost of a nearest neighbors search can be reduced. The KD-Tree (Bentley, 1975) is a binary tree structure which recursively partitions the parameter space along the data axes, dividing it into nested orthotropic regions into which data points are filed. The construction of such a tree requires $\mathcal{O}(dn \log n)$ operation (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 along Cartesian axes, Ball trees partition data in a series of nesting hyper-spheres, making tree construction more costly than KD tree, but results in an efficient data structure even in very high dimensions. In practice, many software libraries contain implementations of KD-tree and Ball-Tree with efficient compression and parallelization (Pedregosa et al., 2011; Johnson et al., 2019).

Remark 2 (Evaluation of $\hat{\varphi}_n^{(i)}(X_i)$)**.** The evaluations of the leave-one-out nearest neighbor estimates can be efficiently computed with nearest neighbor search and masks evaluations. More precisely, let $\mathcal{E} = (\varphi(X_1), \dots, \varphi(X_n))$ denote the vector of evaluations of the integrand. Any query of a nearest neighbor algorithm produces a vector containing the indices of neighbors of the corresponding query points. After fitting a KD-Tree on the particles $X_1, \dots, X_n$, one can query the 2-nearest neighbor of each X_i to produce the vector of indices $\mathcal{I}$ such that $\mathcal{I}_i$ is the index of the nearest neighbor of X_i among $\mathcal{X}_n^{(i)}$. The leave-one-out evaluations $(\hat{\varphi}_n^{(1)}(X_1), \dots, \hat{\varphi}_n^{(n)}(X_n))$ are then simply obtained using the slicing operation on array $\mathcal{E}[\mathcal{I}]$.

Remark 3 (Evaluation of $\mu(\hat{\varphi}_n)$)**.** In the case of a complex probability measure μ, the Voronoi volumes may be hard to compute but can always be approximated. The integral of the nearest neighbor estimate $\mu(\hat{\varphi}_n)$ may be replaced by a Monte Carlo estimate that uses M particles, such as $\mu(\hat{\varphi}_n) \simeq M^{-1} \sum_{i=1}^{M} \hat{\varphi}_n(\tilde{X}_i)$ where the variables $\tilde{X}_i$ are drawn independently from μ. No additional evaluations of φ are required. The error of this naive Monte Carlo approximation is $\mathcal{O}(M^{-1/2})$ meaning that large values of the form $M = n^{1+2/d}$ can be taken to preserve the $\mathcal{O}(n^{-1/2} n^{-1/d})$ convergence of the control neighbors estimate.

Remark 4 (Computing time)**.** When using the standard KD-tree approach, the computing time of our algorithm can be estimated in light of the previous remarks. The different computing 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, the estimation of $\mu(\hat{\varphi}_n)$ requires $\mathcal{O}(M \log n)$ operations. Choosing M as recommended before, the overall complexity is $\mathcal{O}((nd + n^{1+2/d}) \log n)$ operations. Note that when the Voronoï volumes are available, the computing time reduces to $\mathcal{O}(nd \log n)$.

Remark 5 (Voronoi volume when μ is uniform)**.** The quantity $\mu(\hat{\varphi})$ may be written as a sum of the evaluations $\varphi(X_i)$ weighted by the value of the Voronoi volumes associated to the sample points X_i (see Definition 2 in the next section). In case the measure μ is the uniform measure on $[0,1]^d$, one may be able to explicitly compute those volumes. Starting from the pioneering work of Richards (1974) in the context of protein structures, there has been advances to perform efficient Voronoi volume computations using Delaunay triangulations and taking advantage of graphic hardware (Hoff III et al., 1999). For 2D and 3D Voronoi diagrams, one can refer to the software Voro++ (Rycroft, 2009) for computations related to the Voronoi tessellation. However, this type of algorithm is subjected to the curse of dimensionality and might be inefficient when d is large.

Remark 6 (k-NN estimates)**.** A natural variant of the proposed method is obtained by replacing the 1-NN estimate $\hat{\varphi}_n$ in Eq. (2) 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 mentioned in Section 2.1.

Remark 7 (Monte Carlo integration on manifolds)**.** The notion of nearest neighbors on $\mathbb{R}^d$ is defined through the Euclidean distance and can be easily extended to a manifold equipped with a metric. When integrating on a Riemannian manifold $\mathcal{M}$ (see for instance Barp et al., 2022), one can consider the control neighbors estimate with respect to the distance on $\mathcal{M}$ induced by the metric tensor.

3 Nearest Neighbor estimation

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. It is assumed in this section that $X_1, \ldots, X_n$ are independent and identically distributed random vectors in $\mathbb{R}^d$. Let $\|\cdot\|$ be the Euclidean norm on $\mathbb{R}^d$ and let x be a given point in $\mathbb{R}^d$.

Definition 1 (Nearest neighbors and distances)**.** *Given any point $x \in \mathbb{R}^d$ and any numbered collection $X_1, \ldots, X_n$ in $\mathbb{R}^d$, define $\hat{N}_n(x)$ as the nearest neighbor of x among $X_1, \ldots, X_n$ and $\hat{\tau}_n(x)$ the associated distance, i.e.,*

$$\hat{N}_n(x) \in \underset{Y \in \{X_1, \ldots, X_n\}}{\arg\min} \|x - Y\|, \qquad \hat{\tau}_n(x) = \|\hat{N}_n(x) - x\|.$$

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 $k = 1, \ldots, n$, let $\hat{N}_{n,k}(x)$ denote the k-nearest neighbor of x and $\hat{\tau}_{n,k}(x) = \|\hat{N}_{n,k}(x) - x\|$ 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 such that their nearest neighbor is X_i, as detailed below.

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

$$\forall i = 1, \dots, n, \qquad S_{n,i} = \{x \in \mathbb{R}^d : \hat{N}_n(x) = X_i\},$$

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

The 1-NN estimate of φ is defined as $\hat{\varphi}_n(x) = \varphi(\hat{N}_n(x))$ 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 the i-th sample 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. It is therefore useful to introduce the leave-one-out nearest neighbor and the leave-one-out Voronoi cells.

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 $x \in \mathbb{R}^d$ is*

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

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.,*

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

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

The leave-one-out 1-NN predictor used in Section 2 to define the proposed integral estimate is $\hat{\varphi}_n^{(i)}(x) = \varphi(\hat{N}_n^{(i)}(x))$. 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 Lipschitz, their supremum distance is of the same order as the nearest neighbor distance. In terms of the $L^1(\mu)$-norm, their difference is even smaller as the cell $S_{n,i}$ has a small volume. Relevant for our numerical integration problem is that the average of the integrals $\mu(\hat{\varphi}_n^{(i)})$ is close to $\mu(\hat{\varphi}_n)$, as stated in the following lemma.

Lemma 1. *Let $\bar{\varphi}_n(x) = \sum_{i=1}^n \hat{\varphi}_n^{(i)}(x)\, \mathbb{1}_{S_{n,i}}(x)$ then $\sum_{i=1}^n \{\mu(\hat{\varphi}_n^{(i)}) - \mu(\hat{\varphi}_n)\} = \mu(\bar{\varphi}_n - \hat{\varphi}_n)$.*

A central quantity that reflects how much a point is surrounded within the sample is given by enumerating how many times a point, say X_i, is the nearest neighbor of points from the sample $\mathcal{X}_n^{(i)}$. Another important quantity that qualifies the isolation of a point is obtained by summing the Voronoi volumes. 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,*

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

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 the next section 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*

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

4 Main results

This section gathers some technical result on nearest neighbors and the main theoretical properties of the control neighbors estimates (1) and (2) presented in Section 2.2. First, these estimates can be written as simple linear integration rules with weights that only depend on the nearest neighbor estimates and may be efficiently computed in practice. Next, the convergence rate of the error is shown to reach optimality as $\mathbb{E}[|\hat{\mu}_n(\varphi) - \mu(\varphi)|^2]^{1/2} \lesssim n^{-1/2} n^{-1/d}$. Finally, the finite-sample performance of the proposed estimates is studied through a concentration inequality on the probabilistic error $|\hat{\mu}_n(\varphi) - \mu(\varphi)|$, which scales, with probability $(1-\varepsilon)$, at rate $\sqrt{\log(1/\varepsilon)}(\log n)^{1+1/d} n^{-1/2} n^{-1/d}$.

4.1 Technical results on nearest neighbors

For the analysis, we consider the following assumptions which are related to the *strong density assumption* of Audibert and Tsybakov (2007). This condition ensures that the density f of the measure μ has a *regular* support and that it is bounded away from zero and infinity. Let $B(x, r)$ denote the ball with centre $x \in \mathbb{R}^d$ and radius $r > 0$.

(A1) $X, X_1, X_2, \ldots$ are independent and identically distributed random vectors in $\mathbb{R}^d$ with common distribution μ having support $\mathcal{X} \subset \mathbb{R}^d$.

(A2) The measure μ 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

- $\forall x \in \mathcal{X}, \quad b \leq f(x) \leq U$,
- $\forall 0 < r \leq r_0, \forall x \in \mathcal{X}, \quad \lambda_d(\mathcal{X} \cap B(x, r)) \geq c \lambda_d(B(x, r))$.

(A3) The function $\varphi : \mathbb{R}^d \to \mathbb{R}$ is Lipschitz, i.e., there exists $L > 0$ such that

$$\forall x, y \in \mathbb{R}^d, \quad |\varphi(x) - \varphi(y)| \leq L \|x - y\| .$$

Under (A3), the distance $\hat{\tau}_n$ is key in the analysis of the functional approximation problem of φ by the estimate $\hat{\varphi}_n$. Indeed, for every $x \in \mathbb{R}^d$ we have almost surely $|\hat{\varphi}_n(x) - \varphi(x)| \leq L \hat{\tau}_n(x)$. The next lemma is dedicated to control the distance $\hat{\tau}_n(x)$ and follows from standard considerations in the k-NN literature (Biau and Devroye, 2015). It relies on the uniform lower bounds required in (A2). Let Γ denote the Euler gamma function and $V_d = \lambda_d(B(0, 1))$ the volume of the unit ball in $\mathbb{R}^d$.

Lemma 3 (Bounding moments of nearest neighbor distances)**.** *Under (A1) and (A2), we have, for any $q \geq 1$,*

$$\forall x \in \mathcal{X}, \quad \mathbb{E}[\hat{\tau}_n(x)^q] \leq (n V_d b c)^{-q/d} \Gamma(q/d + 1).$$

Furthermore, if $2k \leq n$, we have

$$\forall x \in \mathcal{X}, \quad \mathbb{E}[\hat{\tau}_{n,k}(x)^q] \leq 2^{2q/d+1} \Gamma(q/d + 1)(n V_d b c/k)^{-q/d}.$$

Remark 8. Condition (A2) plays an important role in the analysis of nearest neighbor estimates as it allows a uniform control on the radii of the Voronoi cells. Such a uniform bound on the radius is the key to study the convergence of general k-NN estimates. When dealing with densities having general supports, one can also consider some minimal mass assumption (Gadat et al., 2016) to guarantee that no region is empty. Furthermore, this question of necessary conditions for general uniform bounds remains an active field of research with recent progress for unbounded data (Kohler et al., 2006) and relaxations through tail assumptions (Gadat et al., 2016). Extending the present analysis to such general measures is left for further research.

4.2 Linear integration rules

The control neighbor estimates $\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi)$ in (1) and $\hat{\mu}_n^{(\mathrm{NN})}(\varphi)$ in (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 Definition 4.

Proposition 2 (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*

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

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 2, the proposed estimate $\hat{\mu}_n^{(\mathrm{NN})}(\varphi)$ consists in a simple modification of $\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi)$ by replacing $\hat{c}_{n,j}$, which involves $n-1$ Voronoi volumes, by $nV_{n,i}$. The difference between both is of the order $n^{-1/2-1/d}$ as shown in the next section.

4.3 Convergence rate of the error

The main result of this section provides a finite-sample bound on the mean-squared error of the two control neighbors estimates. The leave-one-out version $\hat{\mu}_n^{(\mathrm{NN-loo})}$, although its rate of convergence matches the optimal rate, involves n additional integrals, which might represent a computational burden. As detailed in Section 2, the proposed estimate $\hat{\mu}_n^{(\mathrm{NN})}$ reduces this to only a single additional integral while remaining close to the leave-one-out estimate as

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

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

Theorem 1. *(Convergence Rate of the Mean-Square Error) Under (A1), (A2) and (A3), if $n \geq 4$, then*

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

where

$$C_{\mathrm{NN-loo}} = 16L\,(V_d bc)^{-1/d}\,(U/bc)^{-1/2},$$
$$C_{\mathrm{NN}} = 39L(V_d bc)^{-1/d}(U/bc)^{-1/2}.$$

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 paper for more details) can reach a better accuracy – measured in terms of mean-squared error – than $\mathcal{O}(n^{-1-2/d})$ when the integration measure is the uniform measure over the unit cube and the function φ is Lipschitz. Theorem 1 states that the optimal rate is in fact achieved by some integration rule in situations where the integration measure's density is not necessarily uniform but only bounded from below and above.

4.4 Non-asymptotic bounds

In order to obtain a finite-sample performance guarantee of the proposed estimates, we use 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 and is itself a minor extension of an inequality due to Combes (2015).

In combination with (A2), Assumption (A3) implies that φ is uniformly bounded on $\mathcal{X}$, since $\mathcal{X}$ has a finite diameter (as it is contained in the union of a finite number of balls of radius r_0). Write $C_\varphi = \frac{1}{2}\{\sup_{x\in\mathcal{X}} \varphi(x) - \inf_{x\in\mathcal{X}} \varphi(x)\}$. The two control neighbors estimates satisfy the following concentration inequalities.

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

$$\left|\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi) - \mu(\varphi)\right| \leq C_1 \Delta_{n,\varepsilon} + \frac{6C_\varphi}{n^{1/2+1/d}}$$
$$\left|\hat{\mu}_n^{(\mathrm{NN})}(\varphi) - \mu(\varphi)\right| \leq C_2 \Delta_{n,\varepsilon} + \frac{6C_\varphi}{n^{1/2+1/d}} + \frac{(2^{3/d+1}+1)L(V_d bc)^{-1/d}}{n^{1+1/d}}$$

where

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

The values of $C_1, C_2 > 0$ depend on (b, c, d, L, U) and are given explicitly in the proof.

5 Numerical experiments

This section gather several experiments that highlight the wide range of applications of the proposed *control neighbors* estimate. To illustrate its finite-sample performance, we first present in Section 5.1 some examples involving integration problems on different spaces: first on the unit cube $[0,1]^d$ and on $\mathbb{R}^d$ with uniform and Gaussian measures respectively; then a more complex integration problem dealing with the moments of random orthogonal matrices; finally a challenging problem dealing with the integration of some functions on a sphere. Next, Section 5.2 presents an application of the method in finance for Monte Carlo exotic option pricing under the standard Black–Scholes model with constant volatility and the more difficult Heston model with stochastic volatility. Finally, Section 5.3 deals with the application of Monte Carlo estimates for optimal transport.

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 $M = n^2$ particles, with n the number of evaluations of φ.

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

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 standard Euclidean spaces such as the unit cube $[0,1]^d$ or the group of orthogonal matrices $O_d(\mathbb{R})$ to non-Euclidean spaces such as the sphere $\mathbb{S}^{d-1}$. In the different settings, the sample size evolves from $n = 10^1$ to $n = 10^4$ and the different 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.

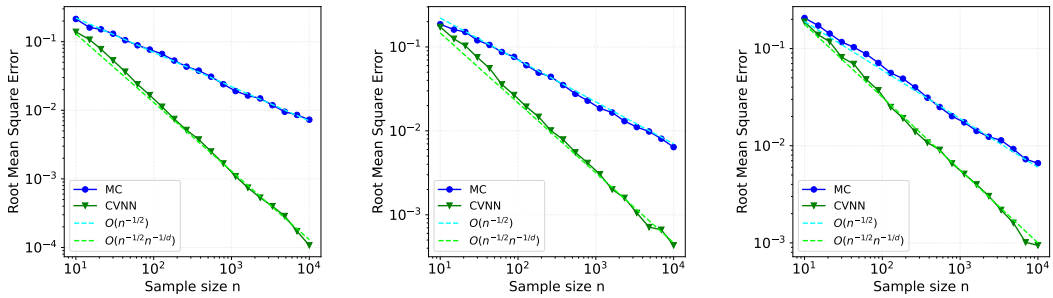
Integration on $[0,1]^d$ and $\mathbb{R}^d$. Consider first the integration problem $\mu(\varphi) = \int \varphi \, \mathrm{d}\mu$ where the measure μ is either the uniform distribution over the unit cube $[0,1]^d$ or the multivariate Gaussian measure on $\mathbb{R}^d$. The goal is to compute $\int \varphi_1(x)\, \mathbb{1}_{[0,1]^d}(x)\, \mathrm{d}x$ and $\int \varphi_2(x)\psi(x)\, \mathrm{d}x$ with

$$\varphi_1(x_1,\dots,x_d) = \sin\left(\pi\left(\tfrac{2}{d}\textstyle\sum_{i=1}^d x_i - 1\right)\right) \quad \text{and} \quad \varphi_2(x_1,\dots,x_d) = \sin\left(\tfrac{\pi}{d}\textstyle\sum_{i=1}^d x_i\right) \tag{3}$$

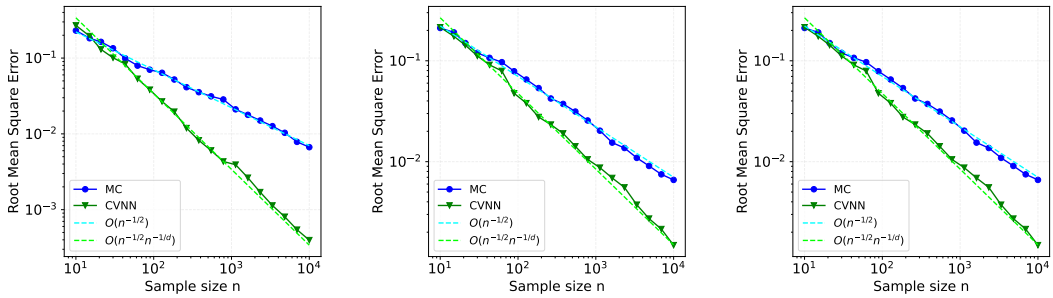
and with $\psi(\cdot)$ the probability density function of the multivariate Gaussian distribution $\mathcal{N}(0, I_d)$. Figure 2 displays the evolution of the root mean squared error for the two integrals in dimension $d \in \{2,3,4\}$. The different error curves confirm the optimal convergence rate $\mathcal{O}(n^{-1/2}n^{-1/d})$ for the control neighbors estimate. In dimension $d = 2$ and $d = 3$, the root mean squared error of the CVNN estimate can be reduced by a factor ten compared to the standard Monte Carlo approach.

Integration on the orthogonal group $O_d(\mathbb{R})$. Consider the group of real orthogonal matrices $O_d(\mathbb{R}) = \{X \in \mathrm{GL}_d(\mathbb{R}) : X^\top X = XX^\top = \mathrm{I}_d\}$ and the moments of the trace of a random orthogonal matrix, i.e., integrands of the form $\varphi_k : O_d(\mathbb{R}) \to \mathbb{R}$ with $\varphi_k(X) = \mathrm{tr}(X)^k$. The goal is to compute the integrals

$$a_k = \mu(\varphi_k) = \int_{O_d(\mathbb{R})} \mathrm{tr}(X)^k \, \mathrm{d}\mu(X), \tag{4}$$



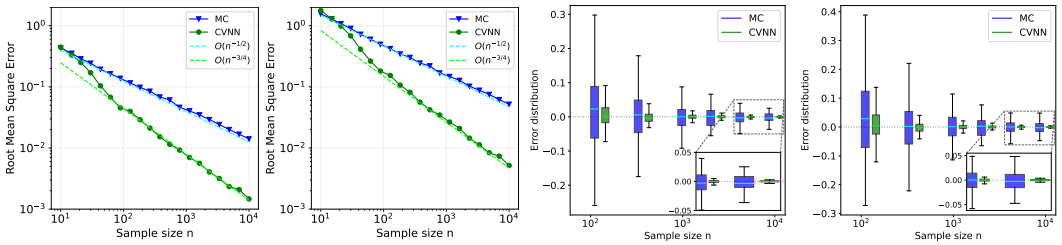
(a) $\varphi_1, d = 2$ (b) $\varphi_1, d = 3$ (c) $\varphi_1, d = 4$



(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. (3) in dimension $d \in \{2, 3, 4\}$ (left to right).

with $\mathrm{d}\mu(X)$ the Haar measure on $O_d(\mathbb{R})$. In the particular case of dimension $d = 2$, one can obtain the closed-form expression of the even moments (Pastur and Vasilchuk, 2004): $a_{2k} = (1/2)\binom{2k}{k}$. In practice, random orthogonal matrices $X_1, \ldots, X_n \in O_d(\mathbb{R})$ 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[1] as in Mezzadri (2007). The nearest neighbors are computed using the norm associated to the Frobenius inner product $\langle X, Y \rangle = \mathrm{tr}(X^\top Y)$ for $X, Y \in O_d(\mathbb{R})$. Figure 3 reports the evolution of the root mean squared error and the boxplots of errors for integrands φ_2 and φ_4 over the group $O_2(\mathbb{R})$. 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.



(a) RSME for a_2 (b) RMSE for a_4 (c) Boxplot for a_2 (d) Boxplot for a_4

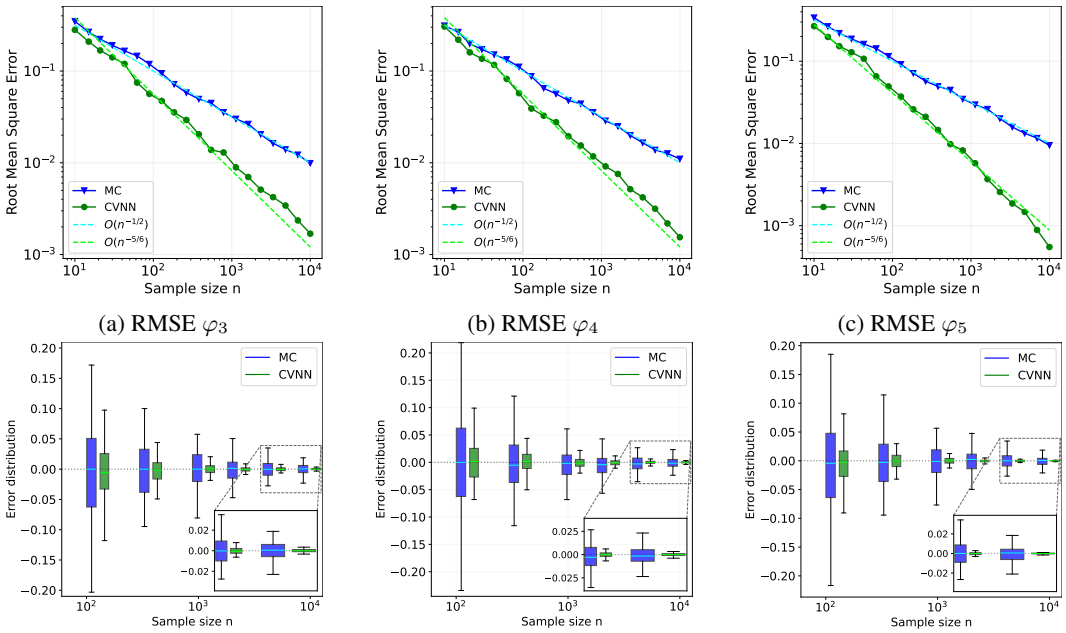
Figure 3: Root mean squared errors (left) and boxplots of the errors (right) obtained over 100 replications for functions φ_2 and φ_4 in Eq. (4) over the group $O_2(\mathbb{R})$.

Integration on the sphere $\mathbb{S}^2$. Consider the sphere $\mathbb{S}^2$ embedded in $\mathbb{R}^3$. This is a non-Euclidean manifold with positive curvature and 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

[1] 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.

$$\left.\begin{aligned}\varphi_3(x,y,z) &= \alpha^{-1}\{1+\tanh(-\alpha(x+y-z))\}\\ \varphi_4(x,y,z) &= \alpha^{-1}\{1-\text{sign}(x+y-z)\}\\ \varphi_5(x,y,z) &= \alpha^{-1}\{1-\text{sign}(\pi x+y)\}\end{aligned}\right\} \tag{5}$$

with $\alpha = 12$ so that $\mu(\varphi_3) = \mu(\varphi_4) = \mu(\varphi_5) = (4\pi)/\alpha = \pi/3$ (Beentjes, 2015). Figure 4 below reports the evolution of the root mean squares error and the boxplots of the errors over the sphere $\mathbb{S}^2$.



(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. (5) when integrating over $\mathbb{S}^2$.

5.2 Monte Carlo Option Pricing

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

$$V(S_0) = \mathbb{E}_Q[e^{-rT}V(S_T)], \tag{6}$$

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

$$\begin{aligned} V_{(\mathrm{UI})}(S) &= (S_T - K)_+ \mathbb{1}\{\max_{t\in[0,T]} S_t \geq H\}, \\ V_{(\mathrm{UO})}(S) &= (S_T - K)_+ \mathbb{1}\{\max_{t\in[0,T]} S_t < H\}. \end{aligned} \tag{7}$$

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. (7): $(V_1, \ldots, V_n)$.

(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.

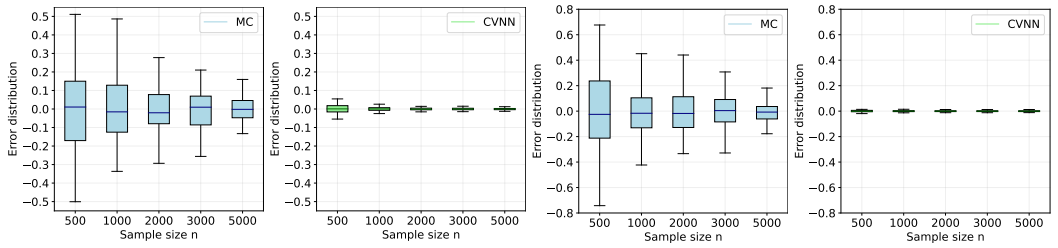
Parameters. Several numerical experiments are performed for the pricing of European Barrier call options "up-in" and "up-out". The number of sampled paths evolves as $n \in \{500, 1\,000, 2\,000, 3\,000, 5\,000\}$ and the granularity of the grid is equal to $m = 240$. Two different mathematical models 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 $\xi = 0.9$.

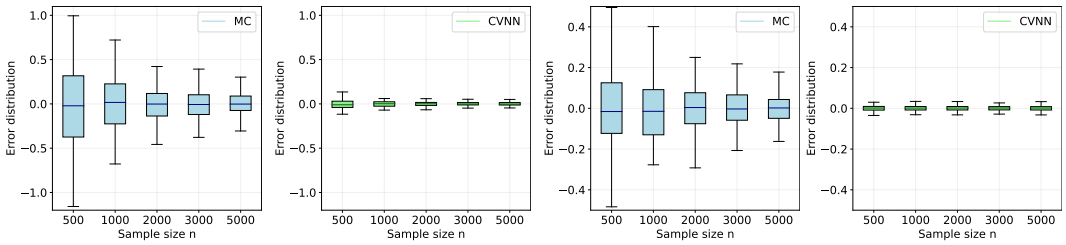
In both cases the fixed parameters are: spot price $S_0 = 100$, interest rate $r = 0.10$, maturity $T = 2$ months, strike price $K = S_0 = 100$ and barrier price $H = 130$.

Results. Figure 5 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 6 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.



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

Figure 5: 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 6: 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.

5.3 Monte Carlo 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 $\mathsf{X} \subseteq \mathbb{R}^d$, denote by $\mathcal{P}(\mathsf{X})$ the set of probability measures supported on X and let $p \in [1, \infty)$. The Wasserstein distance of order p between $\mu, \nu \in \mathcal{P}(\mathsf{X})$ is

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

where $\Pi(\mu, \nu) \subset \mathcal{P}(\mathsf{X} \times \mathsf{X})$ denotes the set of couplings for (μ, ν), i.e., probability measures whose marginals with respect to the first and second variables are μ and ν respectively. While the Wasserstein distance enjoys attractive theoretical properties (Villani, 2009, Chapter 6), it suffers from a high computational cost. When computing $W_p(\mu_m, \nu_m)$ for discrete distributions μ_m and ν_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 $\mu, \nu \in \mathcal{P}(\mathbb{R})$. Indeed, for $\mu_m = (1/m) \sum_{i=1}^m \delta_{x_i}$ and $\nu_m = (1/m) \sum_{i=1}^m \delta_{y_i}$ 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

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

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

Recall that $\mathbb{S}^{d-1} = \{\theta \in \mathbb{R}^d : \|\theta\| = 1\}$ is the unit sphere of $\mathbb{R}^d$ and let $\theta^\star : \mathbb{R}^d \to \mathbb{R}$ denote the linear map $\theta^\star(x) = \langle \theta, x \rangle$ for $x \in \mathbb{R}^d$. Let $\rho \in \mathcal{P}(\mathbb{S}^{d-1})$. The Sliced-Wasserstein (SW) distance (Rabin et al., 2012; Bonneel et al., 2015; Kolouri et al., 2019) of order p based on ρ is defined for $\mu, \nu \in \mathcal{P}(\mathsf{X})$ as

$$SW_p^p(\mu, \nu, \rho) = \mathbb{E}_\theta[W_p^p(\theta^\star_\# \mu, \theta^\star_\# \nu)] = \int_{\mathbb{S}^{d-1}} W_p^p(\theta^\star_\# \mu, \theta^\star_\# \nu) \, \mathrm{d}\rho(\theta), \tag{8}$$

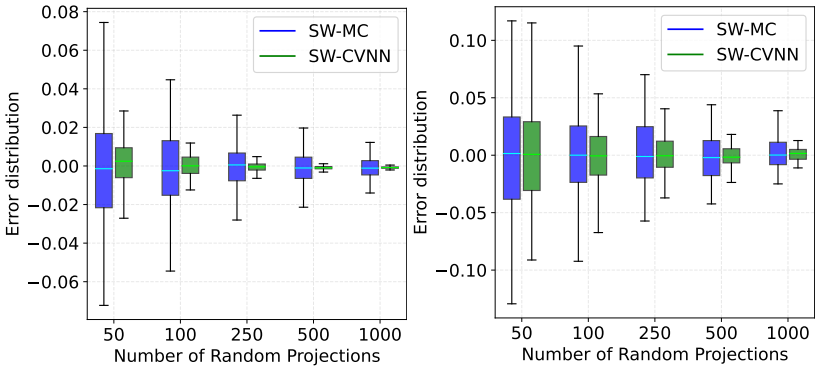
where $f_\# \xi$ is the *push-forward* measure of $\xi \in \mathcal{P}(\mathbb{R}^d)$ by a measurable function f on $\mathbb{R}^d$. For $\theta \in \mathbb{S}^{d-1}$, the measures $\theta^\star_\# \mu$ and $\theta^\star_\# \nu$ are the distributions of the projections $\langle \theta, X \rangle$ and $\langle \theta, Y \rangle$ of random vectors X and Y on $\mathbb{R}^d$ with distributions μ and ν, respectively. In practice, the random directions $\theta_1, \ldots, \theta_n$ are sampled independently from the uniform distribution on the unit sphere $\mathbb{S}^{d-1}$ and the SW-distance of Eq. (8) is approximated using a standard Monte Carlo estimate with n random projections as

$$\widehat{SW}_p^p(\mu_m, \nu_m) = \frac{1}{n} \sum_{i=1}^n W_p^p((\theta_i)^\star_\# \mu_m, (\theta_i)^\star_\# \nu_m).$$

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 Gaussian distributions. More precisely we want to compute $SW_2(\mu, \nu)$ where $\mu = \mathcal{N}_d(m_X, \sigma_X^2 \mathrm{I}_d)$ and $\nu = \mathcal{N}_d(m_Y, \sigma_Y^2 \mathrm{I}_d)$ with $m_X, m_Y \sim \mathcal{N}(0, \mathrm{I}_d)$ and $\sigma_X = 2$ and $\sigma_Y = 5$. We consider the corresponding empirical distributions μ_m and ν_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 $d \in \{3; 6\}$. Figure 7 below shows the error distribution of the different Monte Carlo estimates (SW-MC and SW-CVNN) of the SW_2 distance where the exact value (Nadjahi et al., 2021, Appendix S3.1) is given by

$$SW_2^2(\mathcal{N}_d(m_X, \sigma_X^2 \mathrm{I}_d), \mathcal{N}_d(m_Y, \sigma_Y^2 \mathrm{I}_d)) = \frac{1}{d} \|m_X - m_Y\|_2^2 + (\sigma_X - \sigma_Y)^2.$$

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



(a) SW_2 $(d = 3)$ (b) SW_2 $(d = 6)$

Figure 7: Boxplots of Sliced-Wasserstein estimates SW-MC and SW-CVNN in dimension $d \in \{3; 6\}$. The boxplots are obtained over 100 replications.

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.

Alessandro Barp, Chris J Oates, Emilio Porcu, and Mark Girolami. A Riemann–Stein kernel method. *Bernoulli*, 28(4):2181–2208, 2022.

Casper HL Beentjes. Quadrature on a spherical surface. *Working note available on the website http://people.maths.ox.ac.uk/beentjes/Essays*, 2015. URL https://cbeentjes.github.io/files/Ramblings/QuadratureSphere.pdf.

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

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

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

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

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

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

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. *Advances in neural information processing systems*, 30, 2017.

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

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Arnaud Doucet, Nando De Freitas, Neil James Gordon, et al. *Sequential Monte Carlo methods in practice*, volume 1. Springer, 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 Transactions on Mathematical Software (TOMS)*, 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 *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR, 2018.

Paul Glasserman. *Monte Carlo Methods in Financial Engineering*, volume 53. Springer, 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. *Advances in neural information processing systems*, 30, 2017.

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

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

Seymour Haber. Stochastic quadrature formulas. *Mathematics of Computation*, 23(108):751–764, 1969.

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. *Journal of Physics G: Nuclear and Particle Physics*, 42(3):034009, 2015.

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.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on 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. *Statistics & Probability Letters*, 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. *Journal of Multivariate Analysis*, 97(2): 311–323, 2006.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced Wasserstein distances. *Advances in neural information processing systems*, 32, 2019.

R. Leluc, F. Portier, and J. Segers. Control variate selection for Monte Carlo integration. *Statistics and Computing*, 31, 07 2021. doi: 10.1007/s11222-021-10011-z.

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

Nicholas Metropolis and Stanislaw Ulam. The Monte Carlo method. *Journal of the American Statistical Association*, 44(247):335–341, 1949.

Francesco Mezzadri. How to generate random matrices from the classical compact groups. *Notices of the American Mathematical Society*, 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. *Advances in Neural Information Processing Systems*, 34:12411–12424, 2021.

Nigel J Newton. Variance reduction for simulated diffusions. *SIAM Journal on Applied Mathematics*, 54(6):1780–1805, 1994.

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. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 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.

Art B Owen. Monte carlo theory, methods and examples. 2013.

Leonid Pastur and Vladimir Vasilchuk. On the moments of traces of matrices of classical groups. *Communications in mathematical physics*, 252:149–166, 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. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

F. Portier and J. Segers. Monte Carlo integration with a growing number of control variates. *Journal of Applied Probability*, 56(4):1168–1186, 2019. doi: 10.1017/jpr.2019.78.

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. *Journal of Molecular Biology*, 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. *Statistical science*, 4(4):409–423, 1989.

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

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040–1053, 1982.

Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the royal statistical society: Series B (Methodological)*, 36(2):111–133, 1974.

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

Cédric Villani. *Optimal transport: Old and New*, volume 338. Springer, 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. *Nature 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.

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

Appendix A contains the auxiliary results. Appendix B gathers the technical 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.

A Auxiliary results

Theorem 3 (Extension of McDiarmid's inequality)**.** *Consider independent random variables $X_1, \dots, X_n$ 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 $A \subset \Omega$ 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*

$$|\phi(x) - \phi(x')| \leq c_\ell. \tag{9}$$

Then, for any $t \geq 0$, we have, writing $m = \mathbb{E}[\phi(X) \mid A]$, $p = 1 - P(A)$ 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 (A1), (A2) and (A3), if $n \geq 4$, we have*

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n \mu(\hat{\varphi}_n^{(i)}) - \mu(\hat{\varphi}_n)\right|\right] \leq (2^{3/d+1} + 1)L(V_d bc)^{-1/d}\Gamma(1/d + 1)n^{-1-1/d}$$

and

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n \mu(\hat{\varphi}_n^{(i)}) - \mu(\hat{\varphi}_n)\right|^2\right] \leq 2(2^{6/d+1} + 1)L^2(V_d bc)^{-2/d}\Gamma(2/d + 1)n^{-1-2/d}$$

Lemma 5 (Portier (2023)). *Under (A1) and (A2), for all $n \geq 1$, all $\delta \in (0,1)$ and all $k \in \{1,\dots,n\}$ such that $24d\log(12n/\delta) \leq k \leq r_0^d nbcV_d/2$, 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 measurable function and assume there exists an event $A \subset \Omega$. Denote $m = \mathbb{E}[\phi(X) \mid A]$ and $p = 1 - P(A)$. If $\|\phi\|_\infty := \sup_{x\in\Omega} |\phi(x)| \leq \infty$, then*

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

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

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

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

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

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

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

and the result follows by integrating with respect to μ. □

B.2 Proof of Lemma 2

Because the Voronoi cells define a partition of $\mathbb{R}^d$, we have for any $x \in \mathbb{R}^d$,

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

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

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

B.3 Proof of Lemma 3

First, concerning the moments of 1-NN distance, the proof bears resemblance with the proof of Theorem 2.3 in Biau and Devroye (2015). Let $x \in \mathcal{X}$ and start with

$$\begin{aligned}
\mathbb{P}(|\hat{\tau}_n(x)| > t) &= \mathbb{P}\left(\min_{i=1,\dots,n} \|X_i - x\| > t\right) \\
&= [\mathbb{P}(\|X_1 - x\| > t)]^n \\
&= [1 - \mathbb{P}(B(x,t))]^n \\
&\leq \exp[-n\,\mathbb{P}(B(x,t))] \\
&\leq \exp(-nt^d V_d bc).
\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{E}\left[\hat{\tau}_n(x)^q\right] &= \int_0^\infty \mathbb{P}\left(|\hat{\tau}_n(x)| > t^{1/q}\right) \mathrm{d}t \\
&\leq \int_0^\infty \exp(-nt^{d/q} V_d bc)\, \mathrm{d}t \\
&= (nV_d bc)^{-q/d}\,(q/d) \int_0^\infty \exp(-u) u^{q/d-1}\, \mathrm{d}u \\
&= (nV_d bc)^{-q/d}\,\Gamma(q/d+1).
\end{aligned}$$

Then, concerning the moments of k-NN distance, the proof is based on the one of Theorem 2.4 in Biau and Devroye (2015). Partition the set $X_1, \dots, X_n$ into $2k$ sets of sizes $n_1, \dots, n_{2k}$, with

$$\sum_{j=1}^{2k} n_j = n \quad \text{and} \quad \left\lfloor \frac{n}{2k} \right\rfloor \leq n_j \leq \left\lfloor \frac{n}{2k} \right\rfloor + 1.$$

Let $\hat{N}_{n_j}(x, j)$ be the nearest neighbor of x among all X_i's in the j-th group. Observe that, deterministically,

$$\|\hat{N}_{n,k}(x) - x\| \leq \frac{1}{k} \sum_{j=1}^{2k} \|\hat{N}_{n_j}(x,j) - x\|$$

and, similarly,

$$\|\hat{N}_{n,k}(x) - x\|^q \leq \frac{1}{k} \sum_{j=1}^{2k} \|\hat{N}_{n_j}(x,j) - x\|^q,$$

because at least k of these nearest neighbors have values that are at least $\|\hat{N}_{n,k}(x) - x\|$. This last inequality may be written as

$$\|\hat{N}_{n,k}(x) - x\|^q \leq \frac{1}{k} \sum_{j=1}^{2k} \hat{\tau}_{n_j}(x)^q.$$

Applying the previous upper bound for 1-NN moment gives

$$\begin{aligned}
\mathbb{E}\left[\|\hat{N}_{n,k}(x) - x\|^q\right] &\leq \frac{1}{k} \sum_{j=1}^{2k} (n_j V_d bc)^{-q/d}\, \Gamma(q/d+1) \\
&= \frac{(V_d bc)^{-q/d} \Gamma(q/d+1)}{k} \sum_{j=1}^{2k} \left(\frac{1}{n_j}\right)^{q/d} \\
&= \frac{2^{q/d} (V_d bc)^{-q/d} \Gamma(q/d+1)}{k} \sum_{j=1}^{2k} \left(\frac{1}{2n_j}\right)^{q/d} \\
&\leq \frac{2^{q/d} (V_d bc)^{-q/d} \Gamma(q/d+1)}{k} \sum_{j=1}^{2k} \left(\frac{2k}{n}\right)^{q/d} \\
&= 2^{2q/d+1} \Gamma(q/d+1) (nV_d bc/k)^{-q/d}.
\end{aligned}$$

C Proofs of Propositions

C.1 Proof of Proposition 1

By conditioning on $\tilde{X}_1, \ldots, \tilde{X}_N$, we obtain that

$$\begin{aligned}\mathbb{E}\left[|\hat{\mu}_n^{(CV)}(\varphi) - \mu(\varphi)|^2 | \tilde{X}_1, \ldots, \tilde{X}_N\right] &= n^{-1} \operatorname{var}[(\varphi(X_1) - \hat{\varphi}(X_1))|\tilde{X}_1, \ldots, \tilde{X}_N] \\ &\leq n^{-1} \mathbb{E}[(\varphi(X_1) - \hat{\varphi}(X_1))^2 | \tilde{X}] \\ &= n^{-1} \int (\varphi - \hat{\varphi})^2 d\mu.\end{aligned}$$

and taking the expectation with respect to the $\tilde{X}_1, \ldots, \tilde{X}_N$ leads to the result. □

C.2 Proof of Proposition 2

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.3 Proof of Proposition 3

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

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

Denote $R_n = \left|\frac{1}{n}\sum_{i=1}^n \left[\mu(\hat{\varphi}_n^{(i)}) - \mu(\hat{\varphi}_n)\right]\right|$. Then using the triangle inequality and the definition of $\hat{\varphi}$ gives

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

And because φ is L-Lipschitz, one has

$$\begin{aligned}
R_n &\leq \frac{L}{n}\sum_{i=1}^n \int_{S_{n,i}^\circ} \|\hat{N}_n^{(i)}(x) - X_i\| \mathrm{d}\mu(x) \\
&\leq \frac{L}{n}\sum_{i=1}^n \int_{S_{n,i}^\circ} \left(\|\hat{N}_n^{(i)}(x) - x\| + \|x - X_i\|\right) \mathrm{d}\mu(x) \\
&= \frac{L}{n}\sum_{i=1}^n \int_{S_{n,i}^\circ} \left[\hat{\tau}_n^{(i)}(x) + \hat{\tau}_n(x)\right] \mathrm{d}\mu(x).
\end{aligned}$$

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

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

For $x \in S_{n,i}^\circ$, the nearest neighbor in $\{X_1, \ldots, X_n\}$ is X_i. Hence, for $x \in S_{n,i}^\circ$,

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

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

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

and

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

Consequently, by Lemma 3,

$$\begin{aligned}
\mathbb{E}[R_n] &\leq \frac{L}{n}\int_{\mathcal{X}} \mathbb{E}\left[\hat{\tau}_{n,2}(x)+\hat{\tau}_n(x)\right]\mathrm{d}\mu(x)\\
&\leq \frac{L}{n}\left(\sup_{x\in\mathcal{X}}\mathbb{E}\left[\hat{\tau}_{n,2}(x)\right]+\sup_{x\in\mathcal{X}}\mathbb{E}\left[\hat{\tau}_n(x)\right]\right)\\
&\leq \frac{L}{n}\left(2^{2/d+1}\Gamma(1/d+1)(nV_dbc/2)^{-1/d}+(nV_dbc)^{-1/d}\Gamma(1/d+1)\right)\\
&= (2^{3/d+1}+1)L(V_dbc)^{-1/d}\Gamma(1/d+1)n^{-1-1/d}
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[R_n^2] &\leq \frac{2L^2}{n}\int_{\mathcal{X}} \mathbb{E}\left[\hat{\tau}_{n,2}(x)^2+\hat{\tau}_n(x)^2\right]\mathrm{d}\mu(x)\\
&\leq \frac{2L^2}{n}\left(\sup_{x\in\mathcal{X}}\mathbb{E}\left[\hat{\tau}_{n,2}(x)^2\right]+\sup_{x\in\mathcal{X}}\mathbb{E}\left[\hat{\tau}_n(x)^2\right]\right)\\
&\leq \frac{2L^2}{n}\left(2^{4/d+1}\Gamma(2/d+1)(nV_dbc/2)^{-2/d}+(nV_dbc)^{-2/d}\Gamma(2/d+1)\right)\\
&= 2(2^{6/d+1}+1)L^2(V_dbc)^{-2/d}\Gamma(2/d+1)n^{-1-2/d}.
\end{aligned}$$

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

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

Now it is suitable to 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 $(n-1)$ element 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 $\mathcal{N}^{(1,2)}:\mathbb{R}^d\to\mathbb{R}^d$ such that $\mathcal{N}^{(1,2)}(x)$ is the nearest neighbor to x among the sample $\{X_1,\ldots,X_n\}$ without X_1 and X_2. We write (using that $\mathcal{N}^{(1,2)}(x) = X_i$ whenever $x \in S_{n-1,i}^{(1)}$ for $i \geq 3$),

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

It follows that

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

with $\hat{L}^{(1)}(x) = (\varphi(X_2) - \varphi(\mathcal{N}^{(1,2)}(x)))\, \mathbb{1}_{S^{(1)}_{n-1,2}}(x)$. Therefore,

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

Denote

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

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

$$\begin{aligned}
\mathbb{E}[\{Y_1 - Y_{n,1}\}\{Y_2 - Y_{n,2}\}] = \mathbb{E}[A_1A_2] &+ \mathbb{E}[A_1B_2] + \mathbb{E}[A_1C_2]\\
&+ \mathbb{E}[B_1A_2] + \mathbb{E}[B_1B_2] + \mathbb{E}[B_1C_2]\\
&+ \mathbb{E}[C_1A_2] + \mathbb{E}[C_1B_2] + \mathbb{E}[C_1C_2].
\end{aligned}$$

Since A_1 and A_2 are independent, $\mathbb{E}[A_1A_2] = 0$. This also applies to $\mathbb{E}[A_1C_2]$ and $\mathbb{E}[A_2C_1]$. Considering $\mathbb{E}[A_1B_2]$ gives

$$\begin{aligned}
\mathbb{E}[A_1B_2] &= \mathbb{E}\left[Y_1(\hat{L}^{(2)}(X_2) - \mu(\hat{L}^{(2)}))\right]\\
&= \mathbb{E}\left[\mathbb{E}\left[Y_1(\hat{L}^{(2)}(X_2) - \mu(\hat{L}^{(2)})) \mid X_1, X_3, \ldots, X_n\right]\right]\\
&= \mathbb{E}\left[Y_1\, \mathbb{E}\left[(\hat{L}^{(2)}(X_2) - \mu(\hat{L}^{(2)})) \mid X_1, X_3, \ldots, X_n\right]\right] = 0.
\end{aligned}$$

Due to similar reasoning, $\mathbb{E}[B_1A_2] = 0$, $\mathbb{E}[B_1C_2] = 0$ and $\mathbb{E}[C_1B_2] = 0$. For $\mathbb{E}[C_1C_2]$, we have

$$\begin{aligned}
\mathbb{E}[C_1C_2] &= \mathbb{E}\left[(\hat{\varphi}^{(1,2)}_n(X_1) - \mu(\hat{\varphi}^{(1,2)}_n))(\hat{\varphi}^{(1,2)}_n(X_2) - \mu(\hat{\varphi}^{(1,2)}_n))\right]\\
&= \mathbb{E}\left[\mathbb{E}\left[(\hat{\varphi}^{(1,2)}_n(X_1) - \mu(\hat{\varphi}^{(1,2)}_n))(\hat{\varphi}^{(1,2)}_n(X_2) - \mu(\hat{\varphi}^{(1,2)}_n)) \mid X_3, \ldots, X_n\right]\right] = 0.
\end{aligned}$$

Therefore, we get

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

The use of Cauchy-Schwarz inequality gives $\|(A-B)(C-D)\|_1 \le \|A-B\|_2\|C-D\|_2$ and the fact that B and D are conditional expectation of A and C, respectively, leads to $\|(A-B)(C-D)\|_1 \le \|A\|_2\|C\|_2 = \|A\|_2^2$. As a result,

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

Using the Lipschitz property, we obtain

$$\begin{aligned}
|\hat{L}^{(1)}(x)| &= |\varphi(X_2) - \varphi(\mathcal{N}^{(1,2)}(x))|\, \mathbb{1}_{S^{(1)}_{n-1,2}}(x)\\
&= |\varphi(\mathcal{N}^{(1)}(x)) - \varphi(\mathcal{N}^{(1,2)}(x))|\, \mathbb{1}_{S^{(1)}_{n-1,2}}(x)\\
&\le L\|\mathcal{N}^{(1)}(x) - \mathcal{N}^{(1,2)}(x)\|\, \mathbb{1}_{S^{(1)}_{n-1,2}}(x)\\
&\le 2L\|x - \mathcal{N}^{(1,2)}(x)\|\, \mathbb{1}_{S^{(1)}_{n-1,2}}(x)\\
&= 2L\|x - \mathcal{N}^{(1,2)}(x)\|\, \mathbb{1}_{B(x,\hat{\tau}^{(1)}(x))}(X_2)\\
&\le 2L\|x - \mathcal{N}^{(1,2)}(x)\|\, \mathbb{1}_{B(x,\hat{\tau}^{(1,2)}(x))}(X_2)
\end{aligned}$$

Hence

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

Moreover,

$$\begin{aligned}\mu\{B(x,\hat{\tau}^{(1,2)}(x))\} &= \int_{B(x,\hat{\tau}^{(1,2)}(x))} f(z)\mathrm{d}z \\ &\leq U \int_{B(x,\hat{\tau}^{(1,2)}(x))\cap\mathcal{X}} \mathrm{d}z \\ &\leq U\hat{\tau}^{(1,2)}(x)^d V_d\end{aligned}$$

Using that $\hat{\tau}^{(1,2)}(x) = \|x - \mathcal{N}^{(1,2)}(x)\|$, we obtain that

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

Applying to the term

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

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

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

All this together gives

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

Applying Lemma 3 to $\mathbb{E}[\hat{\tau}_{n-1}(X_1)^2]$ and to $\mathbb{E}[\hat{\tau}_{n-2}(X_1)^{2+d}]$, we get

$$\mathbb{E}[\hat{\tau}_{n-1}(X_1)^2] \leq ((n-1)V_dbc)^{-2/d}\,\Gamma(2/d+1)$$

and

$$\mathbb{E}[\hat{\tau}_{n-2}(X_1)^{2+d}] \leq ((n-2)V_dbc)^{-2/d-1}\,\Gamma(2/d+2).$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[|\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi) - \mu(\varphi)|^2\right] \leq\; & L^2n^{-1}\left((n-1)V_dbc\right)^{-2/d}\Gamma(2/d+1) \\ & + 4UV_dL^2\left((n-2)V_dbc\right)^{-2/d-1}\Gamma(2/d+2).\end{aligned}$$

Rearranging the terms gives

$$\begin{aligned}&\mathbb{E}\left[\left|\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi) - \mu(\varphi)\right|^2\right] \\ &\leq L^2n^{-1}\left((n-1)V_dbc\right)^{-2/d}\Gamma(2/d+1) + 4UV_dL^2\left((n-2)V_dbc\right)^{-2/d-1}\Gamma(2/d+2) \\ &\leq L^2n^{-1}\left((n-1)V_dbc\right)^{-2/d}\Gamma(2/d+1) + 4(U/bc)L^2\left((n-2)V_dbc\right)^{-2/d}(n-2)^{-1}\Gamma(2/d+2) \\ &\leq L^2(n-2)^{-1}\left((n-2)V_dbc\right)^{-2/d}\left[\Gamma(2/d+1) + 4(U/bc)\Gamma(2/d+2)\right]\end{aligned}$$

Since $d \geq 1$, it holds that both $(2/d+1)$ and $(2/d+2)$ are in $[1,4]$. Using that $1 \leq \Gamma(x) \leq 6$ whenever $1 \leq x \leq 4$ we first get

$$\mathbb{E}\left[\left|\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi) - \mu(\varphi)\right|^2\right] \leq L^2(n-2)^{-1}\left((n-2)V_dbc\right)^{-2/d}6(1+4(U/b)).$$

Then, since $n \geq 4$ we have $n - 2 \geq n/2$ and obtain

$$\mathbb{E}\left[\left|\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi) - \mu(\varphi)\right|^2\right] \leq 48L^2n^{-1}\left(nV_dbc\right)^{-2/d}(U/bc)\left[(bc/U)+4\right].$$

Using $(bc/U) \leq 1$ finally gives the stated bound. □

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

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\mu(\hat{\varphi}_n^{(i)}) - \mu(\hat{\varphi}_n)\right|\right] \leq (2^{3/d+1}+1)L(V_d b)^{-1/d}\Gamma(1/d+1)n^{-1-1/d}.$$

Using that $1 \leq \Gamma(x) \leq 2$ for $1 \leq x \leq 3$ and $(2^{6/d+1}+1) \leq 2^7+1 = 129$, we obtain

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\mu(\hat{\varphi}_n^{(i)}) - \mu(\hat{\varphi}_n)\right|^2\right] \leq 516L^2(V_d b)^{-2/d}n^{-1-2/d}, \tag{10}$$

By Minkowski's inequality, we have

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

Conclude by using the bounds of Theorem 1 and (10) along with $\sqrt{516} \leq 23$.

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 $k = \lceil 24d\log(12n/\delta_n)\rceil$,

$$\begin{aligned}\sup_{x\in\mathcal{X}}\hat{\tau}_{n,k}(x) \leq \left(\frac{2k}{nbcV_d}\right)^{1/d} &= \left(\frac{2\lceil 24d\log(12n/\delta_n)\rceil}{nbcV_d}\right)^{1/d} \\ &\leq \left(\frac{49d\log(12n/\delta_n)}{nbcV_d}\right)^{1/d} = \bar{\tau}_n.\end{aligned}$$

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 (9) 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

$$\phi_1 \equiv \phi_1(X_1,\ldots,X_n) = \frac{1}{n}\sum_{i=1}^{n}\{\varphi(X_i) - \hat{\varphi}_n^{(i)}(X_i)\}. \tag{11}$$

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),

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

By the triangle inequality, we get, in hopefully obvious notation,

$$\begin{aligned}
\left\|\hat{\tilde{N}}_n^{(i)}(X_i) - \hat{N}_n^{(i)}(X_i)\right\| &\le \left\|X_i - \hat{\tilde{N}}_n^{(i)}(X_i)\right\| + \left\|X_i - \hat{N}_n^{(i)}(X_i)\right\| \\
&\le \left\|X_i - \hat{\tilde{N}}_n^{(i,\ell)}(X_i)\right\| + \left\|X_i - \hat{N}_n^{(i,\ell)}(X_i)\right\| \\
&\le 2\left\|X_i - \hat{N}_n^{(i,\ell)}(X_i)\right\| \le 2\sup_{x\in\mathcal{X}}\hat{\tau}_{n,3}(x), \qquad (12)
\end{aligned}$$

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,

$$\sum_{\substack{i=1\\ i\neq\ell}}^{n}\mathbb{1}_{\{\hat{\tilde{N}}_n^{(i)}(X_i)=\tilde{X}_\ell \text{ or } \hat{N}_n^{(i)}(X_i)=X_\ell\}} \le 2\psi_d. \qquad (13)$$

Therefore, by (12) and (13),

$$\frac{1}{n}\sum_{\substack{i=1\\ i\neq\ell}}^{n}\left|\hat{\tilde{\varphi}}_n^{(i)}(X_i) - \hat{\varphi}_n^{(i)}(X_i)\right| \le \frac{4L\psi_d}{n}\sup_{x\in\mathcal{X}}\hat{\tau}_{n,3}(x). \qquad (14)$$

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

$$\begin{aligned}
&\frac{1}{n}\left|\{\varphi(X_\ell) - \varphi(\tilde{X}_\ell)\} - \{\hat{\varphi}_n^{(\ell)}(X_\ell) - \hat{\varphi}_n^{(\ell)}(\tilde{X}_\ell)\}\right| \\
&\le \frac{L}{n}\left[\left\|X_\ell - \hat{N}_n^{(\ell)}(X_\ell)\right\| + \left\|\tilde{X}_\ell - \hat{N}_n^{(\ell)}(\tilde{X}_\ell)\right\|\right] \\
&\le \frac{2L}{n}\sup_{x\in\mathcal{X}}\hat{\tau}_{n,3}(x). \qquad (15)
\end{aligned}$$

On the event A, by (14) and (15), we therefore have, since $k \ge 3$ for our choice of k in Step 1, the bound

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

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

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

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

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

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 and (A3), we have, for $x \in S_{n+1,i} \cup S_{n+1,\ell} \cup S_{n+1,n+1}$,

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

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

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

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

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

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

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

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

$$\max_{j=1,\dots,n+1} \mu(S_{n+1,j}) \le U \cdot \bar{\tau}_n^d V_d, \tag{16}$$

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

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

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 $1 - \delta_n$,

$$\begin{aligned} D_{\ell,3} &= \left|\phi_3(X_1, \dots, X_\ell, \dots, X_n) - \phi_3(X_1, \dots, \tilde{X}_\ell, \dots, X_n)\right| \le |D_{\ell,1}| + |D_{\ell,2}| \\ &\le 2Ln^{-1}(2\psi_d + 1)\bar{\tau}_n + 2L\bar{\tau}_n \cdot \left(n^{-1} + 2UV_d\bar{\tau}_n^d \right) \\ &= 2Ln^{-1}(2\psi_d + 2)\bar{\tau}_n + 4LUV_d\bar{\tau}_n^{1+d} \\ &= 4L(\psi_d + 1)n^{-1} \left(\frac{49d \log(12n/\delta_n)}{nbcV_d} \right)^{1/d} + 4LUV_d \left(\frac{49d \log(12n/\delta_n)}{nbcV_d} \right)^{1+1/d} \\ &\le C_1 \left(\frac{\log(12n/\delta_n)}{n} \right)^{1+1/d} = c_{\ell,3} \quad \text{with} \quad C_1 = 4L \left(\frac{49d}{bcV_d} \right)^{1/d} \left(\psi_d + 1 + \frac{49dU}{bc} \right). \end{aligned}$$

We have

$$\sum_{\ell=1}^{n} c_{\ell,3} = n c_{\ell,3} = C_1 n^{-1/d} (\log(12n/\delta_n))^{1+1/d},$$
$$\sum_{\ell=1}^{n} c_{\ell,3}^2 = n c_{\ell,3}^2 = C_1^2 n^{-1-2/d} (\log(12n/\delta_n))^{2+2/d}.$$

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 $t \geq \delta_n \sum_{\ell=1}^{n} c_{\ell,3} = \delta_n C_1 n^{-1/d} (\log(12n/\delta_n))^{1+1/d}$, we get

$$\mathbb{P}\left(|\phi_3 - \mathbb{E}(\phi_3 \mid A)| \geq t\right) \leq \delta_n + 2\exp\left(-\frac{2\left\{t - \delta_n C_1 n^{-1/d} (\log(12n/\delta_n))^{1+1/d}\right\}^2}{C_1^2 n^{-1-2/d} (\log(12n/\delta_n))^{2+2/d}}\right). \quad (17)$$

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

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

provided t is such that the exponential function in (17) is bounded by $\delta/2$, which happens if

$$\begin{aligned} t &\geq \sqrt{\frac{C_1^2 \log(2/\delta)(\log(12n/\delta_n))^{2+2/d}}{2n^{1+2/d}}} + \frac{C_1 \delta_n (\log(12n/\delta_n))^{1+1/d}}{n^{1/d}} \\ &= C_1 \left(n^{-1/2}\sqrt{\log(2/\delta)/2} + \delta_n\right) \frac{(\log(12n/\delta_n))^{1+1/d}}{n^{1/d}}. \end{aligned}$$

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

$$|\phi_3 - \mathbb{E}(\phi_3 \mid A)| < C_1 \left(n^{-1/2}\sqrt{\log(2/\delta)/2} + \delta_n\right) \frac{(\log(12n/\delta_n))^{1+1/d}}{n^{1/d}}$$

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

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

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

$$\begin{aligned} \left|\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi) - \mu(\varphi)\right| &= |\phi_3 - \mathbb{E}(\phi_3)| \\ &\leq |\phi_3 - \mathbb{E}(\phi_3 \mid A)| + |\mathbb{E}(\phi_3 \mid A) - \mathbb{E}(\phi_3)| \\ &\leq C_1 \left(n^{-1/2}\sqrt{\log(2/\delta)/2} + \delta_n\right) \frac{(\log(12n/\delta_n))^{1+1/d}}{n^{1/d}} + 6\delta_n \|\varphi\|_\infty. \end{aligned}$$

Step 4. Fix $\varepsilon > 0$ and choose $\delta_n = \min(\varepsilon/2, n^{-1/2-1/d})$, $\delta = \varepsilon - \delta_n$. Then $\delta = \varepsilon - \min(\varepsilon/2, n^{-1/2-1/d}) = \max(\varepsilon/2, \varepsilon - n^{-1/2-1/d})$. Considering the case when $\varepsilon \leq \frac{2}{n^{1/2+1/d}}$, we get

$$\begin{aligned} &\left|\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi) - \mu(\varphi)\right| \\ &\leq C_1 \left(\frac{\sqrt{\log(4/\varepsilon)/2}}{n^{1/2}} + \frac{\varepsilon}{2}\right) \frac{(\log(24n/\varepsilon))^{1+1/d}}{n^{1/d}} + 3\varepsilon \|\varphi\|_\infty \\ &\leq C_1 \left(\frac{\sqrt{\log(4/\varepsilon)/2}}{n^{1/2}} + \frac{1}{n^{1/2+1/d}}\right) \frac{(\log(24n/\varepsilon))^{1+1/d}}{n^{1/d}} + \frac{6\|\varphi\|_\infty}{n^{1/2+1/d}} \\ &\leq C_1 \left(\sqrt{\log\left(\frac{4}{\varepsilon}\right)\Big/ 2} + \frac{1}{n^{1/d}}\right) \frac{(\log(24n/\varepsilon))^{1+1/d}}{n^{1/2+1/d}} + \frac{6\|\varphi\|_\infty}{n^{1/2+1/d}}. \end{aligned}$$

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

$$\begin{aligned}
&\left|\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi) - \mu(\varphi)\right| \\
&\le C_1 \left(\sqrt{\log\left(\frac{2}{\varepsilon - n^{-1/2-1/d}}\right) \Big/ 2} + \frac{1}{n^{1/d}} \right) \frac{(\log(12n^{3/2+1/d}))^{1+1/d}}{n^{1/2+1/d}} + \frac{6\,\|\varphi\|_\infty}{n^{1/2+1/d}} \\
&\le C_1 \left(\sqrt{\log\left(\frac{4}{\varepsilon}\right) \Big/ 2} + \frac{1}{n^{1/d}} \right) \frac{(3\log(3n))^{1+1/d}}{n^{1/2+1/d}} + \frac{6\,\|\varphi\|_\infty}{n^{1/2+1/d}}.
\end{aligned}$$

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

$$\begin{aligned}
&\left|\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi) - \mu(\varphi)\right| \le \frac{6\,\|\varphi\|_\infty}{n^{1/2+1/d}} \\
&+C_1 \left(\sqrt{\log\left(\frac{4}{\varepsilon}\right) \Big/ 2} + \frac{1}{n^{1/d}} \right) \times \begin{cases} \dfrac{(\log(24n/\varepsilon))^{1+1/d}}{n^{1/2+1/d}}, & \text{for } \varepsilon \le \frac{2}{n^{1/2+1/d}}, \\ \dfrac{(3\log(3n))^{1+1/d}}{n^{1/2+1/d}}, & \text{for } \varepsilon > \frac{2}{n^{1/2+1/d}}. \end{cases}
\end{aligned}$$

Since $\sum_{i=1}^n w_{n,i} = 1$, then 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 $C_\varphi = \frac{1}{2}\{\sup_{x\in\mathcal{X}} \varphi(x) - \inf_{x\in\mathcal{X}} \varphi(x)\}$.

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

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

with ϕ_1 as in Eq. (11) and with

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

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

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

For $i = 1,\dots,n+1$, 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

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

Moreover, by the triangle inequality and the Lipschitz property, we have, for $x \in \mathcal{X} \cap (S_{n+1,\ell} \cup S_{n+1,n+1})$,

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

On the event A, we thus obtain

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

By Eq. (16) in Step 2b in the proof of Theorem 2 for $\hat{\mu}_n^{(\mathrm{NN-loo})}$, we have, on the event A,

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

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

$$|D_{\ell,4}| \le 4LUV_d\bar{\tau}_n^{1+d}. \tag{18}$$

Step 2. By (18) and by the result of Step 1 in the proof of Theorem 2 for $\hat{\mu}_n^{(\mathrm{NN-loo})}$, we have

$$\begin{aligned}
D_{\ell,5} &= \left|\phi_5(X_1,\ldots,X_\ell,\ldots,X_n) - \phi_5(X_1,\ldots,\tilde{X}_\ell,\ldots,X_5)\right| \\
&\le |D_{\ell,1}| + |D_{\ell,4}| \\
&\le 2Ln^{-1}(2\psi_d+1)\bar{\tau}_n + 4LUV_d\bar{\tau}_n^{1+d} \\
&= 2Ln^{-1}(2\psi_d+1)\left(\frac{49d\log(12n/\delta_n)}{nbcV_d}\right)^{1/d} + 4LUV_d\left(\frac{49d\log(12n/\delta_n)}{nbcV_d}\right)^{1+1/d} \\
&\le C_2\left(\frac{\log(12n/\delta_n)}{n}\right)^{1+1/d} = c_{\ell,5} \quad \text{with} \quad C_2 = 2L\left(\frac{49d}{bcV_d}\right)^{1/d}\left(2\psi_d+1+\frac{98dU}{bc}\right).
\end{aligned}$$

We have

$$\begin{aligned}
\sum_{\ell=1}^n c_{\ell,5} &= nc_{\ell,5} = C_2 n^{-1/d}(\log(12n/\delta_n))^{1+1/d}, \\
\sum_{\ell=1}^n c_{\ell,5}^2 &= nc_{\ell,5}^2 = C_2^2 n^{-1-2/d}(\log(12n/\delta_n))^{2+2/d}.
\end{aligned}$$

Thanks to Theorem 3, it follows that, for $t \ge \delta_n \sum_{\ell=1}^n c_{\ell,5} = \delta_n C_2 n^{-1/d}(\log(12n/\delta_n))^{1+1/d}$,

$$\mathbb{P}\left(|\phi_5 - \mathbb{E}(\phi_5 \mid A)| \ge t\right) \le \delta_n + 2\exp\left(-\frac{2\left\{t - \delta_n C_2 n^{-1/d}(\log(12n/\delta_n))^{1+1/d}\right\}^2}{C_2^2 n^{-1-2/d}(\log(12n/\delta_n))^{2+2/d}}\right).$$

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

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

provided

$$t \ge C_2\left(n^{-1/2}\sqrt{\log(2/\delta)/2} + \delta_n\right)\frac{(\log(12n/\delta_n))^{1+1/d}}{n^{1/d}}.$$

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

$$|\phi_5 - \mathbb{E}(\phi_5 \mid A)| < C_2\left(n^{-1/2}\sqrt{\log(2/\delta)/2} + \delta_n\right)\frac{(\log(12n/\delta_n))^{1+1/d}}{n^{1/d}}. \tag{19}$$

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

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

Recall $\mathbb{E}[\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi)] = \mu(\varphi)$. On the event on which (19) holds, we have, in view of Proposition 3, for $n \ge 4$,

$$\begin{aligned}
&\left|\hat{\mu}_n^{(\mathrm{NN})}(\varphi) - \mu(\varphi)\right| \\
\le& |\phi_5 - \mathbb{E}(\phi_5 \mid A)| + |\mathbb{E}(\phi_5 \mid A) - \mathbb{E}(\phi_5)| + \left|\mathbb{E}\left[\hat{\mu}_n^{(\mathrm{NN})}(\varphi)\right] - \mathbb{E}\left[\hat{\mu}_n^{(\mathrm{NN-loo})}(\varphi)\right]\right| \\
\le& C_2\left(n^{-1/2}\sqrt{\log(2/\delta)/2} + \delta_n\right)\frac{(\log(12n/\delta_n))^{1+1/d}}{n^{1/d}} + 6\delta_n\|\varphi\|_\infty \\
&+ \frac{(2^{3/d+1}+1)L(V_d bc)^{-1/d}\Gamma(1/d+1)}{n^{1+1/d}} \\
\le& C_2\left(n^{-1/2}\sqrt{\log(2/\delta)/2} + \delta_n\right)\frac{(\log(12n/\delta_n))^{1+1/d}}{n^{1/d}} + 6\delta_n\|\varphi\|_\infty \\
&+ \frac{(2^{3/d+1}+1)L(V_d bc)^{-1/d}}{n^{1+1/d}}.
\end{aligned}$$

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 $1-\varepsilon$,

$$\left|\hat{\mu}_n^{(\mathrm{NN})}(\varphi) - \mu(\varphi)\right| \leq \frac{6\,\|\varphi\|_\infty}{n^{1/2+1/d}} + \frac{(2^{3/d+1}+1)L(V_d bc)^{-1/d}}{n^{1+1/d}} + C_2\left(\sqrt{\log\left(\frac{4}{\varepsilon}\right)\Big/ 2} + \frac{1}{n^{1/d}}\right) \times \begin{cases} \dfrac{(\log(24n/\varepsilon))^{1+1/d}}{n^{1/2+1/d}}, & \text{for } \varepsilon \leq \frac{2}{n^{1/2+1/d}}, \\ \dfrac{(3\log(3n))^{1+1/d}}{n^{1/2+1/d}}, & \text{for } \varepsilon > \frac{2}{n^{1/2+1/d}}. \end{cases}$$

□

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. □