Sliced-Wasserstein Estimation with Spherical Harmonics as Control Variates

Rémi Leluc [1]   **Aymeric Dieuleveut** [1]   **François Portier** [2]   **Johan Segers** [3]   **Aigerim Zhuman** [3]

Abstract

The Sliced-Wasserstein (SW) distance between probability measures is defined as the average of the Wasserstein distances resulting for the associated one-dimensional projections. As a consequence, the SW distance can be written as an integral with respect to the uniform measure on the sphere and the Monte Carlo framework can be employed for calculating the SW distance. Spherical harmonics are polynomials on the sphere that form an orthonormal basis of the set of square-integrable functions on the sphere. Putting these two facts together, a new Monte Carlo method, hereby referred to as Spherical Harmonics Control Variates (SHCV), is proposed for approximating the SW distance using spherical harmonics as control variates. The resulting approach is shown to have good theoretical properties, e.g., a no-error property for Gaussian measures under a certain form of linear dependency between the variables. Moreover, an improved rate of convergence, compared to Monte Carlo, is established for general measures. The convergence analysis relies on the Lipschitz property associated to the SW integrand. Several numerical experiments demonstrate the superior performance of SHCV against state-of-the-art methods for SW distance computation.

1. Introduction

The Sliced-Wasserstein (SW) distance between two probability measures, introduced in Rabin et al. (2012), shares similar topological properties with the standard Wasserstein distance (Bonnotte, 2013; Bayraktar & Guo, 2021; Nadjahi et al., 2020b) while having better properties in terms of computational complexity, especially for measures defined on spaces with relatively large dimension (Nadjahi et al., 2020b). As such, it has emerged as a powerful framework

[1]CMAP, École Polytechnique, Institut Polytechnique de Paris, France [2]CREST, ENSAI, France [3]ISBA/LIDAM, UCLouvain, Belgium. Correspondence to: Rémi Leluc <remi.leluc@gmail.com>.

for analyzing and solving a wide range of machine learning problems including generative modeling (Deshpande et al., 2018; 2019; Liutkus et al., 2019), autoencoders (Kolouri et al., 2018), Bayesian computation (Nadjahi et al., 2020a), image processing (Bonneel et al., 2015), and computational physics (Carpintero Perez et al., 2024).

Informally, for $p \geq 1$ the SW_p distance between probability distributions μ and ν on $\mathbb{R}^d$ is defined as the integral value, over θ uniformly distributed on the unit sphere $\mathbb{S}^{d-1}$, of the Wasserstein distance W_p between the projections of distributions μ and ν on the direction of θ. It takes advantage of the fact that (a) the Wasserstein distance between one-dimensional measures has a simple analytical form, being the $L^p([0,1])$-distance between their quantile functions (approximated using empirical quantiles), and (b) the integral is approximated by Monte Carlo sampling (uniformly on the unit sphere). As a consequence, two types of error affect the accuracy of most algorithms estimating the SW distance: first, the statistical error, due to the approximation of the distributions μ and ν by their empirical counterparts, which scales as $m^{-1/2}$, with m the size of the samples observed from μ and ν (Nietert et al., 2022); second, the Monte Carlo integration error, which scales as $n^{-1/2}$, with n the number of projections, as demonstrated in Nadjahi et al. (2020b) and Nietert et al. (2022). Importantly, these results establish that the sample complexity to approximate the SW distance does not depend on the dimension. As pointed out in Nadjahi et al. (2020b), even though the SW sample complexity is reduced compared to Wasserstein, it requires an additional Monte Carlo step which might offset the previous benefits if the associated variance is large.

In this paper, we propose and study a new approach to improve SW distance computation by improving the Monte Carlo estimation (step (b) above). The approach follows from a well-known variance reduction principle introducing *control variates* (functions having known integrals) in the Monte Carlo estimate. Following recent papers (Oates et al., 2017; Portier & Segers, 2019; Leluc et al., 2021; South et al., 2023), the use of many control variates can significantly reduce the variance of the Monte Carlo method when the integrand f is well-approximated in some L^2-space by the control variates. In Portier & Segers (2019) for instance, the obtained convergence rate is $n^{-1/2}\sigma_{\mathrm{cv}}$, where σ_{cv}^2 is the variance of the residual function resulting from approx-

imating f in the control variate space. By allowing the number of control variates to grow with n, the residual standard deviation σ_{cv} can vanish asymptotically, leading to an improvement of the $n^{-1/2}$ rate of the Monte Carlo estimate.

As the integral of interest in the SW distance is over the sphere, we propose to use as control variates the well-known *spherical harmonics*, which are homogeneous orthogonal polynomials defined on the sphere. Our approach, referred to as Sliced Wasserstein estimates with Spherical Harmonics as Control Variates (SHCV), leverages the property of spherical harmonics to be a an orthonormal basis of $L_2(\mathbb{S}^{d-1})$ to build a precise and flexible approximation of the integrand leading to a more accurate Monte Carlo estimate.

Related methods. Several methods have been introduced for computing the SW distance and improve upon the standard Monte Carlo estimate. Nguyen & Ho (2024) use a *single* quadratic control variate to reduce the variance. This approach significantly differs from SHCV, which relies on a potentially large number of control variates. Further, for functions defined on $\mathbb{R}^d$, Leluc et al. (2024) use the *leave-one-out* nearest neighbor regression estimate as a control variate. As soon as the integrand is Lipschitz, the rate of convergence is $n^{-1/2-1/d}$, which improves upon the $n^{-1/2}$ rate of the Monte Carlo estimate. Even though this technique has not been studied for other spaces than $\mathbb{R}^d$ (such as the sphere), good practical performance has been reported in Leluc et al. (2024) for approximating the SW distance.

In an attempt to improve the Monte Carlo estimate, Quasi-Monte Carlo (QMC) and Randomized Quasi Monte Carlo (RQMC) methods have been proposed to approximate the SW distance (Nguyen et al., 2024) based on deterministic sequences that are well spread over the integration domain. Finally, we mention the orthogonal approach of Nadjahi et al. (2021) which is based on Gaussian approximation as the dimension d grows, thus targets large dimension, and only applies for $p = 2$.

Contributions. The main contributions of this paper are summarized as follows:

- We propose a novel variance-reduced SW estimate based on control variate with spherical harmonics.

- We derive some theoretical properties as well as the convergence rate in probability on the SHCV integration error, providing rigorous underpinnings for the proposed methodology. Our analysis reveals that SHCV can achieve faster convergence rates than traditional unbiased estimators of the SW distance.

- In numerical experiments, our approach outperforms not only existing Monte Carlo approaches based on different control variates but also some recently introduced quasi Monte Carlo (QMC) methods (see related methods described above), without prohibitive additional cost.

Outline. Section 2 introduces the mathematical background of (Sliced-)Wasserstein distances while Section 3 presents MC integration with control variates. In Section 4, we introduce our estimator, based on spherical harmonics. It is empirically evaluated in Section 5. Section 6 provides a theoretical analysis and Section 7 concludes the paper. The appendices in the supplement contain proofs, algorithms and additional numerical results.

Notation. For a Borel set $\mathsf{X} \subseteq \mathbb{R}^d$, $\mathcal{P}(\mathsf{X})$ is the set of probability measures with support contained in X. Let $p \in [1, \infty)$ and $\mathcal{P}_p(\mathsf{X}) = \{\mu \in \mathcal{P}(\mathsf{X}) : \int_\mathsf{X} \|x\|^p \, \mathrm{d}\mu(x) < +\infty\}$ be the set of probability measures on X with finite moment of order p. The unit sphere on $\mathbb{R}^d$ is $\mathbb{S}^{d-1} = \{\theta \in \mathbb{R}^d : \|\theta\| = 1\}$ with surface area $|\mathbb{S}^{d-1}| = (2\pi^{d/2})/\Gamma(d/2)$, where Γ is the Euler gamma function. $\langle \cdot, \cdot \rangle$ is the Euclidean inner product and $\mathcal{N}(\mathrm{m}, \boldsymbol{\Sigma})$ is the d-Gaussian distribution with mean $\mathrm{m} \in \mathbb{R}^d$ and positive definite $d \times d$ covariance matrix $\boldsymbol{\Sigma}$. The Laplace operator is $\Delta = \partial_1^2 + \cdots + \partial_d^2$ where ∂_i is the i-th partial derivative.

2. Optimal Transport Distances: Wasserstein and Sliced-Wasserstein

2.1. Wasserstein Distance

The Wasserstein distance $\mathrm{W}_p(\mu, \nu)$ of order $p \in [1, \infty)$ between probability measures $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ is defined by

$$\mathrm{W}_p^p(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \, \mathrm{d}\pi(x, y),$$

where $\Pi(\mu, \nu) \subset \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ 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 et al., 2009, Chapter 6), it suffers from a high computational cost. When computing $\mathrm{W}_p(\mu_m, \nu_m)$ 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). However, for *univariate* distributions $\mu, \nu \in \mathcal{P}_p(\mathbb{R})$, the Wasserstein distance can be is expressed through the quantile functions F^{-1}(inverse c.d.f.) as (see Peyré et al., 2019, Section 2)

$$\mathrm{W}_p^p(\mu, \nu) = \int_0^1 |F_\mu^{-1}(t) - F_\nu^{-1}(t)|^p \, \mathrm{d}t.$$

In particular, for discrete measures $\mu_m = m^{-1} \sum_{i=1}^m \delta_{x_i}$ and $\nu_m = m^{-1} \sum_{i=1}^m \delta_{y_i}$ with $x_i, y_i \in \mathbb{R}$, the Wasserstein distance can be computed by sorting the atoms, yielding

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

where $(x_{(i)})_{i=1}^m$ and $(y_{(i)})_{i=1}^m$ are the order statistics of $(x_i)_{i=1}^m$ and $(y_i)_{i=1}^m$ respectively. The number of operations induced by the sorting step is $\mathcal{O}(m \log m)$.

2.2. Sliced-Wasserstein Distances

The Sliced-Wasserstein distance (Rabin et al., 2012; Bonneel et al., 2015; Kolouri et al., 2019) alleviates the high computational of Wasserstein distance, by *slicing* distributions and by taking advantage of the fast computation of the Wasserstein distance between *univariate* distributions (see Equation (1)).

Definition 1 (Sliced-Wasserstein distances). *For $\theta \in \mathbb{S}^{d-1}$, let $\theta^\star : \mathbb{R}^d \to \mathbb{R}$ denote the linear functional $\theta^\star(x) = \langle \theta, x \rangle$ for $x \in \mathbb{R}^d$. Let $\mathrm{P} \in \mathcal{P}(\mathbb{S}^{d-1})$ be a probability distribution on the unit sphere. The SW distance of order $p \in [1, \infty)$ based on P is defined for $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ as*

$$\mathrm{SW}_p^p(\mu, \nu, \mathrm{P}) = \int_{\mathbb{S}^{d-1}} \mathrm{W}_p^p(\theta_\sharp^\star \mu, \theta_\sharp^\star \nu) \, \mathrm{dP}(\theta), \quad (2)$$

where for any measure $\xi \in \mathcal{P}(\mathbb{R}^d)$, $\theta_\sharp^\star \xi$ is the push-forward measure by function $\theta^\star$, i.e. the distribution of the projection $\theta^\star(X)$ of a random vector $X \in \mathbb{R}^d$ with distribution ξ.

We introduce the integrand $f_{\mu,\nu}^{(p)} : \mathbb{S}^{d-1} \to \mathbb{R}$ defined by

$$\forall \theta \in \mathbb{S}^{d-1}, \quad f_{\mu,\nu}^{(p)}(\theta) = \mathrm{W}_p^p(\theta_\sharp^\star \mu, \theta_\sharp^\star \nu). \quad (3)$$

In practice, the SW distance of Equation (2) is approximated using a standard Monte Carlo estimate by *(i)* sampling a number $n \geq 1$ of independent random directions $\theta_1, \ldots, \theta_n$ from P on $\mathbb{S}^{d-1}$ and *(ii)* by averaging the values $(f_{\mu,\nu}^{(p)}(\theta_i))_i$. This is described in Algorithm 1. In the paper, we will often work with the uniform distribution on the sphere, denoted by U or $\mathcal{U}(\mathbb{S}^{d-1})$.

Computing the SW distance between discrete measures $\mu_m = m^{-1} \sum_{i=1}^m \delta_{x_i}$ and $\nu_m = m^{-1} \sum_{i=1}^m \delta_{y_i}$ with atoms $x_i, y_i \in \mathbb{R}^d$ amounts to projecting $(x_i)_{i=1}^m$ and $(y_i)_{i=1}^m$ along the n random directions $\theta_1, \ldots, \theta_n$ followed by computing the one-dimensional Wasserstein distances using Equation (1). This scheme requires $\mathcal{O}(n(dm + m \log m))$ operations which is, in general, faster than computing $\mathrm{W}_p^p(\mu_m, \nu_m)$, especially for large m.

Remark 1 (Gaussian case when $p = 2$). The Wasserstein distance between two Gaussians $\mu = \mathcal{N}(a, \mathbf{A})$ and $\nu = \mathcal{N}(b, \mathbf{B})$ is known in closed form (Dowson & Landau, 1982) as $\mathrm{W}_2^2(\mu, \nu) = \|a - b\|_2^2 + \mathfrak{B}^2(\mathbf{A}, \mathbf{B})$ where $\mathfrak{B}$ is the Bures distance (Bhatia et al., 2019) on positive definite matrices, $\mathfrak{B}^2(\mathbf{A}, \mathbf{B}) = \mathrm{Tr}(\mathbf{A}) + \mathrm{Tr}(\mathbf{B}) - 2\mathrm{Tr}[(\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2})^{1/2}]$. The Sliced-Wasserstein distance between μ and ν relies on the projected distributions $\theta_\sharp^\star \mu \sim \mathcal{N}(\theta^\top a, \theta^\top \mathbf{A}\theta)$ and $\theta_\sharp^\star \nu \sim \mathcal{N}(\theta^\top b, \theta^\top \mathbf{B}\theta)$ via the integration over $\theta \in \mathbb{S}^{d-1}$ of $\mathrm{W}_2^2(\theta_\sharp^\star \mu, \theta_\sharp^\star \nu) = |\theta^\top(a - b)|^2 + (\sqrt{\theta^\top \mathbf{A}\theta} - \sqrt{\theta^\top \mathbf{B}\theta})^2$.

3. Monte Carlo with Control Variates

This section presents the mathematical framework of Monte Carlo integration with control variates.

Algorithm 1 Sliced Wasserstein Monte Carlo

Require: $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, number of random projections n, probability $\mathrm{P} \in \mathcal{P}(\mathbb{S}^{d-1})$
1: Sample random projections $\theta_1, \ldots, \theta_n \sim \mathrm{P}$
2: Compute $f_n = (f_{\mu,\nu}^{(p)}(\theta_i))_{i=1}^n$
3: Return average $\mathrm{MC}_n = (\mathbb{1}_n/n)^\top f_n$

Monte Carlo integration. Consider a square-integrable, real-valued function $f \in L_2(\mathrm{P})$ on a probability space $(\Theta, \mathcal{F}, \mathrm{P})$ of which we would like to compute the integral

$$\mathrm{I}(f) = \int_\Theta f(\theta) \, \mathrm{dP}(\theta). \quad (4)$$

When $\mathrm{I}(f)$ does not admit a closed form or when only calls from f are available, one may rely on Monte Carlo estimates. Let $\theta_1, \ldots, \theta_n \sim \mathrm{P}$ be an independent and identically distributed (i.i.d.) random sample from P. The naive Monte Carlo estimate $\mathrm{I}_n(f)$ of $\mathrm{I}(f)$ is given by the empirical mean

$$\mathrm{I}_n(f) = \frac{1}{n} \sum_{i=1}^n f(\theta_i).$$

The Monte Carlo estimate is unbiased and has variance equal to $\sigma^2(f)/n$ where $\sigma^2(f) = \mathrm{I}[\{f - \mathrm{I}(f)\}^2]$. Increasing the sample size n reduces the variance but at the cost of an increased computation time. Instead, it is possible to find another estimate with smaller variance.

Control Variates and OLSMC. A classical way to reduce the variance of the Monte Carlo estimate consists in incorporating knowledge about some specially chosen functions. Control variates are functions $\varphi_1, \ldots, \varphi_s : \Theta \to \mathbb{R}$ with known expectations; without loss of generality, we assume that $\mathrm{I}(\varphi_j) = 0$ for $j = 1, \ldots, s$. Let $\varphi = (\varphi_1, \ldots, \varphi_s)^\top$ denote the $\mathbb{R}^s$-valued function with s control variates as elements. Since all control variates are centered, adding any linear combination of the form $\beta^\top \varphi$ with $\beta \in \mathbb{R}^s$ gives $\mathrm{I}(f - \beta^\top \varphi) = \mathrm{I}(f)$. Thus, given an i.i.d. sample $\theta_1, \ldots, \theta_n$ from P, any coefficient vector $\beta \in \mathbb{R}^s$ yields an unbiased estimate of $I(f)$, given by

$$\mathrm{I}_n(f, \beta) = \frac{1}{n} \sum_{i=1}^n \left(f(\theta_i) - \beta^\top \varphi(\theta_i)\right), \quad (5)$$

with variance equal to $\sigma^2(f - \beta^\top \varphi)/n$. The optimal coefficient $\beta_\star$ is defined as the one minimizing the variance,

$$\beta_\star \in \arg\min_{\beta \in \mathbb{R}^s} \mathrm{I}[\{f - \mathrm{I}(f) - \beta^\top \varphi\}^2].$$

By the Hilbert projection theorem, $\beta_\star^\top \varphi$ is the $L_2(\mathrm{P})$-projection of $f - \mathrm{I}(f)$ on the linear space of control variates $\mathrm{Span}\{\varphi_1, \ldots, \varphi_s\}$, as illustrated in Figure 1.
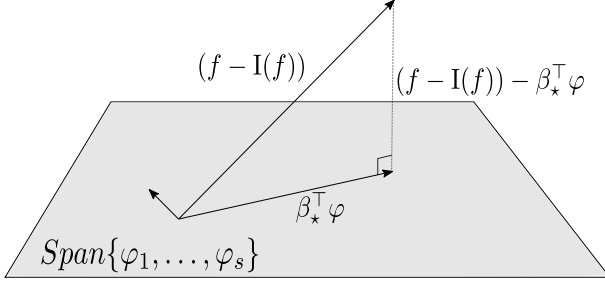
Figure 1. Visualization of the L_2-projection of the integrand f onto the linear space $\text{Span}\{\varphi_1, \ldots, \varphi_s\}$ of control variates.

Different approaches have been proposed to estimate $\beta_\star$, each of which, when used in Equation (5), produces an enhanced Monte Carlo estimate with reduced variability (Glynn & Szechtman, 2002). Following Owen (2013, Section 8.9), or more recently Portier & Segers (2019) and Leluc et al. (2021), we frame the problem as an ordinary least-squares minimization problem. The integral $\text{I}(f)$ appears as the intercept of the linear regression model with explanatory variables $\varphi_1, \ldots, \varphi_s$ and target response f, i.e.,

$$(\text{I}(f), \beta_\star(f)) \in \underset{(\alpha, \beta) \in \mathbb{R} \times \mathbb{R}^s}{\arg\min} \ \text{I}[(f - \alpha - \beta^\top \varphi)^2].$$

In practice, these quantities are estimated using the directions $\theta_1, \ldots, \theta_n$ sampled from P. The Ordinary Least Squares Monte Carlo (OLSMC) estimate $\text{I}_n^{\text{ols}}(f)$ is[1]

$$(\text{I}_n^{\text{ols}}(f), \beta_n(f)) \in \underset{(\alpha, \beta) \in \mathbb{R} \times \mathbb{R}^s}{\arg\min} \ \|f_n - \alpha \mathbb{1}_n - \Phi\beta\|_2^2, \quad (6)$$

where $f_n = (f(\theta_1), \ldots, f(\theta_n))^\top \in \mathbb{R}^n$ is the vector of evaluations of f, $\mathbb{1}_n = (1, \ldots, 1)^\top \in \mathbb{R}^n$ and $\Phi \in \mathbb{R}^{n \times s}$ is the random matrix of control variates $\Phi = (\varphi(\theta_i)^\top)_{i=1}^n$.

Remark 2 (Complexity). The Monte Carlo estimate requires $\mathcal{O}(n\omega_f)$ computations where ω_f is the cost of evaluating the integrand f. The OLSMC estimate has a complexity of $\mathcal{O}(n\omega_f + \omega(\Phi))$ where $\omega(\Phi)$ is the complexity due to the additional step of fitting the optimal control variates. Typically, $\omega(\Phi) = \mathcal{O}(ns^2 + s^3)$ where $\mathcal{O}(ns^2)$ and $\mathcal{O}(s^3)$ operations are needed to compute and invert $\Phi^\top \Phi$, though it can be reduced by using approximate solution via optimization algorithms such as (stochastic) gradient descent algorithms (Zhang, 2004). When f is heavy to compute—which is the case for $f_{\mu_m, \nu_m}^{(p)}$ with discrete measures supported on a large number m of atoms—the extra term $\omega(\Phi)$ becomes negligible compared to $\mathcal{O}(n\omega_f)$.

Remark 3 (Regularization). As the number s of control variates may increase quickly with the dimension, Leluc et al. (2021) rely on the LASSO procedure (Tibshirani, 1996) in place of Equation (6) to achieve estimation and variable selection at the same time in the regime $n \ll s$.

---

[1]It is well-defined if and only if $\mathbb{1}_n \notin \text{Span}(\Phi)$. If ill-defined, one can reduce the number of control variates such that $s < n$.

4. Sliced-Wasserstein Estimate with Spherical Harmonics as Control Variates

4.1. Spherical Harmonics

Originating from the solutions to the Laplace equation on the unit sphere, spherical harmonics (Atkinson & Han, 2012; Dai, 2013) constitute a mathematical framework with applications in computer graphics (Ramamoorthi & Hanrahan, 2001; Basri & Jacobs, 2003; Green, 2003) and machine learning (Cohen et al., 2018; Dutordoir et al., 2020), particularly in tasks involving spherical data representation.

Definition 2 (Polynomial spaces). *Let $\mathscr{P}_\ell^d$ be the space of homogeneous polynomials of degree $\ell \geq 0$ on $\mathbb{R}^d$, i.e., $\mathscr{P}_\ell^d = \text{Span}\{x_1^{a_1} \cdots x_d^{a_d} \mid a_k \in \mathbb{N}, \sum_{k=1}^d a_k = \ell\}$. Let $\mathscr{H}_\ell^d$ be the space of real harmonic polynomials, homogeneous of degree ℓ on $\mathbb{R}^d$, i.e., $\mathscr{H}_\ell^d = \{Q \in \mathscr{P}_\ell^d \mid \Delta Q = 0\}$.*

Definition 3 (Spherical harmonics). *Spherical harmonics of degree (or level) $\ell \geq 0$ are defined as the restriction of elements in $\mathscr{H}_\ell^d$ to the unit sphere $\mathbb{S}^{d-1}$, i.e., restrictions to the sphere of harmonic homogeneous polynomials with d variables of degree ℓ.*

Let N_ℓ^d denote the number of linearly independent spherical harmonics of degree $\ell \geq 0$ in dimension d, that is $N_\ell^d = \dim \mathscr{H}_\ell^d = \dim \mathscr{P}_\ell^d - \dim \mathscr{P}_{\ell-2}^d$ (Dai, 2013, Corollary 1.1.4) where it is agreed that $\dim \mathscr{P}_{\ell-2}^d = 0$ for $\ell = 0, 1$. We have

$$N_\ell^d = \frac{(2\ell + d - 2)(\ell + d - 3)!}{\ell!(d-2)!}.$$

Note that, when restricted to the sphere, all polynomials are linear combinations of spherical harmonics. Moreover, spherical harmonics of different degrees are orthogonal with respect to $\langle f, g \rangle = \int_{\mathbb{S}^{d-1}} f(\theta)g(\theta) \, d\text{U}(\theta)$, where U is the uniform distribution on $\mathbb{S}^{d-1}$ (Dai, 2013, Theorem 1.1.2). In practice, orthonormal bases can be constructed by applying the Gram–Schmidt process and an explicit basis of spherical harmonics can be written in terms of the Gegenbauer polynomials in spherical coordinates (Dai, 2013, Theorem 1.5.1). In the following, let $\varphi_{\ell, k}$ with $\ell \geq 0$ and $k = 1, \ldots, N_\ell^d$ denote a set of spherical harmonics forming an orthonormal basis of $\mathscr{H}_\ell^d$.

Proposition 1 (Hilbert basis). *The spherical harmonics $\{\varphi_{\ell, k}, 0 \leq \ell, 1 \leq k \leq N_\ell^d\}$ form an orthonormal eigenbasis of the Hilbert space $L_2(\mathbb{S}^{d-1}, \text{U})$, so that for every $f \in L_2(\mathbb{S}^{d-1})$ we have*

$$f = \sum_{\ell=0}^\infty \sum_{k=1}^{N_\ell^d} \hat{f}_{\ell, k} \varphi_{\ell, k} \quad \text{where} \quad \hat{f}_{\ell, k} = \int f \varphi_{\ell, k} \, d\text{U}.$$

Spherical harmonics thus constitute an eigenbasis of centered functions, and can therefore be used as control variates

with respect to the uniform distribution on the unit sphere. For $\ell \geq 1$ and $k = 1, \ldots, N_\ell^d$, we have

$$\mathrm{I}(\varphi_{\ell,k}) = \int_{\mathbb{S}^{d-1}} \varphi_{\ell,k}(\theta) \, \mathrm{dU}(\theta) = 0. \tag{7}$$

Orthonormality ensures that the empirical covariance matrix is well-conditioned when solving the OLS problem (6).

Remark 4 (Examples). In most applications, spherical harmonics in two and three variables are used. For $d = 2$, we have $\dim \mathscr{H}_\ell^2 = 2$ for all $\ell \geq 1$ and an orthogonal basis is given by the real and imaginary parts of $(x + \mathrm{i}y)^\ell$. In polar coordinates $(x, y) = (r\cos(t), r\sin(t)) \in \mathbb{R}^2$, this basis is $\varphi_{\ell,1}(x, y) = r^\ell \cos(\ell t)$ and $\varphi_{\ell,2}(x, y) = r^\ell \sin(\ell t)$. By restriction on the circle $\mathbb{S}^1$, spherical harmonics yield the classical Fourier expansions in cosine and sine functions and are commonly referred to as *circular harmonics*. For $d = 3$, we have $\dim \mathscr{H}_\ell^3 = 2\ell + 1$ and a classical orthogonal basis in spherical coordinates is written in terms of the associated Legendre polynomials (see Appendix D.1).

4.2. Spherical Harmonics Control Variates Estimate

With the idea of performing variance reduction when computing SW distances, we incorporate spherical harmonics as natural control variates on $\mathbb{S}^{d-1}$ into the standard Monte Carlo estimate. We propose a novel procedure called *Spherical Harmonics Control Variates* (SHCV) estimate using the ordinary least squares formulation of Equation (6).

For $p \geq 1$, consider $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ and recall the integrand $f_{\mu,\nu}^{(p)} : \theta \mapsto \mathrm{W}_p^p(\theta_\sharp^\star \mu, \theta_\sharp^\star \nu)$ so that $\mathrm{SW}_p^p(\mu, \nu) = \mathrm{I}(f_{\mu,\nu}^{(p)})$. Interestingly, this integrand is even so that for odd functions φ, the inner product vanishes, $\langle f_{\mu,\nu}^{(p)}, \varphi \rangle = 0$. The spherical harmonics expansion of $f_{\mu,\nu}^{(p)}$ (see Proposition 1) is only made of spherical harmonics with even degree. Thus, when finding the optimal control variates among spherical harmonics, we can omit the ones of odd degree as these are uncorrelated with our integrand.

Lemma 1 (Number of control variates). *In dimension d, the dimension of the linear space generated by the spherical harmonics of even degree up to maximal degree $2L$ is*

$$s_{L,d} = \sum_{\ell=1}^L N_{2\ell}^d = \binom{2L + d - 1}{d - 1} - 1.$$

For $d = 2$, the formula specializes to $s_{L,2} = 2L$ which is the number of circular harmonics of degree $2, 4, \ldots, 2L$; For $d = 3$, the formula becomes $s_{L,3} = (2L+3)L$, which is indeed the sum of $(2\ell + 1)$ for $\ell = 2, 4, \ldots, 2L$ (Remark 4).

Definition 4 (SHCV). *For $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, the SHCV estimate of maximum degree $2L$ based on projections $\theta_1, \ldots, \theta_n \in \mathbb{S}^{d-1}$ and control variates $\varphi = (\varphi_j)_{j=1}^{s_{L,d}}$ is defined as the intercept of the OLS problem (6) with the*

Algorithm 2 Spherical Harmonics Control Variate Estimate

Require: $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, number of random projections n, $\mathrm{P} \in \mathcal{P}(\mathbb{S}^{d-1})$, spherical harmonics $\varphi = (\varphi_j)_{j=1}^s$
1: Sample random projections $\theta_1, \ldots, \theta_n \sim \mathrm{P}$
2: Compute $f_n = (f_{\mu,\nu}^{(p)}(\theta_i))_{i=1}^n$, $\Phi = (\varphi(\theta_i)^\top)_{i=1}^n$
3: Solve $(\mathrm{I}_n^{\mathrm{ols}}, \beta_n) \in \arg\min_{\alpha,\beta} \|f_n - \alpha\mathbb{1}_n - \Phi\beta\|_2^2$
4: Return SHCV$_n = \mathrm{I}_n^{\mathrm{ols}}$

---

integrand $f_{\mu,\nu}^{(p)}$ as response and all spherical harmonics of even degree from 2 up to $2L$ as $n \times s_{L,d}$ covariate matrix $\Phi = (\varphi(\theta_i)^\top)_{i=1}^n$, i.e.,

$$\mathrm{SHCV}_{n,L}^p(\mu, \nu) = \mathrm{I}_n^{\mathrm{ols}}(f_{\mu,\nu}^{(p)}). \tag{8}$$

The procedure is described in Algorithm 2. In the algorithm, the number of spherical harmonics is specified in advance. A selection based on a lasso procedure with cross-validation for the penalization parameter would be possible too, but is not pursued here for the sake of computing efficiency.

Remark 5 (Multiple integrals). Computational benefits occur when there are multiple integrands, since the SHCV estimate can be represented as a linear rule $w^\top f_n$ (Portier & Segers, 2019; Leluc et al., 2021), where the weight vector $w \in \mathbb{R}^n$ does not depend on the integrands. If K integrals need to be evaluated, the computation time of SHCV becomes $\mathcal{O}(Kn\omega_f + \omega(\Phi))$ compared to $\mathcal{O}(Kn\omega_f)$ for standard Monte Carlo, and the additional cost $\omega(\Phi)$ becomes negligible for large K. This is particularly relevant for SW-based kernels involving many pairwise computations $\mathrm{SW}_2^2(\mu_i, \nu_j)$; see Algorithms 9 and 10 in Appendix B.4.

Remark 6 (Quadratic control variates). Any symmetric $d \times d$ matrix M yields a quadratic control variate $\varphi_M(\theta) = \theta^\top M\theta - \mathrm{Tr}(M)/d$ as $\mathrm{I}(\varphi_M) = 0$. This is actually the approach used in Nguyen & Ho (2024) where the authors propose two different choices for M leading to two different estimators (see Appendix A). Note that, whenever $2L \geq 2$, these quadratic functions are included in our spherical harmonics space, making our method more general with potentially greater variance reduction (see Section 6).

5. Numerical Experiments

To asses the finite-sample performance of the SHCV estimate, we first present some synthetic data examples involving the computation of $\mathrm{SW}_2^2(\mu, \nu)$ between multivariate Gaussian distributions. We consider different dimensions $d \in \{3; 5; 6; 10; 20\}$ and a varying number of random projections $n \in [10^2; 10^4]$. Then we focus on $\mathrm{SW}_2^2(\mu_m, \nu_m)$ between 3D point clouds from the ShapeNetCore dataset of Chang et al. (2015). In both cases, the ground truth of the Sliced-Wasserstein distance is estimated with a very large number $n = 10^8$ of projections. The different results
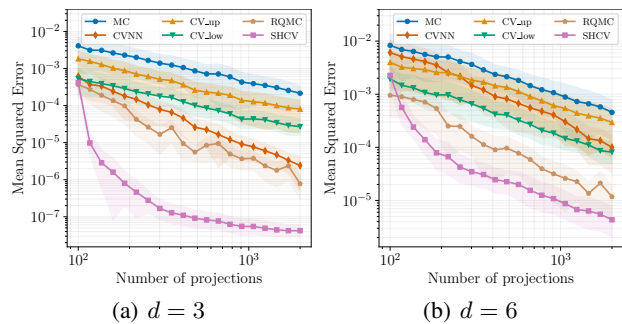
(a) $d = 3$ (b) $d = 6$

Figure 2. MSE curves for sampled Gaussian distributions in dimension $d \in \{3; 6\}$.

Method	$d = 5$		$d = 10$		$d = 20$	
	MSE	Time	MSE	Time	MSE	Time
MC	1.45e-3	81.1 ± 3.5	9.45e-4	80.7 ± 4.4	1.47e-3	81.1 ± 1.8
CV$_{low}$	2.67e-4	79.7 ± 1.1	3.45e-4	80.1 ± 1.4	3.82e-4	80.0 ± 1.0
CV$_{up}$	8.44e-4	83.0 ± 1.2	7.51e-4	83.0 ± 1.7	1.09e-3	83.1 ± 1.5
CVNN	4.29e-4	110 ± 2.2	1.12e-3	122 ± 1.6	2.14e-3	127 ± 1.4
QMC	2.91e-4	100 ± 1.2	2.37e-4	113 ± 1.4	6.60e-4	129 ± 1.4
RQMC	5.80e-5	96.3 ± 2.2	2.75e-4	113 ± 1.2	1.17e-3	130 ± 1.0
SHCV	**2.68e-6**	89.0 ± 6.3	**1.93e-4**	89.0 ± 4.5	**2.95e-4**	88.1 ± 2.8

Table 1. MSE and computing time (ms) for Gaussian distributions in dimension $d \in \{5; 10; 20\}$ based on $n = 500$ projections.

report the mean squared error (MSE) $\mathbb{E}[|\mathrm{SW}_{2,n}^2(\mu, \nu) - \mathrm{SW}_2^2(\mu, \nu)|^2]$ for the different estimates $\mathrm{SW}_{2,n}^2(\mu, \nu)$ where the expectation is computed as an average over 100 independent runs. Finally, in the spirit of Kolouri et al. (2016) and Meunier et al. (2022), we implement a Kernel Support Vector Machine (SVM) classifier for image classification. For ease of reproducibility, numerical details are in the appendix and the code is available here. In particular, we rely on the efficient implementation of Dutordoir et al. (2020) to build the spherical harmonics. The degrees of the spherical harmonics are reported in Table 9 in Appendix D.2.

Methods in competition. In the experiments, the SW distance is computed through different estimates. Along with the standard Monte Carlo estimate and the proposed SHCV estimate, we consider three other baselines based on control variates: the two estimates of Nguyen & Ho (2024) based on lower and upper bounds of a Gaussian approximation and the *control neighbors* estimate of Leluc et al. (2024) based on nearest neighbors estimates acting as control variates. For the sake of completeness, we also include a comparison with (Randomized) Quasi Monte Carlo procedures as in Nguyen et al. (2024) where deterministic QMC sets from $[0,1]^d$ are mapped to the unit sphere and randomly rotated. The methods in competition are: the standard Monte Carlo estimate (MC), the lower-CV (CV$_{low}$) and upper-CV (CV$_{up}$) estimates, the control neighbors estimate (CVNN), the proposed estimate with spherical harmonics as control variates (SHCV), the Quasi Monte Carlo (QMC) and Random Quasi Monte Carlo estimates (RQMC).

5.1. Synthetic Data: Multivariate Gaussians

We compute $\mathrm{SW}_2^2(\mu_m, \nu_m)$ between empirical distributions $\mu_m = m^{-1} \sum_{i=1}^m \delta_{x_i}$ and $\nu_m = m^{-1} \sum_{j=1}^m \delta_{y_j}$ where $x_1, \ldots, x_m \sim \mu = \mathcal{N}(a, \mathbf{A})$, $y_1, \ldots, y_m \sim \nu = \mathcal{N}(b, \mathbf{B})$ with $m = 1000$ samples. The parameters of the Gaussians are $a, b \sim \mathcal{N}(\mathbb{1}_d, I_d)$ and covariance matrices $\mathbf{A} = \Sigma_a \Sigma_a^\top$ and $\mathbf{B} = \Sigma_b \Sigma_b^\top$ where all entries of Σ_a, Σ_b are drawn

according to $\mathcal{N}(0, 1)$. In Appendix B.1, we also consider the *exact case* where we directly evaluate the smooth integrand $f_{\mu,\nu}^{(2)}$ given in Remark 1. In that case, as the original lower-CV (CV$_{low}$) and upper-CV (CV$_{up}$) estimates are written to take discrete measures as inputs, we adapt them with the true means and covariance matrices.

Figure 2 reports the MSE of the different methods in dimension $d \in \{3; 6\}$ with respect to the number of projections, while Figure 5 in Appendix B.2 does so with respect to computing time. Table 1 reports the MSE and computing times (in ms) in higher dimension $d \in \{5; 10; 20\}$. Figure 4 and Table 3 in Appendix B.1 deal with the *exact case* accordingly. In each case, the SHCV estimate gives the best MSE performance with respect to both the number of projections and computing time, beating other baselines by a factor up to 100.

5.2. Empirical Distributions of 3D Point Clouds

Similarly to Nguyen & Ho (2024), we use three point clouds taken at random from the dataset ShapeNetCore (Chang et al., 2015) corresponding to the objects plane, lamp, and bed, each composed of $m = 2048$ points in $\mathbb{R}^3$. Each point cloud $(x_1, \ldots, x_m)$ is encoded as a discrete measure $\mu_m = m^{-1} \sum_{i=1}^m \delta_{x_i}$ with equal masses over the points in the cloud and the goal is to compute $\mathrm{SW}(\mu_m, \nu_m)$ of order $p = 2$ between empirical distributions. Figure 3 reports the boxplots of the error distribution $\widehat{\mathrm{SW}}_n(\mu_m, \nu_m) - \mathrm{SW}(\mu_m, \nu_m)$ for the different SW estimates based on n random projections with $n \in \{100; 250; 500; 1000\}$ obtained over 100 independent runs. Once again, our SHCV estimate provides huge gains in terms of variance reduction and outperforms other baselines by a factor 100; see Table 6 in Appendix B.3.

5.3. Sliced-Wasserstein and Kernel SVM

SW distances are useful in the context of distribution regression (Szabó et al., 2016) where the goal is to learn a real-valued function defined on the space of probability distributions from a sequence of observations. Kolouri et al. (2016) first introduced a SW-based kernel for absolutely
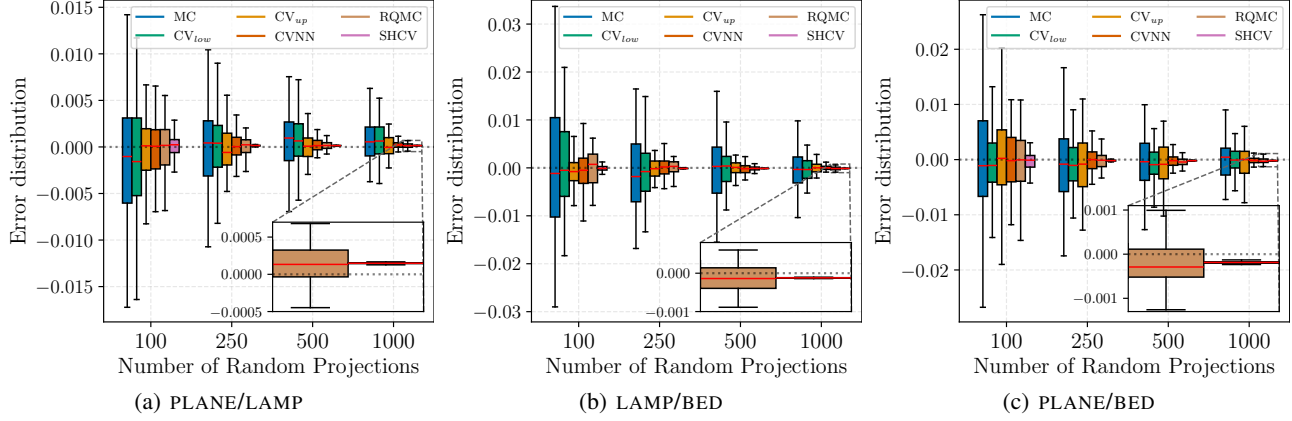
Figure 3. Boxplots of the error distribution $\widehat{\mathrm{SW}}_n(\mu_m, \nu_m) - \mathrm{SW}(\mu_m, \nu_m)$ for different SW estimates based on n random projections with $n \in \{100; 250; 500; 1000\}$ obtained over 100 independent runs.

continuous distributions, which was later on extended to empirical distributions by Meunier et al. (2022). For any $\gamma > 0$, both $e^{-\gamma \mathrm{SW}_2^2(\cdot, \cdot)}$ and $e^{-\gamma \mathrm{SW}_1(\cdot, \cdot)}$ are valid kernels on probability measures. Using the SW kernel of Kolouri et al. (2016) defined by $k(\mu_i, \mu_j) = \exp(-\gamma \mathrm{SW}_2^2(\mu_i, \mu_j))$ with $\gamma > 0$, we consider the image classification task on the digits dataset of Alpaydin & Kaynak (1998). This dataset contains $N = 1\,797$ images of size 8×8 and the goal is to predict the label $z \in \{0, 1, \dots, 9\}$ of the handwritten digit. Similarly to Meunier et al. (2022), we convert each image to a histogram so that the SW kernels have probability distributions as inputs. For image $i \in [N]$, let $(x_k^i, y_k^i)_{k=1}^{n_i}$ denote the coordinates of the n_i active pixels and $(\psi_k^i)_{k=1}^{n_i}$ the associated pixel intensities. The positions are renormalized to the grid $[0, 1]^2$ and we consider the weighted histogram $\mu_i = \sum_{k=1}^{n_i} \alpha_k^i \delta_{(x_k^i, y_k^i)}$ where $\alpha_k^i = \psi_k^i / \sum_{l=1}^{n_i} \psi_l^i$ are the normalized intensities. We use a train/test split of size $80/20$ giving $N_{\mathrm{train}} = 1\,437$ training and $N_{\mathrm{test}} = 360$ testing images, respectively, where each set is balanced.

We implement a Kernel SVM classifier where the SW_2 distance involved in the kernel expression is either computed with the standard Monte Carlo estimate (MC) or with the SHCV estimate (SHCV). The number of random projections n evolves in $\{25; 50; 75; 100\}$. To compare different scenarios, the bandwidth parameter $\gamma > 0$ of the Gaussian kernel is in $\{0.01; 1\}$ and the regularization parameter C of the SVM is optimized with a cross-validation strategy. Table 2 reports the means and standard deviations of the test accuracies of the different methods where the statistics are obtained over 10 independent runs.

6. Theoretical Analysis

This section gathers several theoretical properties of the novel SHCV estimate. We first state some elementary prop-

n	$\gamma = 0.01$		$\gamma = 1$	
	MC	SHCV	MC	SHCV
25	86.0 ± 0.6	$\mathbf{87.3 \pm 0.3}$	92.7 ± 0.6	$\mathbf{93.3 \pm 0.6}$
50	86.7 ± 0.5	$\mathbf{87.6 \pm 0.3}$	93.2 ± 0.4	$\mathbf{94.0 \pm 0.3}$
75	86.8 ± 0.2	$\mathbf{87.8 \pm 0.2}$	93.3 ± 0.4	$\mathbf{94.3 \pm 0.2}$
100	86.9 ± 0.3	$\mathbf{87.7 \pm 0.1}$	93.3 ± 0.3	$\mathbf{94.3 \pm 0.2}$

Table 2. Kernel SVM: test accuracies with means and standard deviations. The statistics are obtained over 10 independent runs.

erties of the underlying integration rule with respect to a general distribution P on the sphere and then provide a convergence rate in probability of the integration error in case of the uniform distribution U on the sphere.

6.1. Elementary Properties

By definition of the OLSMC estimate, the SHCV estimate is exact on the linear span of the control variates (Portier & Segers, 2019). Thus, if $f_{\mu,\nu}^{(p)}$ happens to be a polynomial function, then the SHCV estimate will be exact as soon as the maximal degree of the spherical harmonics is equal to or larger than the degree of $f_{\mu,\nu}^{(p)}$.

Proposition 2 (Exact rule). *If $f_{\mu,\nu}^{(p)}$ is a polynomial function with degree m, considering the SHCV estimate and control variates $\varphi = (\varphi_j)_{j=1}^{s_{L,d}}$, if $2L \geq m$ and $n > s_{L,d}$ then SHCV is exact, i.e.,* $\mathrm{SHCV}_{n,L}^p(\mu, \nu) = \mathrm{SW}_p^p(\mu, \nu)$.

In the Gaussian case of Remark 1, for two proportional dispersion matrices $\mathbf{A}$ and $\mathbf{B}$, say $\mathbf{B} = \gamma \mathbf{A}$ for some $\gamma > 0$, the formula of $f_{\mu,\nu}^{(2)}(\theta) = \mathrm{W}_2^2(\theta_\sharp^\star \mu, \theta_\sharp^\star \nu)$ simplifies to a quadratic homogeneous polynomial. Therefore, the SHCV estimate is exact[2] as soon as the spherical harmonics of de-

---

[2]See numerical comparison in Appendix D.5 where the SHCV estimate is the only one achieving exact integration (zero error).

gree two are included as control variates, since the centered integrand is then contained in the linear space spanned by the control variates. Actually, this remains true in case $p = 2$ if the measures are related by an affine transformation.

Corollary 1 (Affine transform). *In case $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ are related by $X \sim \mu$ and $\alpha X + b \sim \nu$ where $\alpha \in (0, \infty)$ and $b \in \mathbb{R}^d$ then the* SHCV *estimate is exact.*

The SHCV estimator with spherical harmonics of maximal degree two or higher is guaranteed to have an asymptotic variance that is bounded above by the lower- and upper-CV methods of Nguyen & Ho (2024). Indeed, the latter two methods are control variate methods, each based on a *single* control variate, which happens to be a quadratic polynomial.

Corollary 2 (Mean invariance). *For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, the error of the SHCV method is (exactly) invariant under changes of the mean vectors m_μ and m_ν of μ and ν respectively.*

This result implies that we can, without loss of generality, suppose that μ and ν are centered. The reason is that the effect of the two mean vectors being potentially different is the translation term $\langle \theta, \mathrm{m}_Y - \mathrm{m}_X \rangle$ coming from the optimal transport map that pushes $\theta^\star_\sharp \mu$ to $\theta^\star_\sharp \nu$, and this term produces an additional quadratic term $\langle \theta, \mathrm{m}_Y - \mathrm{m}_X \rangle^2$ in the function $f_{\mu,\nu}^{(2)}$. This additional quadratic polynomial in θ is integrated exactly by SHCV; its integral is $\|\mathrm{m}_X - \mathrm{m}_Y\|^2/d$. See Remark 1 for the Gaussian case.

6.2. Asymptotic Error Bound

We provide a convergence rate in probability[3] of the integration error $\mathrm{SHCV}_{n,L}^p(\mu, \nu) - \mathrm{SW}_p^p(\mu, \nu)$ for general probability measures $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ and $p \in [1, \infty)$, where $\mathrm{SW}_p(\mu, \nu)$ is the SW distance with respect to the uniform distribution on the sphere and $\mathrm{SHCV}_{n,L}^p(\mu, \nu)$ is our estimate in Definition 4 based on n random projections on the sphere and using spherical harmonics of even degree up to $2L$ as control variates. The guarantee relies on two properties: first, a general convergence rate of the integration error of the OLSCV estimate in Portier & Segers (2019, Theorem 1), available in Appendix C.2, and second, a bound on the uniform approximation error of a Lipschitz function on the sphere by its orthogonal projection on the space of polynomials (Newman & Shapiro, 1964, Theorem 2). The latter result applies since the integrand $f_{\mu,\nu}^{(p)}$ in Equation (3) is always at least Lipschitz: by Han (2023, Theorem 2.4), we have

$$\left| f_{\mu,\nu}^{(p)}(\theta) - f_{\mu,\nu}^{(p)}(\gamma) \right| \le M_p(\mu, \nu) \, \|\theta - \gamma\| \qquad (9)$$

---

[3]The notation $R_n = \mathcal{O}_{\mathbb{P}}(a_n)$, for random variables R_n and constants $a_n > 0$, means that for every $\delta > 0$ there exists $K_\delta > 0$ such that $\mathbb{P}(|R_n| > K_\delta a_n) < \delta$ for all n.

for $\theta, \gamma \in \mathbb{S}^{d-1}$, where $M_p(\mu, \nu)$ is an explicit function of p and the pth order moments of μ and ν.

The modulus of continuity of $f \in \mathscr{C}(\mathbb{S}^{d-1})$ is defined for $h > 0$ by

$$\omega(h) = \sup\{|f(\theta) - f(\gamma)| : \rho(\theta, \gamma) \le h\}$$

with $\rho(\theta, \gamma) = \arccos\langle \theta, \gamma \rangle$ the length of the arc connecting $\theta, \gamma \in \mathbb{S}^{d-1}$.

Theorem 1 (Newman & Shapiro (1964), Theorem 2). *There exists an absolute constant $A > 0$ such that for each dimension $d \ge 2$, each function $f \in \mathscr{C}(\mathbb{S}^{d-1})$ and each degree L there exists a polynomial $P_{f,L}$ on $\mathbb{S}^{d-1}$ of degree at most L such that*

$$\sup_{\theta \in \mathbb{S}^{d-1}} |f(\theta) - P_{f,L}(\theta)| \le A \, \omega(d/L).$$

The polynomial $P_{f,L}$ in Theorem 1 that yields the uniform approximation is in general not the same as the orthogonal projection $\Pi_{f,L} = \sum_{\ell=0}^{L} \sum_{k=1}^{N_\ell^d} \hat{f}_{\ell,k} \varphi_{\ell,k}$ of f in $L_2(\mathbb{S}^{d-1})$ on the space of spherical harmonics of degree at most L, where we use the notation of Proposition 1. Still, the $L_2(\mathbb{S}^{d-1})$ approximation error of $\Pi_{f,L}$ is bounded by the one of $P_{f,L}$, which is in turn bounded by the uniform approximation error of $P_{f,L}$. Applying this argument to the integrand $f_{\mu,\nu}^{(p)}$, we obtain the following convergence rate in probability of the SHCV estimate.

Theorem 2 (Convergence rate). *Let $d \ge 2$, $p \in [1, \infty)$ and $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$ be fixed. For any degree sequence $L = L_n$ such that $L = o(n^{1/(2(d-1))})$ as $n \to \infty$, the integration error of the SHCV estimate with respect to the uniform distribution on the sphere satisfies*

$$\left| \mathrm{SHCV}_{n,L}^p(\mu, \nu) - \mathrm{SW}_p^p(\mu, \nu) \right| = \mathcal{O}_{\mathbb{P}}(L^{-1} n^{-1/2}). \quad (10)$$

The proof of Theorem 2 is given in Appendix C.3. Writing $L = n^{1/(2(d-1))}/\ell_n$ where $\ell_n > 0$ diverges to infinity but can do so arbitrarily slowly (for instance logarithmically), the convergence rate in Equation (10) becomes

$$\left| \mathrm{SHCV}_{n,L}^p(\mu, \nu) - \mathrm{SW}_p^p(\mu, \nu) \right| = \mathcal{O}_{\mathbb{P}}(\ell_n n^{-\frac{1}{2} \frac{d}{d-1}}).$$

For $d = 3$, this yields the rate $n^{-3/4+o(1)}$ for the SHCV estimate, in comparison to the Monte Carlo rate $n^{-1/2}$.

If the integrand $f_{\mu,\nu}^{(p)}$ has higher-order derivatives, then a faster convergence rate can be expected because of approximation results in Ragozin (1971) improving Theorem 1. This concerns for instance the SW distance between two Gaussian distributions, for which the integrand in Remark 1 has derivatives of all orders. In contrast, for discrete measures, the integrand will not even be continuously differentiable, and the Lipschitz property is the best we can hope for. Both cases are illustrated in Appendix D.4.

7. Conclusion

We have developed a novel method for reducing the variance of Monte Carlo estimation of the SW distance using spherical harmonics as control variates. The excellent practical performance of the SHCV estimate against state-of-the-art baselines is confirmed by theoretical properties and a convergence rate in probability for the integration error.

Using control variates with QMC sequences is usually *not* implemented in the same way as with Monte Carlo sequences (Hickernell et al., 2005) and would require a particular treatment which is beyond the scope of this paper. In statistical inference with parametric probability measures, note that SHCV is compatible with the computation of gradient $\nabla_\phi \mathrm{SW}_p^p(\mu, \nu_\phi)$ and can be used for *generalized* SW flows (Kolouri et al., 2019). While the proposed SHCV estimate focuses on the uniform distribution on $\mathbb{S}^{d-1}$, it can be extended to more general probability distributions by combining control variates with importance sampling techniques as in Leluc et al. (2022).

Acknowledgments

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Alpaydin, E. and Kaynak, C. Optical Recognition of Handwritten Digits. UCI Machine Learning Repository, 1998. DOI: https://doi.org/10.24432/C50P49.

Atkinson, K. and Han, W. *Spherical Harmonics and Approximations on the Unit Sphere: An Introduction*. Lecture Notes in Mathematics. Springer, 2012.

Basri, R. and Jacobs, D. W. Lambertian reflectance and linear subspaces. *IEEE transactions on pattern analysis and machine intelligence*, 25(2):218–233, 2003.

Bayraktar, E. and Guo, G. Strong equivalence between metrics of Wasserstein type. *Electronic Communications in Probability*, 26(none):1 – 13, 2021. doi: 10.1214/ 21-ECP383. URL https://doi.org/10.1214/ 21-ECP383.

Bhatia, R., Jain, T., and Lim, Y. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. Sliced and radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51:22–45, 2015.

Bonnotte, N. *Unidimensional and evolution methods for optimal transportation*. PhD thesis, Université Paris Sud-Paris XI; Scuola normale superiore (Pise, Italie), 2013.

Carpintero Perez, R., Da Veiga, S., Garnier, J., and Staber, B. Gaussian process regression with sliced Wasserstein Weisfeiler-Lehman graph kernels. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238, pp. 1297–1305. PMLR, 2024.

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

Cohen, T. S., Geiger, M., Köhler, J., and Welling, M. Spherical CNNs. In *International Conference on Learning Representations*, 2018. URL https://openreview. net/forum?id=Hkbd5xZRb.

Dai, F. *Approximation theory and harmonic analysis on spheres and balls*. Springer, 2013.

Deshpande, I., Zhang, Z., and Schwing, A. G. Generative modeling using the sliced Wasserstein distance. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3483–3491, 2018.

Deshpande, I., Hu, Y.-T., Sun, R., Pyrros, A., Siddiqui, N., Koyejo, S., Zhao, Z., Forsyth, D., and Schwing, A. G. Max-sliced Wasserstein distance and its use for gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10648–10656, 2019.

Dowson, D. and Landau, B. The Fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

Dutordoir, V., Durrande, N., and Hensman, J. Sparse Gaussian processes with spherical harmonic features. In *International Conference on Machine Learning*, pp. 2793–2802. PMLR, 2020.

Feragen, A., Lauze, F., and Hauberg, S. Geodesic exponential kernels: When curvature and linearity conflict. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3032–3042, 2015.

Glynn, P. W. and Szechtman, R. 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*, pp. 27–49. Springer, 2002.

Green, R. Spherical harmonic lighting: The gritty details. In *Archives of the game developers conference*, volume 56, pp. 4, 2003.

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

Hickernell, F. J., Lemieux, C., and Owen, A. B. Control variates for Quasi-Monte Carlo. *Statistical Science*, 20 (1):1–31, 2005.

Kolouri, S., Zou, Y., and Rohde, G. K. Sliced Wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5258–5267, 2016.

Kolouri, S., Pope, P. E., Martin, C. E., and Rohde, G. K. Sliced Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.

Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., and Rohde, G. Generalized sliced Wasserstein distances. *Advances in neural information processing systems*, 32, 2019.

Leluc, R., Portier, F., and Segers, J. Control variate selection for Monte Carlo integration. *Statistics and Computing*, 31(4):50, 2021.

Leluc, R., Portier, F., Segers, J., and Zhuman, A. A Quadrature Rule combining Control Variates and Adaptive Importance Sampling. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11842–11853. Curran Associates, Inc., 2022.

Leluc, R., Portier, F., Segers, J., and Zhuman, A. Speeding up Monte Carlo integration: control neighbors for optimal convergence. *Bernoulli*, in press, 2024.

Liutkus, A., Simsekli, U., Majewski, S., Durmus, A., and Stöter, F.-R. Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *International Conference on Machine Learning*, pp. 4104–4113. PMLR, 2019.

Meunier, D., Pontil, M., and Ciliberto, C. Distribution regression with sliced Wasserstein kernels. In *International Conference on Machine Learning*, pp. 15501–15523. PMLR, 2022.

Mezzadri, F. How to generate random matrices from the classical compact groups. *Notices of the American Mathematical Society*, 54(5):592–604, 2007.

Nadjahi, K., De Bortoli, V., Durmus, A., Badeau, R., and Şimşekli, U. Approximate Bayesian computation with the sliced-wasserstein distance. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5470–5474. IEEE, 2020a.

Nadjahi, K., Durmus, A., Chizat, L., Kolouri, S., Shahrampour, S., and Simsekli, U. Statistical and topological properties of sliced probability divergences. *Advances in Neural Information Processing Systems*, 33:20802–20812, 2020b.

Nadjahi, K., Durmus, A., Jacob, P. E., Badeau, R., and Simsekli, U. Fast approximation of the sliced-Wasserstein distance using concentration of random projections. *Advances in Neural Information Processing Systems*, 34: 12411–12424, 2021.

Newman, D. J. and Shapiro, H. S. Jackson's Theorem in Higher Dimensions. In *Conference on Approximation Theory (Oberwolfach, 1963)*, pp. 208–219, 1964.

Nguyen, K. and Ho, N. Sliced Wasserstein Estimation with Control Variates. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=StYc4hQAEi.

Nguyen, K., Bariletto, N., and Ho, N. Quasi-Monte Carlo for 3D Sliced Wasserstein. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Wd47f7HEXg.

Nietert, S., Goldfeld, Z., Sadhu, R., and Kato, K. Statistical, robustness, and computational guarantees for sliced Wasserstein distances. *Advances in Neural Information Processing Systems*, 35:28179–28193, 2022.

Oates, C. J., Girolami, M., and Chopin, N. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79 (3):695–718, 2017.

Owen, A. B. *Monte Carlo theory, methods and examples*. https://artowen.su.domains/mc/, 2013.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

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

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

Rabin, J., Peyré, G., Delon, J., and Bernot, M. 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*, pp. 435–446. Springer, 2012.

Ragozin, D. L. Constructive Polynomial Approximation on Spheres and Projective Spaces. *Transactions of the American Mathematical Society*, 162:157–170, 1971.

Ramamoorthi, R. and Hanrahan, P. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pp. 497–500, 2001.

South, L. F., Oates, C. J., Mira, A., and Drovandi, C. Regularized zero-variance control variates. *Bayesian Analysis*, 18(3):865–888, 2023.

Szabó, Z., Sriperumbudur, B. K., Póczos, B., and Gretton, A. Learning theory for distribution regression. *The Journal of Machine Learning Research*, 17(1):5272–5311, 2016.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Villani, C. et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, 17(3):261–272, 2020.

Zhang, T. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 116, 2004.

Supplementary Material:
Sliced-Wasserstein Estimation with Spherical Harmonic as Control Variates

Appendix A gathers the pseudo-code of all the algorithms used in the experiments while Appendix B presents additional details and results for the numerical experiments. Appendix C gathers the proofs of the theoretical results and Appendix D is dedicated to additional results on spherical harmonics with (1) a case-study on the regularity of the integrand $f_{\mu,\nu}^{(p)}$ and (2) a numerical check of the exact integration rule.

A. Algorithms for Sliced-Wasserstein Estimates

This section presents algorithmic details of the different Sliced-Wasserstein estimates in the numerical experiments:

• Algorithm 3 corresponds to the standard Monte Carlo estimate where random projections $\theta_1, \ldots, \theta_n \sim \mathrm{P} = \mathcal{U}(\mathbb{S}^{d-1})$ are used to compute the Wasserstein distance of projected distributions with $f_{\mu,\nu}^{(p)}(\theta) = \mathrm{W}_p^p(\theta_\sharp^\star \mu, \theta_\sharp^\star \nu)$.

• Algorithm 4 describes the proposed SHCV estimate which relies on the OLS procedure with spherical harmonics φ_j for $j = 1, \ldots, s$ as control variates.

• Algorithm 5 and Algorithm 6 describe the two estimates of Nguyen & Ho (2024) which rely on Gaussian approximation with lower and upper control variates.

• Algorithm 7 presents the control neighbors technique of Leluc et al. (2024) and consists in building a control functional estimate with the 1-nearest neighbor estimate of the integrand f.

• Algorithm 8 implements a (Randomized) Quasi-Monte Carlo estimate. A low-discrepancy sequence on the unit cube $[0, 1]^d$ is mapped on $\mathbb{S}^{d-1}$ to obtain a deterministic QMC point set on the unit sphere. Next, rotations from the orthogonal group are applied to randomize the samples.

Algorithm 3 Sliced Wasserstein Monte Carlo

Require: $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, number of random projections n
  1: Sample random projections $\theta_1, \ldots, \theta_n$ uniformly on $\mathbb{S}^{d-1}$
  2: Compute $f_n = (f_{\mu,\nu}^{(p)}(\theta_i))_{i=1}^n$
  3: Return average $\mathrm{MC}_n = (\mathbb{1}_n/n)^\top f_n$

Algorithm 4 Spherical Harmonics Control Variate Estimate

Require: $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, number of projections n, spherical harmonics $\varphi = (\varphi_j)_{j=1}^s$
  1: Sample random projections $\theta_1, \ldots, \theta_n$ uniformly on $\mathbb{S}^{d-1}$
  2: Compute $f_n = (f_{\mu,\nu}^{(p)}(\theta_i))_{i=1}^n$ and $\Phi = (\varphi(\theta_i)^\top)_{i=1}^n$
  3: $(\mathrm{I}_n^{\mathrm{ols}}, \beta_n) \in \arg\min_{\alpha,\beta} \|f_n - \alpha\mathbb{1}_n - \Phi\beta\|_2^2$
  4: Return $\mathrm{SHCV}_n = \mathrm{I}_n^{\mathrm{ols}}$

Algorithm 5 Lower-CV Sliced Wasserstein (Nguyen & Ho, 2024)

Require: $\mu = \sum_{i=1}^{m_X} \alpha_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{m_Y} \beta_j \delta_{y_j}$, number of random projections n
  1: Sample random projections $\theta_1, \ldots, \theta_n$ uniformly on $\mathbb{S}^{d-1}$
  2: Compute $f_n = (f_{\mu,\nu}^{(p)}(\theta_i))_{i=1}^n$
  3: Compute $\mathrm{MC}_n = (\mathbb{1}_n/n)^\top f_n$
  4: Compute $\bar{x} = \sum_{i=1}^{m_X} \alpha_i x_i$, and $\bar{y} = \sum_{j=1}^{m_Y} \beta_j y_j$
  5: $C = (c_1, \ldots, c_n)$ with $c_k = (\theta_k^\top (\bar{x} - \bar{y}))^2$
  6: Compute $b = \|\bar{x} - \bar{y}\|_2^2/d$
  7: Compute $\gamma_{\mathrm{low}} = \frac{\langle f_n - \mathrm{MC}_n, C - b\rangle}{\|C - b\|_2^2}$ and control variate $\mathcal{C} = (\mathbb{1}_n^\top (C - b))/n$
  8: Return $\mathrm{CV}_{\mathrm{low}} = \mathrm{MC}_n - \gamma_{\mathrm{low}}\mathcal{C}$

Algorithm 6 Upper-CV Sliced-Wasserstein (Nguyen & Ho, 2024)

Require: $\mu = \sum_{i=1}^{m_X} \alpha_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{m_Y} \beta_j \delta_{y_j}$, number of random projections n
1: Sample random projections $\theta_1, \dots, \theta_n$ uniformly on $\mathbb{S}^{d-1}$
2: Compute $f_n = (f_{\mu,\nu}^{(p)}(\theta_i))_{i=1}^n$
3: Compute $\mathrm{MC}_n = (\mathbb{1}_n/n)^\top f_n$
4: Compute $\bar{x} = \sum_{i=1}^{m_X} \alpha_i x_i$, and $\bar{y} = \sum_{j=1}^{m_Y} \beta_j y_j$
5: $C = (c_1, \dots, c_n)$ with $c_k = (\theta_k^\top(\bar{x}-\bar{y}))^2 + \sum_{i=1}^{m_X} \alpha_i(\theta_k^\top(x_i-\bar{x}))^2 + \sum_{j=1}^{m_Y} \beta_j(\theta_k^\top(y_j-\bar{y}))^2$
6: Compute $b = \left(\|\bar{x}-\bar{y}\|_2^2 + \sum_{i=1}^{n_X} \alpha_i\|(x_i-\bar{x})\|_2^2 + \sum_{j=1}^{m_Y} \beta_j\|(y_j-\bar{y})\|_2^2 \right)/d$
7: Compute $\gamma_{\mathrm{up}} = \frac{\langle f_n - \mathrm{MC}_n, C-b \rangle}{\|C-b\|_2^2}$ and control variate $\mathcal{C} = (\mathbb{1}_n^\top(C-b))/n$
8: Return $\mathrm{CV}_{up} = \mathrm{MC}_n - \gamma_{\mathrm{up}}\mathcal{C}$

Algorithm 7 Control Neighbors Sliced Wasserstein (Leluc et al., 2024)

Require: $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, number of random projections n, probability $\mathrm{P} \in \mathcal{P}(\mathbb{S}^{d-1})$
1: Sample random projections $\theta_1, \dots, \theta_n \sim \mathrm{P}$
2: Compute $f_n = (f_{\mu,\nu}^{(p)}(\theta_i))_{i=1}^n$
3: Compute $\mathrm{MC}_n = (\mathbb{1}_n/n)^\top f_n$
4: Compute nearest neighbor evaluations $\hat{f}_n = (\hat{f}_n^{(1)}(\theta_1), \dots, \hat{f}_n^{(n)}(\theta_n))$
5: Return $\mathrm{CVNN}_n = \mathrm{MC}_n - [\mathbb{1}_n^\top\{\hat{f}_n - \mathrm{I}(\hat{f}_n)\}]/n$

Algorithm 8 Randomized Quasi-Monte Carlo Sliced Wasserstein (Nguyen et al., 2024)

Require: $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, number of projections n, Φ the cdf of $\mathcal{N}(0,1)$
1: Generate low-discrepancy sequence $U_1, \dots, U_n$ on $[0,1]^d$
2: Compute Gaussian mapping $v_i = \Phi^{-1}(U_i)/\|\Phi^{-1}(U_i)\|_2$, where the standard normal quantile function Φ^{-1} is applied component-wise.
3: Generate a random rotation $R \sim \mathcal{U}(O_d(\mathbb{R}))$ where $O_d(\mathbb{R}) = \{R \in \mathbb{R}^{d \times d} \mid R^\top R = I_d\}$
4: Generate random projections $\theta_i = Rv_i$
5: Compute $f_n = (f_{\mu,\nu}^{(p)}(\theta_i))_{i=1}^n$
6: Return $\mathrm{RQMC}_n = (\mathbb{1}_n/n)^\top f_n$

Remarks and Practical Details on the Algorithms.

• In Algorithm 5, with the notation $z = \bar{x} - \bar{y}$, the control variate is the centered version of the function $\theta \mapsto |\theta^\top z|^2$, which has known integral $\int_{\mathbb{S}^{d-1}} |\theta^\top z|^2 \mathrm{d}\mathrm{U}(\theta) = \|z\|^2/d = b$ (Step 6 in the algorithm) and where U denotes the uniform distribution on $\mathbb{S}^{d-1}$. The Lower-CV estimate is thus an OLSMC estimate with control variate $\varphi(\theta) = |\theta^\top z|^2 - \|z\|^2/d$, a polynomial in θ of degree two.

• Algorithm 6 is actually an OLSMC estimate with a single control variate given by the centered version of $\theta \mapsto \theta^\top M\theta$ where $M = M_0 + M_X + M_Y$ with $M_0 = (\bar{x}-\bar{y})(\bar{x}-\bar{y})^\top$, $M_X = \sum_{i=1}^{m_X} \alpha_i(x_i-\bar{x})(x_i-\bar{x})^\top$ and similarly for M_Y. The integral with respect to the uniform distribution on the sphere is $\int_{\mathbb{S}^{d-1}} \theta^\top M\theta \, \mathrm{d}\mathrm{U}(\theta) = \mathrm{Tr}(M)/d = b$ in Step 6 of the algorithm. The control variate is thus the centered quadratic polynomial $\varphi(\theta) = \theta^\top M\theta - \mathrm{Tr}(M)/d$.

• In Step 4 of Algorithm 7, $\hat{f}_n^{(i)}(\theta_i)$ is a leave-one-out nearest-neighbor estimate computed with the Euclidean metric. In Step 5, computing $\mathrm{I}(\hat{f}_n)$ requires sampling M additional directions θ_i from P and we set $M = n^{1+2/d}$ as recommended by the theory in Leluc et al. (2024).

• In Algorithm 8, random orthogonal matrices $R \in O_d(\mathbb{R})$ 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[4] as in Mezzadri (2007).

---

[4]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.

B. Numerical Details and Additional Results

To assess the finite-sample performance of the SHCV estimate, we first present some synthetic data examples involving the computation of $\text{SW}_2^2(\mu, \nu)$ between multivariate Gaussian distributions. We consider different dimensions $d \in \{3; 5; 6; 10; 20\}$ and a varying number of random projections $n \in [10^2; 10^4]$. We consider the two cases of *exact* distributions in Section B.1 and *sampled* distributions in Section B.2. Then in Section B.3, we focus on the Sliced-Wasserstein distance $\text{SW}_2(\mu_m, \nu_m)$ between 3D point clouds from the ShapeNetCore dataset of Chang et al. (2015). In both cases, the true value of the Sliced-Wasserstein distance is estimated with a very large number $n = 10^8$ of projections. The different results report the mean squared error $\mathbb{E}[|\text{SW}_{2,n}^2(\mu, \nu) - \text{SW}_2^2(\mu, \nu)|^2]$ for the different estimates $\text{SW}_{2,n}^2(\mu, \nu)$ where the expectation is computed as an average over 100 independent runs. The experiments were performed on a laptop Intel Core i7-10510U CPU 1.80GHz × 8.

Methods in competition. Along with the standard Monte Carlo estimate (Algorithm 3) and the proposed SHCV estimate (Algorithm 4), we consider three other baselines based on control variates: the two estimates of Nguyen & Ho (2024) based on lower and upper bounds of a Gaussian approximation (Algorithm 5 and Algorithm 6) and the *control neighbors* estimate of Leluc et al. (2024) based on nearest neighbors estimates acting as control variates (Algorithm 7). For completeness, we also include a comparison with (Randomized) Quasi Monte Carlo procedures as in Nguyen et al. (2024) where deterministic QMC sets from $[0, 1]^d$ are mapped to the unit sphere and randomly rotated (Algorithm 8). The methods in competition are: the standard Monte Carlo estimate (MC), the lower-CV (CV$_{low}$) and upper-CV (CV$_{up}$) estimates, the control neighbors estimate (CVNN), the proposed estimate with spherical harmonics as control variates (SHCV), the Quasi Monte Carlo (QMC) and Random Quasi Monte Carlo estimates (RQMC).

B.1. Multivariate Gaussian: Exact Case with Distributions μ, ν and Smooth Integrands

We compute $\text{SW}_2^2(\mu, \nu)$ between $\mu = \mathcal{N}(a, \mathbf{A})$ and $\nu = \mathcal{N}(b, \mathbf{B})$ with means $a, b \sim \mathcal{N}(\mathbb{1}_d, I_d)$ and covariance matrices $\mathbf{A} = \Sigma_a \Sigma_a^\top$ and $\mathbf{B} = \Sigma_b \Sigma_b^\top$ where all the entries of Σ_a, Σ_b are drawn according to $\mathcal{N}(0, 1)$. The integrand of interest is $f_{\mu,\nu}^{(2)}(\theta) = |\theta^\top(a - b)|^2 + \left(\sqrt{\theta^\top \mathbf{A} \theta} - \sqrt{\theta^\top \mathbf{B} \theta}\right)^2$. Figure 4 reports the MSE of the different methods with respect to both the number of projections and the computing time in dimension $d \in \{3; 6\}$, while Table 3 reports the MSE and computing times (in ms) in dimension $d \in \{5; 10; 20\}$.
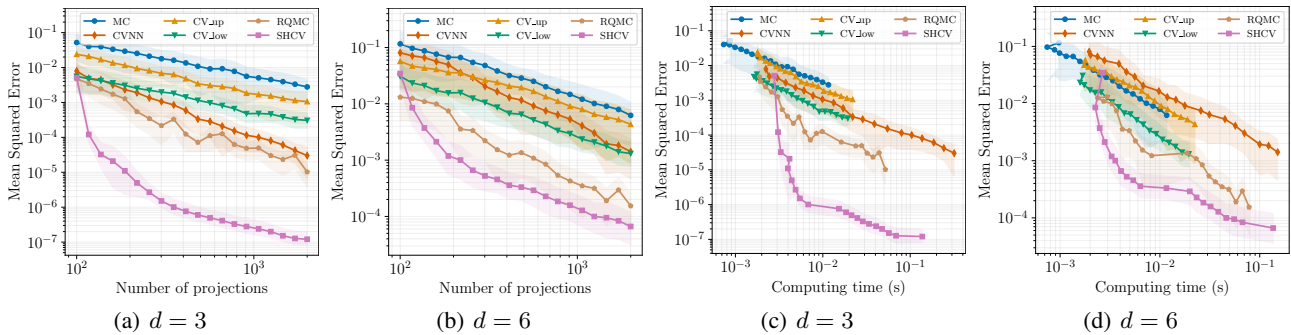

| (a) $d = 3$ | (b) $d = 6$ | (c) $d = 3$ | (d) $d = 6$ |

Figure 4. MSE and computing time for Gaussian distributions, dimension $d \in \{3; 6\}$, obtained over 100 replications.

Method	$d = 5$		$d = 10$		$d = 20$	
	MSE	Time	MSE	Time	MSE	Time
MC	2.6e-2	06.4 ± 1.1	8.4e-3	06.7 ± 0.8	2.0e-2	06.6 ± 0.5
CV$_{low}$	4.7e-3	11.1 ± 3.4	3.0e-3	11.6 ± 3.0	4.4e-3	12.1 ± 2.6
CV$_{up}$	1.5e-2	11.4 ± 3.5	7.0e-3	11.8 ± 3.1	1.6e-2	13.2 ± 2.8
CVNN	7.7e-3	43.4 ± 13	1.0e-2	60.0 ± 16	2.6e-2	70.6 ± 14
QMC	4.9e-3	22.4 ± 3.4	4.1e-3	41.1 ± 11	1.3e-2	60.9 ± 7.1
RQMC	7.5e-4	25.6 ± 6.8	2.4e-3	40.9 ± 11	1.4e-2	66.4 ± 14
SHCV	**4.8e-5**	09.8 ± 5.1	**1.7e-3**	08.9 ± 1.2	**2.7e-3**	14.3 ± 2.0

Table 3. Mean Squared Error (MSE) and computing time (ms) for Gaussian distributions in dimension $d \in \{5; 10; 20\}$ and $n = 500$.

B.2. Multivariate Gaussian: Sampled Case with Discrete Distributions μ_m, ν_m

We compute $\mathrm{SW}_2^2(\mu_m, \nu_m)$ between empirical distributions $\mu_m = m^{-1} \sum_{i=1}^m \delta_{x_i}$ and $\nu_m = m^{-1} \sum_{j=1}^m \delta_{y_j}$,

$$x_1, \ldots, x_m \sim \mu = \mathcal{N}(a, \mathbf{A}), \quad y_1, \ldots, y_m \sim \nu = \mathcal{N}(b, \mathbf{B})$$

with means $a, b \sim \mathcal{N}_d(\mathbb{1}_d, I_d)$ and covariance matrices $\mathbf{A} = \Sigma_a \Sigma_a^\top$ and $\mathbf{B} = \Sigma_b \Sigma_b^\top$ where all the entries of Σ_a, Σ_b are drawn according to $\mathcal{N}(0, 1)$. This time, instead of evaluating the integrand $f_{\mu,\nu}^{(2)}$ directly, we assume that one only has access to data samples from the source and the target. We use $m = 1000$ data samples from the source $x_1, \ldots, x_m \sim \mu$ and the target $y_1, \ldots, y_m \sim \nu$. Figure 5 reports the mean squared errors of the different methods with respect to both the number of projections and the computing time in dimension $d \in \{3; 6\}$. Tables 4 and 5 report the MSE and computing times (in ms) in dimension $d \in \{5; 10; 20\}$.
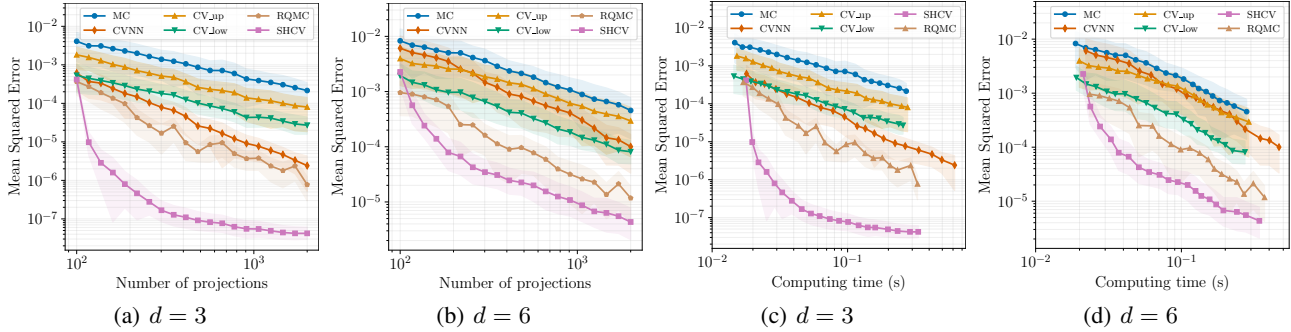

| (a) $d = 3$ | (b) $d = 6$ | (c) $d = 3$ | (d) $d = 6$ |

Figure 5. MSE for sampled Gaussian distributions supported on $m = 1000$ points, dimension $d \in \{3; 6\}$, obtained over 100 replications.

Method	$d = 5$		$d = 10$		$d = 20$	
	$n = 500$	$n = 1000$	$n = 500$	$n = 1000$	$n = 500$	$n = 1000$
MC	1.45e-3	8.05e-4	9.45e-4	5.46e-4	1.47e-3	6.67e-4
$\mathrm{CV}_{\mathrm{low}}$	2.67e-4	1.46e-4	3.45e-4	1.63e-4	3.82e-4	1.73e-4
$\mathrm{CV}_{\mathrm{up}}$	8.44e-4	4.82e-4	7.51e-4	4.76e-4	1.09e-3	4.73e-4
CVNN	4.29e-4	1.79e-4	1.12e-3	4.77e-4	2.14e-3	1.44e-3
QMC	2.91e-4	7.97e-5	2.37e-4	8.35e-5	6.60e-4	1.42e-4
RQMC	5.80e-5	1.91e-5	2.75e-4	1.20e-4	1.17e-3	3.64e-4
SHCV	**2.68e-6**	**5.63e-7**	**1.93e-4**	**6.86e-5**	**2.95e-4**	**1.16e-4**

Table 4. Mean Squared Error (MSE) for sampled Gaussian distributions supported on $m = 1000$ points in dimension $d \in \{5; 10; 20\}$ with $n \in \{500; 1000\}$ projections, obtained over 100 independent runs.

Method	$d = 5$		$d = 10$		$d = 20$	
	$n = 500$	$n = 1000$	$n = 500$	$n = 1000$	$n = 500$	$n = 1000$
MC	81.1 ± 3.5	152 ± 0.9	80.7 ± 4.4	152 ± 1.2	81.1 ± 1.8	155 ± 2.3
$\mathrm{CV}_{\mathrm{low}}$	79.7 ± 1.1	151 ± 1.1	80.1 ± 1.4	150 ± 2.2	80.0 ± 1.0	155 ± 4.0
$\mathrm{CV}_{\mathrm{up}}$	83.0 ± 1.2	157 ± 1.7	83.0 ± 1.7	157 ± 1.6	83.1 ± 1.5	160 ± 3.0
CVNN	110 ± 2.2	208 ± 1.5	122 ± 1.6	244 ± 2.1	127 ± 1.4	275 ± 3.6
QMC	100 ± 1.2	174 ± 1.3	113 ± 1.4	187 ± 1.0	129 ± 1.4	214 ± 3.5
RQMC	96.3 ± 2.2	190 ± 1.4	113 ± 1.2	188 ± 2.4	130 ± 1.0	214 ± 3.8
SHCV	89.0 ± 6.3	175 ± 3.4	89.0 ± 4.5	176 ± 2.4	88.1 ± 2.8	180 ± 5.6

Table 5. Computing time (ms) for sampled Gaussian distributions supported on $m = 1000$ points in dimension $d \in \{5; 10; 20\}$ with $n \in \{500; 1000\}$ projections, obtained over 100 independent runs.

B.3. Empirical Distributions of 3D Point Clouds

We use three point clouds taken at random from the ShapeNetCore dataset of Chang et al. (2015) corresponding to the objects plane, lamp and bed, each composed of $m = 2048$ points of $\mathbb{R}^3$ (see Figure 6 below for visualization). The SHCV estimate uses spherical harmonics up to degree $2L = 16$ leading to $s = 152$ control variates. Figure 3 reports the boxplots of the error distribution $(\widehat{\mathrm{SW}}_n(\mu_m, \nu_m) - \mathrm{SW}(\mu_m, \nu_m))$ for the different SW estimates based on n random projections with $n \in \{100; 250; 500; 1000\}$ obtained over 100 independent runs while Table 6 presents the mean squared errors $\mathbb{E}[|\widehat{\mathrm{SW}}_n(\mu_m, \nu_m) - \mathrm{SW}(\mu_m, \nu_m)|^2]$ for the different SW estimates with $n \in \{100; 500; 1000\}$. Once again, our SHCV estimate provides the best performance with huge gains in terms of variance reduction.
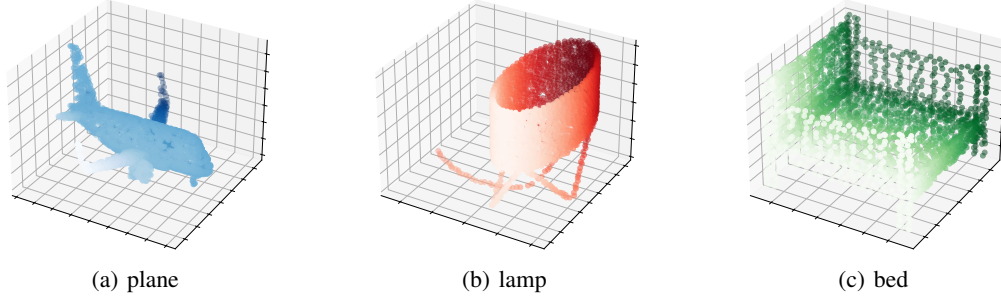

(a) plane   (b) lamp   (c) bed

Figure 6. Visualization of 3D point clouds of plane, lamp, bed.

Method	PLANE/LAMP			LAMP/BED			PLANE/BED		
	$n = 100$	$n = 500$	$n = 1000$	$n = 100$	$n = 500$	$n = 1000$	$n = 100$	$n = 500$	$n = 1000$
MC	5.1e-5	1.1e-5	6.0e-6	2.0e-4	3.8e-5	1.5e-5	1.3e-4	2.6e-5	1.2e-5
CV_{low}	4.8e-5	9.3e-6	5.1e-6	9.3e-5	1.4e-5	6.5e-6	4.4e-5	9.8e-6	5.2e-6
CV_{up}	1.1e-5	2.2e-6	1.3e-6	1.1e-5	2.3e-6	1.0e-6	6.6e-5	1.5e-5	8.7e-6
CVNN	1.3e-5	5.2e-7	1.8e-7	2.4e-5	1.1e-6	3.5e-7	3.2e-5	1.1e-6	3.3e-7
QMC	1.4e-5	6.9e-7	2.8e-8	2.4e-5	3.8e-7	2.9e-7	1.2e-4	1.2e-7	4.9e-7
RQMC	7.7e-6	2.2e-7	8.1e-8	1.5e-5	2.7e-7	1.3e-7	2.6e-5	1.0e-6	3.0e-7
SHCV	**5.3**e-6	**2.1**e-8	**2.1**e-8	**2.5**e-6	**1.5**e-8	**1.4**e-8	**1.3**e-5	**3.7**e-8	**3.6**e-8

Table 6. Mean Squared Error $\mathbb{E}[|\widehat{\mathrm{SW}}_n(\mu_m, \nu_m) - \mathrm{SW}(\mu_m, \nu_m)|^2]$ for different SW estimates based on n random projections with $n \in \{100; 500; 1000\}$ obtained over 100 independent runs.

Method	PLANE/LAMP			LAMP/BED			PLANE/BED		
	$n = 100$	$n = 500$	$n = 1000$	$n = 100$	$n = 500$	$n = 1000$	$n = 100$	$n = 500$	$n = 1000$
MC	36.4 ± 10	184 ± 3	309 ± 3	37.9 ± 10	184 ± 3	308 ± 3	35.6 ± 10	185 ± 3	308 ± 3
CV_{low}	36.8 ± 10	180 ± 1	301 ± 1	38.4 ± 10	181 ± 2	300 ± 2	36.1 ± 11	181 ± 1	300 ± 1
CV_{up}	38.1 ± 10	187 ± 2	316 ± 1	39.7 ± 10	187 ± 1	315 ± 1	37.2 ± 11	188 ± 1	314 ± 1
CVNN	40.4 ± 11	220 ± 1	417 ± 1	42.0 ± 11	220 ± 1	414 ± 2	39.5 ± 11	221 ± 1	414 ± 1
QMC	32.9 ± 12	194 ± 3	324 ± 1	27.9 ± 10	194 ± 2	327 ± 1	38.0 ± 13	195 ± 2	325 ± 1
RQMC	40.6 ± 11	202 ± 2	346 ± 2	42.0 ± 10	203 ± 4	344 ± 2	39.3 ± 11	202 ± 1	343 ± 1
SHCV	37.9 ± 10	185 ± 2	311 ± 1	39.7 ± 10	185 ± 1	310 ± 1	37.1 ± 11	186 ± 1	310 ± 1

Table 7. Computing times (ms) for different SW estimates based on n random projections with $n \in \{100; 500; 1000\}$ obtained over 100 independent runs.

B.4. Kernel Support Vector Machines and Sliced-Wasserstein Distances

Kernel Support Vector Machines. Kernel Support Vector Machines (SVMs) are a powerful class of supervised learning algorithms used for classification and regression tasks. At their core, SVMs aim to find a hyperplane that best separates the data into different classes while maximizing the margin between these classes. However, in cases where the classes are not linearly separable, Kernel SVMs come into play. The key idea behind Kernel SVMs is to map the input data $x \in \mathbb{R}^q$ into a higher-dimensional space via $\phi(x)$, where a linear separation might be achievable (see Figure 7 below).
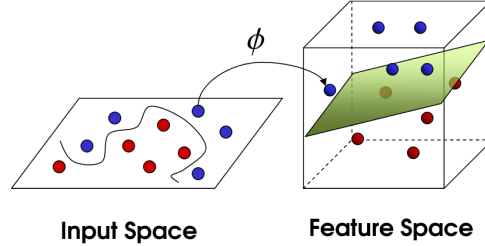

Figure 7. Visualization of Kernel transformation.

The kernel function $k(\cdot, \cdot)$ computes the similarity between pairs of data points in the higher-dimensional space without explicitly calculating the transformed data. Given training data $x_i \in \mathbb{R}^q$ with $i = 1, \ldots, n$ divided into two classes and a vector of labels $y \in \{-1; +1\}$, the goal is to find the parameters $\omega \in \mathbb{R}^q$ and $b \in \mathbb{R}$ such that the prediction given by $\mathrm{sign}(\omega^\top \phi(x_i) + b)$ is correct for most samples.

In Scikit-learn (Pedregosa et al., 2011), the 'SVC class' provides an implementation of Kernel SVMs. The objective function (also known as the loss function) for the SVM is derived from the problem of finding the optimal hyperplane that maximizes the margin between different classes. The formulation includes a regularization term C to balance the desire for a large margin with the goal of minimizing classification errors and the optimization problem of interest is given by

$$\min_{\omega, b, \zeta} \frac{1}{2} \omega^\top \omega + C \sum_{i=1}^{n} \zeta_i \quad \text{subject to } y_i \left(\omega^\top \phi(x_i) + b \right) \geq 1 - \zeta_i \text{ and } \zeta_i \geq 0, i = 1, \ldots, n.$$

The parameter C, which is shared among all Support Vector Machine (SVM) kernels, balances the compromise between misclassifying training examples and maintaining a simple decision surface. A lower C value results in a smoother decision surface, while a higher C value strives to accurately classify all training examples. This problem is usually solved with the dual formulation through the Kernel matrix $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$.
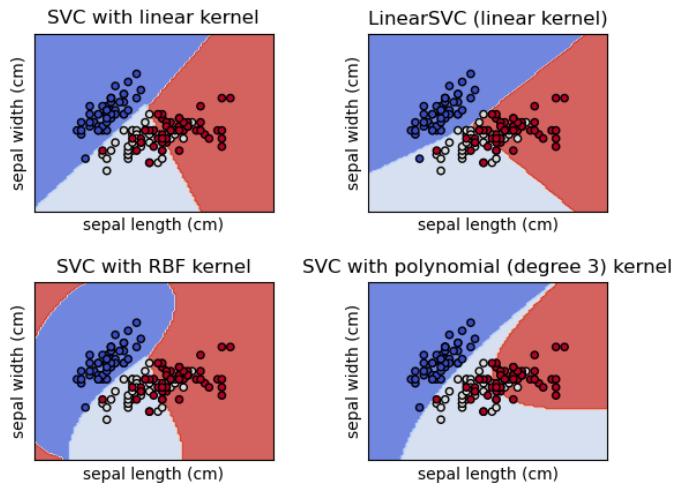

Figure 8. Comparison of different linear SVM classifiers on a 2D projection of the iris dataset.

Sliced-Wasserstein Gaussian Kernels. SW distances inherit topological properties from the original Wasserstein distance such as the weak convergence of probability measures (Nadjahi et al., 2020b). SW distances are useful in the context of distribution regression (Szabó et al., 2016) where the goal is to learn a real-valued function defined on the space of probability distributions from a sequence of observations. The classical Wasserstein distance cannot be used to build a positive definite kernel on the space of distributions as the induced Wasserstein space has curvature (Feragen et al., 2015). To overcome this problem, a SW-based kernel for absolutely continuous distributions was first introduced in Kolouri et al. (2016) and later on extended to empirical distributions by Meunier et al. (2022). For any $\gamma > 0$, both $e^{-\gamma \operatorname{SW}_2^2(\cdot,\cdot)}$ and $e^{-\gamma \operatorname{SW}_1(\cdot,\cdot)}$ are valid kernels on probability measures.

Using the SW kernel of Kolouri et al. (2016) defined by $k(\mu_i, \mu_j) = \exp(-\gamma \operatorname{SW}_2^2(\mu_i, \mu_j))$ with $\gamma > 0$, we consider the image classification task on the digits dataset of Alpaydin & Kaynak (1998). This dataset contains $N = 1\,797$ images of size 8×8 and the goal is to predict the label $z \in \{0, 1, \ldots, 9\}$ of the handwritten digit. Similarly to Meunier et al. (2022), we convert each image to an histogram so that the SW kernels have probability distributions as inputs. We use a train/test split of size $80/20$ giving $N_{\text{train}} = 1\,437$ training and $N_{\text{test}} = 360$ testing images respectively where each set is balanced. The Kernel SVM classification requires to compute the Gram matrices $K = (k(\mu_i, \mu_j))$ with $(i, j) \in [N_{\text{train}}] \times [N_{\text{train}}]$ and with $(i, j) \in [N_{\text{train}}] \times [N_{\text{test}}]$. Computational benefits occur with the SHCV estimate when many SW distances need to be computed. Using the projections $\theta_1, \ldots, \theta_n \sim \text{P}$ and the spherical harmonics $\varphi = (\varphi_j)_{j=1}^s$ to compute $\Pi = \Phi(\Phi^\top \Phi)^{-1} \Phi^\top$ with $\Phi = (\varphi(\theta_i)^\top)_{i=1}^n$, we need to compute the weights $w = [(I - \Pi)\mathbb{1}_n]/[\mathbb{1}_n^\top (I - \Pi)\mathbb{1}_n]$ *only once* in order to obtain a linear integration rule $w_n^\top f_n^{(i,j)}$ with $f_n^{(i,j)} = (f_{\mu_i,\nu_j}^{(p)}(\theta_1), \ldots, f_{\mu_i,\nu_j}^{(p)}(\theta_n))^\top$ to estimate $\operatorname{SW}_p^p(\mu_i, \nu_j)$.

Algorithm 9 SW-Kernel Monte Carlo	**Algorithm 10** SW-Kernel SHCV
Require: $\mathcal{D}_X = (\mu_1, \ldots, \mu_{N_X}), \mathcal{D}_Y = (\nu_1, \ldots, \nu_{N_Y})$, nb random projections $n \in \mathbb{N}, \text{P} \in \mathcal{P}(\mathbb{S}^{d-1})$	**Require:** $\mathcal{D}_X = (\mu_1, \ldots, \mu_{N_X}), \mathcal{D}_Y = (\nu_1, \ldots, \nu_{N_Y})$, $n \in \mathbb{N}, \text{P} \in \mathcal{P}(\mathbb{S}^{d-1})$, control variates $\varphi = (\varphi_j)_{j=1}^s$
1: Initialize Kernel matrix $K \in \mathbb{R}^{N_X \times N_Y}$	1: Sample random projections $\theta_1, \ldots, \theta_n \sim \text{P}$
2: Sample random projections $\theta_1, \ldots, \theta_n \sim \text{P}$	2: Compute $\Phi = (\varphi(\theta_i)^\top)_{i=1}^n$ and SHCV weights w_n
3: **For** $i = 1, \ldots, N_X$ **do**	3: **For** $i = 1, \ldots, N_X$ **do**
4: **For** $j = 1, \ldots, N_Y$ **do**	4: **For** $j = 1, \ldots, N_Y$ **do**
5: $f_n^{(i,j)} = (f_{\mu_i,\nu_j}^{(2)}(\theta_k))_{k=1}^n$	5: $f_n^{(i,j)} = (f_{\mu_i,\nu_j}^{(2)}(\theta_k))_{k=1}^n$
6: $\operatorname{MC}_n = (\mathbb{1}_n/n)^\top f_n^{(i,j)}$	6: $\operatorname{SHCV}_n = w_n^\top f_n^{(i,j)}$
7: $K[i, j] = \exp(-\gamma \operatorname{MC}_n)$	7: $K[i, j] = \exp(-\gamma \operatorname{SHCV}_n)$
8: Return Kernel matrix K	8: Return Kernel matrix K

As mentioned in Remark 5, computational benefits occur when $K > 1$ integrals need to be computed since the SHCV time complexity scales as $\mathcal{O}(Kn\omega_f + \omega(\Phi))$ compared to $\mathcal{O}(Kn\omega_f)$ for the standard Monte Carlo estimate when dealing with K integrals. Thus, when K is large, the extra term $\omega(\Phi)$ becomes negligible. In the experiments we are dealing with images and probability measures $\mu_1, \ldots, \mu_{N_X}, \nu_1, \ldots, \nu_{N_Y}$ in dimension $d = 2$ with a number of Monte Carlo projections $n \in \{25; 50; 75; 100\}$. The total number of integrals to be computed are:

- $K_{XX} = \binom{N_X}{2} = 1437 \times 1436/2 = 1,024,581$ for the 'train/train' kernel matrix.
- $K_{XY} = N_X \times N_Y = 1437 \times 360 = 517,320$ for the 'train/test' kernel matrix.

Below, we compare the time (in μs) to compute the evaluations $f_n^{(i,j)} = (f_{\mu_i,\nu_j}^{(2)}(\theta_k))_{k=1}^n$ for a single pair (i, j) vs. the time to compute the SHCV weights w_n. These weights need to be computed only once for all pairs of images. Interestingly, this extra computing time is of the same order are the time required for the evaluations of the integrands so that in the end it becomes negligible as the SHCV time complexity scales as $\mathcal{O}(Kn\omega_f + \omega(\Phi))$ compared to $\mathcal{O}(Kn\omega_f)$ for the standard Monte Carlo estimate.

MC projections n	evaluations $f_n^{(i,j)}$ in $\mathcal{O}(n\omega_f)$	SHCV weights w_n in $\mathcal{O}(\omega(\Phi))$
25	422 ± 36	207 ± 243
50	700 ± 16	479 ± 516
75	945 ± 17	761 ± 559
100	1182 ± 20	1057 ± 654

Table 8. Comparison of time computations (in μs) for the evaluations of the integrands vs. SHCV weights.

C. Mathematical Proofs

C.1. Proofs of Lemma 1 and Corollary 1

Proof of Lemma 1. Let $\mathscr{P}_\ell^d$ be the space of homogeneous polynomials of degree ℓ. Then $\dim \mathscr{P}_\ell^d = \binom{\ell+d-1}{d-1}$, the number of multi-indices of the form $(\alpha_1, \ldots, \alpha_d)$ for nonnegative integer α_j such that $\sum_{k=1}^d \alpha_k = \ell$. From this space, we need to remove the polynomials that are a multiple of $x_1^2 + \cdots + x_d^2$, since, on the sphere, this factor trivial. Any such multiple is of the form $(x_1^2 + \cdots + x_d^2)p(x)$ with $p \in \mathscr{P}_{\ell-2}^d$. For $\ell \geq 2$, we have (Atkinson & Han, 2012, Proposition 2.16) $N_\ell^d = \dim \mathscr{P}_\ell^d - \dim \mathscr{P}_{\ell-2}^d$. But then we have $\sum_{\ell=2,4,\ldots,2k} N_\ell^d = \sum_{\ell=2,4,\ldots,2k} \left[\binom{\ell+d-1}{d-1} - \binom{\ell-2+d-1}{d-1} \right] = \binom{2k+d-1}{d-1} - 1$.

Proof of Corollary 1. For $\theta \in \mathbb{S}^{d-1}$, we have $\langle \theta, X \rangle \sim \theta_\sharp^\star \mu$ and $\alpha \langle \theta, X \rangle + \langle \theta, b \rangle \sim \theta_\sharp^\star \nu$. The optimal transport map between $\theta_\sharp^\star \mu$ and $\theta_\sharp^\star \nu$ is $x \mapsto \alpha x + \langle \theta, b \rangle$ and $f_{\mu,\nu}^{(2)}(\theta) = \mathbb{E}_\mu[\langle \theta, (1-\alpha)X - b \rangle^2]$, which is a quadratic polynomial in θ.

C.2. Asymptotic Bound for OLSMC

Consider the general set-up of Section 3, in particular the OLSMC estimate in Equation (6).

Theorem 3 (Portier & Segers (2019)). *Let $(\Theta, \mathcal{F}, P)$ be a probability space. Let $\varphi = (\varphi_1, \ldots, \varphi_s)^\top \in (L_2(P))^s$ be a vector of orthonormal control variates, where $s = s_n$ depends on the Monte Carlo sample size n. Define $q_n(\theta) = \varphi(\theta)^\top \varphi(\theta)$ for $\theta \in \Theta$. If $\sup_{\theta \in \Theta} q(\theta) = o(n/s)$ as $n \to \infty$, then $\mathrm{I}_n^{\mathrm{ols}}(f) - \mathrm{I}(f) = \mathcal{O}_\mathbb{P}(\sigma_n/\sqrt{n})$ as $n \to \infty$, where $\sigma_n^2 = \mathrm{I}(\varepsilon_n^2)$, $\varepsilon_n = f - \hat{f}_s$, and $\hat{f}_s = \mathbb{E}[f] + \sum_{i=1}^s \langle f, \varphi_i \rangle \varphi_i$.*

C.3. Proof of Theorem 2

To apply Theorem 3, we consider $q_n(\theta) = \varphi(\theta)^\top \varphi(\theta)$ for $\theta \in \mathbb{S}^{d-1}$. Since $\varphi_1, \ldots, \varphi_s$ are the orthonormal control variates,

$$q_n(\theta) = \sum_{i=1}^s \varphi_i^2(\theta) = \sum_{\ell=2,4,\ldots,2L} \sum_{k=1}^{N_\ell^d} \varphi_{\ell,k}^2(\theta).$$

As a corollary to Theorem 2.9 in Atkinson & Han (2012), we have $\sum_{k=1}^{N_\ell^d} \varphi_{\ell,k}^2(\theta) = N_\ell^d/|\mathbb{S}^{d-1}|$. In view of Lemma 1,

$$q_n(\theta) = |\mathbb{S}^{d-1}|^{-1} \sum_{\ell=2,4,\ldots,2L} N_\ell^d = |\mathbb{S}^{d-1}|^{-1} s_{L,d} \tag{11}$$

with $s = s_{L,d} = \binom{2L+d-1}{d-1} - 1$. Applying Stirling's formula gives $s_{L,d} = \mathcal{O}(L^{d-1})$ as $L \to \infty$. Hence the condition $\sup_{\theta \in \mathbb{S}^{d-1}} q(\theta) = o(n/s)$ in Theorem 3 becomes $s_{L,d}^2 = o(n)$, which is fulfilled as soon as

$$L = o\left(n^{1/(2(d-1))}\right), \qquad n \to \infty. \tag{12}$$

Since the function $\theta \mapsto f(\theta) = f_{\mu,\nu}^{(p)}(\theta) = \mathrm{W}_p^p(\theta_\sharp^\star \mu, \theta_\sharp^\star \nu)$ on $\mathbb{S}^{d-1}$ is Lipschitz (Han, 2023, Theorem 2.4), the modulus of continuity of f evaluated at $h = d/(2L)$ satisfies

$$\omega(d/(2L)) \leq \mathrm{Lip}_\rho(f)\, d/(2L) \leq \mathrm{Lip}_{\|\cdot\|}(f)\, d/(2L), \tag{13}$$

where $\mathrm{Lip}_\rho(f) = \sup_{\theta,\gamma \in \mathbb{S}^{d-1}} |f(\theta) - f(\gamma)|/\rho(\theta,\gamma)$ and $\mathrm{Lip}_{\|\cdot\|}(f) = \sup_{\theta,\gamma \in \mathbb{S}^{d-1}} |f(\theta) - f(\gamma)|/\|\theta - \gamma\|$. More precisely, from Theorem 2.4 in Han (2023), we know that

$$\mathrm{Lip}_{\|\cdot\|}(f) \leq M_p(\mu,\nu) = p2^{p-1} \max\{M_p(\mu), M_p(\nu)\}^{(p-1)/p} \left((M_p(\mu))^{1/p} + (M_p(\nu))^{1/p}\right) \tag{14}$$

with $M_p(\mu) = \int_{\mathbb{R}^d} \|x\|^p \, \mathrm{d}\mu(x)$ and $M_p(\nu) = \int_{\mathbb{R}^d} \|x\|^p \, \mathrm{d}\nu(x)$. Hence, by Theorem 1, the residual function $\varepsilon_n = f - \Pi_{f,L}$ with $\Pi_{f,L} = \sum_{\ell=0}^L \sum_{k=1}^{N_{2\ell}^d} \langle f, \varphi_{2\ell,k} \rangle \varphi_{2\ell,k}$ satisfies

$$\sigma_n = \|\varepsilon_n\|_2 \leq \|f - P_{f,L}\|_2 \leq \|f - P_{f,L}\|_\infty \leq \frac{AM_p(\mu,\nu)d}{2L}. \tag{15}$$

Therefore, by Theorem 3 along with (12) and (15), we have, since $\mathrm{I}(f_{\mu,\nu}^{(p)}) = \mathrm{SW}_p^p(\mu,\nu)$, the rate

$$\left| \mathrm{SHCV}_{n,L}^p(\mu,\nu) - \mathrm{SW}_p^p(\mu,\nu) \right| = \mathcal{O}_\mathbb{P}(L^{-1}n^{-1/2}).$$

D. Additional Results on Spherical Harmonics

D.1. Spherical Harmonics in Small Dimensions

Circular Harmonics on $\mathbb{S}^1$. On the unit circle, the spherical harmonics are often called *circular harmonics* and correspond to the Fourier basis defined in circular coordinates $(x, y) = (\cos(t), \sin(t))$ by $\psi_0(\theta) = 1/\sqrt{2\pi}$ and for $k > 0$

$$\psi_{2k}(\theta) = \frac{\sin(kt)}{\sqrt{\pi}}, \qquad \psi_{2k-1}(\theta) = \frac{\cos((k-1)t)}{\sqrt{\pi}}.$$

The circular harmonics are $\varphi_{1,0} = \psi_0$ and for $\ell \geq 1$, $\varphi_{\ell,1} = \psi_{2\ell-1}$ and $\varphi_{\ell,2} = \psi_{2\ell}$.

Spherical Harmonics on $\mathbb{S}^2$. On the sphere $\mathbb{S}^2$ there exist $N_\ell^3 = 2\ell + 1$ linearly independent complex spherical harmonics $\psi_{\ell,k} : \mathbb{S}^2 \to \mathbb{C}$ of degree ℓ. Using the spherical coordinates $(x, y, z) = (\cos(u)\sin(v), \sin(u)\sin(v), \cos(v))$ with $0 \leq u < 2\pi$ and $0 \leq v < \pi$, the Laplace spherical harmonics are generally defined for all $-\ell \leq k \leq \ell$ by

$$\psi_{\ell,k}(u, v) = \sqrt{\frac{(2\ell+1)(\ell-k)!}{4\pi(\ell+k)!}} P_\ell^k(\cos(v)) e^{iku},$$

where P_ℓ^k is the associated Legendre polynomial defined by $P_\ell^k(x) = (-1)^k(1-x^2)^{k/2}(\mathrm{d}^k/\mathrm{d}x^k)P_\ell(x)$ with P_ℓ the Legendre polynomial of degree ℓ and $P_\ell^{-k}(x) = (-1)^k[(\ell-k)!/(\ell+k)!]P_\ell^k(x)$. This definition ensures orthonormality $\int_{v=0}^\pi \int_{u=0}^{2\pi} \psi_{\ell,k}\psi_{\ell',k'}^\star \mathrm{d}\Omega = \delta_{\ell,\ell'}\delta_{k,k'}$ with $\mathrm{d}\Omega = \sin(v)\mathrm{d}u\mathrm{d}v$. A real basis of spherical harmonics $\varphi_{\ell,k} : \mathbb{S}^2 \to \mathbb{R}$ can be defined in terms of their complex analogues by setting $\varphi_{\ell,0} = \psi_{\ell,0}$ and taking the real and imaginary parts as $\varphi_{\ell,k} = \sqrt{2}(-1)^k \Im(\psi_{\ell,|k|})$ if $k < 0$ and $\varphi_{\ell,k} = \sqrt{2}(-1)^k \Re(\psi_{\ell,k})$ if $k > 0$.

D.2. Number of Spherical Harmonics

Figure 9(a) presents the evolution of the number N_ℓ^d of spherical harmonics of degree ℓ in dimension d with $d \in [3, 10]$ and $\ell \in [1, 10]$ while Figures 9(b) and 9(c) report the evolution of the cumulative sum $\sum_{\ell=1}^L N_\ell^d$ and even cumulative sum $\sum_{\ell=1}^L N_{2\ell}^d$ for maximum degree $L \in [1, 10]$ and $2L \in \{2; 4; 6; 8; 10; 12\}$ respectively.
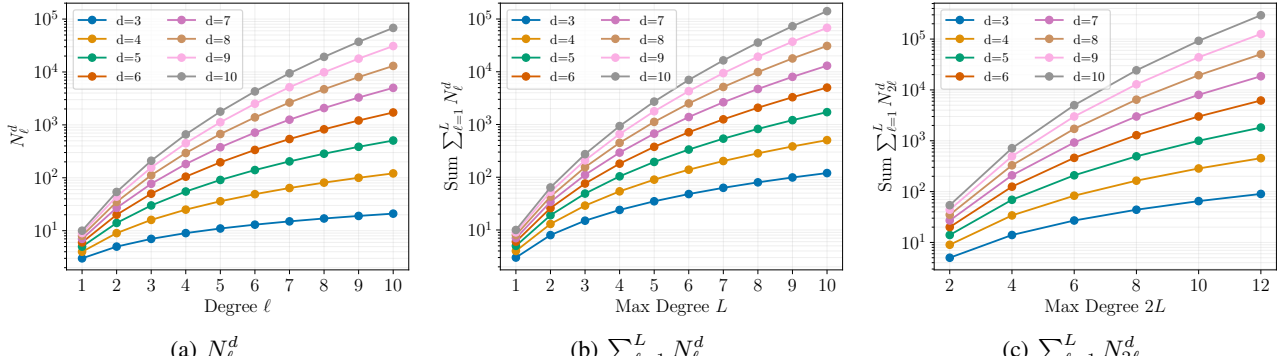

(a) N_ℓ^d (b) $\sum_{\ell=1}^L N_\ell^d$ (c) $\sum_{\ell=1}^L N_{2\ell}^d$

Figure 9. Evolution of the number N_ℓ^d of Spherical Harmonics and (even) cumulative sums according to the degree.

Table 9 reports the maximal degree of spherical harmonics and number of control variates for the considered dimensions in the experiments (See Lemma 1). For $d \in \{10; 20\}$ we only consider a subset of spherical harmonics of degree 4 to obtain a total number of $s = 300$ control variates.

dimension d	3	5	6	10	20
max degree $2L$	16	6	4	4	4
number s of control variates	152	209	125	714	8854

Table 9. Parameter setting of control variates.

D.3. Spherical Harmonics and Gegenbauer Polynomials

For $\alpha > 0$, Gegenbauer polynomials of degree ℓ, denoted by $C_\ell^\alpha : [-1, 1] \to \mathbb{R}$ are orthogonal polynomials with respect to the weight function $z \mapsto \left(1 - z^2\right)^{\alpha - 1/2}$. They are special cases of Jacobi polynomials and can be written as

$$C_\ell^\alpha(z) = \sum_{k=0}^{\lfloor \ell/2 \rfloor} \frac{(-1)^k \Gamma(\ell - k + \alpha)}{\Gamma(\alpha) k! (\ell - 2k)!} (2z)^{\ell - 2k}, \qquad C_\ell^\alpha(1) = \frac{\Gamma(2\alpha + \ell)}{\Gamma(2\alpha)\ell!}.$$

In the following we fix ℓ, set $N = N_\ell^d$, and fix $\{\varphi_1, \dots, \varphi_N\}$ as an orthonormal basis for $\mathscr{H}_\ell^d$. Denote by $Z_\ell(\cdot, \cdot)$ the reproducing kernel of $\mathscr{H}_\ell^d$. In terms of orthonormal basis we have

$$\forall \theta, \eta \in \mathbb{S}^{d-1}, \quad Z_\ell(\theta, \eta) = \sum_{k=1}^N \varphi_k(\theta) \varphi_k(\eta).$$

Interestingly, the spherical harmonics and linked to the Gegenbauer polynomials through the following addition formula

Theorem 4. *(Dai, 2013, Theorem 1.2.6) For $\ell \geq 0$ and $\theta, \eta \in \mathbb{S}^{d-1}$, $d \geq 3$*

$$Z_\ell(\theta, \eta) = \sum_{k=1}^N \varphi_k(\theta) \varphi_k(\eta) = \frac{\ell + \alpha}{\alpha} C_\ell^\alpha(\langle \theta, \eta \rangle), \qquad \alpha = \frac{d-2}{2}.$$

Corollary 3. *(Dai, 2013, Corollary 1.2.8) For $\alpha = \frac{d-2}{2}$, the Gegenbauer polynomials C_ℓ^α satisfy the orthogonality relation*

$$\frac{|\mathbb{S}^{d-2}|}{|\mathbb{S}^{d-1}|} \int_{-1}^1 C_\ell^\alpha(t) C_{\ell'}^\alpha(t) \left(1 - t^2\right)^{\alpha - \frac{1}{2}} \, \mathrm{d}t = h_\ell^\alpha \delta_{\ell, \ell'}, \qquad h_\ell^\alpha = \frac{\alpha}{\ell + \alpha} C_\ell^\alpha(1).$$

Let $\{\theta_1, \dots, \theta_N\}$ be a collection of points on $\mathbb{S}^{d-1}$. The following remarks motivate the definition of a fundamental set. We let $M_1 := \varphi_1(\theta_1)$ and for $k = 2, 3, \dots, N$, define matrices

$$M_k := \begin{bmatrix} \varphi_1(\theta_1) & \dots & \varphi_1(\theta_k) \\ \vdots & \dots & \vdots \\ \varphi_k(\theta_1) & \dots & \varphi_k(\theta_k) \end{bmatrix}, \quad M_k(\theta) := \left[\begin{array}{c|c} M_{k-1} & \begin{matrix} \varphi_1(\theta) \\ \vdots \\ \varphi_{k-1}(\theta) \end{matrix} \\ \hline \varphi_k(\theta_1), \dots \varphi_k(\theta_{k-1}) & \varphi_k(\theta), \end{array} \right]$$

The product of M_N and its transpose $M_N^\top$ can be summed, on applying the addition formula, as $M_N^\top M_N = [Z_\ell(\theta_i, \theta_j)]_{i,j=1}^N$, which shows, in particular, $\det [Z_\ell(\theta_i, \theta_j)]_{i,j=1}^N = (\det M_N)^2 \geq 0$

Definition 5. *(Dai, 2013, Definition 1.3.1) A collection of points $\{\theta_1, \dots, \theta_N\}$ in $\mathbb{S}^{d-1}$ is called a fundamental system of degree ℓ on the sphere $\mathbb{S}^{d-1}$ if*

$$\det [C_\ell^\alpha(\langle \theta_i, \theta_j \rangle)]_{i,j=1}^N > 0, \quad \alpha = \frac{d-2}{2}$$

Theorem 5. *(Dai, 2013, Theorem 1.3.3) If $\{\theta_1, \dots, \theta_N\}$ is a fundamental system of points on the sphere, then $\{C_\ell^\alpha(\langle \cdot, \theta_i \rangle) | i = 1, \dots, N\}$, $\alpha = (d-2)/2$ is a basis of $\mathscr{H}_\ell^d|_{\mathbb{S}^{d-1}}$*

This theorem lies at the heart of the practical implementation used in Dutordoir et al. (2020) which constructs a fundamental system of vectors for efficient evaluation of spherical harmonics. The mathematical aspects of this process involve concepts such as the Gegenbauer polynomial, Cholesky decomposition, optimization, and normalization. The goal is to construct a set of vectors efficiently spanning the space required for evaluating spherical harmonics in the given dimension and degree. The procedure works as follows : Start by defining the Gegenbauer polynomial, denoted as $C_\ell^\alpha(\theta)$, of a specific degree ℓ and a dimension-dependent parameter $\alpha = (d-2)/2$. Select the first vector arbitrarily, often chosen as the north pole or the unit vector along the last dimension. Compute the Cholesky decomposition of the Gegenbauer-gram matrix associated with the current vectors and iteratively add new vectors to to expand the fundamental system. For each new vector, optimize the determinant of the Gegenbauer-gram matrix, measuring the span of the space generated by the vectors. The optimization process involves finding a new vector that maximizes the determinant of the augmented matrix. This is typically done using optimization techniques such as BFGS. After finding the optimal vector, normalize it to ensure it has unit length.

D.4. Simple Case Study for Discrete Measures and Gaussian Measures

To investigate the approximation power of the integrand $\theta \mapsto f^{(p)}_{\mu,\nu}(\theta) = \mathrm{W}^p_p(\theta^\star_\sharp \mu_2, \theta^\star_\sharp \nu_2)$ via spherical harmonics, we consider a case study in dimension $d = 2$ on the unit circle $\mathbb{S}^1$ with circular harmonics $(\varphi_{\ell,k})$ (see Appendix D.1) . Using an Ordinary Least Squares regression, the polynomial approximation $\widehat{f}_{2L}$ of $f^{(p)}_{\mu,\nu}$ can be written as

$$\widehat{f}_{2L}(\theta) = \alpha + \sum_{\ell=2,4,\ldots,2L} \left(\beta^{(1)}_\ell \varphi_{\ell,1}(\theta) + \beta^{(2)}_\ell \varphi_{\ell,2}(\theta) \right)$$

where α and $\beta = (\beta^{(1)}_2, \beta^{(2)}_2, \ldots)$ are estimated by the OLS procedure. We investigate the smoothness properties of the integrand $f^{(p)}_{\mu,\nu}$ by comparing different measures μ and ν. We first consider the case of discrete measures μ_2 and ν_2 each supported on two atoms. In this simple case, the integrand $f^{(p)}_{\mu,\nu}(\theta)$ is only Lipschitz and we study the approximation error of our method. Then we consider the case of Gaussian measures $\mu = \mathcal{N}_2(\mathrm{m}_X, \boldsymbol{\Sigma}_X)$ and $\nu = \mathcal{N}_2(\mathrm{m}_Y, \boldsymbol{\Sigma}_Y)$. In that case, the integrand of interest is infinitely smooth.

Discrete measures. We consider the discrete measures μ_2 and ν_2 associated to the points $x_1 = (0,0)$, $x_2 = (1,0)$ and $y_1 = (0,0)$, $y_2 = (1,1)$ respectively. We have $\mu_2 = \frac{1}{2}\delta_{(0,0)} + \frac{1}{2}\delta_{(1,0)}$ and $\nu_2 = \frac{1}{2}\delta_{(0,0)} + \frac{1}{2}\delta_{(1,1)}$. The integrand $f^{(p)}_{\mu,\nu}$ can be computed exactly as

$$\mathrm{W}^p_p(\theta^\star_\sharp \mu_2, \theta^\star_\sharp \nu_2) = \begin{cases} \frac{1}{2}|\sin(\theta)|^p, & \theta \in [0, \pi/2] \cup [3\pi/4, 3\pi/2] \cup [7\pi/4, 2\pi], \\ \frac{1}{2}(|\cos(\theta)|^p + |\cos(\theta) + \sin(\theta)|^p), & \theta \in [\pi/2, 3\pi/4] \cup [3\pi/2, 7\pi/4]. \end{cases}$$

By focusing on the case of order $p = 2$, we have

$$I_1 = \int_0^{\pi/2} \sin^2(\theta) \mathrm{d}\theta = \frac{\pi}{4}, \qquad\qquad I_2 = \int_{\pi/2}^{3\pi/4} \cos^2(\theta) + (\cos(\theta) + \sin(\theta))^2 \mathrm{d}\theta = \frac{3\pi}{8} - \frac{3}{4},$$

$$I_3 = \int_{3\pi/4}^{3\pi/2} \sin^2(\theta) \mathrm{d}\theta = \frac{3\pi}{8} - \frac{1}{4}, \qquad I_4 = \int_{3\pi/2}^{7\pi/4} \cos^2(\theta) + (\cos(\theta) + \sin(\theta))^2 \mathrm{d}\theta = \frac{3\pi}{8} - \frac{3}{4},$$

$$I_5 = \int_{7\pi/4}^{2\pi} \sin^2(\theta) \mathrm{d}\theta = \frac{\pi}{8} - \frac{1}{4}.$$

Therefore,

$$\mathrm{SW}^2_2(\mu_2, \nu_2) = \frac{1}{2\pi} \int_0^{2\pi} W^2_2(\theta^\star_\sharp \mu_2, \theta^\star_\sharp \nu_2) \mathrm{d}\theta = \frac{1}{4\pi}(I_1 + I_2 + I_3 + I_4 + I_5) = \frac{3}{8} - \frac{1}{2\pi}.$$

Recall that $\mathrm{SHCV}^p_{n,L}(\mu_2, \nu_2) = \mathrm{I}^{\mathrm{ols}}_n$ where $(\mathrm{I}^{\mathrm{ols}}_n, \beta_n) \in \arg\min_{\alpha,\beta} \|f_n - \alpha \mathbb{1}_n - \Phi\beta\|^2_2$.

Figure 10(a) reports the evolution of $W^2_2(\theta^\star_\sharp \mu_2, \theta^\star_\sharp \nu_2)$ (black line) and its polynomial approximation $\widehat{f}_{2L}(\theta)$ with different degrees of harmonics $2L \in \{2; 4; 8; 10\}$. Figure 10(b) displays the approximation error, i.e., the difference ε_n between $W^2_2(\theta^\star_\sharp \mu_2, \theta^\star_\sharp \nu_2)$ and its polynomial approximation $\widehat{f}_{2L}(\theta)$.

Gaussian measures. Let $\mu = \mathcal{N}_d(\mathrm{m}_X, \boldsymbol{\Sigma}_X)$ and $\nu = \mathcal{N}_d(\mathrm{m}_Y, \boldsymbol{\Sigma}_Y)$. Then, for $\theta^\star_\sharp \mu = \mathcal{N}_1(\theta^\top \mathrm{m}_X, \theta^\top \boldsymbol{\Sigma}_X \theta)$ and $\theta^\star_\sharp \nu = \mathcal{N}_1(\theta^\top \mathrm{m}_Y, \theta^\top \boldsymbol{\Sigma}_Y \theta)$, we have $W^2_2(\theta^\star_\sharp \mu, \theta^\star_\sharp \nu) = \left|\theta^\top(\mathrm{m}_X - \mathrm{m}_Y)\right|^2 + (\sqrt{\theta^\top \boldsymbol{\Sigma}_X \theta} - \sqrt{\theta^\top \boldsymbol{\Sigma}_Y \theta})^2$. We consider $\mathrm{m}_X = (0,0)^\top, \mathrm{m}_Y = (1,1)^\top, \boldsymbol{\Sigma}_X = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1 \end{bmatrix}, \boldsymbol{\Sigma}_Y = \begin{bmatrix} 10 & 3 \\ 3 & 10 \end{bmatrix}$

Figure 11(a) reports the evolution of $W^2_2(\theta^\star_\sharp \mu_2, \theta^\star_\sharp \nu_2)$ (black line) and its polynomial approximation $\widehat{f}_{2L}(\theta)$ with different degrees of harmonics $2L \in \{2; 4; 6\}$. Observe that in this case the integrand is very smooth so that the polynomial approximations seem to overlap. Figure 11(b) displays the approximation error, *i.e.*, the difference ε_n between $W^2_2(\theta^\star_\sharp \mu_2, \theta^\star_\sharp \nu_2)$ and its polynomial approximation $\widehat{f}_{2L}(\theta)$.

Figure 12(a) presents the boxplots of the error distribution $(\mathrm{SHCV}^2_{n,L} - \mathrm{SW}^2_2)$ both for discrete measures μ_2, ν_2 and for Gaussian measures with different degrees of harmonics $2L \in \{2; 4; 6; 8\}$ and Monte Carlo sample size $n = 10^4$. The Sliced-Wasserstein distance for Gaussian measures is obtained by the naive Monte Carlo method with $n = 10^8$. The boxplots are obtained over 100 replications.
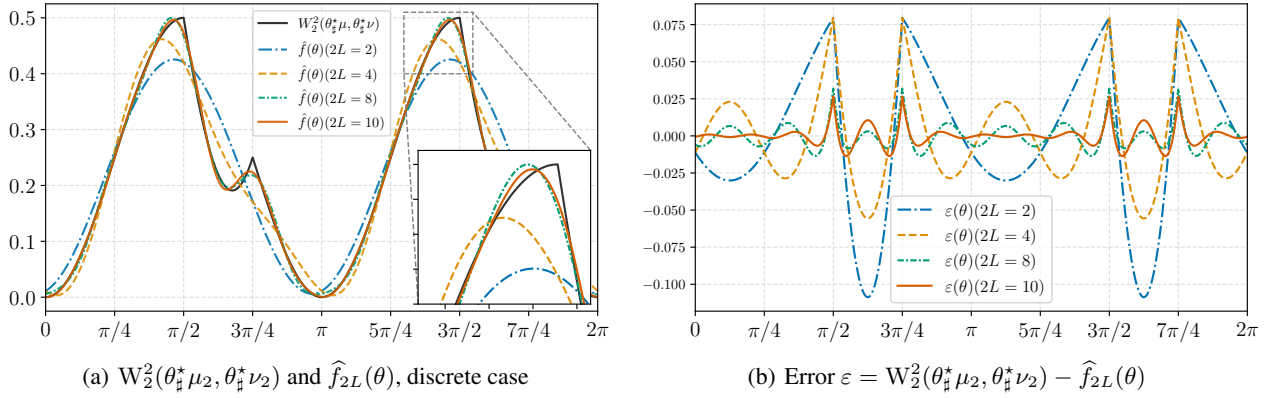
(a) $W_2^2(\theta_\sharp^\star\mu_2, \theta_\sharp^\star\nu_2)$ and $\widehat{f}_{2L}(\theta)$, discrete case

(b) Error $\varepsilon = W_2^2(\theta_\sharp^\star\mu_2, \theta_\sharp^\star\nu_2) - \widehat{f}_{2L}(\theta)$

Figure 10. $W_2^2(\theta_\sharp^\star\mu_2, \theta_\sharp^\star\nu_2)$, polynomial approximation $\widehat{f}_{2L}(\theta)$ and errors $\varepsilon = W_2^2(\theta_\sharp^\star\mu_2, \theta_\sharp^\star\nu_2) - \widehat{f}_{2L}(\theta)$ for two discrete measures.

(a) $W_2^2(\theta_\sharp^\star\mu_2, \theta_\sharp^\star\nu_2)$ and $\widehat{f}_{2L}(\theta)$, gaussian case
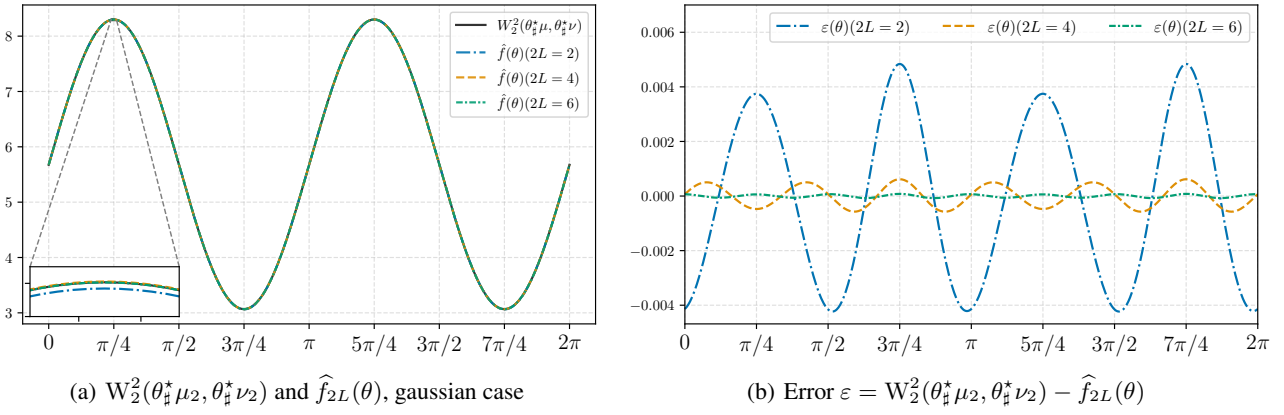
(b) Error $\varepsilon = W_2^2(\theta_\sharp^\star\mu_2, \theta_\sharp^\star\nu_2) - \widehat{f}_{2L}(\theta)$

Figure 11. $W_2^2(\theta_\sharp^\star\mu_2, \theta_\sharp^\star\nu_2)$, polynomial approximation $\widehat{f}_{2L}(\theta)$ and errors $\varepsilon = W_2^2(\theta_\sharp^\star\mu_2, \theta_\sharp^\star\nu_2) - \widehat{f}_{2L}(\theta)$ for two Gaussian measures.

(a) Boxplots Gaussian vs Discrete

(b) Boxplot Gaussian

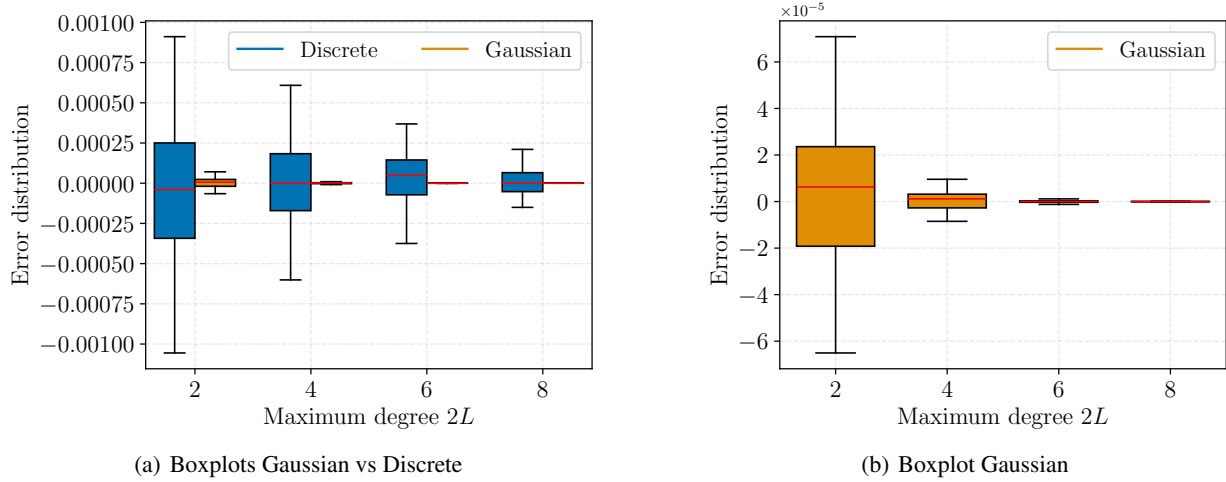*Figure 12.* Error distribution of $(\mathrm{SHCV}_{n,L}^2 - \mathrm{SW}_2^2)$.

D.5. Numerical Check for Exact Integration Rule

We compute $\mathrm{SW}_2^2(\mu, \nu)$ for $\mu = \mathcal{N}(a, \mathbf{A})$ and $\nu = \mathcal{N}(b, \mathbf{B})$ with means $a, b \sim \mathcal{N}(\mathbb{1}_d, I_d)$ and proportional (see Section 6.1) covariance matrices $\mathbf{A} = \Sigma_a \Sigma_a^\top$ and $\mathbf{B} = \gamma \mathbf{A}$ where all the entries of Σ_a are drawn according to $\mathcal{N}(0, 1)$. We consider dimensions $d \in \{3; 5\}$ and a number of projections $n \in [10^2; 10^4]$. Observe that in this case, the integrand of interest is a quadratic polynomial in θ given by

$$f_{\mu,\nu}^{(2)}(\theta) = |\theta^\top (a - b)|^2 + \left(\sqrt{\theta^\top \mathbf{A} \theta} - \sqrt{\theta^\top \mathbf{B} \theta} \right)^2$$
$$= |\theta^\top (a - b)|^2 + (1 - \sqrt{\gamma})^2 (\theta^\top \mathbf{A} \theta),$$

and $\mathrm{SW}_2^2(\mu, \nu)$ is given in close-form by integrating over $\theta \in \mathbb{S}^{d-1}$ such that

$$\mathrm{SW}_2^2(\mu, \nu) = \frac{\|a - b\|_2^2}{d} + \frac{(1 - \sqrt{\gamma})^2}{d} \mathrm{Tr}(\mathbf{A}).$$

Figure 13 reports the mean squared errors for the different estimates $\widehat{\mathrm{SW}}_n(\mu, \nu)$ where the expectation is computed as an average over 100 independent runs with parameter $\gamma = 2$. Observe that, while the standard Monte Carlo estimate and other control variate-based methods achieve a small MSE in the range $[10^{-7}; 10^{-3}]$, our SHCV estimate obtains a MSE of 10^{-30} which is virtually zero up to machine precision. This numerical comparison empirically validates the exact integration rule of Proposition 2.
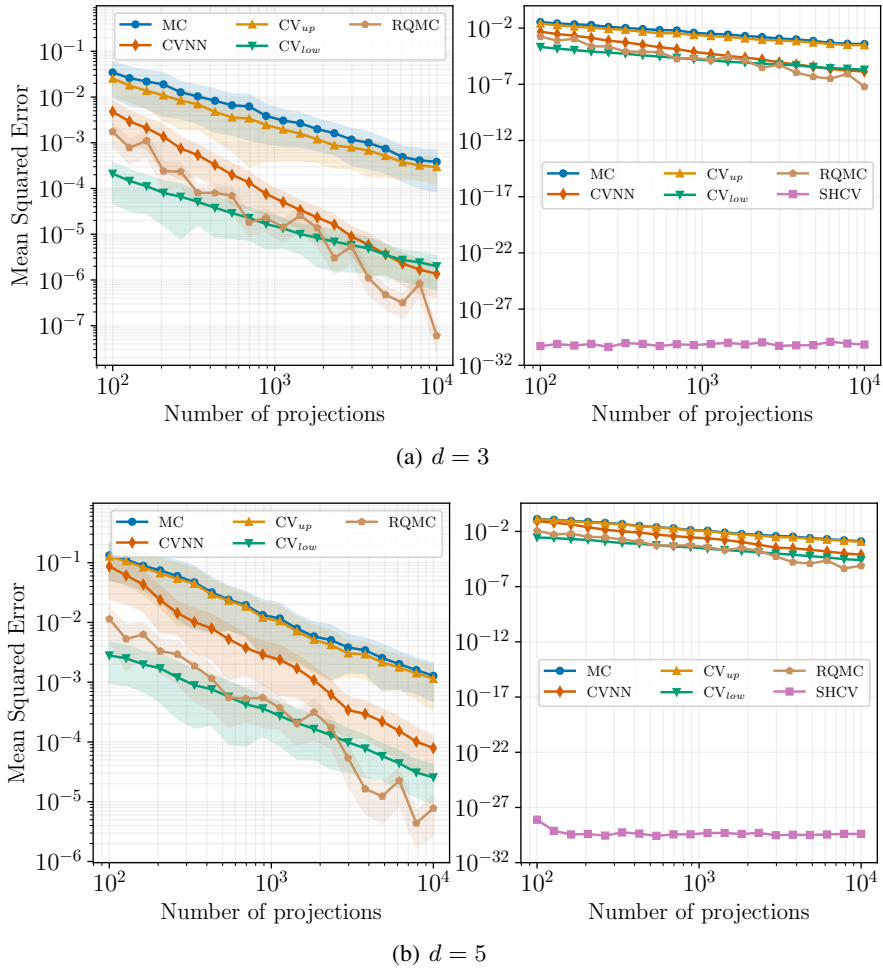

(a) $d = 3$

(b) $d = 5$

Figure 13. MSE for Gaussian distributions with proportional dispersion matrices, dimension $d \in \{3; 5\}$, obtained over 100 replications.