Efficient Nonparametric Smoothness Estimation

Shashank Singh
Carnegie Mellon University
`sss1@andrew.cmu.edu`

Simon S. Du
Carnegie Mellon University
`ssdu@cs.cmu.edu`

Barnabás Póczos
Carnegie Mellon University
`bapoczos@cs.cmu.edu`

Abstract

Sobolev quantities (norms, inner products, and distances) of probability density functions are important in the theory of nonparametric statistics, but have rarely been used in practice, due to a lack of practical estimators. They also include, as special cases, L^2 quantities which are used in many applications. We propose and analyze a family of estimators for Sobolev quantities of unknown probability density functions. We bound the finite-sample bias and variance of our estimators, finding that they are generally minimax rate-optimal. Our estimators are significantly more computationally tractable than previous estimators, and exhibit a statistical/computational trade-off allowing them to adapt to computational constraints. We also draw theoretical connections to recent work on fast two-sample testing and empirically validate our estimators on synthetic data.

1 Introduction

L^2 quantities (i.e., inner products, norms, and distances) of continuous probability density functions are important information theoretic quantities with many applications in machine learning and signal processing. For example, L^2 norm estimates can be used for goodness-of-fit testing [7], image registration and texture classification [12], and parameter estimation in semi-parametric models [36]. L^2 inner products estimates can generalize linear or polynomial kernel methods to inputs which are distributions rather than numerical vectors. [28] L^2 distance estimators are used for two-sample testing [1, 25], transduction learning [30], and machine learning on distributional inputs [27]. [29] gives applications of L^2 quantities to adaptive information filtering, classification, and clustering.

L^2 quantities are a special case of less-well-known *Sobolev quantities*. Sobolev norms measure *global smoothness* of a function in terms of integrals of squared derivatives. For example, for a non-negative integer s and a function $f : \mathbb{R} \to \mathbb{R}$ with an sth derivative $f^{(s)}$, the s-order Sobolev norm $\| \cdot \|_{H^s}$ is given by $\|f\|_{H^s} = \int_{\mathbb{R}} \left(f^{(s)}(x) \right)^2 dx$ (when this quantity is finite). See Section 2 for more general definitions, and see [21] for an introduction to Sobolev spaces.

Estimation of general Sobolev norms has a long history in nonparametric statistics (e.g., [32, 13, 10, 2]) This line of work was motivated by the role of Sobolev norms in many semi- and non-parametric problems, including density estimation, density functional estimation, and regression, (see [35], Section 1.7.1) where they dictate the convergence rates of estimators. Despite this, to our knowledge, these quantities have *never* been studied in real data, leaving an important gap between the theory and practice of nonparametric statistics. We suggest this is in part due a lack of *practical* estimators for these quantities. For example, the only one of the above estimators that is statistically minimax-optimal [2] is extremely difficult to compute in practice, requiring numerical integration over each of $O(n^2)$ different kernel density estimates, where n denotes the sample size. We know of no estimators previously proposed for Sobolev inner products and distances.

The **main goal of this paper** is to propose and analyze a family of computationally and statistically efficient estimators for Sobolev inner products, norms, and distances. Our specific contributions are:

1. We propose families of nonparametric estimators for all Sobolev quantities (Section 4).

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

2. We analyze the estimators' bias and variance. Assuming the underlying density functions lie in a Sobolev class of smoothness parametrized by s', we show the estimator for Sobolev quantities of order $s < s'$ converges to the true value at the "parametric" rate of $O(n^{-1})$ in mean squared error when $s' \geq 2s + D/4$, and at a slower rate of $O\left(n^{\frac{8(s-s')}{4s'+D}}\right)$ otherwise. (Section 5).

3. We validate our theoretical results on simulated data. (Section 8).

4. We derive asymptotic distributions for our estimators, and we use these to derive tests for the general statistical problem of two-sample testing. We also draw theoretical connections between our test and the recent work on nonparametric two-sample testing. (Section 9).

In terms of mean squared error, minimax lower bounds matching our convergence rates over Sobolev or Hölder smoothness classes have been shown by [16] for $s = 0$ (i.e., L^2 quantities), and [3] for Sobolev norms with integer s. Since these lower bounds intuitively "span" the space of relevant quantities, it is a small step to conjecture that our estimators are minimax rate-optimal for all Sobolev quantities and $s \in [0, \infty)$.

As described in Section 7, our estimators are computable in $O(n^{1+\varepsilon})$ time using only basic matrix operations, where n is the sample size and $\varepsilon \in (0,1)$ is a tunable parameter trading statistical and computational efficiency; the smallest value of ε at which the estimator continues to be minimax rate-optimal approaches 0 as we assume more smoothness of the true density.

2 Problem setup and notation

Let $\mathcal{X} = [-\pi, \pi]^D$ and let μ denote the Lebesgue measure on $\mathcal{X}$. For D-tuples $z \in \mathbb{Z}^D$ of integers, let $\psi_z \in L^2 = L^2(\mathcal{X})$ [1] defined by $\psi_z(x) = e^{-i\langle z, x\rangle}$ for all $x \in \mathcal{X}$ denote the zth element of the L^2-orthonormal Fourier basis, and, for $f \in L^2$, let $\widetilde{f}(z) := \langle \psi_z, f \rangle_{L^2} = \int_{\mathcal{X}} \psi_z(x) \overline{f(x)} \, d\mu(x)$ denote the zth Fourier coefficient of f. [2] For any $s \in [0, \infty)$, define the Sobolev space $H^s = H^s(\mathcal{X}) \subseteq L^2$ of order s on $\mathcal{X}$ by [3]

$$H^s = \left\{ f \in L^2 : \sum_{z \in \mathbb{Z}^D} z^{2s} \left| \widetilde{f}(z) \right|^2 < \infty \right\}. \tag{1}$$

Fix a known $s \in [0, \infty)$ and a unknown probability density functions $p, q \in H^s$, and suppose we have n IID samples $X_1, ..., X_n \sim p$ and $Y_1, ..., Y_n \sim q$ from each of p and q. We are interested in estimating the inner product

$$\langle p, q \rangle_{H^s} := \sum_{z \in \mathbb{Z}^D} z^{2s} \widetilde{p}(z) \overline{\widetilde{q}(z)} \quad \text{defined for all} \quad p, q \in H^s. \tag{2}$$

Estimating the inner product gives an estimate for the (squared) induced norm and distance, since [4]

$$\|p\|_{H^s}^2 := \sum_{z \in \mathbb{Z}^D} z^{2s} |\widetilde{p}(z)|^2 = \langle p, p \rangle_{H^s} \quad \text{and} \quad \|p - q\|_{H^s}^2 = \|p\|_{H^s}^2 - 2\langle p, q \rangle_{H^s} + \|q\|_{H^s}^2. \tag{3}$$

Since our theoretical results assume the samples from p and q are independent, when estimating $\|p\|_{H^s}^2$, we split the sample from p in half to compute two independent estimates of $\widetilde{p}$, although this may not be optimal in practice.

For a more classical intuition, we note that, in the case $D = 1$ and $s \in \{0, 1, 2, \dots\}$, (via Parseval's identity and the identity $\widetilde{f^{(s)}}(z) = (iz)^s \widetilde{f}(z)$), that one can show the following: H^s includes the

---

[1] We suppress dependence on $\mathcal{X}$; all function spaces are over $\mathcal{X}$ except as discussed in Section 2.1.

[2] Here, $\langle \cdot, \cdot \rangle$ denotes the dot product on $\mathbb{R}^D$. For a complex number $c = a + bi$, $\overline{c} = a - bi$ denotes the complex conjugate of c, and $|c| = \sqrt{c\overline{c}} = \sqrt{a^2 + b^2}$ denotes the modulus of c.

[3] When $D > 1$, $z^{2s} = \prod_{j=1}^{D} z_j^{2s}$. For $z < 0$, z^{2s} should be read as $(z^2)^s$, so that $z^{2s} \in \mathbb{R}$ even when $2s \notin \mathbb{Z}$. In the L^2 case, we use the convention that $0^0 = 1$.

[4] $\|p\|_{H^s}$ is *pseudonorm* on H^s because it fails to distinguish functions identical almost everywhere up to additive constants; a combination of $\|p\|_{L^2}$ and $\|p\|_{H^s}$ is used when a proper norm is needed. However, since probability densities integrate to 1, $\| \cdot - \cdot \|_{H^s}$ *is* a proper metric on the subset of (almost-everywhere equivalence classes of) probability density functions in H^s, which is important for two-sample testing (see Section 9). For simplicity, we use the terms "norm", "inner product", and "distance" for the remainder of the paper.

2

Functional Name	Functional Form	References
L^2 norms	$\|p\|_{L^2}^2 = \int (p(x))^2 \, dx$	[32, 6]
(Integer) Sobolev norms	$\|p\|_{H^k}^2 = \int \left(p^{(k)}(x)\right)^2 \, dx$	[2]
Density functionals	$\int \varphi(x, p(x)) \, dx$	[18, 19]
Derivative functionals	$\int \varphi(x, p(x), p'(x), \ldots, p^{(k)}(x)) \, dx$	[3]
L^2 inner products	$\langle p_1, p_2 \rangle_{L^2} = \int p_1(x) p_2(x) \, dx$	[16, 17]
Multivariate functionals	$\int \varphi(x, p_1(x), \ldots, p_k(x)) \, dx$	[34, 14]

Table 1: Some related functional forms for which estimators for which nonparametric estimators have been developed and analyzed. $p, p_1, ..., p_k$ are unknown probability densities, from each of which we draw n IID samples, φ is a known real-valued measurable function, and k is a non-negative integer.

subspace of L^2 functions with at least s derivatives in L^2 and, if $f^{(s)}$ denotes the sth derivative of f

$$\|f\|_{H^s}^2 = 2\pi \int_{\mathcal{X}} \left(f^{(s)}(x)\right)^2 \, dx = 2\pi \left\|f^{(s)}\right\|_{L^2}^2, \quad \forall f \in H^s. \tag{4}$$

In particular, when $s = 0$, $H^s = L^2$, $\|\cdot\|_{H^s} = \|\cdot\|_{L^2}$, and $\langle \cdot, \cdot \rangle_{H^s} = \langle \cdot, \cdot \rangle_{L^2}$. As we describe in the supplement, equation (4) and our results generalizes trivially to weak derivatives, as well as to non-integer $s \in [0, \infty)$ via a notion of fractional derivative.

2.1 Unbounded domains

A notable restriction above is that p and q are supported in $\mathcal{X} := [-\pi, \pi]^D$. In fact, our estimators and tests are well-defined and valid for densities supported on arbitrary subsets of $\mathbb{R}^D$. In this case, they act on the 2π-periodic summation $p_{2\pi} : [-\pi, \pi]^D \to [0, \infty]$ defined for $x \in \mathcal{X}$ by $p_{2\pi}(x) := \sum_{z \in \mathbb{Z}^D} p(x + 2\pi z)$, which is itself a probability density function on $\mathcal{X}$. For example, the estimator for $\|p\|_{H^s}$ will instead estimate $\|p_{2\pi}\|_{H^s}$, and the two-sample test for distributions p and q will attempt to distinguish $p_{2\pi}$ from $q_{2\pi}$. Typically, this is not problematic; for example, for most realistic probability densities, p and $p_{2\pi}$ have similar orders of smoothness, and $p_{2\pi} = q_{2\pi}$ if and only if $p = q$. However, there are (meagre) sets of exceptions; for example, if q is a translation of p by exactly 2π, then $p_{2\pi} = q_{2\pi}$, and one can craft a highly discontinuous function p such that $p_{2\pi}$ is uniform on $\mathcal{X}$. [39] These exceptions make it difficult to extend theoretical results to densities with arbitrary support, but they are fixed, in practice, by randomly rescaling the data (as in [4]). If the densities have (known) bounded support, they can simply be shifted and scaled to be supported on $\mathcal{X}$.

3 Related work

There is a large body of work on estimating nonlinear functionals of probability densities, with various generalizations in terms of the class of functionals considered. Table 1 gives a subset of such work, for functionals related to Sobolev quantities. As shown in Section 2, the functional form we consider is a strict generalization of L^2 norms, Sobolev norms, and L^2 inner products. It overlaps with, but is neither a special case nor a generalization of the remaining functional forms in the table.

Nearly all of the above approaches compute an optimally smoothed kernel density estimate and then perform bias corrections based on Taylor series expansions of the functional of interest. They typically consider distributions with densities that are β-Hölder continuous and satisfy periodicity assumptions of order β on the boundary of their support, for some constant $\beta > 0$ (see, for example, Section 4 of [16] for details of these assumptions). The Sobolev class we consider is a strict superset of this Hölder class, permitting, for example, certain "small" discontinuities. In this regard, our results are slightly more general than most of these prior works.

Finally, there is much recent work on estimating entropies, divergences, and mutual informations, using methods based on kernel density estimates [14, 17, 16, 24, 33, 34] or k-nearest neighbor statistics [20, 23, 22, 26]. In contrast, our estimators are more similar to orthogonal series density estimators, which are computationally attractive because they require no pairwise operations between samples. However, they require quite different theoretical analysis; unlike prior work, our estimator

3

is constructed and analyzed entirely in the frequency domain, and then related to the data domain via Parseval's identity. We hope our analysis can be adapted to analyze new, computationally efficient information theoretic estimators.

4 Motivation and construction of our estimator

For a non-negative integer parameter Z_n (to be specified later), let

$$p_n := \sum_{\|z\|_\infty \leq Z_n} \widetilde{p}(z)\psi_z \quad \text{and} \quad q_n := \sum_{\|z\|_\infty \leq Z_n} \widetilde{q}(z)\psi_z \quad \text{where} \quad \|z\|_\infty := \max_{j \in \{1,...,D\}} z_j \quad (5)$$

denote the L^2 projections of p and q, respectively, onto the linear subspace spanned by the L^2-orthonormal family $\mathcal{F}_n := \{\psi_z : z \in \mathbb{Z}^D, |z| \leq Z_n\}$. Note that, since $\widetilde{\psi_z}(y) = 0$ whenever $y \neq z$, the Fourier basis has the special property that it is orthogonal in $\langle \cdot, \cdot \rangle_{H^s}$ as well. Hence, since p_n and q_n lie in the span of $\mathcal{F}_n$ while $p - p_n$ and $q - q_n$ lie in the span of $\{\psi_z : z \in \mathbb{Z}\} \backslash \mathcal{F}_n$, $\langle p - p_n, q_n \rangle_{H^s} = \langle p_n, q - q_n \rangle_{H^s} = 0$. Therefore,

$$\langle p, q \rangle_{H^s} = \langle p_n, q_n \rangle_{H^s} + \langle p - p_n, q_n \rangle_{H^s} + \langle p_n, q - q_n \rangle_{H^s} + \langle p - p_n, q - q_n \rangle_{H^s}$$
$$= \langle p_n, q_n \rangle_{H^s} + \langle p - p_n, q - q_n \rangle_{H^s}. \quad (6)$$

We propose an unbiased estimate of $S_n := \langle p_n, q_n \rangle_{H^s} = \sum_{\|z\|_\infty \leq Z_n} z^{2s} \widetilde{p}_n(z)\overline{\widetilde{q}_n(z)}$. Notice that Fourier coefficients of p are the expectations $\widetilde{p}(z) = \mathbb{E}_{X \sim p}[\psi_z(X)]$. Thus, $\hat{p}(z) := \frac{1}{n}\sum_{j=1}^n \psi_z(X_j)$ and $\hat{q}(z) := \frac{1}{n}\sum_{j=1}^n \psi_z(Y_j)$ are independent unbiased estimates of $\widetilde{p}$ and $\widetilde{q}$, respectively. Since S_n is bilinear in $\widetilde{p}$ and $\widetilde{q}$, the plug-in estimator for S_n is unbiased. That is, **our estimator** for $\langle p, q \rangle_{H^s}$ is

$$\hat{S}_n := \sum_{\|z\|_\infty \leq Z_n} z^{2s} \hat{p}(z)\overline{\hat{q}(z)}. \quad (7)$$

5 Finite sample bounds

Here, we present our main theoretical results, bounding the bias, variance, and mean squared error of our estimator for finite n.

By construction, our estimator satisfies

$$\mathbb{E}\left[\hat{S}_n\right] = \sum_{\|z\|_\infty \leq Z_n} z^{2s}\, \mathbb{E}\left[\hat{p}(z)\right]\overline{\mathbb{E}\left[\hat{q}(z)\right]} = \sum_{\|z\|_\infty \leq Z_n} z^{2s}\widetilde{p}_n(z)\overline{\widetilde{q}_n(z)} = S_n.$$

Thus, via (6) and Cauchy-Schwarz, the bias of the estimator $\hat{S}_n$ satisfies

$$\left|\mathbb{E}\left[\hat{S}_n\right] - \langle p, q \rangle_{H^s}\right| = |\langle p - p_n, q - q_n \rangle_{H^s}| \leq \sqrt{\|p - p_n\|_{H^s}^2 \|q - q_n\|_{H^s}^2}. \quad (8)$$

$\|p - p_n\|_{H^s}$ is the error of approximating p by an order-Z_n trigonometric polynomial, a classic problem in approximation theory, for which Theorem 2.2 of [15] shows:

$$\text{if } p \in H^{s'} \text{ for some } s' > s, \quad \text{then} \quad \|p - p_n\|_{H^s} \leq \|p\|_{H^{s'}} Z_n^{s-s'}. \quad (9)$$

In combination with (8), this implies the following bound on the bias of our estimator:

Theorem 1. (Bias bound) *If $p, q \in H^{s'}$ for some $s' > s$, then, for $C_B := \|p\|_{H^{s'}} \|q\|_{H^{s'}}$,*

$$\left|\mathbb{E}\left[\hat{S}_n\right] - \langle p, q \rangle_{H^s}\right| \leq C_B Z_n^{2(s-s')} \quad (10)$$

Hence, the bias of $\hat{S}_n$ decays polynomially in Z_n, with a power depending on the "extra" $s' - s$ orders of smoothness available. On the other hand, as we increase Z_n, the number of frequencies at which we estimate $\hat{p}$ increases, suggesting that the variance of the estimator will increase with Z_n. Indeed, this is expressed in the following bound on the variance of the estimator.

4

Theorem 2. (Variance bound) *If $p, q \in H^{s'}$ for some $s' \geq s$, then*

$$\mathbb{V}\left[\hat{S}_n\right] \leq 2C_1 \frac{Z_n^{4s+D}}{n^2} + \frac{C_2}{n}, \quad \text{where} \quad C_1 := \frac{2^D \Gamma(4s+1)}{\Gamma(4s+D+1)} \|p\|_{L^2} \|q\|_{L^2} \quad (11)$$

and $C_2 := \left(\|p\|_{H^s} + \|q\|_{H^s}\right) \|p\|_{W^{2s,4}} \|q\|_{W^{2s,4}} + \|p\|_{H^s}^4 \|q\|_{H^s}^4$ are the constants (in n)

The proof of Theorem 2 is perhaps the most significant theoretical contribution of this work. Due to space constraints, the proof is given in the supplement. Combining Theorems 1 and 2 gives a bound on the mean squared error (MSE) of $\hat{S}_n$ via the usual decomposition into squared bias and variance:

Corollary 3. (Mean squared error bound) *If $p, q \in H^{s'}$ for some $s' > s$, then*

$$\mathbb{E}\left[\left(\hat{S}_n - \langle p, q \rangle_{H^s}\right)^2\right] \leq C_B^2 Z_n^{4(s-s')} + 2C_1 \frac{Z_n^{4s+D}}{n^2} + \frac{C_2}{n}. \quad (12)$$

If, furthermore, we choose $Z_n \asymp n^{\frac{2}{4s'+D}}$ (optimizing the rate in inequality 12), then

$$\mathbb{E}\left[\left(\hat{S}_n - \langle p, q \rangle_{H^2}\right)^2\right] \asymp n^{\max\left\{\frac{8(s-s')}{4s'+D}, -1\right\}}. \quad (13)$$

Corollary 3 recovers the phenomenon discovered by [2]: when $s' \geq 2s + \frac{D}{4}$, the minimax optimal MSE decays at the "semi-parametric" n^{-1} rate, whereas, when $s' \in \left(s, 2s + \frac{D}{4}\right)$, the MSE decays at a slower rate. Also, the estimator is L^2-consistent if $Z_n \to \infty$ and $Z_n n^{-\frac{2}{4s+D}} \to 0$ as $n \to \infty$. This is useful in practice, since s is known but s' is not.

Finally, it is worth reiterating that, by (3), these finite sample rates extend, with additional constant factors, to estimating Sobolev norms and distances.

6 Asymptotic distributions

In this section, we derive the asymptotic distributions of our estimator in two cases: (1) the inner product estimator and (2) the distance estimator in the case $p = q$. These results provide confidence intervals and two-sample tests without computationally intensive resampling. While (1) is more general in that it can be used with (3) to bound the asymptotic distributions of the norm and distance estimators, (2) provides a more precise result leading to a more computationally and statistically efficient two-sample test. Proofs are given in the supplementary material.

Theorem 4 shows that our estimator has a normal asymptotic distribution, assuming $Z_n \to \infty$ slowly enough as $n \to \infty$, and also gives a consistent estimator for its asymptotic variance. From this, one can easily estimate asymptotic confidence intervals for inner products, and hence also for norms.

Theorem 4. (Asymptotic normality) *Suppose that, for some $s' > 2s + \frac{D}{4}$, $p, q \in H^{s'}$, and suppose $Z_n n^{\frac{1}{4(s-s')}} \to \infty$ and $Z_n n^{-\frac{1}{4s+D}} \to 0$ as $n \to \infty$. Then, $\hat{S}_n$ is asymptotically normal with mean $\langle p, q \rangle_{H^s}$. In particular, for $j \in \{1, \ldots, n\}$ and $z \in \mathbb{Z}^D$ with $\|z\|_\infty \leq Z_n$, define $W_{j,z} := z^s e^{izX_j}$ and $V_{j,z} := z^s e^{izY_j}$, so that W_j and V_j are column vectors in $\mathbb{R}^{(2Z_n)^D}$. Let $\overline{W} := \frac{1}{n} \sum_{j=1}^n W_j, \overline{V} := \frac{1}{n} \sum_{j=1}^n V_j \in \mathbb{R}^{(2Z_n)^D}$,*

$$\Sigma_W := \frac{1}{n} \sum_{j=1}^n (W_j - \overline{W})(W_j - \overline{W})^T, \quad \text{and} \quad \Sigma_V := \frac{1}{n} \sum_{j=1}^n (V_j - \overline{V})(V_j - \overline{V})^T \in \mathbb{R}^{(2Z_n)^D \times (2Z_n)^D}$$

denote the empirical means and covariances of W and V, respectively. Then, for

$$\hat{\sigma}_n^2 := \begin{bmatrix} \overline{V} \\ \overline{W} \end{bmatrix}^T \begin{bmatrix} \Sigma_W & 0 \\ 0 & \Sigma_V \end{bmatrix} \begin{bmatrix} \overline{V} \\ \overline{W} \end{bmatrix}, \quad \text{we have} \quad \sqrt{n}\left(\frac{\hat{S}_n - \langle p, q \rangle_{H^s}}{\hat{\sigma}_n}\right) \xrightarrow{D} \mathcal{N}(0, 1),$$

where $\xrightarrow{D}$ denotes convergence in distribution.

Since distances can be written as a sum of three inner products (Eq. (3)), Theorem 4 might suggest an asymptotic normal distribution for Sobolev distances. However, extending asymptotic normality

5

from inner products to their sum requires that the three estimates be independent, and hence that we split data between the three estimates. This is inefficient in practice and somewhat unnatural, as we know, for example, that distances should be non-negative. For the particular case $p = q$ (as in the null hypothesis of two-sample testing), the following theorem [5] provides a more precise asymptotic (χ^2) distribution of our Sobolev distance estimator, after an extra decorrelation step. This gives, for example, a more powerful two-sample test statistic (see Section 9 for details).

Theorem 5. (Asymptotic null distribution) *Suppose that, for some $s' > 2s + \frac{D}{4}$, $p, q \in H^{s'}$, and suppose $Z_n n^{\frac{1}{4(s-s')}} \to \infty$ and $Z_n n^{-\frac{1}{4s+D}} \to 0$ as $n \to \infty$. For $j \in \{1, \ldots, n\}$ and $z \in \mathbb{Z}^D$ with $\|z\|_\infty \leq Z_n$, define $W_{j,z} := z^s \left(e^{-izX_j} - e^{-izY_j} \right)$, so that W_j is a column vector in $\mathbb{R}^{(2Z_n)^D}$. Let*

$$\overline{W} := \frac{1}{n} \sum_{j=1}^{n} W_j \in \mathbb{R}^{(2Z_n)^D} \quad and \quad \Sigma := \frac{1}{n} \sum_{j=1}^{n} \left(W_j - \overline{W} \right) \left(W_j - \overline{W} \right)^T \in \mathbb{R}^{(2Z_n)^D \times (2Z_n)^D}$$

denote the empirical mean and covariance of W, and define $T := n \overline{W}^T \Sigma^{-1} \overline{W}$. Then, if $p = q$, then

$$Q_{\chi^2((2Z_n)^D)}(T) \xrightarrow{D} \mathrm{Uniform}([0, 1]) \quad as \quad n \to \infty,$$

where $Q_{\chi^2(d)} : [0, \infty) \to [0, 1]$ denotes the quantile function (inverse CDF) of the χ^2 distribution $\chi^2(d)$ with d degrees of freedom.

Let $\hat{M}$ denote our estimator for $\|p - q\|_{H^s}$ (i.e., plugging $\hat{S}_n$ into (3)). While Theorem 5 immediately provides a *valid* two-sample test of desired level, it is not immediately clear how this relates to $\hat{M}$, nor is there any suggestion of why the test statistic ought to be a good (i.e., consistent) one. Some intuition is as follows. Notice that $\hat{M} = \overline{W}^T \overline{W}$. Since, by the central limit theorem, $\overline{W}$ has a normal asymptotic distribution, if the components of $\overline{W}$ were uncorrelated (and Z_n were fixed), we would expect $n\hat{M}$ to have an asymptotic χ^2 distribution with $(2Z_n)^D$ degrees of freedom. However, because we use the same data to compute each component of $\hat{M}$, they are *not* typically uncorrelated, and so the asymptotic distribution of $\hat{M}$ is difficult to derive. This motivates the statistic

$T = \left(\sqrt{\Sigma_W^{-1}} \overline{W} \right)^T \sqrt{\Sigma_W^{-1}} \overline{W}$, since the components of $\sqrt{\Sigma_W^{-1}} \overline{W}$ *are* (asymptotically) uncorrelated.

7 Parameter selection and statistical/computational trade-off

Here, we give statistical and computational considerations for choosing the smoothing parameter Z_n.

Statistical perspective: In practice, of course, we do not typically know s', so we cannot simply set $Z_n \asymp n^{\frac{2}{4s'+D}}$, as suggested by the mean squared error bound (13). Fortunately (at least for ease of parameter selection), when $s' \geq 2s + \frac{D}{4}$, the dominant term of (13) is C_2/n for $Z_n \asymp n^{-\frac{1}{4s+D}}$. Hence if we are willing to assume that the density has at least $2s + \frac{D}{4}$ orders of smoothness (which may be a mild assumption in practice), then we achieve *statistical optimality* (in rate) by setting $Z_n \asymp n^{-\frac{1}{4s+D}}$, which depends only on known parameters. On the other hand, the estimator can continue to benefit from additional smoothness *computationally*.

Computational perspective An attractive property of the estimator discussed is its computational simplicity and efficiency, in low dimensions. Most competing nonparametric estimators, such as kernel-based or nearest-neighbor methods, either take $O(n^2)$ time or rely on complex data structures such as k-d trees or cover trees [31] for $O(2^D n \log n)$ time performance. Since computing the estimator takes $O(nZ_n^D)$ time and $O(Z_n^D)$ memory (that is, the cost of estimating each of $(2Z_n)^D$ Fourier coefficients by an average), a statistically optimal choice of Z_n gives a runtime of $O\left(n^{\frac{4s'+2D}{4s'+D}} \right)$. Since the estimate requires only a vector outer product, exponentiation, and averaging, constants involved are small and computations parallelize trivially over frequencies and data.

Under severe computational constraints, for very large data sets, or if D is large relative to s', we can reduce Z_n to trade off statistical for computational efficiency. For example, if we want an estimator

---

[5]This result is closely related to Proposition 4 of [4]. However, in their situation, $s = 0$ and the set of test frequencies is fixed as $n \to \infty$, whereas our set is increasing.

6

(a) 1D Gaussians with different means.

(b) 1D Gaussians with different variance.

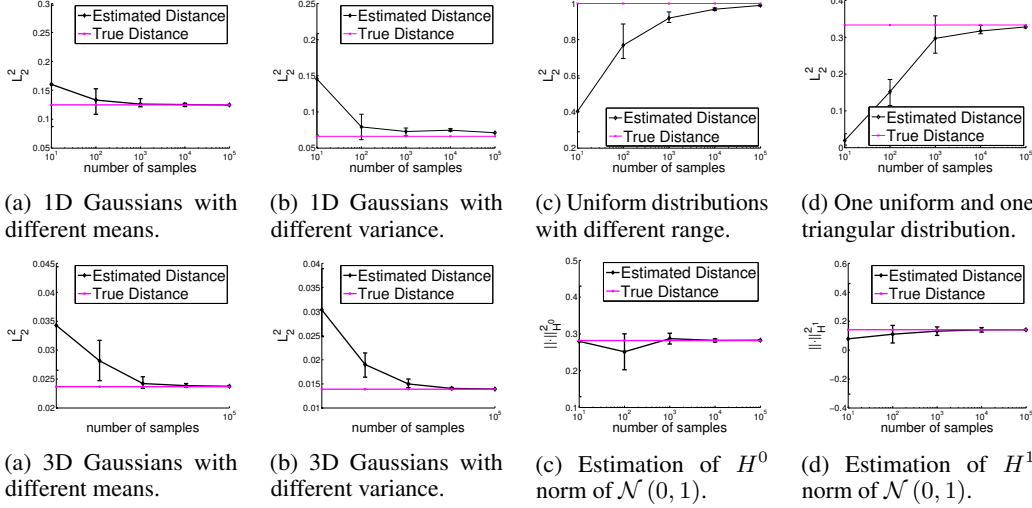
(c) Uniform distributions with different range.

(d) One uniform and one triangular distribution.

(a) 3D Gaussians with different means.

(b) 3D Gaussians with different variance.

(c) Estimation of H^0 norm of $\mathcal{N}(0,1)$.

(d) Estimation of H^1 norm of $\mathcal{N}(0,1)$.

with runtime $O(n^{1+\theta})$ and space requirement $O(n^\theta)$ for some $\theta \in \left(0, \frac{2D}{4s'+D}\right)$, setting $Z_n \asymp n^{\theta/D}$ still gives a consistent estimator, with mean squared error of the order $O\left(n^{\max\{\frac{4\theta(s-s')}{D}, -1\}}\right)$.

Kernel- or nearest-neighbor-based methods, including nearly all of the methods described in Section 3, tend to require storing previously observed data, resulting in $O(n)$ space requirements. In contrast, orthogonal basis estimation requires storing only $O(Z_n^D)$ estimated Fourier coefficients. The estimated coefficients can be incrementally updated with each new data point, which may make the estimator or close approximations feasible in streaming settings.

8 Experimental results

In this section, we use synthetic data to demonstrate effectiveness of our methods. [6] All experiments use $10, 10^2, \ldots, 10^5$ samples for estimation.

We first test our estimators on 1D L_2 distances. Figure 1a shows estimated distance between $\mathcal{N}(0,1)$ and $\mathcal{N}(1,1)$; Figure 1b shows estimated distance between $\mathcal{N}(0,1)$ and $\mathcal{N}(0,4)$; Figure 1c shows estimated distance between Unif$[0,1]$ and Unif$[0.5, 1.5]$; Figure 1d shows estimated distance between $[0,1]$ and a triangular distribution whose density is highest at $x = 0.5$. Error bars indicate asymptotic 95% confidence intervals based on Theorem 4. These experiments suggest 10^5 samples is sufficient to estimate L_2 distances with high confidence. Note that we need fewer samples to estimate Sobolev quantities of Gaussians than, say, of uniform distributions, consistent with our theory, since (infinitely differentiable) Gaussians are smoother than (discontinuous) uniform distributions.

Next, we test our estimators on L_2 distances of multivariate distributions. Figure 2a shows estimated distance between $\mathcal{N}([0,0,0], \mathbf{I})$ and $\mathcal{N}([1,1,1], \mathbf{I})$; Figure 2b shows estimated distance between $\mathcal{N}([0,0,0], \mathbf{I})$ and $\mathcal{N}([0,0,0], 4\mathbf{I})$. These experiments show that our estimators can also handle multivariate distributions. Lastly, we test our estimators for H^s norms. Figure 2c shows estimated H^0 norm of $\mathcal{N}(0,1)$ and Figure 2d shows H^1 norm of $\mathcal{N}(0,1)$. Additional experiments with other distributions and larger values of s are given in the supplement.

9 Connections to two-sample testing

We now discuss the use of our estimator in two-sample testing. From the large literature on nonparametric two-sample testing, we discuss only some recent approaches closely related to ours.

Let $\hat{M}$ denote our estimate of the Sobolev distance, consisting of plugging $\hat{S}$ into equation (3). Since $\| \cdot - \cdot \|_{H^s}$ is a metric on the space of probability density functions in H^s, computing $\hat{M}$ leads naturally to a two-sample test on this space. Theorem 5 suggests an asymptotic test, which is computationally preferable to a permutation test. In particular, for a desired Type I error rate $\alpha \in (0,1)$ our test rejects the null hypothesis $p = q$ if and only if $Q_{\chi^2(2Z_n^D)}(T) < \alpha$.

---

[6]MATLAB code for these experiments is available at `https://github.com/sss1/SobolevEstimation`.

7

When $s = 0$, this approach is closely related to several two-sample tests in the literature based on comparing empirical characteristic functions (CFs). Originally, these tests [11, 5] computed the same statistic T with a *fixed* number of random $\mathbb{R}^D$-valued frequencies instead of deterministic $\mathbb{Z}^D$-valued frequencies. This test runs in linear time, but is not generally consistent, since the two CFs need not differ almost everywhere. Recently, [4] suggested using *smoothed CFs*, i.e., the convolution of the CF with a universal smoothing kernel k. This is computationally easy (due to the convolution theorem) and, when $p \neq q$, $(\widetilde{p} * k)(z) \neq (\widetilde{q} * k)(z)$ for almost all $z \in \mathbb{R}^D$, reducing the need for carefully choosing test frequencies. Furthermore, this test is almost-surely consistent under very general alternatives. However, it is not clear what sort of assumptions would allow finite sample analysis of the power of their test. Indeed, the convergence as $n \to \infty$ can be arbitrarily slow, depending on the random test frequencies used. Our analysis instead uses the assumption $p, q \in H^{s'}$ to ensure that small, $\mathbb{Z}^D$-valued frequencies contain most of the power of $\widetilde{p}$. [7]

These fixed-frequency approaches can be thought of as the extreme point $\theta = 0$ of the computational/statistical trade-off described in section 7: they are computable in linear time and (with smoothing) are strongly consistent, but do not satisfy finite-sample bounds under general conditions.

At the other extreme ($\theta = 1$) are MMD-based tests of [8, 9], which use the entire spectrum $\widetilde{p}$. These tests are statistically powerful and have strong guarantees for densities in an RKHS, but have $O(n^2)$ computational complexity. [8] The computational/statistical trade-off discussed in Section 7 can be thought of as an interpolation (controlled by θ) of these approaches, with runtime in the case $\theta = 1$ approaching quadratic for large D and small s'.

10   Conclusions and future work

In this paper, we proposed nonparametric estimators for Sobolev inner products, norms and distances of probability densities, for which we derived finite-sample bounds and asymptotic distributions.

A natural follow-up question to our work is whether estimating smoothness of a density can guide the choice of smoothing parameters in nonparametric estimation. When analyzing many nonparametric estimators, Sobolev norms appear as the key unknown term in error bounds. While theoretically optimal smoothing parameter values are often suggested based on optimizing these error bounds, our work may suggest a practical way of mimicking this procedure by plugging *estimated* Sobolev norms into these bounds. For some problems, such as estimating functionals of a density, this may be especially useful, since no error metric is typically available for cross-validation. Even when cross-validation is an option, as in density estimation or regression, estimating smoothness may be faster, or may suggest an appropriate range of parameter values.

Acknowledgments
This material is based upon work supported by a National Science Foundation Graduate Research Fellowship to the first author under Grant No. DGE-1252522.

References

[1] N. H. Anderson, P. Hall, and D. M. Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50(1):41–54, 1994.

[2] P. J. Bickel and Y. Ritov. Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 381–393, 1988.

[3] L. Birgé and P. Massart. Estimation of integral functionals of a density. *The Annals of Statistics*, pages 11–29, 1995.

[4] K. P. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast two-sample testing with analytic representations of probability measures. In *NIPS*, pages 1972–1980, 2015.

[5] T. Epps and K. J. Singleton. An omnibus test for the two-sample problem using the empirical characteristic function. *Journal of Statistical Computation and Simulation*, 26(3-4):177–203, 1986.

[6] E. Giné and R. Nickl. A simple adaptive estimator of the integrated square of a density. *Bernoulli*, pages 47–61, 2008.

---

[7]Note that smooth CFs can be used in our test by replacing $\hat{p}(z)$ with $\frac{1}{n}\sum_{j=1}^{n} e^{-izX_j} k(x)$, where k is the inverse Fourier transform of a characteristic kernel. However, smoothing seems less desirable under Sobolev assumptions, as it spreads the power of the CF away from small $\mathbb{Z}^D$-valued frequencies where our test focuses.

[8]Fast MMD approximations have been proposed, including the Block MMD, [37] FastMMD, [38] and $\sqrt{n}$ sub-sampled MMD, but these lack the statistical guarantees of MMD.

8

[7] M. N. Goria, N. N. Leonenko, V. V. Mergel, and P. L. N. Inverardi. A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *J. Nonparametric Stat.*, 17:277–297, 2005.

[8] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520, 2006.

[9] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[10] P. Hall and J. S. Marron. Estimation of integrated squared density derivatives. *Statistics & Probability Letters*, 6(2):109–115, 1987.

[11] C. Heathcote. A test of goodness of fit for symmetric random variables. *Australian Journal of Statistics*, 14(2):172–181, 1972.

[12] A. O. Hero, B. Ma, O. J. J. Michel, and J. Gorman. Applications of entropic spanning graphs. *IEEE Signal Processing Magazine*, 19(5):85–95, 2002.

[13] I. Ibragimov and R. Khasminskii. On the nonparametric estimation of functionals. In *Symposium in Asymptotic Statistics*, pages 41–52, 1978.

[14] K. Kandasamy, A. Krishnamurthy, B. Poczos, L. Wasserman, et al. Nonparametric von Mises estimators for entropies, divergences and mutual informations. In *NIPS*, pages 397–405, 2015.

[15] H.-O. Kreiss and J. Oliger. Stability of the Fourier method. *SIAM Journal on Numerical Analysis*, 16(3): 421–433, 1979.

[16] A. Krishnamurthy, K. Kandasamy, B. Poczos, and L. A. Wasserman. Nonparametric estimation of renyi divergence and friends. In *ICML*, pages 919–927, 2014.

[17] A. Krishnamurthy, K. Kandasamy, B. Poczos, and L. A. Wasserman. On estimating L_2^2 divergence. In *AISTATS*, 2015.

[18] B. Laurent. *Efficient estimation of integral functionals of a density*. Université de Paris-sud, Département de mathématiques, 1992.

[19] B. Laurent et al. Efficient estimation of integral functionals of a density. *The Annals of Statistics*, 24(2): 659–681, 1996.

[20] N. Leonenko, L. Pronzato, V. Savani, et al. A class of rényi information estimators for multidimensional densities. *The Annals of Statistics*, 36(5):2153–2182, 2008.

[21] G. Leoni. *A first course in Sobolev spaces*, volume 105. American Math. Society, Providence, RI, 2009.

[22] K. Moon and A. Hero. Multivariate f-divergence estimation with confidence. In *Advances in Neural Information Processing Systems*, pages 2420–2428, 2014.

[23] K. R. Moon and A. O. Hero. Ensemble estimation of multivariate f-divergence. In *Information Theory (ISIT), 2014 IEEE International Symposium on*, pages 356–360. IEEE, 2014.

[24] K. R. Moon, K. Sricharan, K. Greenewald, and A. O. Hero III. Improving convergence of divergence functional ensemble estimators. *arXiv preprint arXiv:1601.06884*, 2016.

[25] L. Pardo. *Statistical inference based on divergence measures*. CRC Press, 2005.

[26] B. Póczos and J. G. Schneider. On the estimation of alpha-divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 609–617, 2011.

[27] B. Póczos, L. Xiong, and J. Schneider. Nonparametric divergence estimation with applications to machine learning on distributions. *arXiv preprint arXiv:1202.3758*, 2012.

[28] B. Póczos, L. Xiong, D. J. Sutherland, and J. Schneider. Nonparametric kernel estimators for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2989–2996. IEEE, 2012.

[29] J. C. Principe. *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media, 2010.

[30] N. Quadrianto, J. Petterson, and A. J. Smola. Distribution matching for transduction. In *Advances in Neural Information Processing Systems*, pages 1500–1508, 2009.

[31] P. Ram, D. Lee, W. March, and A. G. Gray. Linear-time algorithms for pairwise statistical problems. In *Advances in Neural Information Processing Systems*, pages 1527–1535, 2009.

[32] T. Schweder. Window estimation of the asymptotic variance of rank estimators of location. *Scandinavian Journal of Statistics*, pages 113–126, 1975.

[33] S. Singh and B. Póczos. Generalized exponential concentration inequality for renyi divergence estimation. In *Proceedings of The 31st International Conference on Machine Learning*, pages 333–341, 2014.

[34] S. Singh and B. Póczos. Exponential concentration of a density functional estimator. In *Advances in Neural Information Processing Systems*, pages 3032–3040, 2014.

[35] A. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519, 9780387790510.

[36] E. Wolsztynski, E. Thierry, and L. Pronzato. Minimum-entropy estimation in semi-parametric models. *Signal Processing*, 85(5):937–949, 2005.

[37] W. Zaremba, A. Gretton, and M. Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In *Advances in neural information processing systems*, pages 755–763, 2013.

[38] J. Zhao and D. Meng. Fastmmd: Ensemble of circular discrepancy for efficient two-sample test. *Neural computation*, 27(6):1345–1372, 2015.

[39] A. Zygmund. *Trigonometric series*, volume 1. Cambridge university press, 2002.

9