Efficient Interpolation of Density Estimators

Paxton Turner
Massachusetts Institute of Technology

Jingbo Liu
University of Illinois at Urbana–Champaign

Philippe Rigollet
Massachusetts Institute of Technology

Abstract

We study the problem of space and time efficient evaluation of a nonparametric estimator that approximates an unknown density. In the regime where consistent estimation is possible, we use a piecewise multivariate polynomial interpolation scheme to give a computationally efficient construction that converts the original estimator to a new estimator that can be queried efficiently and has low space requirements, all without adversely deteriorating the original approximation quality. Our result gives a new statistical perspective on the problem of fast evaluation of kernel density estimators in the presence of underlying smoothness. As a corollary, we give a succinct derivation of a classical result of Kolmogorov—Tikhomirov on the metric entropy of Hölder classes of smooth functions.

1 INTRODUCTION

The fast evaluation of kernel density estimators has been well-studied including approaches based on the fast Gauss transform (Greengard & Strain, 1991), hierarchical space decompositions (Greengard & Rokhlin, 1987), locality sensitive hashing (Charikar & Siminelakis, 2017; Backurs *et al.*, 2018; Siminelakis *et al.*, 2019; Backurs *et al.*, 2019), and binning (Scott & Sheather, 1985), as well as interpolation (Jones, 1989; Kogure, 1998), our main technique in this work. Typically these techniques carefully leverage the structure of the kernel under consideration, and many of them operate in a worst-case framework over the dataset. In this work, we consider the problem of fast evaluation of a density estimator $\hat{f}$ in a statistical setting where $\hat{f}$ gives a good pointwise approximation to an unknown density $f : [0,1]^d \to \mathbb{R}$ that lies in a Hölder class of smooth functions. We show that a pointwise approximation guarantee alone, without assuming any specific structure of the estimator $\hat{f}$, is enough to construct a new estimator $\tilde{f}$ that can be stored and queried cheaply, and whose approximation error is similar to that of the original estimator. Our approach is based on a multivariate polynomial interpolation scheme of Nicolaides (1972) (see also Chung & Yao, 1977) and provides an explicit formula for $\tilde{f}$ in terms of some judiciously chosen queries of the original estimator.

1.1 Background and related work

Density estimation is the task of estimating an unknown density f given an i.i.d. sample $X_1, \dots, X_n \sim \mathbb{P}_f$, where $\mathbb{P}_f$ is the probability distribution associated to f. A popular choice of density estimator is the kernel density estimator (KDE)

$$\hat{f}(y) := \frac{1}{nh^d} \sum_{j=1}^{n} K\left(\frac{X_i - y}{h}\right). \tag{1}$$

With proper setting of the bandwidth parameter h and choice of kernel K, the KDE $\hat{f}$ is a minimax optimal estimator over the L-Hölder smooth densities $\mathcal{P}_{\mathcal{H}}(\beta, L)$ of order β (see e.g. Tsybakov, 2009, Theorem 1.2):

$$\inf_{\hat{f}} \sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{E}_f \|\hat{f} - f\|_2 = \Theta_{\beta,d,L}(n^{-\frac{\beta}{2\beta+d}}). \tag{2}$$

Despite its statistical utility, the KDE (1) has the computational drawback that it naively requires $\Omega(n)$ time to evaluate a query. The problem of improving on these computational aspects has thus received a lot of attention.

Motivated by multi-body problems, Greengard & Strain (1991) developed the fast Gauss transform to rapidly evaluate sums of the form (1) when $K(x) = \exp(-|x|_2^2)$ is the Gaussian kernel. Their work is posed a worst-case batch setting where $\hat{f}$ is to be evaluated at

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

m points $y_1, \ldots, y_m$ specified in advance and the locations $X_1, \ldots, X_n$ lie in a box. Their techniques use hierarchical space decompositions and series expansions to show that (1) may be evaluated at $y_1, \ldots, y_m$ with precision ε in time $h^{-d}(\log \frac{1}{\varepsilon})^d(n+m)$. These results apply to any kernel that has a rapidly converging Hermite expansion (see also Greengard & Rokhlin, 1987). There are also follow up works on the improved fast Gauss transform and tree-based methods that use related ideas (Yang *et al.*, 2003; Lee *et al.*, 2006).

More recently, several works (Charikar & Siminelakis, 2017; Backurs *et al.*, 2018; Siminelakis *et al.*, 2019; Backurs *et al.*, 2019; Coleman & Shrivastava, 2020) are devoted to the problem of fast evaluation of (1) in high dimension using locality sensitive hashing. In these works, the dataset is carefully reweighted for importance sampling such that a randomly drawn datapoint X_r's corresponding kernel value $K(X_r - y)$ gives a good approximation to $\hat{f}(y)$. This sampling procedure can be executed efficiently using hashing-based methods. For example, Backurs *et al.* (2019) show that for the Laplace and Exponential kernels with bandwidth $h = 1$, e.g., the value $\hat{f}(y)$ can be computed with multiplicative $1 \pm \varepsilon$ error in time $O(\frac{d}{\sqrt{\tau}\varepsilon^2})$ even in worst case over the dataset, where τ is a uniform lower bound on the KDE.

Another effective approach to this problem in high dimensions is through coresets (Agarwal *et al.*, 2005; Clarkson, 2010; Phillips & Tai, 2018a,b). A coreset is a representative subset $\{X_i\}_{i \in S}$ of a dataset such that

$$\hat{f}(y) \approx \frac{1}{nh^d} \sum_{i \in S} K\left(\frac{X_i - y}{h}\right).$$

When $h = O(1)$, for example, the results of Phillips & Tai (2018b) give a polynomial time algorithm in n, d such that the coreset KDE yields an additive ε approximation to $\hat{f}$ using a coreset of size $\tilde{O}(\frac{\sqrt{d}}{\varepsilon})$. Their results hold in worst case over the dataset and for a variety of popular kernels. The methods of Phillips & Tai (2018b) are powered by state-of-the-art algorithms from discrepancy theory (Bansal *et al.*, 2018) (see Matoušek, 1999; Chazelle, 2000, for a comprehensive exposition on discrepancy).

Our approach is most closely related to prior work on the interpolation of kernel density estimators due to Jones (1989) and Kogure (1998). Motivated by visualization and computational aspects, Jones (1989) studies binned and piecewise linearly interpolated univariate kernel density estimators and provides precise bounds on the mean-integrated squared error. Kogure (1998) extends this work and constructs higher order piecewise polynomial interpolants of multivariate kernel density estimators, and shows that for very smooth densities, this procedure improves the mean-integrated squared error. In addition, we note the recent work of Belkin *et al.* (2019); Liang *et al.* (2020) demonstrating the perhaps surprising effectiveness of interpolation in nonparametric regression. We also remark that nonparametric estimators based on multivariate piecewise polynomials are well-studied in statistics (see e.g. Györfi *et al.*, 2006, Chapter 10), and there is a line of related literature in computer science on fast estimation of univariate densities that are well-approximated by piecewise polynomials (Chan *et al.*, 2014; Acharya *et al.*, 2017; Hao *et al.*, 2020).

Our work differs from Kogure (1998) in a few important respects. We do not assume $\hat{f}$ to be a KDE in the first place, but rather give a general method for effectively interpolating a minimax density estimator. Also, our results hold for the entire range of the smoothness parameter β and dimension d, while Kogure (1998) requires the density to be at least qd times differentiable when interpolating KDEs with kernels of order q (Tsybakov, 2009, Definition 1.3). On the other hand, our method increases the mean squared by a multiplicative factor $\tilde{O}(c_{\beta,d})$, while Kogure's approach improves the mean squared error (though our focus here is the L^∞ norm). Finally, we use a different interpolation scheme as detailed in Section 2.1.

1.2 Results

We seek to impose minimal requirements on a density estimator $\hat{f}$ of an unknown smooth density f so that it can be converted to a new estimator $\tilde{f}$ that performs well on the following criteria.

1. (**Minimax**) $\tilde{f}$ is a minimax estimator for f
2. (**Space-efficient**) $\tilde{f}$ can be stored efficiently
3. (**Fast querying**) $\tilde{f}$ can be evaluated efficiently
4. (**Fast preprocessing**) $\tilde{f}$ can be constructed efficiently

In this work, we focus on near-minimax estimation in the L^∞ norm, motivated by the aforementioned works on efficient evaluation of kernel density estimators. Since we impose that the unknown density f is supported on $[0,1]^d$, such a guarantee also implies upper bounds on the L^p error for all $p \geq 1$.

In the statistical setup where typically $\beta, d = O(1)$, by *efficient* we mean requiring only polynomial time or space in the sample size n. In particular for fixed β, by (2) consistent estimation is only possible when $d \ll \log n$. In what follows we indicate dependencies on the parameters β and d for clarity.

The requirement that we place on the estimator $\hat{f}$ to be converted is the following assumption.

Assumption 1. *For all $y \in [0,1]^d$ and $1 \geq t \geq \varepsilon$, we have*

$$\sup_{f \in \mathcal{P}_{\mathcal{H}}(\beta,L)} \mathbb{P}_f \left[\left| \hat{f}(y) - f(y) \right| > t \right] \leq 2 \exp\left(-\frac{t^2}{\varepsilon^2} \right),$$

where $\varepsilon := c^* \, n^{-\beta/(2\beta+d)}$ is the minimax rate of estimating L-Hölder smooth densities $\mathcal{P}_{\mathcal{H}}(\beta, L)$ of order β and $c^* = c_{\beta,d,L} > 0$.

The formal definition of the Hölder class $\mathcal{P}_{\mathcal{H}}(\beta, L)$ we consider is given in Section 1.3. In particular Assumption 1 is satisfied if the pointwise error is a sub-Gaussian random variable with parameter ε that captures the minimax rate of estimation. For the KDE built from a kernel K of order $\ell := \lfloor \beta \rfloor$ (Tsybakov, 2009, Definition 1.3) and bandwidth $h = n^{-\frac{1}{2\beta+d}}$, this assumption follows from a standard bias-variance trade-off and an application of Bernstein's inequality (see Section 4).

Under Assumption 1, we have our main result.

Theorem 1. *Let $f : [0,1]^d \to \mathbb{R}$ denote a probability density function, and let $\hat{f}$ denote an estimator satisfying Assumption 1 for some $\beta > 0$ and $d \geq 1$. Let Q denote the amount of time it takes to query $\hat{f}$. Set $\ell = \lfloor \beta \rfloor$. Then there exists an estimator $\tilde{f}$ that can be constructed in time $c_{\mathsf{con}} \, Q \, n^{\frac{d}{2\beta+d}}$, that requires at most $c_{\mathsf{sto}} \, n^{\frac{d}{2\beta+d}} \log n$ bits to store, that can be queried in time $c_{\mathsf{que}} \log n$, and that satisfies*

$$\mathbb{E}_f \|\tilde{f} - f\|_\infty < c_{\mathsf{err}} (\log n)^{1/2} n^{-\frac{\beta}{2\beta+d}}.$$

In Theorem 1, we may take

$$\begin{aligned}
c_{\mathsf{con}} &= \binom{\ell+d}{\ell}, \\
c_{\mathsf{sto}} &= 5d(\ell+1)(\log L), \binom{\ell+d}{\ell} \\
c_{\mathsf{que}} &= 14(d+\ell)^2 \binom{\ell+d}{d}, \text{ and} \\
c_{\mathsf{err}} &= 8c^* L d^{\frac{3}{2}\ell+2} \ell^\ell \binom{\ell+d}{\ell} \sqrt{\log 2 \binom{\ell+d}{\ell}}.
\end{aligned}$$

In particular, for $\beta, d = O(1)$, we can evaluate queries to $\tilde{f}$ in nearly constant time, and the estimator $\tilde{f}$ can be stored using sublinear space. Moreover, $\tilde{f}$ can be preprocessed in subquadratic time, assuming that the evaluation time of the original estimator $\hat{f}$ is $O_d(n)$, which holds for the KDE (1). We also note that $\tilde{f}$ is a near-minimax estimator in the sup norm, up to logarithmic factors, and thus by our domain assumption is also near-minimax in the L^p norms, again up to logarithmic factors. Finally, our construction in Section 2.1 yields an explicit formula for $\tilde{f}$ in terms of a sublinear number of initial queries of $\hat{f}$ on a judiciously chosen mesh. Specifically, the estimator $\tilde{f}$ is a piecewise multivariate interpolation of the estimator $\hat{f}$ on this mesh.

Though our focus is on density estimation, our method is not limited to this setting. The next result holds under a modified version of Assumption 1 and is derived by following the proof of Theorem 1. We omit the argument as it is very similar.

Theorem 2. *Let $f : [0,1]^d \to \mathbb{R}$ denote an L-smooth Hölder function of order β, and suppose that one has query access to a function $\hat{f}$ where $\|\hat{f} - f\|_\infty \leq \varepsilon$. Then by first computing $c_{\mathsf{con}} \, \varepsilon^{-\frac{d}{\beta}}$ initial queries of $\hat{f}$, one can construct a new function $\tilde{f}$ that satisfies $\|f - \tilde{f}\|_\infty \leq c_{\mathsf{err}} \, \varepsilon$, that can be stored using $c_{\mathsf{sto}} \, \varepsilon^{-\frac{d}{\beta}} \log \varepsilon^{-1}$ bits, and that can be queried in time $c_{\mathsf{que}} \log \varepsilon^{-1}$.*

Theorem 2 is useful when it is possible to design a procedure for estimating a smooth function f pointwise, but that procedure cannot necessarily be carried out efficiently per query. For example in nonparametric regression, Nadaraya–Watson estimators are known to be accurate pointwise (Tsybakov, 2009) but naively require evaluation time that is linear in the number of data points. One can also imagine a numerical or experimental setting where it is only possible to gather a limited number of accurate measurements of a smooth response, and one wants to graph the underlying function efficiently and accurately over the entire domain.

1.3 Setup and notation

Fix an integer $d \geq 1$. For any multi-index $s = (s_1, \ldots, s_d) \in \mathbb{Z}^d_{\geq 0}$, let $|s| = s_1 + \cdots + s_d$ and for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, define $s! = s_1! \cdots s_d!$ and $x^s = x_1^{s_1} \cdots x_d^{s_d}$. Let D^s denote the differential operator

$$D^s = \frac{\partial^{|s|}}{\partial x_1^{s_1} \cdots \partial x_d^{s_d}}.$$

Fix a positive real number β, and let $\lfloor \beta \rfloor$ denote the maximal integer *strictly* less than β. We reserve the notation $\|\cdot\|_p$ for the L^p norm and $|\cdot|_p$ for the ℓ^p norm.

Given $L > 0$ we let $\mathcal{H}(\beta, L)$ denote the space of Hölder functions $f : \mathbb{R}^d \to \mathbb{R}$ that are supported on the cube $[0,1]^d$, are $\lfloor \beta \rfloor$ times differentiable, and satisfy

$$|D^s f(x) - D^s f(y)| \leq L \, |x - y|_2^{\beta - \lfloor \beta \rfloor},$$

for all $x, y \in \mathbb{R}^d$ and for all multi-indices s such that $|s| = \lfloor \beta \rfloor$.

Let $\mathcal{P}_{\mathcal{H}}(\beta, L)$ denote the set of probability density functions contained in $\mathcal{H}(\beta, L)$. For $f \in \mathcal{P}_{\mathcal{H}}(\beta, L)$, let $\mathbb{P}_f$ (resp. $\mathbb{E}_f$) denote the probability distribution (resp. expectation) associated to f.

The parameter L will be fixed in what follows, so typically we write $\mathcal{P}_{\mathcal{H}}(\beta) := \mathcal{P}_{\mathcal{H}}(\beta, L)$. The constants $c, c_{\beta,d}, c_L$, etc. vary from line to line and their subscripts indicate parameter dependences.

2 EFFICIENT INTERPOLATION OF DENSITY ESTIMATORS

The important implication of Assumption 1 is that we can query $\hat{f}$ at a polynomial number of data points such that for each query y, $\hat{f}(y) \approx f(y)$, where f is the unknown density.

Lemma 1. *Let $A > 0$ and set $N = \Delta n^A$ with $\Delta \geq 1$. Let $y_1, \ldots, y_N \subset [0,1]^d$ denote a set of points. Then with probability at least $1 - n^{-2}$,*

$$\left|\hat{f}(y_i) - f(y_i)\right| \leq \sqrt{\log(2\Delta n^{A+2})}\, \varepsilon$$

for all $1 \leq i \leq N$, where $\varepsilon = c^* n^{-\beta/(2\beta+d)}$ is the minimax rate.

Proof. Set $t = \sqrt{\log(2\Delta n^{A+2})}\varepsilon \geq \varepsilon$ and apply Assumption 1 to y_i. Then by the union bound,

$$\mathbb{P}\left[\exists y_i \,:\, \left|\hat{f}(y_i) - f(y_i)\right| > t\right] \leq 2\Delta n^A e^{-\frac{t^2}{\varepsilon^2}} \leq n^{-2}.$$

□

We now describe our construction of $\tilde{f}$. Define $\ell := \lfloor \beta \rfloor$ and $M = \binom{\ell+d}{\ell}$.

Construction of $\tilde{f}$ (*informal*):

1. PARTITION: Divide $[0,1]^d$ into h^{-d} sub-cubes $\{I_{\vec{j}}\} \subset [0,1]^d$ of side-length $h = n^{-1/(2\beta+d)}$ where $\vec{j} \in \mathbb{Z}^d_{\geq 0}$ and $I_{\vec{j}} := [0,h]^d + h\vec{j}$.
2. MESH: For each $\vec{j}$, construct a mesh consisting of $M = \binom{\ell+d}{\ell}$ points $U_1^{\vec{j}}, \ldots, U_M^{\vec{j}} \in I_{\vec{j}}$.
3. INTERPOLATE: In each sub-cube $I_{\vec{j}}$, construct a multivariate polynomial interpolant $\hat{q}_{\vec{j}}$ on the M points $(U_1^{\vec{j}}, \hat{f}(U_1^{\vec{j}})), \ldots, (U_M^{\vec{j}}, \hat{f}(U_M^{\vec{j}}))$.

Return: $\tilde{f} : [0,1]^d \to \mathbb{R}$ defined by

$$\tilde{f}(y) = \sum_{\vec{j}} \hat{q}_{\vec{j}}(y) \mathbb{1}(y \in I_{\vec{j}}).$$

We first give some intuition for why $\tilde{f}$ is an accurate estimator. On each sub-cube $I_{\vec{j}}$, the true density $f \in \mathcal{P}_{\mathcal{H}}(\beta, L)$ is approximated up to the minimax error by a polynomial $q_{\vec{j}}$ of degree at most ℓ by the properties of Hölder functions. Upon setting $\Delta = M$ and $A = d/(2\beta + d)$ in Lemma 1, this guarantees that for all points $U_k^{\vec{j}}$ in the mesh, $\hat{f}(U_k^{\vec{j}}) \approx f(U_k^{\vec{j}}) \approx q_{\vec{j}}(U_k^{\vec{j}})$ with high probability. By studying the stability of the resulting polynomial system of equations, we can show that this construction yields a good approximation to the 'true' interpolation polynomial $q_{\vec{j}}$ on the sub-cube $I_{\vec{j}}$. This argument, carried out formally later in this section, yields the estimation bound of Theorem 1.

Next, we comment on the remaining guarantees of Theorem 1. As we show later, there is an explicit formula for $\hat{q}_{\vec{j}}$, so the main preprocessing bottleneck is the evaluation of $\hat{f}$ on the $Mn^{d/(2\beta+d)}$ points in the mesh, which naively takes $QMn^{d/(2\beta+d)}$ time. For the space requirement, it suffices to store the values $\{\hat{f}(U_k^{\vec{j}})\}$ up to polynomial precision as well as the elements of the mesh. Querying $\hat{f}$ at a point $y \in [0,1]^d$ requires checking which sub-cube y belongs to by scanning its d coordinates and then evaluating $\hat{q}_{\vec{j}}(y)$, which is a d-variate polynomial of degree $\lfloor \beta \rfloor$. By a careful consideration of the numerical precision required to perform these steps in Section 2.2.2, we obtain the computational guarantees of Theorem 1.

2.1 Interpolation on the principal lattice

To construct our interpolant, we refer to the next definition and theorem which are classical in finite element analysis (Nicolaides, 1972; Chung & Yao, 1977). The lattice $\mathcal{P}_\ell \subset [0,1]^d$, dubbed the ℓ-th principle lattice, has the special property that every function defined on $\mathcal{P}_\ell$ admits a unique polynomial interpolant of degree at most ℓ. This property is known to be equivalent to a combinatorial geometric condition referred to as GC in Chung & Yao (1977). A set of points $\mathcal{P}$ is called GC if every point $x \in \mathcal{P}$ has an associated set $\mathcal{H}_x$ consisting of ℓ affine hyperplanes whose union contains $\mathcal{P} \backslash x$ and such that none of these hyperplanes contain x.

Definition 1 (ℓ-th principal lattice of Δ_d)**.** *Let $\Delta_d \subset [0,1]^d$ denote the simplex on the points $\{0\} \cup \{e_i\}_{i=1}^d \subset \mathbb{R}^d$, where e_i denotes the i-th standard basis vector in $\mathbb{R}^d$. Label the vertices of Δ_d to be $v_0 = 0, v_i = e_i$ for $1 \leq i \leq d$. For all $x \in \mathbb{R}^d$, there exists a unique vector $(\lambda_0(x), \ldots, \lambda_d(x))$ with entries summing to one such that*

$$x = \sum_{i=0}^{d} \lambda_i(x) v_i \,.$$

Let $\Lambda : \mathbb{R}^d \to \mathbb{R}^{d+1}$ denote the function such that $\Lambda(x) = (\lambda_0(x), \ldots, \lambda_d(x))$. For $\ell \geq 1$, the ℓ-th princi-

pal lattice $\mathcal{P}_\ell$ *of* Δ_d *is defined to be*

$$\mathcal{P}_\ell = \left\{ x \in \Delta_d : \ell\Lambda(x) \in \mathbb{Z}_{\geq 0}^{d+1} \right\}. \tag{3}$$

We also define $\mathcal{P}_0 = 0 \in \mathbb{R}^d$.

Given a point $x \in \mathcal{P}_\ell$, the associated set of affine hyperplanes satisfying the GC condition is

$$\mathcal{H}_x = \bigcup_{\substack{t=0 \\ \lambda_t(x)>0}}^{d} \bigcup_{r=0}^{\ell\lambda_t(x)-1} \left\{ \sum_{i=0}^{d} \alpha_i v_i \Big| \ \ell\alpha_t = r, \sum_{i=0}^{d} \alpha_i = 1 \right\}.$$

Given a set of hyperplanes satisfying this combinatorial condition, it is straightforward to write down a Lagrangian-type interpolation formula, as was first computed for the principal lattice by Nicolaides (1972).

Theorem 3 (Nicolaides (1972),Chung & Yao (1977))**.** *Write* $\mathcal{P}_\ell = \{U_1, \dots, U_M\} \subset \Delta_d$ *and let* $g : \mathcal{P}_\ell \to \mathbb{R}$ *denote a function defined on this lattice. For* $\ell \geq 1$*, define the polynomial*

$$p_i(x) = \prod_{\substack{t=0 \\ \lambda_t(U_i)>0}}^{d} \prod_{r=0}^{\ell\lambda_t(U_i)-1} \frac{\lambda_t(x) - \frac{r}{\ell}}{\lambda_t(U_i) - \frac{r}{\ell}}, \tag{4}$$

where we recall that $\lambda_t(x)$ *is from Definition 1. If* $\ell = 0$*, then* $M = 1$*, and we simply define* $p_1(x) \equiv 1$*. Then*

$$p(x) := \sum_{i=1}^{M} p_i(x) g(U_i)$$

satisfies $p(U_i) = g(U_i)$ *for all* $U_i \in \mathcal{P}_\ell$*. Moreover, this is the unique polynomial of degree at most* ℓ *with this property.*

Since $\lambda_t(x)$ is linear in $x \in \mathbb{R}^d$, it is easy to see that $p_i(x)$ is a polynomial of degree ℓ, and moreover $p_i(U_j) = 1$ if $i = j$ and zero otherwise.

We are now ready to give a precise description of the construction of $\tilde{f}$. The idea is to generate the mesh for interpolation using a shifted and rescaled version of the ℓ-th principal lattice on $\Delta_d \subset [0,1]^d$. Recall that $\hat{f}$ is a density estimator that satisfies Assumption 1.

Construction of $\tilde{f}$ (*formal version*):

1. Partition: Divide $[0,1]^d$ into h^{-d} sub-cubes $\{I_{\vec{j}}\} \subset [0,1]^d$ of side-length $h = n^{-1/(2\beta+d)}$ where $\vec{j} \in \mathbb{Z}_{\geq 0}^d$ and $I_{\vec{j}} := [0,h]^d + h\vec{j}$.
2. Mesh: For each $\vec{j}$, construct a mesh on $I_{\vec{j}}$ consisting of $M = \binom{\ell+d}{\ell}$ points given by the shifted and rescaled principal lattice $\mathcal{P}_\ell^{\vec{j}} := \{h(x + \vec{j}) : x \in \mathcal{P}_\ell\} \subset I_{\vec{j}}$. Let $U_1^{\vec{j}}, \dots, U_M^{\vec{j}}$ denote the points in $\mathcal{P}_\ell^{\vec{j}}$.
3. Interpolate: In each sub-cube $I_{\vec{j}}$, construct a multivariate polynomial interpolant $\hat{q}_{\vec{j}}$ through the M points $(U_1^{\vec{j}}, \hat{f}(U_1^{\vec{j}})), \dots, (U_M^{\vec{j}}, \hat{f}(U_M^{\vec{j}}))$ given by $\hat{q}_{\vec{j}}(y) = p_{\vec{j}}(y/h - \vec{j})$, where p is the polynomial interpolant from Theorem 3 given by

$$p_{\vec{j}}(x) = \sum_{k=1}^{M} p_k(x) \hat{f}(U_k^{\vec{j}}).$$

Return: $\tilde{f} : [0,1]^d \to \mathbb{R}$ defined by

$$\tilde{f}(y) = \sum_{\vec{j}} \hat{q}_{\vec{j}}(y) \mathbb{1}(y \in I_{\vec{j}}).$$

2.2 Proof of Theorem 1

We prove Theorem 1 in two parts, first by studying the estimation error $\|\tilde{f} - f\|_\infty$ in Section 2.2.1 and second by proving the storage and time complexity upper bounds in Section 2.2.2.

2.2.1 Estimation error

First, we quantify the error in the approximation of the values of $q_{\vec{j}}$ on the mesh points. Let $f_{z,\ell}$ denote the degree ℓ polynomial given by the Taylor expansion of $f \in \mathcal{P}_\mathcal{H}(\beta)$ at z. Since $f \in \mathcal{P}_\mathcal{H}(\beta)$, by a standard fact (see Lemma 5) it holds that

$$|f(y) - f_{z,\ell}(y)| \leq \frac{L d^{\ell/2}}{\ell!} |y - z|_2^\beta,$$

where $f_{z,\ell}$ is the degree-ℓ Taylor expansion of the function f at $z \in \mathbb{R}^d$.

For $\vec{j} \in \{0, \dots, h^{-1} - 1\}^d$, define $q_{\vec{j}} := f_{z_{\vec{j}},\ell}$, where $z_{\vec{j}}$ is the vertex of $I_{\vec{j}}$ closest to the origin. Then for all $y \in I_{\vec{j}}$, it holds that

$$\begin{aligned} |f(y) - q_{\vec{j}}(y)| &\leq \left(\frac{L d^\beta}{\ell!} \right) h^\beta \\ &= \left(\frac{L d^\beta}{\ell!} \right) n^{-\beta/(2\beta+d)} \\ &=: \hat{c} n^{-\beta/(2\beta+d)} \end{aligned} \tag{5}$$

Note that the right-hand side is the minimax rate of estimation in (2) up to constant factors.

Next, by Lemma 1 (setting $\Delta = M$ and $A = \frac{d}{2\beta+d}$) and (5) it holds with probability at least $1 - n^{-2}$ that

$$\begin{aligned} \left| q_{\vec{j}}(U_k^{\vec{j}}) - \hat{f}(U_k^{\vec{j}}) \right| &\leq (c^* \sqrt{4 \log 2M} + \hat{c})(\log n)^{\frac{1}{2}} n^{-\frac{\beta}{2\beta+d}} \\ &=: \breve{c} (\log n)^{\frac{1}{2}} n^{-\frac{\beta}{2\beta+d}} \end{aligned} \tag{6}$$

for all $\vec{j} \in \{0, \dots, h^{-1}-1\}^d$ and $k \in [M]$. Using this fact, we can show that the polynomial interpolant built on $\{(U_k^{\vec{j}}, \hat{f}(U_k^{\vec{j}}))\}_{k=1}^M$ provides a good approximation for $q_{\vec{j}}$ on the interval $I_{\vec{j}}$, which is our next task. The following lemma establishes stability of the polynomial approximation.

Lemma 2. *Let $\hat{q}_{\vec{j}}$ denote the unique polynomial of degree at most ℓ that passes through the points $\{(U_k^{\vec{j}}, \hat{f}(U_k^{\vec{j}}))\}_{k=1}^M$. Then with probability at least $1 - n^{-2}$, for all $\vec{j}$ and all $x \in I_{\vec{j}}$,*

$$\left| q_{\vec{j}}(x) - \hat{q}_{\vec{j}}(x) \right| \le c_{\beta,d,L} (\log n)^{\frac{1}{2}} n^{-\frac{\beta}{2\beta+d}}. \tag{7}$$

Proof. Define $\hat{g}_{\vec{j}}(x) = \hat{q}_{\vec{j}}(h(x+\vec{j}))$ and $g_{\vec{j}}(x) = q_{\vec{j}}(h(x+\vec{j}))$ to be polynomials restricted to the domain $[0,1]^d$. Recall that $\hat{g}$ and g are given by formulas as in Theorem 3. It holds by (6) that for all $1 \le k \le M$,

$$\left| \hat{g}(U_k^{\vec{j}}) - g(U_k^{\vec{j}}) \right| \le \check{c} (\log n)^{1/2} n^{-\frac{\beta}{2\beta+d}}.$$

Let $y \in [0,1]^d$, and observe that by Theorem 3 and the triangle inequality,

$$\begin{aligned} |\hat{g}(y) - g(y)| &\le M \sup_{\substack{x \in [0,1]^d \\ 1 \le k \le M}} \left| p_k(x) \left(\hat{g}(U_k^{\vec{j}}) - g(U_k^{\vec{j}}) \right) \right| \\ &\le M \check{c} (\log n)^{1/2} n^{-\frac{\beta}{2\beta+d}} \sup_{\substack{x \in [0,1]^d \\ 1 \le k \le M}} |p_k(x)|. \end{aligned} \tag{8}$$

Observe that for $x \in [0,1]^d$, we have $|\lambda_0(x)| = |1 - \sum x_i| \le d$, and for $1 \le t \le d$, we have $|\lambda_t(x)| = |x_i| \le 1$. Therefore, by the definition of p_k and $U_k^{\vec{j}}$,

$$|p_k(x)| \le \ell^\ell d.$$

By this bound, (8), and translation and scale invariance of $\|\cdot\|_\infty$, Lemma 2 follows with $c_{\beta,d,L} = \check{c} M d \ell^\ell$. □

Define $\tilde{f}(x) = \sum_{\vec{j}} \hat{q}_{\vec{j}}(x) \mathbb{1}(x \in I_{\vec{j}})$, and observe that Theorem 1 follows from (5), Lemma 2, and the triangle inequality. Though we have derived a high probability bound, the expectation claimed in Theorem 1 follows using the uniform boundedness of Hölder functions as stated in Lemma 4. Tracing constants above yields the expression for c_{err}.

2.2.2 Time and space requirements

Recall that $M = \binom{\ell+d}{\ell}$ where $\ell = \lfloor \beta \rfloor$. For the space requirement, we store the principal lattices and the values of $\hat{f}$ on these lattice points, and note that each query is at most $L d^{O(\beta+1)}$ in magnitude by Lemma 4. The queries per sub-cube can thus be stored with $M(\log L d^{O(\beta+1)} + \log n)$ bits. The extra $\log n$ bits are required so that the interpolating polynomials can be queried with sufficient precision. The lattices are composed of rational points in $\mathbb{R}^d$, so we need at most $Md \log(\beta+1)$ bits per sub-cube to store them. Since there are $n^{\frac{d}{2\beta+d}}$ sub-cubes, the space requirement of Theorem 1 follows and is a conservative estimate for simplicity.

Next we characterize the time complexity. Assume first that $\ell \ge 1$. For $1 \le k \le M$, it holds that

$$|p_k(y) - p_k(y')| \le (d+1) 2^\ell \ell^{\ell+1} |y - y'|_\infty$$

because by expanding the product in the formula in Theorem 3, p_i is a sum of at most $2^\ell(d+1)$ terms, each having coefficients of size at most ℓ^ℓ, and moreover for $|\alpha| \le \ell$, the monomial y^α is ℓ-Lipschitz with respect to $|\cdot|_\infty$ over the cube. Therefore, it also holds that

$$\left| \hat{q}_{\vec{j}}(y) - \hat{q}_{\vec{j}}(y') \right| \le M L d^{\frac{3}{2}\beta + \frac{1}{2}} (d+1) 2^\ell \ell^{\ell+1} |y - y'|_\infty$$

by the formula in the interpolation step of $\tilde{f}$, noting that without loss of generality, $\left| \hat{f}(U_k^{\vec{j}}) \right| = L d^{O(\beta+1)}$ by Lemma 4. By the form of c_{err}, given a query y it suffices to round its coordinates to $B := \ell + \log d + \log n$ bits to compute $\hat{q}_{\vec{j}}(y)$ with the required level of accuracy.

Next, the number of arithmetic operations needed to evaluate $\hat{q}_{\vec{j}}(y)$ is bounded conservatively by $6(d+\ell)M$. To identify which sub-cube contains y requires time at most $2d \log n$. Hence, the total complexity is upper bounded by

$$6(d+\ell)MB + 2d \log n \le 16(d+\ell)^2 M \log n =: c_{\mathsf{que}}$$

This bound also holds conservatively when $\ell = 0$ since in that case, to evaluate $\hat{f}(y)$, we just need to match the given query y to the sub-cube $I_{\vec{j}}$ containing it and output $\hat{f}(U_1^{\vec{j}})$.

3 A RESULT OF KOLMOGOROV AND TIKHOMIROV

Given a function class $\mathcal{F}$, let $N(\mathcal{F}, \delta)$ denote the minimal number of L^∞ balls of radius δ that cover $\mathcal{F}$, and define $H(\mathcal{F}, \delta) = \log N(\mathcal{F}, \delta)$ to be the metric entropy. Let $\mathcal{H}(\beta) = \mathcal{H}(\beta, L)$ denote the class of Hölder functions supported on $[0,1]^d$ as defined in Section 1.3. A classical result of Kolmogorov & Tikhomirov (1993) shows that

$$H(\mathcal{H}(\beta), \delta) \le c_{\beta,d,L}\, \delta^{-\frac{d}{\beta}}. \tag{9}$$

Their proof strategy is conceptually similar to our piecewise multivariate polynomial approximation scheme in that they subdivide the cube as we do here, approximate f by its Taylor polynomial in each cube, and then discretize the coefficients. We show now that our techniques imply a slightly weaker version of the bound (9).

Define a mesh as in steps 1 and 2 of our formal construction of $\tilde{f}$ as in Section 2.1, but now for a general parameter $h > 0$ to be set later. This mesh has Mh^{-d} points that we denote by $\{U_k^{\vec{j}}\}_{\vec{j},k}$. Let $f, g \in \mathcal{H}(\beta)$ be such that for all $\vec{j}, k$ it holds that

$$\left| f(U_k^{\vec{j}}) - g(U_k^{\vec{j}}) \right| \le h^\beta.$$

By the Hölder condition and Lemma 5, there exists a degree $\ell = \lfloor \beta \rfloor$ polynomial $q_{\vec{j}}$ approximating f in $I_{\vec{j}}$ and a degree $\ell = \lfloor \beta \rfloor$ polynomial $r_{\vec{j}}$ approximating g in $I_{\vec{j}}$, each with error h^β pointwise. We conclude that

$$\left| q_{\vec{j}}(U_k^{\vec{j}}) - r_{\vec{j}}(U_k^{\vec{j}}) \right| \le c_{\beta,d,L}\, h^\beta$$

for all $\vec{j}, k$. Following the proof of Lemma 2, this implies that for all $x \in I_{\vec{j}}$,

$$\left| q_{\vec{j}}(x) - r_{\vec{j}}(x) \right| \le c_{\beta,d,L}\, h^\beta.$$

Hence we conclude that for all $x \in [0,1]^d$,

$$|f(x) - g(x)| \le c_{\beta,d,L}\, h^\beta.$$

The Hölder functions are uniformly bounded by some constant $c_{\beta,d,L}$ (see Lemma 4). Hence setting $\delta = c_{\beta,d,L}\, h^\beta$ and rounding the values of each function at each point $U_k^{\vec{j}}$ to multiples of h^β, we see that there exists a δ-net of size at most

$$\left(\frac{c_{\beta,d,L}}{\delta}\right)^{M c'_{\beta,d,L} \delta^{-d/\beta}}.$$

Therefore

$$H(\mathcal{H}(\beta), \delta) \le c_{\beta,d,L}\, \delta^{-\frac{d}{\beta}} \log \frac{1}{\delta},$$

a mildly weaker bound than (9).

4 KDEs satisfy Assumption 1

In this section, for completeness we verify that for appropriate kernels, the standard KDE satisfies Assumption 1.

Proposition 1. *Let $K(\cdot)$ denote a kernel of order $\lfloor \beta \rfloor$ satisfying*

$$\|K\|_\infty < \infty, \ \int K^2(x)\mathrm{d}x < \infty, \ \int |x^\alpha K(x)|\,\mathrm{d}x < \infty$$

for all multi-indices $\alpha \in \mathbb{R}^d_{\ge 0}$ with $|\alpha| = \beta$. Then Assumption 1 is satisfied for the KDE $\hat{f}$ with bandwidth $h = cn^{-1/(2\beta+d)}$.

Proof. For brevity, c denotes a constant that varies from line to line and can depend on β, d, L and K. Fix $y \in [0,1]^d$. It is well-known that under the conditions of Proposition 1 (see e.g. Tsybakov, 2009),

$$b = b(y) := \left| \mathbb{E} f(y) - \hat{f}(y) \right| \le ch^\beta,$$

and for a data point $X_i \sim \mathbb{P}_f$,

$$\tau^2 = \tau^2(y) := \mathsf{Var}\, K_h(X_i - y) \le \frac{c}{h^d}.$$

By the triangle inequality and Bernstein's inequality for bounded random variables (Vershynin, 2018),

$$\Pr\left(\left| \hat{f}(y) - f(y) \right| > t \right) \le \exp\left(-\frac{n(t-b)^2}{2\tau^2 + 2\|K_h\|_\infty (t-b)/3} \right). \quad (10)$$

Let $h = cn^{-1/(2\beta+d)}$. Note that $\|K_h\|_\infty = h^{-d}\|K\|_\infty$ and $(nh^d)^{-1} = cn^{-\beta/(2\beta+d)}$. Then we recover Assumption 1 by setting $t \ge cn^{-\beta/(2\beta+d)}$ in (10).

□

5 PROPERTIES OF HÖLDER FUNCTIONS

For completeness, we provide proofs of standard facts about the class of Hölder functions.

Lemma 3 (Inclusion)**.** *Let $\mathcal{H}(\beta, d, L)$ denote the class of Hölder functions supported on $[0,1]^d$ in dimension d. If $\beta > 1$, then it holds that $\mathcal{H}(\lfloor \beta \rfloor, d, L) \subset \mathcal{H}(\lfloor \beta \rfloor - 1, d, d^{3/2}L)$.*

Proof. Let $f \in \mathcal{H}(\beta, d, L)$. Since f is supported on $[0,1]^d$ and smooth on $\mathbb{R}^d$, we have that

$$|D^s f(x)| \le L\,|x|_2 \le L\sqrt{d} \quad (11)$$

for all $|s| = \lfloor \beta \rfloor$.

Fix $x, y \in [0,1]^d$, and define for $1 \le i \le d+1$ the point $z^i \in [0,1]^d$ to be

$$z^i_j = \begin{cases} x_j & \text{if } j \ge i \\ y_j & \text{if } j < i. \end{cases}$$

Observe that $z^1 = x$ and $z^{d+1} = y$.

Let t denote a multi-index with $|t| = \lfloor\beta\rfloor - 1$. By the fundamental theorem of calculus and the Hölder condition,

$$|D^t f(x) - D^t f(y)| \le \sum_{i=1}^{d} \left|D^t f(z^i) - D^t f(z^{i+1})\right|$$
$$= \sum_{i=1}^{d} \left| \int_{x_i}^{y_i} \frac{\partial}{\partial x_i} D^t f(x_1, \ldots, z, y_{i+1}, ..., y_d)\, \mathrm{d}z \right|.$$

Using (11), the expression in the second line is bounded above by $Ld^{3/2}$, which proves the lemma.

□

Lemma 4 (Uniform boundedness)**.** *The class $\mathcal{H}(\beta)$ is uniformly bounded. In particular,*

$$\sup_{f\in\mathcal{H}(\beta)} \|f\|_\infty \le d^{3\lfloor\beta\rfloor/2+1/2}\, L.$$

Proof. Suppose first that $f \in \mathcal{H}(\beta)$ for $\beta > 1$. By repeated application of Lemma 3, f is $(d^{3\lfloor\beta\rfloor/2}L)$-Lipschitz. Since f is supported on $[0,1]^d$,

$$|f(x)| = |f(x) - f(0)| \le d^{3\lfloor\beta\rfloor/2} L\, |x|_2 \le d^{3\lfloor\beta\rfloor/2+1/2} L.$$

If $\beta \le 1$, then arguing as in the previous display, we see that $|f(x)| \le L\sqrt{d}$ for all $x \in \mathbb{R}^d$. □

Lemma 5 (Taylor approximation)**.** *Given $f \in \mathcal{H}(\beta)$, let $f_{x,\lfloor\beta\rfloor}$ denote its Taylor polynomial of degree $\lfloor\beta\rfloor$ at a point $x \in \mathbb{R}^d$,*

$$f_{x,\lfloor\beta\rfloor}(y) = \sum_{|s|\le\lfloor\beta\rfloor} \frac{(y-x)^s}{s!} D^s f(x)\,, \quad y \in \mathbb{R}^d\,.$$

Then it holds that

$$\left|f(y) - f_{x,\lfloor\beta\rfloor}(y)\right| \le \frac{Ld^{\lfloor\beta\rfloor/2}}{\lfloor\beta\rfloor!} |x-y|_2^\beta\,, \quad x, y \in \mathbb{R}^d\,.$$

Proof. By Taylor's theorem with remainder (see, eg., Folland, 1999)

$$\left|f(y) - f_{x,\lfloor\beta\rfloor}(y)\right| =$$
$$\left| \sum_{|s|=\lfloor\beta\rfloor} \frac{1}{s!} \left[D^s f(x + c(y-x)) - D^s f(x)\right] (y-x)^s \right|$$

for some constant $c \in (0,1)$. By the triangle inequality and the Hölder condition, the expression in the second line is bounded above by

$$\sum_{|s|=\lfloor\beta\rfloor} \frac{L\,|x-y|_2^{\beta-\lfloor\beta\rfloor}}{s!} |(y-x)^s| =$$
$$\frac{L\,|x-y|_2^{\beta-\lfloor\beta\rfloor}}{\lfloor\beta\rfloor!} \left(\sum_{i=1}^{d} |x_i - y_i|\right)^{\lfloor\beta\rfloor},$$

where the equality is by the multinomial theorem. In turn, this last expression is bounded above by

$$\frac{Ld^{\lfloor\beta\rfloor/2}}{\lfloor\beta\rfloor!} |x-y|_2^\beta$$

using Cauchy–Schwarz. □

Acknowledgments

We thank the anonymous reviewers for their many helpful comments and suggestions. Philippe Rigollet was supported by NSF awards IIS-1838071, DMS-1712596, DMS-1740751, and DMS-2022448.

References

Acharya, Jayadev, Diakonikolas, Ilias, Li, Jerry, & Schmidt, Ludwig. 2017. Sample-optimal density estimation in nearly-linear time. *Pages 1278–1289 of: Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM.

Agarwal, Pankaj K, Har-Peled, Sariel, & Varadarajan, Kasturi R. 2005. Geometric approximation via coresets. *Combinatorial and computational geometry*, **52**, 1–30.

Backurs, Arturs, Charikar, Moses, Indyk, Piotr, & Siminelakis, Paris. 2018. Efficient density evaluation for smooth kernels. *Pages 615–626 of: 2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE.

Backurs, Arturs, Indyk, Piotr, & Wagner, Tal. 2019. Space and Time Efficient Kernel Density Estimation in High Dimensions. *Pages 15799–15808 of: Advances in Neural Information Processing Systems*.

Bansal, Nikhil, Dadush, Daniel, Garg, Shashwat, & Lovett, Shachar. 2018. The gram-schmidt walk: a cure for the Banaszczyk blues. *Pages 587–597 of: Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2018, Los Angeles, CA, USA, June 25-29, 2018*.

Belkin, Mikhail, Rakhlin, Alexander, & Tsybakov, Alexandre B. 2019. Does data interpolation contradict statistical optimality? *Pages 1611–1619 of: The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR.

Chan, Siu-On, Diakonikolas, Ilias, Servedio, Rocco A, & Sun, Xiaorui. 2014. Efficient density estimation via piecewise polynomial approximation. *Pages 604–613 of: Proceedings of the forty-sixth annual ACM symposium on Theory of computing*.

Charikar, Moses, & Siminelakis, Paris. 2017. Hashing-based-estimators for kernel density in high dimensions. *Pages 1032–1043 of: 2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE.

Chazelle, B. 2000. *The Discrepancy Method: Randomness and Complexity*. Cambridge: Cambridge University Press.

Chung, K. C., & Yao, T. H. 1977. On Lattices Admitting Unique Lagrange Interpolations. *SIAM Journal on Numerical Analysis*, **14**(4), 735–743.

Clarkson, Kenneth L. 2010. Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. *ACM Transactions on Algorithms (TALG)*, **6**(4), 1–30.

Coleman, Benjamin, & Shrivastava, Anshumali. 2020. Sub-linear race sketches for approximate kernel density estimation on streaming data. *Pages 1739–1749 of: Proceedings of The Web Conference 2020*.

Folland, Gerald B. 1999. *Real analysis: modern techniques and their applications*. Vol. 40. John Wiley & Sons.

Greengard, Leslie, & Rokhlin, Vladimir. 1987. A fast algorithm for particle simulations. *Journal of computational physics*, **73**(2), 325–348.

Greengard, Leslie, & Strain, John. 1991. The fast Gauss transform. *SIAM Journal on Scientific and Statistical Computing*, **12**(1), 79–94.

Györfi, László, Kohler, Michael, Krzyzak, Adam, & Walk, Harro. 2006. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

Hao, Yi, Jain, Ayush, Orlitsky, Alon, & Ravindrakumar, Vaishakh. 2020. SURF: A Simple, Universal, Robust, Fast Distribution Learning Algorithm. *arXiv preprint arXiv:2002.09589*.

Jones, M Chris. 1989. Discretized and interpolated kernel density estimates. *Journal of the American Statistical Association*, **84**(407), 733–741.

Kogure, Atsuyuki. 1998. Effective interpolations for kernel density estimators. *Journal of Nonparametric Statistics*, **9**(2), 165–195.

Kolmogorov, A. N., & Tikhomirov, V. M. 1993. *ε-Entropy and ε-Capacity of Sets In Functional Spaces*. Dordrecht: Springer Netherlands. Pages 86–170.

Lee, Dongryeol, Moore, Andrew W, & Gray, Alexander G. 2006. Dual-tree fast Gauss transforms. *Pages 747–754 of: Advances in Neural Information Processing Systems*.

Liang, Tengyuan, Rakhlin, Alexander, *et al.* 2020. Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, **48**(3), 1329–1347.

Matoušek, J. 1999. *Geometric Discrepancy: an Illustrated Guide*. New York: Springer.

Nicolaides, R. A. 1972. On a Class of Finite Elements Generated by Lagrange Interpolation. *SIAM Journal on Numerical Analysis*, **9**(3), 435–445.

Phillips, Jeff M., & Tai, Wai Ming. 2018a. Improved coresets for kernel density estimates. *Pages 2718–2727 of: Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM.

Phillips, Jeff M., & Tai, Wai Ming. 2018b. Near-Optimal Coresets of Kernel Density Estimates. *Pages 66:1–66:13 of: 34th International Symposium on Computational Geometry, SoCG 2018, June 11-14, 2018, Budapest, Hungary*.

Scott, David W, & Sheather, Simon J. 1985. Kernel density estimation with binned data. *Communications in Statistics-Theory and Methods*, **14**(6), 1353–1359.

Siminelakis, Paris, Rong, Kexin, Bailis, Peter, Charikar, Moses, & Levis, Philip. 2019. Rehashing kernel evaluation in high dimensions. *Pages 5789–5798 of: International Conference on Machine Learning*.

Tsybakov, Alexandre B. 2009. *Introduction to Nonparametric Estimation*. Springer series in statistics. Springer.

Vershynin, Roman. 2018. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press.

Yang, Changjiang, Duraiswami, Ramani, Gumerov, Nail A, & Davis, Larry. 2003. Improved Fast Gauss Transform and Efficient Kernel Density Estimation. *Page 464 of: Proceedings of the Ninth IEEE International Conference on Computer Vision-Volume 2*.