# Hashing-Based-Estimators for Kernel Density in High Dimensions* 

Moses Charikar<br>Stanford University<br>Department of Computer Science<br>moses@cs.stanford.edu

Paris Siminelakis<br>Stanford University<br>Department of Electrical Engineering<br>psimin@stanford.edu

September 3, 2018


#### Abstract

Given a set of points $P \subset \mathbb{R}^{d}$ and a kernel $k$, the Kernel Density Estimate at a point $x \in \mathbb{R}^{d}$ is defined as $\operatorname{KDE}_{P}(x)=\frac{1}{|P|} \sum_{y \in P} k(x, y)$. We study the problem of designing a data structure that given a data set $P$ and a kernel function, returns approximations to the kernel density of a query point in sublinear time. We introduce a class of unbiased estimators for kernel density implemented through locality-sensitive hashing, and give general theorems bounding the variance of such estimators. These estimators give rise to efficient data structures for estimating the kernel density in high dimensions for a variety of commonly used kernels. Our work is the first to provide data-structures with theoretical guarantees that improve upon simple random sampling in high dimensions.


[^0]
## Contents

1 Introduction ..... 1
1.1 Kernel Density Estimation ..... 1
1.2 Our Contribution ..... 2
1.2.1 Hashing-Based-Estimators (HBE) ..... 2
1.2.2 Kernel Density Estimation through Locality Sensitive Hashing ..... 3
1.2.3 Lower Bounds for KDE problem ..... 4
1.3 Related Work ..... 4
1.4 Open questions ..... 5
1.5 Outline of the paper ..... 5
2 Preliminaries ..... 5
2.1 $V$-bounded Estimators and Median-of-Means ..... 6
2.2 Bounding the Variance ..... 6
2.2.1 Baseline estimators ..... 6
2.2.2 Upper bounds via Hölder-type inequalities ..... 7
3 Importance Sampling through Hashing Based Estimators ..... 7
3.1 Moments of HBE ..... 7
3.2 Variance Bounds for general HBE ..... 9
3.3 Variance Bounds for Scale-free HBE ..... 10
4 Adaptive Estimation through Mean Relaxation ..... 11
4.1 Proof of Theorem 4.1 ..... 12
4.1.1 Case $\mu \geqslant \tau$ ..... 12
4.1.2 Case $\mu<\tau$ ..... 13
4.2 Proof of Lemma 4.3 ..... 13
5 KDE through Euclidean LSH ..... 14
5.1 Euclidean LSH ..... 14
5.2 Scale-free Estimators for Exponential and $t$-Student Kernel ..... 14
5.3 An estimator for Gaussian KDE ..... 15
5.4 Pointwise Bounds for Euclidean LSH ..... 17
6 Scale-free estimator based on Andoni-Indyk LSH ..... 17
6.1 Ball-Carving LSH ..... 17
6.2 Scale-free estimator for Gaussian Kernel ..... 18
6.3 Pointwise bounds for Andoni-Indyk LSH ..... 18
6.3.1 Upper bound ..... 19
6.3.2 Lower Bound ..... 20
6.3.3 Concentration of $\chi^{2}$-random variables ..... 21
7 Fast Kernel-Matrix Vector Multiplication ..... 22
8 Lower Bound for Kernel Density Estimation ..... 23
9 Hölder Inequalities ..... 26
A Chernoff Bounds ..... 32
B Random Fourier Features Estimator ..... 32

## 1 Introduction

A fundamental question in Statistics and Learning Theory is the following: given a set of points $P \subset \mathbb{R}^{d}$ sampled from some unknown distribution $\mathcal{D}$ estimate the probability at an arbitrary point $x \in \mathbb{R}^{d}$. This problem is known as density estimation and different ways to formalize it lead to very different statistical and computational tasks. In the past two decades the problem has attracted significant interest in theoretical computer science [12, 1]. Some of the most important problems that have been studied are learning discrete distributions [26], learning mixture models [41], and more recently the topic of robust estimation in high dimensions [17, 27]. In this paper, we focus on Kernel Density Estimation (KDE), one of the most widely developed methods in non-parametric estimation.

### 1.1 Kernel Density Estimation

In this approach, given a set of $n$ points $P$, starting from the empirical distribution $\tilde{\mu}(x)=\frac{1}{n} \sum_{y \in P} \delta_{y}$, one obtains a smooth distribution by "convolving" it with a kernel function $k$, whose smoothness is typically controlled by a parameter $\sigma>0$ called the bandwidth.

Definition 1.1 (Kernel Density). Given a kernel function $k_{\sigma}: \mathbb{R}^{d} \times \mathbb{R}^{d} \rightarrow[0,1]$ and a dataset $P \subset \mathbb{R}^{d}$ we define the Kernel Density (KD) of $P$ at a point $x \in \mathbb{R}^{d}$ as:

$$
\begin{equation*}
\operatorname{KDE}_{P}(x):=\frac{1}{|P|} \sum_{y \in P} k_{\sigma}(x, y) \tag{1.1}
\end{equation*}
$$

This is a natural way to extend the function smoothly from a discrete set of points to the whole space that is independent of any particular parametric assumption on the underlying distribution of the data. Selecting the kernel and bandwidth are intensively studied subjects in the literature of non-parametric estimation [16] for which there is still ongoing theoretical research [21]. The kernel function $k(x, y)$ is typically a function of only $x-y$ (shift invariant kernels) or just the euclidean distance $\|x-y\|$ (radial kernels). One of the most prominent functions is the Gaussian kernel

$$
\begin{equation*}
k_{\sigma}(x, y)=\exp \left(-\frac{\|x-y\|^{2}}{\sigma^{2}}\right) \tag{1.2}
\end{equation*}
$$

The importance of KDE lies in that it gives a simple and general way of approximating the underlying probability distribution that can be subsequently used to perform more complex and computationally intensive tasks. Examples include mode estimation [7], outlier detection [38], local regression [19], reproducing kernel Hilbert spaces [37], density based clustering [36], and topological data analysis [24, 14]. Kernel Density Estimation is consequently an important primitive that is a building block in many applications.

In all of the above settings, at some point, the following problem is solved: given $P \subset \mathbb{R}^{d}$, $z \in \mathbb{R}^{n}$, compute $\operatorname{KDE}_{P}^{z}(x):=\sum_{i=1}^{n} k\left(x, y_{i}\right) z_{i}$. This can be computed exactly in linear time, but this is prohibitively slow for large data sets, especially since it is needed repeatedly in applications of interest.

The problem has been studied extensively in the batch setting, where given a set of $n$ points, the goal is to compute, for each of the points, a sum of contributions due to all the points, i.e. $n$ queries of the above form. Such computations are prevalent in the field of scientific computing and involve computing approximations to $y=K z$ where $K$ is an $n \times n$ kernel matrix. In low dimensions, Fast Multipole Methods (combining hierarchical space partitions with Taylor approximations) were developed [28] to reduce the trivial $O\left(n^{2}\right)$ runtime to $O(n \log n)$ (and $O(n)$ in some cases). The fast multipole algorithm [22] has been enormously influential in numerical analysis and scientific computing; it was named as one of the top 10 algorithms of the 20th century by the editors of

Computing in Science and Engineering [18]. For this work, Greengard and Rokhlin received the 2001 Steele Prize. The KDE problem thus lies in the core of both scientific computing as well as machine learning.

In this paper, we study the problem of approximately computing the KDE. For most of the paper we fix $k_{\sigma}$ to be the Gaussian Kernel and define the following computational problem:

Definition 1.2 (KDE Problem). Given a dataset $P \subset \mathbb{R}^{d}$ of $n$ points, and $\varepsilon, \delta, \tau \in[0,1]$ construct a data structure that given a query $x \in \mathbb{R}^{d}$ with $\operatorname{KDE}_{P}(x)=\mu \in[\tau, 1]$ returns a number $\hat{\mu}$ such that $\mathbb{P}[|\hat{\mu}-\mu| \geqslant \varepsilon \cdot \mu] \leqslant \delta$. We call this problem the $(\mu, \varepsilon, \delta)$-KDE problem.

### 1.2 Our Contribution

The starting point of our work is the old and tested idea of importance sampling. Given non-negative weights $w_{1}, \ldots, w_{n}$ and a distribution $Q$ over $\left[n\right.$ ] (inducing probabilities $q_{1}, \ldots, q_{n}$ ), an unbiased estimator for $\mu:=\frac{1}{n} \sum_{i=1}^{n} w_{i}$ is given by sampling $I \in[n]$ according to $Q$ and returning $Z=w_{I} /\left(q_{I} n\right)$. The minimum variance estimator is obtained by setting $q_{i}^{*}=w_{i} / \sum_{j} w_{j}$ for which the variance is zero. What precludes us from obtaining such probabilities is that in our setting of $K D E$, the weights $w_{i}=w_{i}(x):=k\left(x, x_{i}\right)$ depend on the query $x$, and thus the sampling distribution $Q$ needs to be adaptive to the query. Furthermore, having an ideal distribution $Q^{*}$ indirectly involves knowing the normalizing constant $\sum_{i} w_{i}=n \mu$, the very quantity we wish to estimate. Thus, the main challenge in turning the idea of importance sampling into an algorithm is to have an efficient way to define an adaptive sampling distribution $Q(x)$ that has low variance. We next present our methods at an abstract level and subsequently show their implications for Kernel Density Estimation.

### 1.2.1 Hashing-Based-Estimators (HBE)

Our main contribution is to introduce a hashing-based framework to succinctly define an adaptive distribution $Q(x)$ and to provide sharp tools bounding the variance of the resulting unbiased estimator. Our estimators are formed by:

- Preprocessing: given a hash function $h$ sampled from a hash family $\mathcal{H}$, where the colision probability of $x, y$ is $p(x, y)$, we evaluate $h$ on $x_{1}, \ldots, x_{n}$ and form the corresponding hash table $H$.
- Querying: let $H(x) \subseteq P$ denote the cell where the query $x$ falls into and let $y \in P$ be a random element of $H(x)$, we return $Z_{h}(x)=\frac{1}{n} \frac{k(x, y)}{p(x, y)}|H(x)|$ (or 0 if $H(x)$ is empty).

We say that a HBE has complexity $T$ if the evaluation time (resp. space usage/preprocessing) is bounded by $T$ (resp. $T \cdot n$ ). This two-level sampling procedure induces probabilities that depend both on $P$ and $\mathcal{H}$. Although $|H(x)|$ is a priori a random variable, it becomes known to us through the preprocessing step. This sidesteps the issue of computing the normalizing constant separately for each query and is at the core of our approach. By sampling a number of hash functions $h_{1}, \ldots, h_{m}$ and creating hash tables $H_{1}, \ldots, H_{m}$, at query time we can produce $m$ independent samples that are used by the median-of-means procedure to produce an accurate estimate of the mean $\mu$.

The challenge in fully implementing this scheme is to bound the second moment of our estimator. In this regard, picking a random element from $H(x)$ turns out to be crucial, as it is this step that allows us to get an analytical handle on the second moment of $Z_{h}$. We provide two general theorems that bound the variance of Hashing-Based-Estimators. The first theorem applies to any HBE.

Theorem 1.3 (Two-points suffice). Up to absolute constants the variance of a HBE is maximized by datasets where there are only two values for the weights and sampling probabilities.

This characterization is extremely useful as given $p_{i}(x):=p\left(x, x_{i}\right)$ and $w_{i}(x)$, it reduces bounding the variance to a simple case analysis. The bound depends on the compatibility between the probability and weights, and is captured by the maximum element of an explicitly defined matrix. Going beyond the general case we identify a natural class of HBE that induce sampling probabilities that vary as a power of the weights.

Definition 1.4. An HBE is $(\beta, M)$-scale free for parameters $\beta \in(0,1]$ and $M \geqslant 1$ if, for all $i \in[n]$ and $x, M^{-1} \cdot w_{i}^{\beta}(x) \leqslant p_{i}(x) \leqslant M \cdot w_{i}^{\beta}(x)$.

The second theorem provides a refined analysis of the variance of scale-free HBE that is able to capture additional structure when one exists. In particular, the upper bound on the variance improves when most of the contribution to $\mu=\mu(x)$ comes from relatively large weights $w_{i}$. We state here the weakest bound that assumes nothing about the weights.

Theorem 1.5. Let $Z$ be an unbiased $(\beta, M)$-scale free $H B E$ for $\beta \in\left[\frac{1}{2}, 1\right]$. Then $\mathbb{E}\left[Z^{2}\right] \leqslant \mu^{2} \cdot\left(4 M^{3} / \mu^{\beta}\right)$.
The main technical tools behind the analysis are two Hölder-type inequalities that we develop. The first one (Lemma [2.3) is a simple two-sided (matrix) extension of Hölder's inequality that bounds a quadratic form over the intersection of two weighted $\ell_{1}$-balls. A clever application of this inequality gives us the proof of Theorem 1.3. The second (Lemma 2.4) is a non-trivial Höldertype inequality for monotone vectors that in combination with some easy consequences of Hölder's inequality gives us the refined analysis of scale-free estimators. The proof also shows the possibility of exploiting other structural assumptions of the data. This is important in the statistical setting (where the data set is sampled from a distribution) or for parameter tuning in practice.

Our theorem shows that under no structural assumptions, the optimal choice is $\beta^{*}=1 / 2$ and results in an estimator with relative variance $\frac{\operatorname{Var}[Z]}{(\mathbb{E}[Z])^{2}} \leqslant \mathrm{~V}(\mu)=4 M^{3} \cdot \mu^{-1 / 2}$. Using the Median-of-means (MoM) technique, we can estimate $\mu$ within a multiplicative accuracy of ( $1 \pm \varepsilon$ ) using $O\left(M^{3} \frac{1}{\varepsilon^{2}} \frac{1}{\sqrt{\mu}} \log (1 / \delta)\right)$ independent samples. Observe that this does not directly imply an algorithm to estimate $\mu$ using this many samples, as setting the number of samples (sufficient for accurate estimation) requires approximate knowledge of $\mu$, the very quantity we are aiming to estimate.

We resolve this issue by proposing an adaptive procedure that uses $O(1)$ times additional samples to get a constant factor approximation to $\mu$. We start with an overestimate of $\mu$ and iteratively decrease it until we get close enough to the truth where a consistency check is satisfied. The resulting algorithm, Adaptive Mean Relaxation, is applicable to settings where one has an unbiased estimator whose (upper bound on) variance is a non-decreasing function of the mean $\mu$ and the relative variance is a decreasing function of the mean.

### 1.2.2 Kernel Density Estimation through Locality Sensitive Hashing

We next turn to address the ( $\mu, \varepsilon, \delta$ )-KDE problem. To provide intuition about the problem we first analyze two simple randomized estimators, the (uniform) Random Sampling (RS) estimator and an estimator based on Random Fourier Features (RFF) [35]. We show that in the worst case, the first has variance bounded by $\mu$ whereas the RFF estimator has constant variance. Using the MoM framework and our adaptive procedure, one immediately gets an algorithm to solve the KDE problem using $O\left(\min \left\{\frac{1}{\varepsilon^{2}} \frac{1}{\mu} \log (1 / \delta), n\right\}\right)$ samples that is polynomial in $(\varepsilon, \mu, \log (1 / \delta))$.

In order to improve upon this simple bound, we employ the framework of Hashing-BasedEstimators instantiated with Locality Sensitive Hashing schemes. All our results follow a similar theme: (a) we obtain pointwise upper and lower bounds on the collision probabilities of the hash functions, and (b) we bound the variance of the estimator by invoking either Theorem 1.3 or Theorem 1.5 in cases where we are able to obtain scale-free estimators (cf. Table 1.2.2).

Table 1: Scale free estimators for KDE using LSH

| Kernel | $(\beta, M)$ | Complexity $T$ |
| :--- | :---: | :---: |
| $e^{-\\|x-y\\|^{2}}$ | $\left(\beta, e^{O\left(R^{\frac{4}{3}} \log \log n\right)}\right)$ | $e^{O\left(R^{\frac{4}{3}} \log \log n\right)}$ |
| $e^{-\\|x-y\\|}$ | $(\beta, \sqrt{e})$ | $O\left(d R^{2}\right)$ |
| $\frac{1}{1+\\|x-y\\|_{2}^{p}}$ | $\left(\frac{q}{p}, 3^{q}\right)$ | $O(d p)$ |

Theorem 1.6 (Informal). There exist scale-free HBE for the Gaussian, Exponential and $t$-Student kernels.
Using our theorem for scale-free estimators and the adaptive algorithm, we arrive at our main result for the KDE problem.
Theorem 1.7. For a kernel $k$ and dataset $P$ for which there exists $a\left(\frac{1}{2}, M\right)$-scale-free estimator with complexity $T$, there exists a data structure that solves the $(\mu, \varepsilon, \delta)$-KDE problem $\forall \mu \in[\tau, 1]$ using $O\left(M^{3} \frac{1}{\varepsilon^{2}} \frac{1}{\sqrt{\mu}} \log (1 / \delta) T\right)$ time and $O\left(M^{3} \frac{1}{\varepsilon^{2}} \frac{1}{\sqrt{\tau}} \log (1 / \delta) T \cdot n\right)$ space.

As an application, we show that one can use such a data structure to get an approximate vectormatrix multiplication algorithm for Kernel matrices using time that is adaptive to the vector and is always bounded by $\frac{n^{1+o(1)}}{\sqrt{\tau^{\prime}}} \frac{1}{\varepsilon^{2}}$ where $\varepsilon, \tau^{\prime}$ indicate respectively the relative and additive error per coordinate (cf. Theorem 7.1). This result is important as it improves on the main bottleneck of Kernel Ridge Regression [8], that is, multiplying a dense Kernel matrix with a vector and requires time $O\left(n^{2}\right)$ in general.

### 1.2.3 Lower Bounds for KDE problem

We also complement our results by providing a reduction between hard instances for the Approximate Nearest Neighbor Search problem to the ( $\mu, \varepsilon, \delta$ )-KDE problem. We show that the latter is at least as hard as the $(r, c)$-ANNS with $n=\frac{1}{\mu}$ points and $c=O\left(\frac{\log (n)}{\log (1 / \varepsilon)}\right)$ under the Hamming distance. Combined with the results of Panigrahy, Talwar, Wieder [32] and Andoni et al [5], we get non-trivial lower bounds in the cell-probe model with a single probe that captures an interesting class of algorithms based on adaptive coresets.

### 1.3 Related Work

The problem of Kernel Density Estimation although widely studied in low-dimensions [23] has largely been unexplored in high dimensions [29]. In recent parallel and independent work, Spring and Shrivastava [39] introduced the idea of using locality sensitive hashing as a sampling scheme to estimate the partition function of log-linear models, albeit without theoretical guarantees.
Coresets. The problem of KDE has mostly been investigated in the context of coresets. The first theoretical work we are aware of is that of Phillips [34]. Given a kernel $k$ and a set $P \subset \mathbb{R}^{d}$, a subset $S \subset P$ is an $\varepsilon$-coreset if $\left|\operatorname{KDE}_{P}(x)-\operatorname{KDE}_{S}(x)\right| \leqslant \varepsilon$ for all $x \in \mathbb{R}^{d}$. Phillips uses techniques from discrepancy theory to show that one can construct an $\varepsilon$-coreset of size $O\left(1 / \varepsilon^{2-\frac{4}{d+2}} \log ^{1-\frac{2}{d+2}}(1 / \varepsilon \delta)\right)$ with probability at least $1-\delta$. For large $d$ the bound deteriorates and becomes similar to what one gets by simple random sampling $O\left(\frac{1}{\varepsilon^{2}} \log (1 / \delta)\right)$ which is known to be tight. For relative error, recent work [42] uses random sampling to give a similar guarantee that roughly requires $O\left(\frac{1}{\varepsilon^{2}} \frac{1}{\mu}\right)$ samples. Our lower bound against the ( $m, w, 1$ )-cell probe models encompasses algorithms of this kind, and shows that any set $S$ (not necessarily a subset of $P$ ) that can be used to answer the $(\mu, \varepsilon, \delta)$-problem must have size at least $\Omega\left(\frac{1}{\mu}\right)$. The implication is that in terms of constructing a coreset for KDE in high dimensions, Random Sampling is essentially optimal.

Kernel Matrix-Vector Multiplication. A closely related idea to that of coresets, is that of Nyström approximation. In this method, given a kernel matrix $K$ a set of $s$ columns (points) is selected and subsequently $K$ is projected on their span. Musco and Musco [31] propose a method based on recursive leverage-score sampling that, using $s=\Theta(k \log k)$ points, $\tilde{O}\left(n k^{2}\right)$ time and $\tilde{O}(n k)$ space, outputs a matrix $\tilde{K}$ such that $\|\tilde{K}-K\|_{2} \leqslant \lambda$ with $\lambda=\frac{1}{k} \sum_{i=k+1}^{n} \sigma_{i}(K)$. One can use this algorithm to obtain an approximate Kenrel-Matrix Vector Multiplication algorithm $\hat{y} \approx K z$. For kernel matrices of large rank, like the ones corresponding to the equilateral metric of $r=\sqrt{n}=\tau^{-1}$ clusters consisting of $\sqrt{n}$ identical points, their algorithm requires $\tilde{O}\left(n^{2}\right)$ time and space $\tilde{O}\left(n^{3 / 2}\right)$ whereas our algorithm requires $\tilde{O}\left(n^{\frac{5}{4}}\right)$ time and space for the Exponential and Polynomial kernel, and $n^{\frac{5}{4}+o(1)}$ for the Gaussian kernel. However, the algorithm of Musco and Musco applies to any kernel and gives guarantees that hold simultaneously for any test vector $z$, whereas our method applies for a single (non-negative) vector $z$.

### 1.4 Open questions

Our work leaves open a few intriguing directions.
Data-dependent Hashing. There is a recent line of work [6, 5] that designs data-dependent LSH schemes that are optimal within a certain class of hashing-based algorithms. We believe that modifications of such schemes can be used for the purposes of KDE.

Batch Setting. The study of the offline or batch setting for Nearest Neighbor Search has received renewed interest over the past years and has brought a wealth of techniques into light [40, 2]. Given the connection between KDE and the Nearest Neighbor Search problem, it would be of practical and theoretical interest to design data-structures that offer provable speedups for the offline setting of the KDE problem. In the regime where $\varepsilon=\exp \left(-\omega\left(\log ^{2}(n)\right)\right)$ there is recent work [9] that shows that under SETH no significant improvements can be made beyond quadratic time.

Applications of HBE. It would be interesting to explore other extensions of HBE, e.g. get theoretical guarantees for estimating the partition function of log-linear models [39] or analyze the performance of multi-probe schemes.

### 1.5 Outline of the paper

In Section 2, we present some preliminaries and describe the general approach of deriving unbiased estimators and bounding their variance through Hölder-type inequalities, whose proofs are presented in Section 9 . In Section 3, we introduce Hashing-Based-Estimators and give two general tools to bound their variance. In Section 4, we present a framework for adaptively estimating the mean of a random variable when we have a known upper bound on the variance that depends on the unknown mean. In Sections5and 6, we use Locality Sensitive Hashing schemes to derive unbiased estimators for the Exponential, $t$-Student and Gaussian Kernels that when instantiated within the framework of Section 4 give rise to efficient data structures for the KDE problem. Using these data structures we then show in Section $\overline{7}$ how to perform approximate Fast Kernel-Matrix Vector multiplication. In Section 8, we present a lower bound for KDE for the Gaussian kernel in the cell-probe model.

## 2 Preliminaries

Notation. For a vector $x \in \mathbb{R}^{n}$ let $\|x\|_{p}^{p}:=\sum_{i=1}^{n}\left|x_{i}\right|^{p}$ for $p \geqslant 1$ denote the $\ell_{p}$-norm. For a strictly positive vector $w \in \mathbb{R}_{++}^{n}$, we denote by $\|x\|_{w, 1}:=\sum_{i} w_{i}\left|x_{i}\right|$ the weighted $\ell_{1}$-norm. Given a number $\tau>0$ let $\left(x_{i}\right)_{\tau}:=\max \left\{x_{i}, \tau\right\}$. Given a probability distribution $v$, we write $Y \sim v$ to denote that $Y$ is sampled from $v$. For a set $S, U(S)$ denotes the uniform distribution over $S$ and $S^{\otimes k}$ denotes
the Cartesian product of $S$ with itself $k \geqslant 1$ times. Similarly, $v^{\otimes k}$ denotes the product distribution $v \times \ldots \times v$. We use $N\left(0, I_{d}\right)$ to denote the standard multivariate normal distribution and $\Phi(\cdot)$ its CCDF. Throughout the paper $P=\left\{x_{1}, \ldots, x_{n}\right\} \subset \mathbb{R}^{d}$ will denote a set of $n$ points from $\mathbb{R}^{d}$ and we shall assume that $\operatorname{diam}(P \cup\{x\}) \leqslant R$ for any query $x \in \mathbb{R}^{d}$ where $R=R(n)>0$.

## 2.1 $V$-bounded Estimators and Median-of-Means

For $\varepsilon, \delta>0$, given a query point $x \in \mathbb{R}^{d}$ and weights $w_{1}(x), \ldots, w_{n}(x)$ induced by a set $P$, our goal is to estimate $\mu(x):=\frac{1}{n} \sum_{i=1}^{n} w_{i}(x)$ within a multiplicative $(1 \pm \varepsilon)$ accuracy with probability at least $1-\delta$. When it is clear from the context we will often drop the dependence on $x$ and simply write $w_{i}$ and $\mu$. An estimator $Z \sim v$ is called unbiased for our problem if $\mathbb{E}[Z]=\mu$. The basis of our approach is to design an unbiased estimator that has small variance relative to $\mu$.

Definition 2.1. Given a non-increasing function $V: \mathbb{R} \rightarrow \mathbb{R}_{+}$, we call an unbiased estimator $Z$, $V$-bounded if $\mathbb{E}\left[Z^{2}\right] \leqslant \mu^{2} \cdot V(\mu)$ and $\mu^{2} V(\mu)$ is non-decreasing.

The function $V$ is intimately related to the number of samples need to estimate $\mu$ and is often referred to as (a bound on) the relative variance. The principal approach to use such an estimator is the Median-of-Means (MoM) technique [3], that allows one to get an estimate $Z_{\varepsilon, \delta}$ such that $\mathbb{P}\left[\left|Z_{\varepsilon, \delta}-\mu\right| \geqslant \varepsilon \cdot \mu\right] \leqslant \delta$ using $O\left(\frac{1}{\varepsilon^{2}} V(\mu) \log \left(\frac{1}{\delta}\right)\right)$ samples.

```
Algorithm 1 Median-of-Means (MoM)
    Input: Estimator \(Z \sim v, V \geqslant 0\), accuracy \(\varepsilon \in(0,1)\), success prob. \(\delta \in(0,1)\).
    \(m(\varepsilon, V) \leftarrow\left\lceil\frac{6}{\varepsilon^{2}} V\right\rceil, L(\delta) \leftarrow\lceil 9 \log (1 / \delta)\rceil\).
    \(Z_{j}^{(i)} \stackrel{i i d}{\sim} v\) for \(j=1, \ldots, m\), and \(i=1, \ldots, L\).
    \(Z^{(i)} \leftarrow \operatorname{mean}\left\{\mathrm{Z}_{1}^{(i)}, \ldots, \mathrm{Z}_{m}^{(i)}\right\}\) for \(i=1, \ldots, L\).
    Output: \(Z_{\varepsilon, \delta} \leftarrow \operatorname{median}\left\{Z^{(1)}, \ldots, Z^{(L)}\right\}\)
```

Most of our effort in this paper goes into obtaining efficient $V$-bounded estimators with $V$ being of the form $V(\mu)=n^{o(1)} \cdot \mu^{-\Delta}$ for some $0 \leqslant \Delta \leqslant 2$. Once we have such an estimator, we will be able to combine it with the MoM technique and eventually get an estimation algorithm. Hence, the main challenge is bounding the variance.

### 2.2 Bounding the Variance

Given an unbiased estimator $Z$ in order to bound the variance we first obtain a simple datadependent upper bound on $\mathbb{E}\left[Z^{2}\right] \leqslant F(w, P)$ for some function $F$, and then for a class of datasets $\mathcal{P}$, we aim to show that $\sup _{P \in \mathcal{P}}\left\{F(w, P) \mid \sum_{i} w_{i}=n \mu\right\} \leqslant \mu^{2} V(\mu)$.

### 2.2.1 Baseline estimators

As a warm up, we present two simple unbiased estimators for the Kernel Density Problem and bound their variance. Let $Y \sim U[P]$, the random sampling estimator is given by $\mathrm{Z}_{\mathrm{RS}}(x)=k(x, Y)$. Let $\theta \sim U[0, \pi]$ and $g \sim N\left(0, I_{d}\right)$, the Random Fourier Features [35] estimator is given by $Z_{R F F}(x)=$ $\frac{2}{|P|} \sum_{y \in P}\left(\cos \left(g^{\top} x+\theta\right) \cos \left(g^{\top} y+\theta\right)\right)$.

Proposition 2.2. The RS and RFF estimators are unbiased and satisfy respectively $\mathbb{E}\left[Z_{R S}^{2}\right] \leqslant \mu^{2} \cdot \mu^{-1}$ and $\mathbb{E}\left[Z_{R F F}^{2}\right] \leqslant \mu^{2} \cdot 4 \mu^{-2}$. Moreover, the bounds are tight up to constants in the worst case.

Proof Sketch. The fact that the RS estimator is unbiased is trivial, whereas the fact that RFF is unbiased was shown by Rahimi and Recht [35] and follows from Bochner's theorem and trigonometric identities. We next bound the second moment

$$
\begin{aligned}
& \mathbb{E}\left[Z_{R S}^{2}\right] \leqslant \max _{y \in P}\{k(x, y)\} \cdot \frac{1}{|P|} \sum_{y \in P} K(x, y) \leqslant \mu^{2} \cdot \mu^{-1} \\
& \mathbb{E}\left[Z_{R F F}^{2}\right] \leqslant\left(\frac{2}{|P|} \sum_{y \in P} 1\right)^{2}=4 \leqslant \mu^{2} \cdot 4 \mu^{-2}
\end{aligned}
$$

To see that these bounds are tight up to constants consider for the RS estimator a dataset with $n \mu$ points located at $x$ and the rest $n(1-\mu)$ points at distance $\sqrt{\log (1 / \mu)}$ from $x$. For the RFF estimator the worst case is when all points are all located at the same point $y_{0}$ at distance $\sqrt{\log (1 / \mu)}$.

We expand more on the RFF estimator in Appendix B The result on Random Sampling shows that KDE problem is solvable in time $O\left(\min \left\{\frac{1}{\varepsilon^{2}} \frac{1}{\mu} \log \left(\frac{1}{\delta}\right), n\right\}\right)$.

### 2.2.2 Upper bounds via Hölder-type inequalities

Despite the simplicity of the above result, some salient features of the problem are revealed. Firstly, the quality of the initial upper bound $F(w, P)$ can differ dramatically between two different estimators. Secondly, despite the simplicitly of the analysis, often the resulting bounds are tight up to constants. Lastly, in both cases we used Hölder's inequality to get the bound. This is going to be a general theme as behind all our bounds on the variance are increasingly sophisticated consequences of Hölder's inequality. We state below the two main inequalities used to bound the variance of Hashing-Based-Estimators.

Lemma 2.3 (two-sided Holder). Let $v, w \in \mathbb{R}^{n}$ be strictly positive vectors, then for any two sets $S, S^{\prime} \subseteq$ [ $n$ ]:

$$
\begin{equation*}
\sum_{i \in S, j \in S^{\prime}} A_{i j} x_{i} x_{j} \leqslant\|x\|_{v, 1}\|x\|_{w, 1} \cdot \max _{i \in S, j \in S^{\prime}}\left\{\frac{\left|A_{i j}\right|}{v_{i} w_{j}}\right\} \tag{2.1}
\end{equation*}
$$

Lemma 2.4 (Monotone Hölder). $\forall n \geqslant 1, \beta \in\left[\frac{1}{2}, 1\right]$, and $\forall x \in \mathbb{R}^{n}$ such that $\left|x_{1}\right| \geqslant\left|x_{2}\right| \geqslant \ldots \geqslant\left|x_{n}\right|$, we have $\sum\left|x_{i}\right|^{\frac{2-\beta}{\beta}}\left(i+\sum_{j>i} \left\lvert\, \frac{\left|x_{j}\right|}{\left|x_{i}\right|}\right.\right) \leqslant n^{\beta} .\left(\sum_{i=1}^{n}\left|x_{i}\right|^{\frac{1}{\beta}}\right)^{2-\beta}$ with equality holding for $x^{*}=c \mathbf{1}$ for any $c \neq 0$.

## 3 Importance Sampling through Hashing Based Estimators

Given a distribution $v$ over a collection of hash functions $\mathcal{H}$, let $h \in \mathcal{H}$ be an element sampled according to $v$. For a point $x \in \mathbb{R}^{d}$ let $H(x):=\{y \in P: h(y)=h(x)\}$ be the set of elements in $P$ that have the same hash value as $x$. Also, let $I(x)$ be a uniform random element out of $H(x)$ or $\perp$ if $H(x)$ is empty. Further, define $p_{i}:=\mathbb{P}[i \in H(x)]$ and $\tilde{P}:=\left\{i \in P \mid p_{i}>0\right\}$. We also set $p_{\perp}:=1$ and $w_{\perp}:=0$. An $(\mathcal{H}, v)$-hashing based estimator (HBE) is defined as $Z_{h}=Z_{h}(x):=\frac{w_{I(x)}}{p_{I(x)}} \cdot \frac{|H(x)|}{n}$.

### 3.1 Moments of HBE

Lemma 3.1 (Moments). Let $Z_{h}$ be $a(\mathcal{H}, v)-H B E$ and $P$ a set of points. Let $p_{1} \geqslant p_{2} \geqslant \ldots \geqslant p_{n}$, then

$$
\begin{equation*}
\mathbb{E}\left[Z_{h}\right]=\frac{1}{n} \sum_{i \in \tilde{P}} w_{i} \quad \text { and } \quad \mathbb{E}\left[Z_{h}^{2}\right] \leqslant \frac{1}{n^{2}} \sum_{i \in \tilde{P}} \frac{w_{i}^{2}}{p_{i}}\left(i+\sum_{j>i} \frac{p_{j}}{p_{i}}\right) \tag{3.1}
\end{equation*}
$$

Proof. The proof is based on applying Bayes rule and total probability law.

$$
\begin{aligned}
\mathbb{E}\left[Z_{h}\right] & =\frac{1}{n} \sum_{i \in \tilde{P}, k \geqslant 1} \frac{w_{i}}{p_{i}} k \mathbb{P}(\tilde{I}=i,|H(q)|=k) \\
& =\frac{1}{n} \sum_{i \in \tilde{P}, k \geqslant 1} \frac{w_{i}}{p_{i}} k \mathbb{P}(I=i|i \in H(q),|H(q)|=k) \mathbb{P}(i \in H(q),|H(q)|=k) \\
& =\frac{1}{n} \sum_{i \in \tilde{P}} \frac{w_{i}}{p_{i}} \sum_{k \geqslant 1} \mathbb{P}(i \in H(q),|H(q)|=k) \\
& =\frac{1}{n} \sum_{i \in \tilde{P}} \frac{w_{i}}{p_{i}} \mathbb{P}(i \in H(q))
\end{aligned}
$$

Similarly, we show for the variance

$$
\begin{aligned}
\mathbb{E}\left[Z_{h}^{2}\right] & =\frac{1}{n^{2}} \mathbb{E}\left[\left(w_{\tilde{I}} \frac{|H(q)|}{p_{I}}\right)^{2}\right] \\
& =\frac{1}{n^{2}} \sum_{i, k}\left(w_{i} \frac{|H(q)|}{p_{i}}\right)^{2} \mathbb{P}(H(q)=k, I=i) \\
& =\frac{1}{n^{2}} \sum_{i, k} \frac{w_{i}^{2} k^{2}}{p_{i}^{2}} \mathbb{P}(H(q)=k, I=i \mid i \in H(q)) \mathbb{P}(i \in H(q)) \\
& =\frac{1}{n^{2}} \sum_{i, k} \frac{w_{i}^{2} k^{2}}{p_{i}} \mathbb{P}(I=i \mid H(q)=k, i \in H(q)) \mathbb{P}(H(q)=k \mid i \in H(q)) \\
& =\frac{1}{n^{2}} \sum_{i, k \geqslant 1} \frac{w_{i}^{2}}{p_{i}} \sum_{k \geqslant 1} k \mathbb{P}(H(q)=k \mid i \in H(q)) \\
& =\frac{1}{n^{2}} \sum_{i} \frac{w_{i}^{2}}{p_{i}} \mathbb{E}[\mid H(q) \| i \in H(q)] \\
& =\frac{1}{n^{2}} \sum_{i} \frac{w_{i}^{2}}{p_{i}}\left(1+\sum_{j \neq i} \frac{\mathbb{P}(i, j \in H(q))}{p_{i}}\right)
\end{aligned}
$$

Finally, we obtain the upper bound by using the fact that $\mathbb{P}(A, B) \leqslant \min \{\mathbb{P}(A), \mathbb{P}(B)\}$ for any two events.

$$
\begin{align*}
\mathbb{E}\left[Z_{h}^{2}\right] & =\frac{1}{n^{2}} \sum_{i} \frac{w_{i}^{2}}{p_{i}}\left(1+\sum_{j \neq i} \frac{\mathbb{P}(i, j \in H(q))}{p_{i}}\right)  \tag{3.2}\\
& \leqslant \frac{1}{n^{2}} \sum_{i} \frac{w_{i}^{2}}{p_{i}}\left(1+\sum_{j \neq i} \frac{\min \left\{p_{i}, p_{j}\right\}}{p_{i}}\right)  \tag{3.3}\\
& =\frac{1}{n^{2}} \sum_{i} \frac{w_{i}^{2}}{p_{i}}\left(i+\frac{1}{p_{i}} \sum_{j>i} p_{j}\right) \tag{3.4}
\end{align*}
$$

Observe that the estimator is unbiased iff $w_{i}>0 \Rightarrow p_{i}>0$ for all $i \in[n]$. The quantity in the parenthesis plays the role of $F(w, P)$ and expresses a pessimistic upper bound on $\mathbb{E}[|H(x)| \mid i \in H(x)]$. This is perhaps the single most important aspect of the method in that it allows us to express the variance purely in terms of known collision probabilities $p_{i}$ that are amenable to analytic manipulations.

### 3.2 Variance Bounds for general HBE

In this section, we prove Theorem 1.3 that provides a characterization of the variance of HBE under worst case assumptions. The main tool we employ is the following inequality.
Lemma 3.2. Given a positive vector $w \in \mathbb{R}_{++}^{n}$ and a parameter $\mu>0$, define $f_{i}^{*}:=\min \left\{1, \frac{\mu}{w_{i}}\right\}$. For any matrix $A \in \mathbb{R}^{n \times n}$ :

$$
\begin{equation*}
\sup _{\|f\|_{w, 1} \leqslant \mu,\|f\|_{1} \leqslant 1}\left\{f^{\top} A f\right\} \leqslant 4 \sup _{i j \in[n]}\left\{f_{i}^{*}\left|A_{i j}\right| f_{j}^{*}\right\} \tag{3.5}
\end{equation*}
$$

Proof. Let $S_{+}(\mu)=\left\{i \in[n] \mid w_{i} \geqslant \mu\right\}$ and $S_{-}(\mu)=\left\{i \in[n] \mid w_{i}<\mu\right\}$ and define $A_{ \pm \pm}:=A_{S_{ \pm}(\mu), S_{ \pm}(\mu)}$, where $A_{U, V}$ is the matrix with elements $A_{i j}, i \in U, j \in V$. We have the decomposition:

$$
\begin{equation*}
f^{\top} A f=f_{+}^{\top} A_{++} f_{+}+f_{+}^{\top} A_{+-} f_{-}+f_{-}^{\top} A_{-+} f_{+}+f_{-}^{\top} A_{--} f_{-} \tag{3.6}
\end{equation*}
$$

Using this decomposition and Lemma 2.3 we have

$$
\begin{align*}
\sup _{\|f\|_{w, 1} \leqslant \mu,\|f\|_{1} \leqslant 1}\left\{f^{\top} A f\right\} \leqslant & \sup _{\left\|f_{+}\right\|\left\|_{w, 1} \leqslant \mu,\right\| f_{+} \|_{w, 1} \leqslant \mu}\left\{f_{+}^{\top} A_{++} f_{+}\right\}+\sup _{\left\|f_{+}\right\|\left\|_{w, 1} \leqslant \mu,\right\| f f_{-} \|_{1} \leqslant 1}\left\{f_{+}^{\top} A_{+-} f_{-}\right\}  \tag{3.7}\\
& +\sup _{\left\|f_{-}\right\|_{1} \leqslant 1,\left\|f_{+}\right\| \|_{w, 1} \leqslant \mu}\left\{f_{-}^{\top} A_{-+} f_{+}\right\}+\sup _{\left\|f_{-}\right\|_{1} \leqslant 1,\left\|f_{-}\right\|_{1} \leqslant 1}\left\{f_{-}^{\top} A_{--} f_{-}\right\}  \tag{3.8}\\
\leqslant & \sup _{i \in S_{+}, j \in S_{+}}\left\{\frac{\mu}{w_{i}}\left|A_{i j}\right| \frac{\mu}{w_{j}}\right\}+\sup _{i \in S_{+}, j \in S_{-}}\left\{\frac{\mu}{w_{i}}\left|A_{i j}\right| \frac{1}{1}\right\}  \tag{3.9}\\
& +\sup _{i \in S_{-}, j \in S_{+}}\left\{\frac{1}{1}\left|A_{i j}\right| \frac{\mu}{w_{j}}\right\}+\sup _{i \in S_{-}, j \in S_{-}}\left\{\frac{1}{1}\left|A_{i j}\right| \frac{1}{1}\right\} \tag{3.10}
\end{align*}
$$

To complete the proof note that $\min \left\{1, \frac{\mu}{w_{i}}\right\}=\frac{\mu}{w_{i}} \forall i \in S_{+}(\mu)$ and $\min \left\{1, \frac{\mu}{w_{j}}\right\}=1 \forall j \in S_{-}(\mu)$.
Proof of Theorem 1.3 Fix $n$ distinct weights $w_{1}, \ldots, w_{n}$ and collision probabilities $p_{1}, \ldots, p_{n}$ that might be considered. Without loss of generality we assume that the probabilities are in decreasing order. Let $f_{i}$ be the fraction of points that are assigned to weight $i$. We have the following:

$$
\begin{aligned}
\mathbb{E}\left[Z_{h}^{2}\right] & \leqslant \frac{1}{n^{2}} \sum_{i}\left(f_{i} n\right) \frac{w_{i}^{2}}{p_{i}}\left(n \sum_{j \leqslant i} f_{j}+\frac{1}{p_{i}} \sum_{j>i} p_{j}\left(n f_{j}\right)\right) \\
& =\sum_{i, j} f_{i} f_{j}\left(\frac{w_{i}^{2}}{p_{i}} \mathbb{I}_{j \leqslant i}+\mathbb{I}_{j>i} \frac{w_{i}^{2}}{p_{i}^{2}} p_{j}\right) \\
& \leqslant \sup _{\substack{\|f\|\|, j \leq \mu\\
\| f \|_{1} \leqslant 1}}\left\{f^{\top} A f\right\}
\end{aligned}
$$

where in the last step we set $A_{i j}:=\frac{w_{i}^{2}}{p_{i}} \mathbb{I}_{j \leqslant i}+\mathbb{I}_{j>i} \frac{w_{i}^{2}}{p_{i}^{2}} p_{j}$. Invoking Lemma 3.2]shows that for any set of weights $\left\{w_{i}\right\}_{i \in[n]}$ and probabilities $\left\{p_{i}\right\}_{i \in[n]}$ there exist two indices $i^{*}, j^{*} \in[n]$ that realize up to a constant the worst case variance for the specific density $\mu$.

The theorem assumes nothing about HBE besides unbiasedness and therefore is applicable in an arbitrary setting. If more structure is assumed stronger statements can be made.

### 3.3 Variance Bounds for Scale-free HBE

We give a refined analysis of scale-free estimators (Definition 1.4) based on whether a large fraction of the density comes from points with relatively large weights.

Definition 3.3. For a query $x$ and $\tau \in[\mu, 1]$ let $B_{\tau, \mu}(x):=\left\{i \in P \left\lvert\, w_{i} \geqslant \frac{\mu}{\tau}\right.\right\}$. For $\gamma \in[0,1]$ the query is said to be $(\tau, \gamma)$-localized if $\sum_{i \in B_{\tau, \mu}} w_{i} \geqslant(1-\gamma) \sum_{i} w_{i}$.

The intuition behind this definition is that for kernels that are decreasing with the distance, the sets of points that would induce large (small) weights are contained in (outside of) a ball of a certain radius around the query. If the total contribution of the points outside of the ball is negligible, our estimator behaves essentially as if we are at a "higher" density regime.
Theorem 3.4 (Restatement of Theorem 1.5). Let $Z_{h}$ be a $(\beta, M)$-scale free estimator with $\beta \in[1 / 2,1]$. For every $(\tau, \gamma)$-localized query $x$,

$$
\mathbb{E}\left[Z_{h}^{2}\right] \leqslant \mu^{2} \cdot M^{3}\left\{2 \tau^{\beta}+\gamma^{2-\beta}+\tau^{2 \beta-1} \gamma^{\beta}\right\} \mu^{-\beta}
$$

The optimal choice of $\beta$ in the case where no assumptions are made $(\gamma=\tau=1)$ is $\beta^{*}=\frac{1}{2}$. Even if $\beta$ is not selected depending on the structure of the query, significant gains can be obtained, in cases where the parameters $\tau, \gamma$ are small enough. As an example, for $\beta=\frac{1}{2}, \tau=\mu^{\frac{1}{2}}, \gamma=\mu^{\delta}$ the theorem implies that the variance is bounded by $\mu^{2} \cdot 4 M^{3} \mu^{-\Delta(\delta)}$ where $\Delta(\delta)=\max \left\{\frac{1}{4}-\delta, \frac{1-3 \delta}{2}\right\}<\frac{1}{2}$ for all $\delta>0$.

Proof. We start by using the scale-free property to simplify (3.1):

$$
\begin{equation*}
\mathbb{E}\left[Z_{h}^{2}\right] \leqslant M^{3} \cdot \frac{1}{n^{2}} \underbrace{\sum_{i=1}^{n} w_{i}^{2-\beta}\left(i+\sum_{j>i}\left(\frac{w_{j}}{w_{i}}\right)^{\beta}\right)}_{A} \tag{3.11}
\end{equation*}
$$

To bound $A$, we break up the terms according to $B_{\tau, \mu}$. Let $J$ be the maximal index in $B_{\tau, \mu}$.

$$
\begin{align*}
& A=\sum_{i \leqslant J} w_{i}^{2-\beta}\left(i+\sum_{J \geqslant j>i}\left(\frac{w_{j}}{w_{i}}\right)^{\beta}\right)+\sum_{i>J} w_{i}^{2-\beta}\left(i-J+\sum_{j>i}\left(\frac{w_{j}}{w_{i}}\right)^{\beta}\right)  \tag{3.12}\\
&+\sum_{i \leqslant J} w_{i}^{2-2 \beta} \cdot \sum_{j>J} w_{j}^{\beta}+\left|B_{\tau, \mu}\right| \sum_{i>J} w_{i}^{2-\beta} \tag{3.13}
\end{align*}
$$

We next use Hölder-type inequalities to bound each term. For the first two terms we invoke Lemma 2.4(Monotone Hölder) to get:

$$
\begin{align*}
\sum_{i \leqslant J} w_{i}^{2-\beta}\left(i+\sum_{J \geqslant j>i}\left(\frac{w_{j}}{w_{i}}\right)^{\beta}\right) & \leqslant(n \tau)^{\beta}(n \mu)^{2-\beta}  \tag{3.14}\\
\sum_{i>J} w_{i}^{2-\beta}\left(i-J+\sum_{j>i}\left(\frac{w_{j}}{w_{i}}\right)^{\beta}\right) & \leqslant n^{\beta}(\gamma n \mu)^{2-\beta} \tag{3.15}
\end{align*}
$$

where we also used the fact that $\left|B_{\tau, \mu}\right| \leqslant \tau n$. Let $w_{\leqslant}=\left(w_{1}, \ldots, w_{J}\right)$ and $w_{>}:=\left(w_{J+1}, \ldots, w_{n}\right)$. To bound the remaining terms, we invoke Corollary 9.2 in total three times to get:

$$
\begin{equation*}
\sum_{i \leqslant J} w_{i}^{2-2 \beta} \cdot \sum_{j>J} w_{j}^{\beta} \leqslant\left\|w_{\leqslant}\right\|_{1}^{2-2 \beta}\left|B_{\tau, \mu}\right|^{1-(2-2 \beta)} \cdot\left\|w_{>}\right\|_{1}^{\beta} n^{1-\beta} \tag{3.16}
\end{equation*}
$$

$$
\begin{equation*}
\leqslant(n \mu)^{2-2 \beta}(\tau n)^{1-(2-2 \beta)} \cdot(\gamma n \mu)^{\beta} n^{1-\beta} \tag{3.17}
\end{equation*}
$$

and

$$
\begin{equation*}
\left|B_{\tau, \mu}\right| \sum_{i>J} w_{i}^{2-\beta} \leqslant n \tau \cdot\left\|w_{>}\right\|_{\infty}^{1-\beta} \cdot\left\|w_{>}\right\|_{1}^{1} \leqslant n \tau\left(\frac{\mu}{\tau}\right)^{1-\beta}(\gamma n \mu) \tag{3.18}
\end{equation*}
$$

By combining all the inequalities, we arrive at $A \leqslant n^{2} \mu^{2-\beta} \cdot\left\{\tau^{\beta}+\gamma^{2-\beta}+\tau^{2 \beta-1} \gamma^{\beta}+\tau^{\beta} \gamma\right\}$.
This concludes the presentation of the general framework of Hashing-Based-Estimators, that given a set of weights $\left\{w_{i}\right\}$ and collisions probabilities $\left\{p_{i}\right\}$ computes a function $V$ such that the resulting HBE is $V$-bounded.

## 4 Adaptive Estimation through Mean Relaxation

Our goal in this section is, given a dataset $P \in \mathbb{R}^{d}$ and a $V$-bounded estimator, to build a data structure that can efficiently approximate the mean $\mu(x)$ for a given query $x \in \mathbb{R}^{d}$. Thus, presenting a complete algorithmic framework that can be instantiated for different problems.

We first address the problem of obtaining a constant factor approximation to the mean. We exploit two facts: monotonicity of the variance in terms of $\mu$ and concentration of measure. Monotonicity suggests that we can start with an over-estimate of the mean and keep refining it until we come very close to the truth. Concentration of measure allows us to come up with a simple consistency check that recognizes when our estimate is close enough to the true mean. Based on this we propose the following adaptive algorithm for which we get strong guarantees.

```
Algorithm 2 Adaptive Mean Relaxation (AMR)
    Input: V-bounded unbiased estimator \(Z \sim v\), query \(x \in \mathbb{R}^{d}\), accuracy \(\alpha \in(0,1]\), threshold
    \(\tau \in(0,1)\), failure prob. \(\chi \in(0,1)\).
    \(\varepsilon \leftarrow \frac{2}{7} \alpha, c \leftarrow \frac{\varepsilon}{2}, \gamma \leftarrow \frac{\varepsilon}{7}, \delta \leftarrow \frac{2 \alpha}{49 \log (1 / \tau)} \chi, i \leftarrow-1\).
    repeat
        \(i \leftarrow i+1, \mu_{i} \leftarrow(1-\gamma)^{i}\),
        \(\mathrm{Z}_{i} \leftarrow \operatorname{MoM}_{\frac{\varepsilon}{3}, \delta}\left(v, V\left(\mu_{i}\right)\right)\)
    until \(\left|Z_{i}-\mu_{i}\right| \leqslant c \cdot \mu_{i}\) or \(i>\frac{\log (\tau /(1-(c+\varepsilon)))}{\log (1-\gamma)}\).
    Output: if \(i \leqslant \frac{7^{2} \log (1 / \tau)}{2 \alpha}\) return \(Z_{i}\) else return 0 .
```

In fact, there is no need for the samples used in different calls of the MoM routine to be independent and we can implement the MoM routing by keeping $L(\delta)=\lceil 9 \log (1 / \delta)\rceil$ running sums. We call the resulting algorithm AMR*.

Theorem 4.1 (Mean Relaxation). Let $Z \sim v$ be a $V$-bounded estimator with $\mathbb{E}[Z]=\mu \in(0,1]$ and $\hat{Z}$ be the output of the $A M R^{*}$ algorithm with parameters $(\alpha, \tau, \chi)$. If $\mu \geqslant \tau$ then $\mathbb{P}[|\hat{Z}-\mu| \leqslant \alpha \mu] \geqslant 1-\chi$ otherwise if $\mu<\tau, \mathbb{P}[\hat{Z}=0] \geqslant 1-\chi$. The total number of samples used is bounded by $O\left(\alpha^{-2} \log \left(\frac{\log (1 / \tau)}{\chi \alpha}\right) \cdot V\left((\mu)_{\tau}\right)\right)$.

With this algorithm in hand we are ready to state our main result in the abstract setting of HBE.
Theorem 4.2 (Main Result). Given a $V$-bounded HBE with complexity $T$, there exists a data structure that can answer any query in time $O\left(\frac{1}{\varepsilon^{2}} V\left((\mu)_{\tau}\right) \log \left(\frac{1}{\chi}\right) T\right)$ using space $O\left(\frac{1}{\varepsilon^{2}} V(\tau) \log \left(\frac{1}{\chi}\right) \cdot n T\right)$ with success probability at least $1-\chi$ for $\chi \leqslant 1 / \log (1 / \tau)$.

Proof. We begin by describing the preprocessing phase. We sample $N=O\left(\log (1 / \chi) \frac{1}{\varepsilon^{2}} V(\tau)\right)$ hash functions $h_{1}, \ldots, h_{N} \stackrel{i . i . d .}{\sim} v$ from $\mathcal{H}$ and evaluate them on the dataset $P$. This can be done in $N T \cdot n$ time and space. The query algorithm interacts with the data-structure by making calls to hash functions. The data-structure always keeps the index of the last hash function called and increments it in a cyclic fashion after each call, thus for a given query it never evaluates the same hash function twice and the samples obtained are independent. When the query arrives, the query algorithm first runs a stage of the adaptive mean relaxation algorithm with $\alpha=1$ and probability $\chi / 2$. Every time a sample is needed a call is made to the data-structure. After $O\left(V\left((\mu)_{\tau} \log (1 / \chi)\right)\right)$ calls with probability at least $1-\chi / 2$ we either have a constant factor approximation or we know that $\mu<\tau$ (if AMR* outputs 0). In the first case, we apply one level of MoM algorithm using an underestimate of $\mu$ that uses $O\left(\frac{1}{\varepsilon^{2}} \log (1 / \chi) V\left((\mu)_{\tau}\right)\right)$ more calls and gets a $(1 \pm \varepsilon)$ multiplicative approximation with probability at least $1-\chi$. In the latter, case we simply output 0 .

The proof of Theorem 1.7 follows by invoking Theorem 4.2 for specific $V$-bounded estimators that we derive in Sections 5 and 6 using Locality Sensitive Hashing schemes. In the rest of this section we give the proof of the intermediate results. We use the following key lemma.
Lemma 4.3. Given a $V$-bounded estimator $Z \sim v$, let $Z_{i}:=\operatorname{MoM}_{\frac{\varepsilon}{3}}\left(v, V\left(\mu_{i}\right)\right)$, with $\mu_{i} \in(0,1]$ then
(i) for all $i$ such that $\mu_{i} \geqslant(1-(c+\varepsilon))^{-1} \mu$, it holds that $\mathbb{P}\left[\left|Z_{i}-\mu_{i}\right| \leqslant c \cdot \mu_{i}\right] \leqslant \delta$.
(ii) for all $\frac{1+\frac{\varepsilon}{3}}{1+c} \mu \leqslant \mu_{i} \leqslant \mu$, it holds that $\mathbb{P}\left[\left|Z_{i}-\mu_{i}\right| \leqslant c \cdot \mu_{i}\right] \geqslant 1-\delta$.

### 4.1 Proof of Theorem 4.1

### 4.1.1 Case $\mu \geqslant \tau$

If $\mu \geqslant \tau$, we say that the algorithm succeeds when the algorithm terminates with $\frac{1+\varepsilon / 3}{1+c} \mu \leqslant \mu_{i} \leqslant$ $\frac{1}{1-(c+\varepsilon)} \mu$. By construction of the algorithm this would imply $\left|Z_{i}-\mu\right| \leqslant \alpha \mu$. To show this we consider two cases. The first case is when $\mu_{i}$ is in the subinterval with elements larger than $\mu$ :

$$
\left.\begin{array}{c}
\mu \leqslant \mu_{i} \leqslant(1-(c+\varepsilon))^{-1} \mu  \tag{4.1}\\
\left|Z-\mu_{i}\right| \leqslant c \mu_{i}
\end{array}\right\} \Rightarrow\left|Z-\mu_{i}\right| \leqslant\left((c+\varepsilon)+\frac{c}{1-(c+\varepsilon)}\right) \mu \leqslant \alpha \mu
$$

where in the last step we used that $\varepsilon=2 c=2 \alpha / 7$ and $1-3 \alpha / 7 \geqslant 1 / 2$. The other case is when:

$$
\left.\begin{array}{l}
\frac{1+\varepsilon / 3}{1+c} \mu \leqslant \mu_{i} \leqslant \mu  \tag{4.2}\\
\left|Z-\mu_{i}\right| \leqslant c \mu_{i}
\end{array}\right\} \Rightarrow\left|Z-\mu_{i}\right| \leqslant \frac{2 \varepsilon}{3} \mu \leqslant \alpha \mu
$$

Thus, what is left is to bound the probability that the algorithm terminates for some $i$ that satisfies $\frac{1+\varepsilon / 3}{1+c} \mu \leqslant \mu_{i} \leqslant(1-(c+\varepsilon))^{-1} \mu$.
Claim 4.4. There exists $0 \leqslant i^{*} \leqslant \frac{\log (\mu /(1-(c+\varepsilon)))}{\log (1-\gamma)}$ such that $\mu \leqslant \mu_{i^{*}} \leqslant(1-(c+\varepsilon))^{-1} \mu$.
Proof. If $\mu>(1-(c+\varepsilon))$ then this is true for $i^{*}=0$, otherwise there exist $i \geqslant 1$ such that $\mu_{i-1} \geqslant \frac{1}{1-(c+\varepsilon)} \mu$. Since we start with $\mu \leqslant \mu_{0}$, we show that by slowly decreasing $\mu_{i}$ (at a rate of $1-\gamma$ ) we never go from the case $\mu_{i-1} \geqslant(1-(c+\varepsilon))^{-1} \mu$ to the case $\mu_{i}<\mu$. Assuming that was possible we obtain a contradiction.

$$
\begin{equation*}
(1-\gamma)=\frac{\mu_{i}}{\mu_{i-1}}<\frac{\mu}{(1-(c+\varepsilon))^{-1} \mu}=1-(c+\varepsilon)<1-\gamma \tag{4.3}
\end{equation*}
$$

Since the sequence $\mu_{i}=(1-\gamma)^{i}$ there must be an index $i \leqslant \log (\mu /(1-c-\varepsilon)) / \log (1-\gamma)$ with the required property.

Claim 4.5. There exists $j^{*}>i^{*}$ such that $\frac{1+\varepsilon / 3}{1+c} \mu \leqslant \mu_{j^{*}} \leqslant \mu$.
Proof. Let $i^{*}$ be as above, then $\mu \leqslant \mu_{i^{*}}$. We show that for $j>i^{*}$ we cannot have $\mu \leqslant \mu_{j-1}$ and $\mu_{j}<\frac{1+\varepsilon / 3}{1+c} \mu$ as in that case $1-\gamma=\frac{\mu_{j}}{\mu_{j-1}}<\frac{1+\varepsilon / 3}{1+c} \Rightarrow \gamma>\frac{\varepsilon}{6(1+c)}>\frac{\varepsilon}{7}=\gamma$. As the sequence $\mu_{i}$ is decreasing there must exist a $j^{*}$ with the required property and $j^{*} \leqslant i^{*}+\left\lceil\frac{\log \left(\frac{1+c}{(1-c+t)(1+\varepsilon / 3)}\right)}{-\log (1-\gamma)}\right\rceil=O\left(i^{*}\right)$

The algorithm thus succeeds if the condition $\left|Z_{i}-\mu_{i}\right|<c \mu_{i}$ is satisfied for any index $k \in$ $\left\{i^{*}, \ldots, j^{*}\right\}$ and is not satisfied for any index $k<i^{*}$. Let $A_{i}$ be the event that condition is satisfied at step $i \geqslant 0$. Lemma 4.3 shows that $\mathbb{P}\left[A_{i}\right] \leqslant \delta$ for all $i<i^{*}$ and $\mathbb{P}\left[A_{j^{*}}\right] \geqslant 1-\delta$. The probability of success is then at least:

$$
\begin{equation*}
\mathbb{P}\left[\bar{A}_{0} \cap \ldots \cap \bar{A}_{i^{*}-1} \cap A_{j^{*}}\right]=1-\mathbb{P}\left[A_{0} \cup A_{i^{*}-1} \cup \bar{A}_{j^{*}}\right] \geqslant 1-\left(i^{*}+1\right) \delta \geqslant 1-\chi \tag{4.4}
\end{equation*}
$$

The number of samples required for the $\mathrm{AMR}^{*}$ procedure are given by:

$$
\begin{equation*}
\left\lceil\frac{54}{\varepsilon^{2}} V\left(\mu_{j^{*}}\right)\right\rceil\lceil 9 \log (1 / \delta)\rceil=O\left(\frac{1}{\alpha^{2}} V\left((\mu)_{\tau}\right) \cdot \log \left(\frac{\log (1 / \tau)}{\chi \alpha}\right)\right) \tag{4.5}
\end{equation*}
$$

### 4.1.2 Case $\mu<\tau$

When $\mu<\tau$ we are always in the case where $\mu_{i} \geqslant(1-(c+\varepsilon))^{-1} \mu$ and by Lemma 4.3 the probability that the algorithm terminates before step $\log (\tau /(1-(c+\varepsilon))) / \log (1-\gamma)$ is upper bounded by $\log (\tau /(1-(c+\varepsilon))) / \log (1-\gamma) \delta \leqslant \chi$. Thus, with probability at least $1-\chi$ we have $\hat{Z}=0$.

### 4.2 Proof of Lemma 4.3

To prove the lemma, we are going to show that the event $\left|Z_{i}-\mu_{i}\right| \leqslant c \mu_{i}$ is related to an event that expresses the deviation of $Z_{i}$ from its true mean. We will then appeal to concentration of measure for the median-of-means estimator to show the required bounds. We have that for all $\mu_{i} \geqslant(1-(c+\varepsilon))^{-1} \mu$

$$
\left|Z_{i}-\mu_{i}\right| \leqslant c \mu_{i} \Rightarrow\left|Z_{i}-\mu\right| \geqslant\left[\left(1-\frac{\mu}{\mu_{i}}\right)-c\right] \mu_{i} \Rightarrow\left|Z_{i}-\mu\right| \geqslant \varepsilon \mu_{i}
$$

By monotonicity this implies that $\mathbb{P}\left[\left|Z_{i}-\mu_{i}\right| \leqslant c \mu_{i}\right] \leqslant \mathbb{P}\left[\left|Z_{i}-\mu\right| \geqslant \varepsilon \mu_{i}\right]$. But $Z_{i}=$ $\operatorname{MoM}_{\frac{\varepsilon}{3}, \delta}\left(v, V\left(\mu_{i}\right)\right)$, is the median of $L(\delta)=\lceil 9 \log (1 / \delta)\rceil$, averages of $54 V\left(\mu_{i}\right) / \varepsilon^{2}$ samples. We compute the probability that any of the $L$ averages deviates from the true mean:

$$
\mathbb{P}\left[|\tilde{Z}-\mu| \geqslant \varepsilon \mu_{i}\right] \leqslant \frac{\mu^{2} \cdot V(\mu)}{\varepsilon^{2} \mu_{i}^{2} \frac{54 V\left(\mu_{i}\right)}{\varepsilon^{2}}} \leqslant \frac{1}{54} \cdot \frac{\mu^{2} V(\mu)}{\mu_{i}^{2} V\left(\mu_{i}\right)} \leqslant \frac{1}{54}
$$

where the last step follows by monontonicity of $\mu^{2} V(\mu)$ and the assumption that $\mu_{i} \geqslant(1-(c+\varepsilon))^{-1} \mu$. Then, by Chernoff bounds (Appendix ) we get that the probability that the median of such averages has the deviation in question is less than $\delta$. Similarly, we have that in the case that $\frac{1+\frac{\varepsilon}{3}}{1+c} \mu \leqslant \mu_{i} \leqslant \mu$ :

$$
\begin{equation*}
\left|Z_{i}-\mu\right| \leqslant \frac{\varepsilon}{3} \mu \Rightarrow\left|Z_{i}-\mu_{i}\right| \leqslant c \mu_{i} \tag{4.6}
\end{equation*}
$$

As before let $\tilde{Z}$ denote any of the $L$ averages of which we output the median. Then,

$$
\mathbb{P}\left[\left|\tilde{Z}-\mu_{i}\right| \leqslant c \mu_{i}\right] \geqslant \mathbb{P}\left[|\tilde{Z}-\mu| \leqslant \frac{\varepsilon}{3} \mu\right] \geqslant 1-\mathbb{P}\left[|\tilde{Z}-\mu| \geqslant \frac{\varepsilon}{3} \mu\right] \geqslant 1-\frac{9 \mu^{2} V(\mu)}{\varepsilon^{2} \mu^{2} \frac{54 V\left(\mu_{i}\right)}{\varepsilon^{2}}} \geqslant 1-\frac{1}{6} \frac{V(\mu)}{V\left(\mu_{i}\right)} \geqslant \frac{5}{6}
$$

Again, by Chernoff bounds we get that the probability of the deviation for the median is less than $\delta$.

## 5 KDE through Euclidean LSH

In this section, we instantiate the framework of HBE for the problem of KDE using Euclidean LSH of Datar et al. [15]. At a high level, given a kernel $k$, the goal is to design a hashing scheme such that the probability that two points hash at the same bucket is as similar as possible to $k(x, y)$. We consider three such kernels: the Exponential, the Generalized $t$-Student (polynomial) and the Gaussian kernel. We show that this specific LSH scheme can be used to construct scalefree estimators for the first two, while for the Gaussian case, although we cannot get a scale-free estimator, we are still able to analyze the variance of the estimator using Theorem 1.3.

### 5.1 Euclidean LSH

The family of hash functions is given by:

$$
\mathcal{H}_{1}(w):=\left\{h(x)=\left\lceil\frac{g^{\top} x+\beta}{w} \| g \in \mathbb{R}^{d}, \beta \in[0, w]\right\}\right.
$$

for some fixed $w>0$. We define a distribution $v_{1}$ over $\mathcal{H}_{1}$ by sampling $g \sim \mathcal{N}\left(0, I_{d}\right)$ and $\beta \sim U[0, w]$. The important quantity to control is the collision probability $p_{1}(c):=\mathbb{P}_{h \sim v}[h(x)=h(y)]$ of two points $x, y$ at distance $\|x-y\|=c w$ and is given by [15]:

$$
\begin{equation*}
p_{1}(c)=1-2 \Phi\left(c^{-1}\right)-\sqrt{\frac{2}{\pi}} c\left(1-\exp \left\{-\frac{c^{-2}}{2}\right\}\right) \tag{5.1}
\end{equation*}
$$

Lemma 5.1 (Pointwise bounds). For all $c>0$

$$
\begin{equation*}
p_{1}(c)=\sqrt{\frac{2}{\pi}} \sum_{k=0}^{\infty} \frac{(-1)^{k}}{2^{k} k!(2 k+2)(2 k+1)} \frac{1}{c^{2 k+1}} \tag{5.2}
\end{equation*}
$$

while for $\delta \leqslant \frac{1}{2}$ and $c \leqslant \min \left\{\delta, \frac{1}{\sqrt{2 \ln (1 / \delta)}}\right\}$

$$
\begin{equation*}
e^{-\sqrt{\frac{2}{\pi}}(1+\delta) \cdot c} \leqslant p_{1}(c) \leqslant e^{-\sqrt{\frac{2}{\pi}}\left(1-\delta^{3}\right) \cdot c} \tag{5.3}
\end{equation*}
$$

We see that this family of hash functions gives collision probability that is exponentially decreasing with the distance for small values of $c \ll 1$ whereas the decay is inverse polynomial for large values values of $c \gg 1$. We exploit these properties to derive scale free estimators for the Exponential and $t$-Student kernels.

### 5.2 Scale-free Estimators for Exponential and $t$-Student Kernel

Theorem 5.2 (Exponential Kernel). For $\beta \in(0,1]$ there exists $a(\beta, \sqrt{e})$-scale free $H B E$ for the exponential kernel $e^{-\|x-y\|}$ that has complexity $O\left(d R^{2}\right)$.
Proof. Set $D=\lceil\sqrt{2 \pi} R\rceil^{2}, w=\frac{D}{\beta \sqrt{\frac{\pi}{2}}}$ and consider the HBE resulting from the family $\mathcal{H}_{1}^{\otimes D}(w)$ with probability measure $v_{1}^{\otimes D}$. We first see that for a pair of points that are distance $\|x-y\|=r$ apart, we have that $c \leqslant \delta=\frac{R}{w}=\beta \sqrt{\frac{\pi}{2}} \frac{R}{[\sqrt{2 \pi} R\rceil^{2}} \leqslant \frac{\beta}{2} \frac{1}{\lceil\sqrt{2 \pi} R\rceil} \leqslant \frac{1}{2}$. Additionally, for these parameters it holds that $e^{-\sqrt{\frac{2}{\pi}} c D}=e^{-\beta \cdot r}$. Using the second part of Lemma 5.1] with $\delta=\beta \sqrt{\frac{\pi}{2}} \frac{R}{[\sqrt{2 \pi} R]^{2}}$ we get

$$
\begin{equation*}
e^{-\sqrt{\frac{2}{\pi}}\left(\frac{\beta}{2} \frac{1}{[\sqrt{2 \pi} R]}\right)^{2} D} \leqslant \frac{p_{1}^{D}(c)}{e^{-\beta \cdot r}} \leqslant e^{+\sqrt{\frac{2}{\pi}}\left(\frac{\beta}{2} \frac{1}{[\sqrt{2 \pi} R]}\right)^{4} D} \tag{5.4}
\end{equation*}
$$

Which gives us that $\frac{1}{\sqrt{e}} \cdot e^{-\beta r} \leqslant p_{1}^{D}(c) \leqslant \sqrt{e} \cdot e^{-\beta r}$. To implement the estimator we require $O\left(R^{2}\right)$ hash functions that each can be evaluate in time $O(d)$ and requires space $O(d+n)$.

Although, this is to be expected given the derived pointwise bounds on the collision probabilities, and seems to suggest a correspondence between kernels and hashing schemes, we show next that one can construct a scale-free estimator for another very different kernel using the same LSH scheme.

Theorem 5.3 (Generalized $t$-Student Kernel). For integers $p, q \geqslant 1$, there exists a $\left(\frac{q}{p}, 3^{q}\right)$-scale free $H B E$ for the kernel $\frac{1}{1+\|x-y\|^{p}}$ that has complexity $T=O(d p)$.

Proof. Set $w=\sqrt{2 \pi}$ and consider the HBE resulting from the family $\mathcal{H}_{1}^{\otimes p}(w)$. We obtain bounds for the collision probabilities for two points at distance $\|x-y\|=r=c \cdot w$ apart. For $c>1$, using the first part of Lemma 5.1 and dropping terms appropriately

$$
\begin{equation*}
\frac{1}{\sqrt{2 \pi}}\left(1-\frac{1}{12 c^{2}}\right) \frac{1}{c} \leqslant p_{1}(c) \leqslant \frac{1}{\sqrt{2 \pi}} \frac{1}{c} \tag{5.5}
\end{equation*}
$$

Setting $\beta=\frac{q}{p}$ and raising to the $q$-th power we get that for all $c>1$ :

$$
\begin{equation*}
\left(\frac{11}{12}\right)^{q} \frac{\left(1+r^{p}\right)^{\beta}}{r^{q}} \leqslant \frac{p_{1}^{q}(c)}{\frac{1}{\left(1+r^{p}\right)^{\beta}}} \leqslant \frac{\left(1+r^{p}\right)^{\beta}}{r^{q}} \tag{5.6}
\end{equation*}
$$

whereas for $c \leqslant 1$ we have from (5.5) and monotonicity of $p_{1}(c)$ :

$$
\begin{equation*}
\left(\frac{11}{12 \sqrt{2 \pi}}\right)^{q}\left(1+r^{p}\right)^{\beta} \leqslant \frac{p_{1}^{q}(c)}{\frac{1}{\left(1+r^{p}\right)^{\beta}}} \leqslant\left(1+r^{p}\right)^{\beta} \tag{5.7}
\end{equation*}
$$

Since $2^{-1} \geqslant 11 /(12 \sqrt{2 \pi}) \geqslant 3^{-1}$ this shows that the resulting estimator is $\left(\frac{q}{p}, 3^{q}\right)$-scale-free for the Generalized $t$-student Kernel for all distances.

These two results combined with Theorem 3.4 and Theorem 4.2 can be used to construct a data structure for the KDE problem under the exponential or $t$-Student kernel. More importantly, this is done effortlessly by appealing to the the general case of HBE and only required showing that our hashing schemes produces scale-free estimators.

### 5.3 An estimator for Gaussian KDE

We show next that the framework of HBE can be useful beyond the ideal scenarios where a scalefree estimator can be derived. We do so by showing that one can use the exponential drop-off of the collision probabilities to simulate the Gaussian kernel and then appeal to the more general Theorem 1.3 to bound its variance.

Theorem 5.4. For any $t \in[1, R]$ there exists a $H B E Z_{t}$ for the Gaussian kernel $e^{-\|x-y\|^{2}}$ with $\mathbb{E}\left[Z_{t}^{2}\right] \leqslant$ $\mu^{2} \cdot 4 e^{\frac{3}{2}} \mu^{-\gamma^{2}+\gamma-1}$ where $\gamma(t, \mu):=t / \sqrt{\log (1 / \mu)}$, that has complexity $T=O\left(d t^{2} R^{2}\right)$.

Proof. Fix $t \geqslant 1$ and set $D=3\lceil t R\rceil^{2}, w=\frac{D}{t} \sqrt{\frac{2}{\pi}}$. We consider once more the HBE resulting from $\mathcal{H}_{1}^{D}(w)$ and probability measure $v_{1}^{\otimes D}$. For a pair of points at distance $r=c \cdot w \leqslant R$ we have that $c \leqslant \delta=\frac{R}{w} \leqslant \frac{t R}{3[t R]^{2}} \sqrt{\frac{\pi}{2}} \leqslant \frac{1}{2}$. By utilizing Lemma 5.1 one again we obtain $\frac{1}{\sqrt{e}} \cdot e^{-r \cdot t} \leqslant p_{1}^{D}(c) \leqslant$ $\sqrt{e} \cdot e^{-r \cdot t}$. To bound the variance due to Theorem 1.3 we only need to consider what happens for datasets supported only on two points. It will be useful to consider that one point is at distance
$r_{1}=\sqrt{\alpha \log (1 / \mu)}$ away from the query and the other at $r_{2}=\sqrt{\alpha^{\prime} \log (1 / \mu)}$ with $0 \leqslant \alpha \leqslant \alpha^{\prime}$. Setting $\gamma(t, \mu)=\frac{t}{\sqrt{\log (1 / \mu)}}$ the pointwise bounds become

$$
\begin{equation*}
\frac{1}{\sqrt{e}} \cdot \mu^{\gamma \sqrt{\alpha}} \leqslant p_{1}^{D}(c) \leqslant \sqrt{e} \cdot \mu^{\gamma \sqrt{\alpha}} \tag{5.8}
\end{equation*}
$$

The weights (contribution) of each point in this case are given by $w_{1}=\mu^{\alpha}$ and $w_{2}=\mu^{\alpha^{\prime}}$. To bound the variance we pefrom a case analysis:

- Case $\alpha \leqslant \alpha^{\prime} \leqslant 1$ : we first bound the term

$$
\begin{equation*}
\tilde{A}_{1,2}=\frac{\mu}{w_{1}} A_{1,2} \frac{\mu}{w_{2}}=\mu^{2} \frac{w_{1}}{p_{1}^{2}} \frac{p_{2}}{w_{2}} \leqslant \mu^{2} e^{3 / 2} \cdot \mu^{\alpha-2 \gamma \sqrt{\alpha}-\alpha^{\prime}+\sqrt{\alpha^{\prime}} \gamma} \leqslant \mu^{2} e^{3 / 2} \cdot \mu^{-\gamma^{2}+\gamma-1} \tag{5.9}
\end{equation*}
$$

where in the last step we optimized for $0 \leqslant \alpha \leqslant \alpha^{\prime} \leqslant 1$ getting values $\alpha^{\prime}=1$ and $\alpha=\gamma^{2}$.

$$
\begin{equation*}
\tilde{A}_{2,1}=\frac{\mu}{w_{2}} A_{2,1} \frac{\mu}{w_{2}}=\mu^{2} \frac{w_{2}}{p_{2} w_{1}} \leqslant \mu^{2} \sqrt{e} \cdot \mu^{\alpha^{\prime}-\gamma \sqrt{\alpha^{\prime}}-\alpha} \leqslant \mu^{2} \sqrt{e} \cdot \mu^{-\gamma} \tag{5.10}
\end{equation*}
$$

where in the last step we set $\alpha=\alpha^{\prime}=1$. The same bound applies to the term $\tilde{A}_{2,2}$. For completeness, we also bound:

$$
\begin{equation*}
\tilde{A}_{1,1}=\left(\frac{\mu}{w_{1}}\right)^{2} A_{1,1}=\frac{1}{p_{1}} \mu^{2} \leqslant \mu^{2} \sqrt{e} \cdot \mu^{-\gamma} \tag{5.11}
\end{equation*}
$$

Hence, the variance is bounded by $\mu^{2} 4 e^{3 / 2} \cdot \mu^{-\gamma^{2}+\gamma-1}$.

- Case $\alpha \leqslant 1<\alpha^{\prime}$ : we only need to bound the cross-terms

$$
\begin{equation*}
\tilde{A}_{1,2}=\frac{\mu}{w_{1}} \cdot A_{1,2} \cdot 1=\mu \frac{w_{1} p_{2}}{p_{1}^{2}} \leqslant \mu^{2} e^{3 / 2} \cdot \mu^{\alpha-2 \gamma \sqrt{\alpha}+\sqrt{\alpha^{\prime}} \gamma-1}<\mu^{2} e^{3 / 2} \cdot \mu^{-\gamma^{2}+\gamma-1} \tag{5.12}
\end{equation*}
$$

where again we optimized for $\alpha \leqslant 1<\alpha^{\prime}$. The other term is

$$
\begin{equation*}
\tilde{A}_{2,1}=1 \cdot A_{2,1} \frac{\mu}{w_{1}}=\mu \frac{w_{2}^{2}}{p_{2} w_{1}} \leqslant \mu \sqrt{e} \cdot \mu^{2 \alpha^{\prime}-\gamma \sqrt{\alpha^{\prime}}-\alpha} \leqslant \mu^{2} \sqrt{e} \cdot \mu^{-\gamma} \tag{5.13}
\end{equation*}
$$

where in the last step we set $\alpha=\alpha^{\prime}=1$. Additionally,

$$
\begin{equation*}
\tilde{A}_{1,1}=\left(\frac{\mu}{w_{1}}\right)^{2} A_{1,1}=\frac{1}{p_{1}} \mu^{2} \leqslant \mu^{2} \sqrt{e} \cdot \mu^{-\gamma} \tag{5.14}
\end{equation*}
$$

Hence, in this case the variance is bounded by $\mu^{2} 4 e^{3 / 2} \cdot \mu^{-\gamma^{2}+\gamma-1}$.

- Case $1<\alpha \leqslant \alpha^{\prime}$ : following the same logic we get that the variance is bounded by $\mu^{2} 4 e^{3 / 2} \mu^{-\gamma}$.

Thus, overall we see that the worst-case distances of the two points are given by $\alpha=\gamma^{2}$ and $\alpha^{\prime}=1$. For which, the upper bound on the second moment becomes $\mu^{2} \cdot 4 e^{\frac{3}{2}} \cdot \mu^{-\gamma^{2}+\gamma-1}$.

Even though that we were not able to get the ideal behavior using this hashing scheme for the Gaussian Kernel, still the estimator has improved variance compared to random sampling for all $0<\gamma<1$ and achieves it's best performance when $t=\frac{1}{2} \sqrt{\log (1 / \mu)} \Rightarrow \gamma=\frac{1}{2}$.

### 5.4 Pointwise Bounds for Euclidean LSH

Proof of Lemma 5.1. To prove the first part we are going to use Taylor expansion of the exponential:

$$
\begin{align*}
1-2 \Phi\left(\rho^{-1}\right) & =\int_{-\rho^{-1}}^{\rho^{-1}} \frac{1}{\sqrt{2 \pi}} \exp \left\{-\frac{t^{2}}{2}\right\} d t  \tag{5.15}\\
& =\frac{1}{\sqrt{2 \pi}} \sum_{k=0}^{\infty} \frac{(-1)^{k}}{2^{k} k!} \int_{-\rho^{-1}}^{\rho^{-1}} t^{2 k} d t  \tag{5.16}\\
& =\sqrt{\frac{2}{\pi}} \sum_{k=0}^{\infty} \frac{(-1)^{k}}{(2 k+1) 2^{k} k!} \rho^{-(2 k+1)} \tag{5.17}
\end{align*}
$$

Similarly we have that $\sqrt{\frac{2}{\pi}} \rho \exp \left\{-\rho^{-2} / 2\right\}=\sqrt{\frac{2}{\pi}} \sum_{k=0}^{\infty} \frac{(-1)^{k}}{2^{k} k!} \rho^{-(2 k-1)}$. Putting it all together we get:

$$
C(\rho)=\sqrt{\frac{2}{\pi}} \sum_{k=0}^{\infty} \frac{(-1)^{k}}{2^{k} k!(2 k+2)(2 k+1)} \rho^{-(2 k+1)}
$$

To prove the second part of the lemma we use the following bounds for gaussian tails.
Lemma 5.5 ([20]). Let $X \sim \mathcal{N}(0,1)$ and $\Phi(x):=\mathbb{P}[X \geqslant x]$ for all $x>0$, then:

$$
\begin{equation*}
\frac{1}{\sqrt{2 \pi}} \frac{1}{x}\left\{1-\frac{1}{x^{2}}\right\} e^{-\frac{x^{2}}{2}} \leqslant \Phi(x) \leqslant \frac{1}{\sqrt{2 \pi}} \frac{1}{x}\left\{1-\left(\frac{1}{x^{2}}-\frac{1}{x^{4}}\right)\right\} e^{-\frac{x^{2}}{2}} \tag{5.18}
\end{equation*}
$$

Starting from (5.1) and using Lemma 5.5 for $x=c^{-1}$ we arrive at:

$$
\begin{equation*}
1-\sqrt{\frac{2}{\pi}} c\left(1-c^{2}\left(1-3 c^{2}\right) e^{-c^{-2} / 2}\right) \leqslant p_{1}(c) \leqslant 1-\sqrt{\frac{2}{\pi}} c\left(1-c^{2} e^{-c^{-2} / 2}\right) \tag{5.19}
\end{equation*}
$$

For the upper bound, the assumption that $c \leqslant \min \left\{\frac{1}{\sqrt{2 \ln (1 / \delta)}}, \delta\right\}$ implies that $1-c^{2} e^{-c^{-2} / 2} \geqslant 1-\delta^{3}$ and consequently $p_{1}(c) \leqslant e^{-\sqrt{\frac{2}{\pi}}\left(1-\delta^{3}\right) c}$. For the lower bound we use the fact that the function $f(c)=c^{2}\left(1-3 c^{2}\right) e^{-c^{2} / 2}$ is decreasing for all $c \in[0,1 / 2]$ to obtain $p_{1}(c) \geqslant 1-\sqrt{\frac{2}{\pi}} c$. The lower bound then follows by the inequalities $1-x \geqslant e^{-\left(1+\frac{\varepsilon}{2}\right) x}$ for $x \in\left[0, \frac{\varepsilon}{1+\varepsilon}\right]$ and $\sqrt{\frac{2}{\pi}} c \leqslant \delta \leqslant \frac{2 \delta}{1+2 \delta}$ for $\delta \leqslant 1 / 2$.

## 6 Scale-free estimator based on Andoni-Indyk LSH

Our framework of scale-free esimators is a natural desideratum when trying to approximate the ideal function for importance sampling. In this section, we show how to use the "Ball-Carving" LSH introduced by Andoni and Indyk [4] for Euclidean distance to get a scale-free estimator for the Gaussian Kernel.

### 6.1 Ball-Carving LSH

Andoni and Indyk [4] introduced a family of hash functions $\mathcal{H}_{t}(w)$ parametrized by an integer $t \geqslant 2$, and a width $w>0$ such that the evaluation cost is bounded by $U_{t}=d 2^{O(t \log t)} \log n$ and space usage by $O\left(U_{t}+d n\right)$. Their scheme partitions the space by randomly projecting into $t$-dimensions and then carving out balls of radius $w$ centered at random points. We refer to the resulting probability measure as $v_{t}$. Using a similar analysis as the one in [4] we show the following bounds on the collision probability $p_{t}(c):=\mathbb{P}_{h \sim v_{t}}[h(x)=h(y)]$ of two points $x, y$ at distance $c w$.

Lemma 6.1 (Pointwise Bounds). The function $p_{t}(c)$ is non-increasing for all $c \geqslant 0, t \geqslant 1$. Furthermore, for all $t \geqslant 12$ and $\frac{16}{t+7} \leqslant c^{2} \leqslant 1$

$$
\begin{align*}
& p_{t}(c) \geqslant\left\{\frac{1}{4 \sqrt{t}} \frac{1}{c}\left(1-2 e^{-\frac{9 t}{100}}\right) e^{-\frac{t-1}{8} \frac{1}{2-c^{2}} c^{4}}\right\} e^{-\frac{t-1}{8} c^{2}}  \tag{6.1}\\
& p_{t}(c) \leqslant\left\{\frac{3}{\sqrt{t}} \frac{1}{c}\left(1+\frac{\sqrt{t} c}{3} e^{-\frac{9}{64} t}\right) e^{\frac{t-1}{8^{2}} c^{4}}\right\} e^{-\frac{t-1}{8} c^{2}} \tag{6.2}
\end{align*}
$$

Observe that for small $c=O(1 / \operatorname{poly}(\log n))$ and large enough $t=\operatorname{poly}(\log n)$, the dominant term is $e^{-\frac{t-1}{8} c^{2}}$ and drops exponentially with the squared of the distance as desired. However, in order for the time $U_{t}$ to compute the hash function on query to be $n^{o(1)}$, we must have $t=o(\log (n))$. This is the main bottleneck that complicates a bit the application of the Andoni-Indyk hashing scheme for our purposes.

### 6.2 Scale-free estimator for Gaussian Kernel

Theorem 6.2 (Gaussian Kernel). For all $\beta \in(0,1]$ there exists $a\left(\beta, e^{O\left(R^{\frac{4}{3}} \log \log n\right)}\right)$-scale free HBE for the Gaussian kernel $e^{-\|x-y\|^{2}}$ that has complexity $T=e^{O\left(R^{\frac{4}{3}} \log \log n\right)}$.

Proof. Set $t:=\max \left\{\left\lceil R^{\frac{4}{3}}\right\rceil, 12\right\}, D=\left\lceil\frac{8 \sqrt{t} R^{2}}{t-1}\right\rceil, w=\sqrt{\frac{t-1}{8 \beta} D} \geqslant \sqrt[4]{t} R$ and consider the HBE resulting from using the family $\mathcal{H}_{t}^{D}(w)$. The HBE can be evaluated in time $D 2^{O(t \log \log (n))}$ and takes space $D 2^{O(t \log \log (n))} \cdot n$. Next, we will show that for this selection of parameters we get a scale-free estimator for the Gaussian kernel. Consider two points at distance $r=c \cdot w$. First, we see that $c=\frac{r}{w} \leqslant \frac{R}{w} \leqslant \frac{1}{\sqrt[4]{t}} \leqslant 1$. For distances $r \geqslant \frac{4}{\sqrt[4]{t}} R \Rightarrow c^{2} \geqslant \frac{16}{t}$, our selection of $D$ results in the dominant term of the bounds given by Lemma 6.1] being $e^{-\frac{t-1}{8 w^{2}} D \cdot r^{2}}=e^{-\beta r^{2}}$. Moreover,

$$
\begin{equation*}
e^{-\Theta(D \log (t))} e^{-\beta r^{2}} \leqslant p_{t}^{D}(c) \leqslant e^{\Theta(D \log (t))} e^{-\beta r^{2}} \tag{6.3}
\end{equation*}
$$

Thus, we see that for the range $4 R / \sqrt[4]{t} \leqslant r \leqslant R$ we have a scale-free estimator with $M=e^{O(D \log (t))}$. To get a bound for $0 \leqslant r \leqslant R / \sqrt[4]{t}$ we use monotonicity of $p_{t}(c)$ to obtain:

$$
\begin{equation*}
M^{-1} \cdot e^{-16 \beta \frac{R^{2}}{\sqrt{t}}} \leqslant p_{t}^{D}(c) \leqslant e^{\beta r^{2}} \cdot e^{-\beta r^{2}} \tag{6.4}
\end{equation*}
$$

Since $D=\Theta\left(R^{2} / \sqrt{t}\right)$ we see that we have constructed an $\left(\beta, e^{O\left(\log \log (n) R^{2} / \sqrt{t}\right)}\right.$-scale free estimator. Our selection of $t=R^{4 / 3}$ balances the complexity of evaluating the hashing function with the deviation from the ideal collision probabilities.

The above theorem shows that as long as $R=O\left(\log ^{\gamma}(n)\right)$ with $0<\gamma<\frac{3}{4}$ the estimator is ( $\beta, n^{o(1)}$ )-scale free. The regime of most interest is when $\gamma=1 / 2$, where polynomially small values of $\mu=n^{-\Omega(1)}$ are permissible. In this case, our estimator is $\left(\beta, e^{O\left(\log \frac{2}{3}(n) \log \log n\right)}\right)$-scale free.

### 6.3 Pointwise bounds for Andoni-Indyk LSH

Let $I(r, w)$ be ratio between the cap volume at relative distance $\alpha:=r / w$ from the origin of a $d$-dimensional sphere relative of radius $w$ and the volume of the sphere.

Lemma 6.3 (Corollary $3.2[11])$. Let $\alpha:=\frac{r}{w} \in[0,1]$,
(i) for $\alpha \geqslant \frac{1}{\sqrt{d+1}}$ we have

$$
\begin{equation*}
\frac{\sqrt{d}}{3 \sqrt{d+1}} \frac{1}{\sqrt{2 \pi d}} \frac{1}{\alpha}\left(1-\alpha^{2}\right)^{d / 2} \leqslant I(r, w) \leqslant \frac{1}{\sqrt{2 \pi d}} \frac{1}{\alpha}\left(1-\alpha^{2}\right)^{d / 2} \tag{6.5}
\end{equation*}
$$

(ii) for $\alpha \leqslant \frac{1}{\sqrt{d+1}}, \frac{1}{2 e \sqrt{2 \pi}} \leqslant I(r, w) \leqslant \frac{1}{2}$

We also require the following lemma on concentration properties of $\chi^{2}$-random variables.
Lemma 6.4. Let $Z_{1}, \ldots, Z_{t}$ be independent $\chi^{2}$-random variables, then:
(i) $\mathbb{P}\left[\sum_{i=1}^{t} Z_{i} \leqslant \alpha \cdot t\right] \leqslant e^{-(\alpha+\ln (1 / \alpha)-1) \frac{t}{2}}$ for all $\alpha<1$.
(ii) $\mathbb{P}\left[\sum_{i=1}^{t} Z_{i}>\alpha \cdot t\right] \leqslant e^{-(a-\ln (a)-1) \frac{t}{2}}$ for all $\alpha>1$.

Proof of Lemma6.1 To bound the collision probability we follow the analysis of Andoni and Indyk [4]. Given two points $p, q \in \mathbb{R}^{d}$ that are within distance $c \cdot w$ find the probability that they hash to the same hash bucket.

$$
\begin{align*}
p_{t}(c) & =\int_{0}^{\infty} \mathbb{P}\left[h(p)=h(q) \left\lvert\,\left\|p^{\prime}-q^{\prime}\right\|=\sqrt{\frac{x}{t}} c w\right.\right] P_{\chi_{t}^{2}}(x) d x  \tag{6.6}\\
& =\int_{0}^{\frac{4 t}{c^{2}}} P_{\chi_{t}^{2}}(x) \frac{I\left(\frac{1}{2} \sqrt{\frac{x}{t}} c w, w\right)}{1-I\left(\frac{1}{2} \sqrt{\frac{x}{t}} c w, w\right)} d x  \tag{6.7}\\
& =\int_{0}^{\frac{4}{c^{2}(t+1)} t} P_{\chi_{t}^{2}}(x) \frac{I\left(\frac{1}{2} \sqrt{\frac{x}{t}} c w, w\right)}{1-I\left(\frac{1}{2} \sqrt{\frac{x}{t}} c w, w\right)} d x+\int_{\frac{4}{c^{2}(t+1)} t}^{\frac{4}{c^{2}} t} P_{\chi_{t}^{2}}(x) \frac{I\left(\frac{1}{2} \sqrt{\frac{x}{t}} c w, w\right)}{1-I\left(\frac{1}{2} \sqrt{\frac{x}{t}} c w, w\right)} d x \tag{6.8}
\end{align*}
$$

where $P_{\chi_{t}^{2}}(x)=\frac{x^{t / 2-1}}{\Gamma(t / 2))^{t / 2}} e^{-x / 2}$ is the density function of a $\chi^{2}$-random variable with $t$-degrees of freedom. We obtain upper and lower bounds for each term separately. Let $A$ and $B$ denote the first and second term respectively. Using Lemma 6.3 and Lemma 6.4 for $\alpha=\frac{4}{c^{2}} \frac{1}{t+1}$, we get the following bounds for the first term:

$$
0 \leqslant A \leqslant\left\{\begin{array}{cc}
e^{-\frac{1}{4} t} & \text { for } c^{2} \geqslant \frac{16}{t+1}  \tag{6.9}\\
1 & \text { otherwise }
\end{array}\right.
$$

### 6.3.1 Upper bound

Using the fact that $I(r, w) \leqslant \frac{1}{2}$ for all $0 \leqslant r \leqslant w$ and Lemma 6.3 we get

$$
\begin{align*}
B & \leqslant 2 \int_{\frac{4}{c^{2}(t+1)} t}^{\frac{4 t}{c^{2}}} P_{\chi_{t}^{2}}(x) \cdot I\left(\frac{1}{2} \sqrt{\frac{x}{t}} c w, w\right) d x  \tag{6.10}\\
& \leqslant 2 \int_{\frac{4}{c^{2}(t+1)} t}^{\frac{4 t}{c^{2}}} P_{\chi_{t}^{2}}(x) \cdot \frac{1}{\sqrt{2 \pi t}} \frac{1}{\frac{1}{2} \sqrt{\frac{x}{t}} c}\left(1-\left(\frac{1}{2} \sqrt{\frac{x}{t}} c\right)^{2}\right)^{t / 2} d x  \tag{6.11}\\
& \leqslant \frac{4}{\sqrt{2 \pi} c} \int_{\frac{4}{c^{2}(t+1)} t}^{\frac{4 t}{c^{2}}} P_{\chi_{t}^{2}}(x) \cdot \frac{1}{\sqrt{x}} \exp \left(-\frac{1}{2} x \frac{c^{2}}{4}\right) d x \tag{6.12}
\end{align*}
$$

$$
\begin{align*}
& =\frac{4}{\sqrt{2 \pi} c} \int_{\frac{4}{c^{2}(t+1)}}^{\frac{4 t}{c^{2}}} t \frac{x^{t / 2-1}}{\Gamma(t / 2) 2^{t / 2}} e^{-\frac{x}{2}} \cdot \frac{1}{\sqrt{x}} \exp \left(-\frac{1}{2} x \frac{c^{2}}{4}\right) d x  \tag{6.13}\\
& \leqslant \frac{2 \Gamma((t-1) / 2)}{\sqrt{\pi} c \Gamma(t / 2)} \frac{1}{\left(1+\frac{c^{2}}{4}\right)^{(t-1) / 2}} \int_{0}^{\frac{4 t}{c^{2}\left(c^{2} / 4+1\right)}} P_{\chi_{t-1}^{2}}(x)  \tag{6.14}\\
& \leqslant \frac{2 \Gamma((t-1) / 2)}{\sqrt{\pi} c \Gamma(t / 2)} \frac{1}{\left(1+\frac{c^{2}}{4}\right)^{(t-1) / 2}}  \tag{6.15}\\
& \leqslant \frac{8 \cdot 2}{7 \sqrt{\pi \frac{t-1}{2}} c} \frac{1}{\left(1+\frac{c^{2}}{4}\right)^{(t-1) / 2}} \tag{6.16}
\end{align*}
$$

where we used the fact that $\frac{8}{7} \frac{1}{\sqrt{\alpha}} \geqslant \frac{\Gamma(\alpha)}{\Gamma\left(\alpha+\frac{1}{2}\right)} \geqslant \frac{1}{\sqrt{\alpha}}$ for $\alpha \geqslant 1$. Further, due to the inequalities $1-x \leqslant \frac{1}{1+x} \leqslant e^{-\frac{2 x}{2+x}}$ for $x \geqslant 0$ we get:

$$
\begin{equation*}
\frac{1}{\left(1+\frac{c^{2}}{4}\right)^{\frac{t-1}{2}}} \leqslant e^{-\frac{t-1}{2} \frac{2}{2+\frac{c^{2}}{4}} \frac{c^{2}}{4}}=e^{-\frac{t-1}{8} c^{2} \frac{1}{1+c^{2} / 8}} \leqslant e^{-\frac{t-1}{8}\left(1-\frac{c^{2}}{8}\right) c^{2}} \leqslant e^{+\frac{t-1}{64} c^{4}} \cdot e^{-\frac{t-1}{8} c^{2}} \tag{6.17}
\end{equation*}
$$

Substituting the last bound back to (6.16) and using $\frac{16 \sqrt{2}}{7 \sqrt{\pi}} \frac{1}{\sqrt{t-1}} \leqslant \frac{3}{\sqrt{t}}$ for $t \geqslant 2$, results in

$$
\begin{align*}
p_{t}(c) & \leqslant e^{-\frac{t}{4}}+\frac{3}{\sqrt{t} c} e^{\frac{t-1}{4} c^{4}} \cdot e^{-\frac{t-1}{8} c^{2}}  \tag{6.18}\\
& \leqslant\left(1+\frac{\sqrt{t} c}{3} e^{-\frac{t-1}{64} c^{4}+\frac{t-1}{8} c^{2}-\frac{t}{4}}\right) \frac{3}{\sqrt{t} c} e^{\frac{t-1}{64} c^{4}} \cdot e^{-\frac{t-1}{8} c^{2}} \tag{6.19}
\end{align*}
$$

Noting that for $c \leqslant 1$ the exponent in the parenthesis is at most $e^{-9 t / 64}$ results in the desired bound.

### 6.3.2 Lower Bound

Using the fact that $1-I(r, w) \leqslant 1$ for all $0 \leqslant r \leqslant w$ and Lemma 6.3 we get:

$$
\begin{align*}
B & \geqslant \int_{\frac{4}{c^{2}(t+1)} t}^{\frac{4 t}{c^{2}}} P_{\chi_{t}^{2}}(x) \cdot I\left(\frac{1}{2} \sqrt{\frac{x}{t}} c, t\right) d x  \tag{6.20}\\
& \geqslant \frac{\sqrt{t}}{3 \sqrt{t+1}} \int_{\frac{4}{c^{2}(t+1)} t}^{\frac{4 t}{c^{2}}} P_{\chi_{t}^{2}}(x) \cdot \frac{1}{\sqrt{2 \pi t}} \frac{1}{\frac{1}{2} \sqrt{\frac{x}{t}} c}\left(1-\left(\frac{1}{2} \sqrt{\frac{x}{t}} c\right)^{2}\right)^{t / 2} d x  \tag{6.21}\\
& =\frac{\sqrt{t}}{3 \sqrt{t+1}} \frac{2}{\sqrt{2 \pi} c} \int_{\frac{4}{c^{2}(t+1)} t}^{\frac{4 t}{c^{2}}} P_{\chi_{t}^{2}}(x) \cdot \frac{1}{\sqrt{x}}\left(1-\frac{x}{4 t} c^{2}\right)^{t / 2} d x  \tag{6.22}\\
& =\frac{\sqrt{t}}{3 \sqrt{t+1}} \frac{2}{\sqrt{2 \pi} c} \int_{\frac{4}{c^{2}(t+1)} t}^{\frac{4 t}{c^{2}}} P_{\chi_{t}^{2}}(x) \cdot \frac{1}{\sqrt{x}}\left(\frac{1}{1+\frac{\frac{x}{4 t} c^{2}}{1-\frac{x}{4 t} c^{2}}}\right)^{t / 2} d x  \tag{6.23}\\
& \geqslant \frac{\sqrt{t}}{3 \sqrt{t+1}} \frac{2}{\sqrt{2 \pi} c} \int_{\frac{4}{c^{2}(t+1)} t}^{\frac{4 t}{c^{2}}} P_{\chi_{t}^{2}}(x) \cdot \frac{1}{\sqrt{x}} \exp \left(-\frac{t}{2} \frac{\frac{x}{4 t} c^{2}}{1-\frac{x}{4 t} c^{2}}\right) d x  \tag{6.24}\\
& \geqslant \frac{\sqrt{t}}{3 \sqrt{t+1}} \frac{2}{\sqrt{2 \pi} c} \int_{\frac{4}{c^{2}(t+1)} t}^{\frac{R-1}{c^{2}}} P_{\chi_{t}^{2}}(x) \cdot \frac{1}{\sqrt{x}} \exp \left(-\frac{1}{2} x \frac{c^{2}}{4} R\right) d x \tag{6.25}
\end{align*}
$$

$$
\begin{equation*}
=\frac{\sqrt{t}}{3 \sqrt{t+1}} \frac{\sqrt{2} \Gamma((t-1) / 2)}{\sqrt{\pi} c \Gamma(t / 2)} \frac{1}{\left(1+\frac{c^{2}}{4} R\right)^{(t-1) / 2}} \int_{\frac{1}{t+1}\left(R+\frac{4}{c^{2}}\right) t}^{(R-1)\left(1+\frac{4}{R c^{2}}\right) t} P_{\chi_{t-1}^{2}}(x) d x \tag{6.26}
\end{equation*}
$$

where in the penultimate inequality we used $1-x \leqslant 1+R x$ for $x \leqslant(R-1) / R$ and $R>1$. To bound the integral we write:

$$
\begin{equation*}
\int_{\frac{1}{t+1}\left(R+\frac{4}{c^{2}}\right) t}^{(R-1)\left(1+\frac{4}{R c^{2}}\right) t} P_{\chi_{t-1}^{2}}(x) d x=1-\int_{0}^{\frac{1}{t+1}\left(R+\frac{4}{c^{2}}\right) t} P_{\chi_{t-1}^{2}}(x) d x-\int_{\frac{R-1}{R} \frac{4 t}{c^{2}}}^{\infty} P_{\chi_{t-1}^{2}}(x) d x \tag{6.27}
\end{equation*}
$$

We set $R=1+\delta(c)=1+\frac{2}{\frac{4}{c^{2}-2}}$ and require that:

$$
\begin{align*}
(R-1) \frac{4}{R c^{2}} \geqslant 2 & \Rightarrow \delta(c) \geqslant \frac{2}{\frac{4}{c^{2}}-2}  \tag{6.28}\\
\frac{1}{t+1}\left(R+\frac{4}{c^{2}}\right) & \leqslant \frac{1}{2} \Rightarrow \frac{16}{t+3+\sqrt{t^{2}-10 t-7}} \leqslant c^{2} \leqslant \frac{16}{t+3-\sqrt{t^{2}-10 t-7}} \tag{6.29}
\end{align*}
$$

For $t \geqslant 12$ the inequalities are satisfied for all $\frac{16}{t+7} \leqslant c \leqslant 1$. For such $c$, we apply Lemma 6.4 for $\alpha_{>}=2$ and $\alpha_{<}=1 / 2$ and obtain the bound:

$$
\begin{equation*}
\int_{\frac{1}{t+1}\left(R+\frac{4}{c^{2}}\right) t}^{(R-1)\left(1+\frac{4}{R c^{2}}\right) t} P_{\chi_{t-1}^{2}}(x) d x \geqslant 1-2 e^{-\frac{9 t}{100}} \tag{6.30}
\end{equation*}
$$

Next, we lower bound the main term $\left(1+c^{2} R / 4\right)^{-(t-1) / 2}$ using $(1+x)^{-1} \geqslant e^{-x}$ :

$$
\begin{equation*}
\frac{1}{\left(1+c^{2} R / 4\right)^{-(t-1) / 2}} \geqslant e^{-\frac{t-1}{2} \frac{c^{2}}{4}(1+\delta(c))}=e^{\frac{t-1}{\delta}\left(\frac{2}{c^{2}}-1\right)^{-1} c^{2}} \cdot e^{-\frac{t-1}{\delta} c^{2}} \tag{6.31}
\end{equation*}
$$

To complete the proof, we note that $\frac{\sqrt{2 t}}{3 \sqrt{(t+1) \pi}} \frac{\Gamma((t-1) / 2)}{\Gamma(t / 2)} \geqslant \frac{1}{4 \sqrt{t}}$.

### 6.3.3 Concentration of $\chi^{2}$-random variables

Proof of Lemma6.4 The property we use is that for all $\lambda<1 / 2, \mathbb{E}\left[e^{\lambda Z_{i}}\right]=(1-2 \lambda)^{-1 / 2}$. Let $\lambda_{-}=\frac{t}{2}(\alpha-1)>0$ for $\alpha>1$ :

$$
\begin{align*}
\mathbb{P}\left[\sum_{i=1}^{t} Z_{i} \leqslant \alpha t\right] & =\mathbb{P}\left[\frac{1}{t} \sum_{i=1}^{t}\left(\alpha-Z_{i}\right) \geqslant 0\right]  \tag{6.32}\\
& =\mathbb{P}\left[\exp \left(\left\{\sum_{i=1}^{t} \frac{\lambda_{-}}{t}\left(\alpha-Z_{i}\right)\right)\right\} \geqslant 1\right]  \tag{6.33}\\
& \leqslant \mathbb{E}\left[\exp \left(\frac{\lambda_{-}}{t}\left(\alpha-Z_{1}\right)\right)\right]^{t}  \tag{6.34}\\
& =e^{\lambda-\alpha-\frac{t}{2} \log \left(1+2 \frac{\lambda_{-}}{t}\right)}  \tag{6.35}\\
& =e^{-\left(\alpha+\log \left(\frac{1}{\alpha}\right)-1\right) \frac{t}{2}} \tag{6.36}
\end{align*}
$$

Similarly, for $\alpha>1$ and $\lambda_{+}=\frac{t}{2}\left(1-\frac{1}{\alpha}\right)<t / 2$ :

$$
\begin{equation*}
\mathbb{P}\left[\sum_{i=1}^{t} Z_{i} \geqslant \alpha t\right]=\mathbb{P}\left[\frac{1}{t} \sum_{i=1}^{t}\left(Z_{i}-\alpha\right) \geqslant 0\right] \tag{6.37}
\end{equation*}
$$

$$
\begin{align*}
& =\mathbb{P}\left[\exp \left(\left\{\sum_{i=1}^{t} \frac{\lambda_{+}}{t}\left(Z_{i}-\alpha\right)\right)\right\} \geqslant 1\right]  \tag{6.38}\\
& \leqslant \mathbb{E}\left[\exp \left(\frac{\lambda_{+}}{t}\left(Z_{1}-\alpha\right)\right)\right]^{d^{\prime}}  \tag{6.39}\\
& =e^{-\alpha^{2} \lambda_{+}-\frac{t}{2} \log \left(1-2 \frac{\lambda_{+}}{t}\right)}  \tag{6.40}\\
& =e^{-(\alpha-\log (\alpha)-1) \frac{t}{2}} \tag{6.41}
\end{align*}
$$

## 7 Fast Kernel-Matrix Vector Multiplication

Given a kernel function $k: \mathbb{R}^{d} \times \mathbb{R}^{d}$ and a set of points $P=\left\{x_{1}, \ldots, x_{n}\right\}$, let $K=\left\{K\left(x_{i}, x_{j}\right)\right\}_{i, j \leqslant n}$ denote the matrix with the pairwise evaluations of a kernel. Given a vector $z \in \mathbb{R}^{n}$, the problem of approximate Kernel Matrix-Vector Multiplication (aKMVM) is to obtain an approximation to $y=K z$. Due to linearity, we can always rescale the vectors without changing the problem, thus we may assume without loss of generality that $\|z\|_{1}=1$. This problem is important as many machine learning applications involve the mutltiplication of a vector with a dense Kernel matrix. Very often this operation is the computational bottleneck.

Observe that if $z=\frac{1}{n} \mathbf{1}$ then the problem is equivalent to estimating the kernel density of all the points in the dataset $P$, a problem that we can solve relatively fast. We show how one can adapt the techniques from this paper to provide a solution to the aKMVM problem.

Theorem 7.1. Given a $V$-bounded HBE for a kernel $k$ with complexity $T$, there exists an algorithm that given a dataset $P$ and a non-negative weight vector $z$ such that $\|z\|_{1}=1$ can compute a vector $\hat{y}$ in time $\tilde{O}\left(\frac{1}{\varepsilon^{2}} V(\varepsilon \tau) \cdot n T\right)$ using space $\tilde{O}\left(\frac{1}{\varepsilon^{2}} V(\varepsilon \tau) \cdot n T\right)$ such that with probability at least $1-n^{-1}$ for all $i \in[n]$ it holds $\left|\hat{y}_{i}-y_{i}\right| \leqslant 3 \varepsilon \tau+\varepsilon\left|y_{i}\right|$ and

$$
\begin{equation*}
\|\hat{y}-y\|_{p} \leqslant \varepsilon\left(3 \tau n^{1 / p}+\|y\|_{p}\right) \tag{7.1}
\end{equation*}
$$

Remark 7.2. Let $z_{+}=(z)_{0}$ and $z_{-}=(-z)_{0}$ be the positive and negative "parts" of vector $z$. We can apply this algorithm separately to estimate $y_{+}=K z_{+}$and $y_{-}=K z_{-}$to obtain a vector $\hat{y}=\hat{y}_{+}-\hat{y}_{-}$ that satisfies: $\|\hat{y}-K z\|_{p} \leqslant \varepsilon\left(6 \tau n^{1 / p}+\left\|y_{-}\right\|_{p}+\left\|y_{+}\right\|_{p}\right)$.
Proof. We consider intervals $I_{\ell}=\left[2^{-\ell}, 2^{-\ell+1}\right]$ for $\ell=1, \ldots, L=\log _{2}\left(n / \tau^{\prime}\right)$, where $\tau^{\prime}=\varepsilon \tau$, and we set $I_{0}=\left[0, \frac{\tau^{\prime}}{n}\right.$ ). We partition points according to which interval their corresponding "weights" $z_{i}$ belong to. Let $S_{0}, S_{1}, \ldots, S_{L}$ be the corresponding sets. For any point $i$, we have

$$
\begin{align*}
y_{i}^{*} & =\sum_{\ell=0}^{L} \sum_{j \in S_{\ell}} K\left(x_{i}, x_{j}\right) z_{j} \\
& =\sum_{\ell=1}^{L} Z_{\ell}\left(\sum_{j \in S_{\ell}} k\left(x_{i}, x_{j}\right) \frac{z_{j}}{\sum_{s \in S_{\ell}} z_{s}}\right)+\delta_{i} \tag{7.2}
\end{align*}
$$

where $\delta_{i}=\sum_{j \in S_{0}} K\left(x_{i}, x_{j}\right) z_{j} \leqslant \tau^{\prime}$ and $Z_{\ell}=\sum_{s \in S_{\ell}} z_{s}$. By construction it holds that $\left|S_{\ell}\right| 2^{-\ell} \leqslant Z_{\ell} \leqslant$ $\left|S_{\ell}\right| 2^{-\ell+1}$ and within each set $S_{\ell}$ the weights differ by at most a factor of two. The numbers $Z_{\ell}$ can be computed with a linear pass on the data. Thus, our problem can be expressed as a weighted version of $L$ KDE problems, where the $\ell$-th problem asks to compute an approximation to :

$$
\begin{equation*}
\operatorname{KDE}_{S_{\ell}}^{z}(x)=\frac{1}{Z_{\ell}} \sum_{y \in S_{\ell}} k(x, y) z_{y} \tag{7.3}
\end{equation*}
$$

Given a hashing scheme $\mathcal{H}$ with collision probabilities $p_{i}$ we evaluate the hash function during the preprocessing step only on $S_{\ell}$ and then for a query point $x \in P$ we define the following estimator.

$$
\begin{equation*}
Z_{h}=\frac{z_{i}}{Z_{\ell}} \frac{k\left(x, y_{I}\right)}{p_{I}}|H(x)| \tag{7.4}
\end{equation*}
$$

where as before I is a random index from $H(x) \subseteq S_{\ell}$. The first two moments of the estimator are given by:

$$
\begin{align*}
\mathbb{E}[Z] & =\frac{1}{Z_{\ell}} \sum_{i \in S_{\ell}} k\left(x, y_{i}\right) z_{i}  \tag{7.5}\\
\mathbb{E}\left[Z^{2}\right] & =\frac{1}{Z_{\ell}^{2}} \sum_{i \in S_{\ell}} \frac{k\left(x, y_{i}\right)^{2}}{p_{i}} z_{i}^{2} \mathbb{E}[\mid H(x) \| i \in H(x)] \tag{7.6}
\end{align*}
$$

Using the properties of the set $S_{\ell}$ we obtain the following upper bound on the variance of this estimator

$$
\mathbb{E}\left[Z^{2}\right] \leqslant \frac{2^{2 \ell}}{\left|S_{\ell}\right|^{2}} 2^{-2 \ell+2} \sum_{i \in S_{\ell}} \frac{k\left(x, y_{i}\right)^{2}}{p_{i}} \mathbb{E}[\mid H(x) \| i \in H(x)]
$$

Now observe that the variance is at most 4 times larger than if we would be trying to estimate $\operatorname{KDE}_{S_{\ell}}(x)$. Hence, given any $V$-bounded HBE for the kernel $k$ and set $S_{\ell}$ we can use the above modification to get a $4 V$-bounded HBE for $\operatorname{KDE}_{S_{\ell}}^{z}(x)$. Invoking Theorem 4.2 with parameters $\left(\varepsilon, \tau^{\prime}, \chi /(n L)\right)$ we can get a data structure that can estimate $\operatorname{KDE}_{S_{\ell}}^{z}\left(x_{i}\right)$ with probability at least $1-\chi$ either within multiplicative accuracy $\varepsilon$ (when AMR ${ }^{*}$ has non-zero output) or with absolute accuracy $\tau^{\prime}$ (when AMR ${ }^{*}$ outputs 0 ) for all $i \in[n]$ (by union bound). Let $\mathcal{L}$ denote the set of indices of $[L]$ such that $Z_{\ell} \geqslant \frac{\tau^{\prime}}{|L|}$. For all $\ell \in \mathcal{L}$ we instantiate the data structure given by Theorem 4.2 for the set $S_{\ell}$ and use it to query all points in $P$. For each query point thus we get estimates $z_{\ell}\left(x_{i}\right)$ for $\ell \in \mathcal{L}$ and set $z_{\ell}\left(x_{i}\right)=0$ for $\ell \notin \mathcal{L}$. The overhead per-query of the whole process is at most a multiplicative factor $L=O\left(\log \left(n / \tau^{\prime}\right)\right)$ compared to the case that we were creating a single data structure for the same problem. For any query $x_{i}$, we aggregate the estimates $z_{\ell}\left(x_{i}\right)$ in the following manner $Z\left(x_{i}\right)=\sum_{\ell \in \mathcal{L}} Z_{\ell} \cdot z_{\ell}\left(x_{i}\right)$. For all $z_{\ell}\left(x_{i}\right)$ with probability at least $1-\chi$ it holds

$$
\left|z_{\ell}\left(x_{i}\right)-\operatorname{KDE}_{S_{\ell}}^{z}\left(x_{i}\right)\right| \leqslant \max \left\{\tau^{\prime}, \varepsilon \cdot \operatorname{KDE}_{S_{\ell}}^{z}\left(x_{i}\right)\right\}
$$

This implies the following bounds:

$$
\begin{aligned}
\left|Z\left(x_{i}\right)-y_{i}^{*}\right| & \leqslant \sum_{\ell \in[L]} Z_{\ell} \cdot\left|z_{\ell}\left(x_{i}\right)-\operatorname{KDE}_{S_{\ell}}^{z}\left(x_{i}\right)\right|+\delta_{i} \\
& \leqslant \sum_{\ell \notin \mathcal{L}} Z_{\ell}+\varepsilon \sum_{\ell \in \mathcal{L}} Z_{\ell}\left(\tau+\operatorname{KDE}_{S_{\ell}}\left(x_{i}\right)\right)+\delta_{i} \\
& \leqslant 3 \varepsilon \tau+\varepsilon\left|y_{i}^{*}\right|
\end{aligned}
$$

Summing over all indices and using triangle inequality gives $\left\|\hat{y}-y^{*}\right\|_{p} \leqslant \varepsilon\left(3 \tau n^{1 / p}+\left\|y^{*}\right\|_{p}\right)$.

## 8 Lower Bound for Kernel Density Estimation

The nature of the KDE problem is quite analytical as it involves a simple summation of $n$ terms that depend smoothly on the distances of the query from the dataset. The Gaussian kernel, though a smooth function of the distances, is rapidly decreasing and can be thought of being an approximation to the indicator function $\mathbb{I}\left\{\|x-y\|^{2} \leqslant \sigma^{2}\right\}$. In particular, for two distances
$r_{1}=\sigma, r_{2}>\sqrt{C} \sigma$ the kernel value varies from $e^{-1}=\Omega(1)$ to $e^{-C}=o(1)$ for any $C=\omega(1)$. This is the basic observation motivating reducing the Approximate Nearest Neighbor Search problem to the KDE problem.

Definition $8.1((r, c)$-ANNS). Given a metric space $\mathcal{M}$ and parameters $c>1$ and $r>0$, and a dataset of $n$ points $x_{1}, \ldots, x_{n}$, our goal is to distinguish between the case that $d\left(x_{i}, y\right) \leqslant r$ for some $i \in[n]$ and the case where for all $i \in[n], d\left(x_{i}, y\right) \geqslant c r$. The query algorithm must output 1 in the former case, 0 in the latter case and may report anything if neither of the two cases hold.

The complexity of ANNS has been a topic of ongoing research over the past two decades [32, [5]. The most popular model of computation to prove lower bounds for is the cell-probe model. In this model, we are allowed an arbitrary amount of preprocessing but only allowed to keep $m$ cells each with $w$ bits of information. The query algorithm then queries (adaptively) $t$ cells and is required to produce the output. We will refer to this model as the ( $m, w, t$ )-cell probe model. For ANNS, lower bounds in this model impose constraints on $m, w, t$ depending on $n, c$. The most general result in this area is given by the following theorem proved by Panigrahy, Talwar and Wieder [32], whose estimates where improved by Andoni et al. [5].

Theorem 8.2 ([32, 5]). There exists a distribution over ( $r, c)$-ANNS instances and $\gamma \in[0,1]$ such that any randomized algorithm in the ( $m, w, t$ )-cell probe model which is correct with probability at least a half on these instances, satisfies:

$$
\frac{m^{t} w}{n} \geqslant \sup _{(q-1)(p-1)=\left(1-\frac{1}{c}\right)^{2}, p, q \geqslant 1}\left\{\left(\frac{\gamma}{t}\right)^{q} m^{t\left(1+\frac{q}{p}-q\right)}\right\}
$$

The quality of those bounds deteriorate as $t$ the number of probes increases. The bound for $t=1$ is optimal and recently Andoni et al.[5] gave an optimal bound up to sub-polynomial factors for $t=2$ in the regime of $w=O(\log n)$. The distribution for which the lower bounds are proved is the following.

Definition 8.3 ((n,d,c)-random instance). Let $P$ be $n$ random points from the boolean hypercube $\{0,1\}^{d}$ with $d=\omega(\log n)$. The query point $x$ is generated by picking a random point $y \in P$ and then generating a $\rho$-correlated point $y_{\rho}$ by keeping each bit of $y$ independently with probability $\rho=1-\frac{1}{c}$.

Our strategy of providing lower bounds for KDE is to show that for the specific distribution over instances used by Panigrahy et al. one can use an algorithm for KDE that would solve the ANNS problem with more than $1 / 2$ probability.

Theorem 8.4 (ANNS to KDE). For $\mu \in\left(0, \frac{1}{4}\right], \varepsilon \in\left[\frac{\mu}{1+2 \mu}, \frac{1}{5}\right]$ and $\delta \in\left(\frac{1}{2}+\mu, 1\right]$, any algorithm that solves the $(\mu, \varepsilon, \delta)$-KDE problem solves also the $(n, d, c)$-random instance of ANNS with $n=\left\lceil\frac{1}{\mu}\right\rceil, d=\Theta\left(\log ^{3}(n)\right)$, and $c=\Theta\left(\frac{\log (n)}{\log (1 / \varepsilon)}\right)$ with probability at least $1 / 2$.

Proof. We define the following parameters: $\varepsilon_{1}=\varepsilon, \varepsilon_{2}=2 \varepsilon, r=\log \left(\frac{n}{\sqrt{\left(1-\varepsilon_{1}\right)\left(1-\varepsilon_{2}\right)}}\right), d=\left\lceil\frac{18 \log (n)}{\ln ^{2}\left(\sqrt{\frac{1-\varepsilon_{1}}{1-\varepsilon_{2}}}\right)} r^{2}\right\rceil$, $\sigma^{2}=\frac{1}{r} \frac{d}{2}, u=\sqrt{\frac{\log ^{2}\left(\frac{1-\varepsilon_{1}}{1-\varepsilon_{2}}\right)}{4 \log \left(\frac{1}{4 \varepsilon}\right)}} \leqslant \frac{1}{3}$ and $c=\frac{(1+u)}{\log (1 / 4 \varepsilon)} 2 r$. Using these parameters we bound the values of the kernel for different pair of points. Let $P$ be a collection of $n=\left\lceil\frac{1}{\mu}\right\rceil$ random points on the $d$-dimensional hypercube and $y_{\rho}$ denotes as before a $\rho$-correlated point. For simplicity we assume that $1 / \mu$ is an integer.

Claim 8.5. For all $y \neq z \in P,\left(1-\varepsilon_{2}\right) \mu \leqslant k_{\sigma}\left(y_{\rho}, z\right) \leqslant\left(1-\varepsilon_{1}\right) \mu$ and $k_{\sigma}\left(y_{\rho}, y\right) \geqslant 4 \varepsilon$ with probability at least $1-n^{-1}$ over the randomness in $P$ and $y_{\rho}$.

Proof. For all $z \in P \backslash\{y\}$ let $X_{i}=\mathbb{I}\left[\left(y_{\rho}\right)_{i} \neq z_{i}\right]$ with $\mathbb{E}\left[X_{i}\right]=\frac{1}{2}$ and define $X=\sum_{i=1}^{d} X_{i}$. In this case, $k_{\sigma}\left(y_{\rho}, z\right)=\exp \left(-\left\|z-y_{\rho}\right\|^{2} / \sigma^{2}\right)=\exp \left(-\sum_{i=1}^{d} \mathbb{I}\left[\left(y_{\rho}\right)_{i} \neq z_{i}\right] / \sigma^{2}\right)=\exp \left(-X / \sigma^{2}\right)$.

$$
\begin{equation*}
\mathbb{P}\left[k_{\sigma}\left(y_{\rho}, z\right)>\left(1-\varepsilon_{1}\right) \mu\right]=\mathbb{P}\left[X<\left\{\frac{2 \sigma^{2}}{d} \log \left(\frac{1}{\left(1-\varepsilon_{1}\right) \mu}\right)\right\} \cdot \frac{d}{2}\right] \tag{8.1}
\end{equation*}
$$

where $\frac{2 \sigma^{2}}{d} \log \left(\frac{1}{\left(1-\varepsilon_{1}\right) \mu}\right)=\frac{\log n+\log \left(\frac{1}{1-\varepsilon_{1}}\right)}{\log n+\frac{1}{2} \log \left(\frac{1}{1-\varepsilon_{1}}\right)+\frac{1}{2} \log \left(\frac{1}{1-\varepsilon_{2}}\right)}=1-\frac{\log \left(\frac{1-\varepsilon_{1}}{1-\varepsilon_{2}}\right)}{2 \log \left(\frac{n}{\left.\sqrt{\left(1-\varepsilon_{1}\right)\left(1-\varepsilon_{2}\right)}\right)}\right.}$. By Chernoff bounds we get:

$$
\begin{equation*}
\mathbb{P}\left[k_{\sigma}\left(y_{\rho}, z\right)>\left(1-\varepsilon_{1}\right) \mu\right] \leqslant \exp \left\{-\frac{\log ^{2}\left(\frac{1-\varepsilon_{1}}{1-\varepsilon_{2}}\right)}{3 \cdot 4 \log ^{2}\left(\frac{n}{\sqrt{\left(1-\varepsilon_{1}\right)\left(1-\varepsilon_{2}\right)}}\right)} \frac{d}{2}\right\} \leqslant n^{-3} \tag{8.2}
\end{equation*}
$$

Similarly, we have that $\mathbb{P}\left[k_{\sigma}\left(y_{\rho}, z\right)>\left(1-\varepsilon_{1}\right) \mu\right]=\mathbb{P}\left[X>\left\{\frac{2 \sigma^{2}}{d} \log \left(\frac{1}{\left(1-\varepsilon_{2}\right) \mu}\right)\right\} \cdot \frac{d}{2}\right]$ where $\frac{2 \sigma^{2}}{d} \log \left(\frac{1}{\left(1-\varepsilon_{2}\right) \mu}\right)=1+\frac{\log \left(\frac{1-\varepsilon_{1}}{1-\varepsilon_{2}}\right)}{2 \log \left(\frac{\sqrt{n}}{\left.\sqrt{\left(1-\varepsilon_{1}\right)\left(1-\varepsilon_{2}\right)}\right)}\right.}$. By Chernoff bounds we get:

$$
\begin{equation*}
\mathbb{P}\left[k_{\sigma}\left(y_{\rho}, z\right)>\left(1-\varepsilon_{2}\right) \mu\right] \leqslant \exp \left\{-\frac{\log ^{2}\left(\frac{1-\varepsilon_{1}}{1-\varepsilon_{2}}\right)}{3 \cdot 4 \log ^{2}\left(\frac{n}{\sqrt{\left(1-\varepsilon_{1}\right)\left(1-\varepsilon_{2}\right)}}\right)} \frac{d}{2}\right\} \leqslant n^{-3} \tag{8.3}
\end{equation*}
$$

Next, let $Y_{i}=\mathbb{I}\left[\left(y_{\rho}\right)_{i} \neq y_{i}\right]$ with $\mathbb{E}\left[Y_{i}\right]=\frac{1}{c}$ and define $Y=\sum_{i=1}^{d} Y_{i}$. Then,

$$
\mathbb{P}\left[k_{\sigma}\left(y_{\rho}, y\right)<4 \varepsilon\right]=\mathbb{P}\left[Y>\left\{\frac{c \sigma^{2}}{d} \log \left(\frac{1}{4 \varepsilon}\right)\right\} \frac{d}{c}\right]
$$

where $\frac{c \sigma^{2}}{d} \log \left(\frac{1}{4 \varepsilon}\right)=\frac{1+u}{\log (1 / 4 \varepsilon)} \frac{\sigma^{2}}{d} \log \left(\frac{1}{4 \varepsilon}\right) 2 r=1+u$. By Chernoff bounds we get for all $n \geqslant 4$

$$
\begin{equation*}
\mathbb{P}\left[k_{\sigma}\left(y_{\rho}, z\right)<4 \varepsilon\right] \leqslant \exp \left\{-\frac{u^{2}}{3} \frac{d}{c}\right\} \leqslant e^{-\frac{9}{4} \log ^{2}(n)} \leqslant n^{-3} \tag{8.4}
\end{equation*}
$$

Taking union bound for all $2\binom{n}{2}+n=n^{2}$ events completes the proof.
Conditionally on these events, a no-instance of $c$-ANN problem, has kernel density at most $\left(1-\varepsilon_{1}\right) \mu$, whereas a yes-instance has density at least $\left.\left(1-\varepsilon_{2}\right) \frac{n-1}{n} \mu+\frac{1}{n} 4 \varepsilon=\left(1+2 \varepsilon\left(1+\frac{1}{n}\right)-\frac{1}{n}\right)\right) \mu \geqslant$ $(1+\varepsilon) \mu$ as long as $\varepsilon\left(1+\frac{2}{n}\right) \geqslant \frac{1}{n}$. We therefore, get that any data structure for the $(\mu, \varepsilon, \delta)-$ KDE problem can also solve the $c$-ANN problem for random instances with probability at least $1-\delta-n^{-1}$.

Corollary 8.6. For $\mu \in\left(0, \frac{1}{4}\right], \varepsilon \in\left[\frac{\mu}{1+2 \mu}, \frac{1}{5}\right]$ and $\delta \in\left(\frac{1}{2}+\mu, 1\right]$, any algorithm that solves the $(\mu, \varepsilon, \delta)-K D E$ problem in the $(m, 1, w)$-cell probe model must satisfy $(m \cdot w) \geqslant \Omega\left(\frac{1}{\mu}\right)$.

Proof. Using the parameters defined in Theorem 8.4, we invoke Theorem 8.2 with $p=\frac{1}{1-\rho^{2}} \geqslant 1$, $q=2-\rho^{2} \geqslant 1$ where $\rho^{2}=\left(1-\frac{1}{c}\right)^{2}$ to obtain:

$$
\begin{equation*}
m \geqslant\left(\frac{1}{\mu w}\right)^{\frac{p}{p-1} \frac{1}{q}} \gamma^{\frac{p}{p-1}} \tag{8.5}
\end{equation*}
$$

$$
\begin{align*}
& =\left(\frac{1}{\mu w}\right)^{\frac{1}{\rho^{2}} \frac{1}{2-\rho^{2}}} \gamma^{\frac{1}{\rho^{2}}}  \tag{8.6}\\
& =\left(\frac{1}{\mu w}\right)^{\frac{1}{1-\frac{1}{c}+\frac{1}{c^{2}}} \frac{1}{1+\frac{2}{c}-\frac{1}{c^{2}}}} \gamma^{\frac{1}{\rho^{2}}}  \tag{8.7}\\
& =\left(\frac{1}{\mu w}\right)^{\frac{1}{1-\left(\frac{2}{c}-\frac{1}{c^{2}}\right)^{2}}} \gamma^{\frac{1}{\rho^{2}}}  \tag{8.8}\\
& \geqslant \frac{1}{\mu w} \gamma^{\frac{1}{\rho^{2}}} \tag{8.9}
\end{align*}
$$

As $\gamma$ is a constant and $1 / \rho^{2}=O(1)$ the statement follows.

Lower Bounds on Adaptive Core-sets. The above bound becomes more interesting if one considers the following class of estimators for KDE. Given $P$, we may perform any amount of preprocessing and only store $m$ sets $S_{1}, \ldots, S_{m}$ (of arbitrary points) each of size at most $\frac{w}{d}$. Given a query $x$, our estimator picks one of those sets $i(x)$ and produces an approximation to the Kernel Density by only a function of $\left(S_{i(x)}, x\right)$, an example of such an estimate could also be something as simple as $\operatorname{KDE}_{S_{i(x)}}(x)$, or weighted versions such as $\sum_{y \in S_{i(x)}} W(x, y) K(x, y)$. The above lower bound shows that either we must have many such sets (large space) or each set must be large itself (query time). Of particular interest is the case $m=1$ where there is a single set. In that case, we have a lower bound on the size of a core set. Hence, even for $t=1$ the lower bound gives non-trivial results for an interesting class of estimation algorithms.

## 9 Hölder Inequalities

Lemma 9.1 (Hölder's Inequality). Let $p, q>0$ such that $\frac{1}{p}+\frac{1}{q}=1$, then $\left|\sum_{i} x_{i} y_{i}\right| \leqslant\|x\|_{p}\|x\|_{q}$.
Corollary 9.2. Let $\beta \in[0,1]$ and $p \geqslant q>0$, then for all $x \in \mathbb{R}^{n}$

$$
\begin{align*}
\|x\|_{\beta}^{\beta} & \leqslant\|x\|_{1}^{\beta} \cdot n^{1-\beta}  \tag{9.1}\\
\|x\|_{p}^{p} & \leqslant\|x\|_{q}^{q} \cdot\|x\|_{\infty}^{p-q} \tag{9.2}
\end{align*}
$$

Proof. The first inequality is proved by fixing $\|x\|_{1}=a>0$, performing the change of variables $u_{j}=\left(\frac{x_{j}}{a}\right)^{\beta}$, so that $\|u\|_{\frac{1}{\beta}}^{\frac{1}{\beta}}=1$, and finally applying Hölder's inequality for $p=\frac{1}{\beta}, q=\frac{1}{1-\beta}$. The second inequality follows by applying Hölder on $\sum_{i=1}^{n}\left|x_{i}\right|^{p-q}\left|x_{i}\right|^{q}$.

Lemma 9.3 (two-sided Holder). Let $v, w \in \mathbb{R}^{n}$ be strictly positive vectors, then for any two sets $S, S^{\prime} \subseteq$ [ $n$ ]:

$$
\begin{equation*}
\sum_{i \in S, j \in S^{\prime}} A_{i j} x_{i} x_{j} \leqslant\|x\|_{v, 1}\|x\|_{w, 1} \cdot \max _{i \in S, j \in S^{\prime}}\left\{\frac{\left|A_{i j}\right|}{v_{i} w_{j}}\right\} \tag{9.3}
\end{equation*}
$$

Proof. We start by setting w.l.o.g. $\|x\|_{w, 1}=\alpha$ and $\|x\|_{v, 1}=\beta$ with $\alpha, \beta>0$.

$$
\sup _{\substack{\|x\|_{w, 1}=\alpha \\\|x\|_{v, 1}=\beta}}\left\{x_{S}^{\top} A_{S, S^{\prime}} x_{S^{\prime}}\right\}=\sup _{\substack{\|x\|_{w, 1}=\alpha \\\|y\|_{v, 1}=\beta \\ y=x}}\left\{x_{S}^{\top} A y_{S^{\prime}}\right\} \leqslant \sup _{\substack{\|x\|_{w, 1}=\alpha \\\|y\|_{v, 1}=\beta}}\left\{x_{S}^{\top} A y_{S^{\prime}}\right\}
$$

We next peform the change of variables $\tilde{x}=\frac{w_{i}}{\alpha} x_{i}$ and $\tilde{y}=\frac{v_{i}}{\beta} y_{i}$ so that $\|\tilde{x}\|_{1}=\|\tilde{y}\|_{1}=1$. We get

$$
\begin{aligned}
\sup _{\substack{\|x\|_{0,1}=\alpha \\
\|y\| v_{,}=\beta}}\left\{x_{S}^{\top} A y_{S^{\prime}}\right\} & =\alpha \beta \cdot \sup _{\substack{\|\tilde{x}\|_{1}=1 \\
\| \tilde{y_{1}}=1}}\left\{\sum_{i \in S} \tilde{x}_{i}\left(\sum_{j \in S^{\prime}} \tilde{y}_{j} \frac{A_{i j}}{w_{i} v_{j}}\right)\right\} \\
& \leqslant \alpha \beta \cdot \sup _{\substack{\|\tilde{x}\|_{1}=1 \\
\| \tilde{y_{1}=1}}}\left\{\left\|\tilde{y}_{S^{\prime}}\right\|_{1} \sum_{i \in S} \tilde{x}_{i} \max _{j \in S^{\prime}}\left(\frac{\left|A_{i j}\right|}{w_{i} v_{j}}\right)\right\} \\
& \leqslant \alpha \beta \cdot \sup _{\|\tilde{x}\|_{1}=1}\left\{\left\|\tilde{x}_{S}\right\|_{1} \max _{i \in S, j \in S^{\prime}}\left(\frac{\left|A_{i j}\right|}{w_{i} v_{j}}\right)\right\} \\
& =\alpha \beta \cdot \max _{i \in S, j \in S^{\prime}}\left\{\frac{\left|A_{i j}\right|}{w_{i} v_{j}}\right\}
\end{aligned}
$$

where we applied Hölder's inequality in the second and third inequality.
Lemma 9.4 (Monotone Hölder). For every integer $n \geqslant 1, \beta \in\left[\frac{1}{2}, 1\right]$, and for all $x \in \mathbb{R}^{n}$ such that $\left|x_{1}\right| \geqslant\left|x_{2}\right| \geqslant \ldots \geqslant\left|x_{n}\right|$, we have

$$
\begin{equation*}
\sum\left|x_{i}\right|^{\frac{2-\beta}{\beta}}\left(i+\sum_{j>i} \frac{\left|x_{j}\right|}{\left|x_{i}\right|}\right) \leqslant n^{\beta} \cdot\left(\sum_{i=1}^{n}\left|x_{i}\right|^{\frac{1}{\beta}}\right)^{2-\beta} \tag{9.4}
\end{equation*}
$$

where equality holds for $x^{*}=c \mathbf{1}$ for any $c \neq 0$.
Proof. Without loss of generality we may assume that $x \in \mathbb{R}_{+}^{n}$ and that $\|x\|_{\frac{1}{\beta}}^{\frac{1}{\beta}}=a$ for $a>0$. The claim is equivalent to proving the following inequality for all $n \geqslant 1$ :

$$
\begin{gather*}
f_{n}^{\beta}(a):=\sup _{\|x\|_{\frac{1}{\beta}}^{\frac{1}{\beta}}=a}\left\{\sum x_{i}^{\frac{2-\beta}{\beta}}\left(i+\sum_{j>i} \frac{x_{j}}{x_{i}}\right)\right\} \leqslant n^{\beta} a^{2-\beta}  \tag{9.5}\\
x_{1} \geqslant \ldots \geqslant x_{n} \geqslant 0
\end{gather*}
$$

We will use induction on $n$ to prove our claim. The base case is trivial. We next perform the inductive step.

$$
\left.\begin{array}{rl}
f_{n+1}^{\beta}(a)= & \sup _{\|x\|_{\frac{1}{\beta}}^{\frac{1}{\beta}}=a}\left\{x_{1}^{\frac{2-\beta}{\beta}}+x_{1}^{\frac{2(1-\beta)}{\beta}} \sum_{j=1}^{n} x_{j+1}+\sum_{i=1}^{n} x_{i+1}^{\frac{2(1-\beta)}{\beta}}\left(i+1+\sum_{j>i}\left(\frac{x_{j+1}}{x_{i+1}}\right)\right)\right\} \\
& x_{1} \geqslant \ldots \geqslant x_{n+1} \\
1 \geqslant x_{i} \geqslant 0
\end{array}\right\}
$$

where in the last inequality we used Hölder's inequality (twice) and the inductive hypothesis. We perform the change of variables $x_{1}=(a \gamma)^{\beta}$ to obtain

$$
\begin{aligned}
f_{n+1}^{\beta}(a) \leqslant(n+1)^{\beta} a^{2-\beta} & \sup _{\gamma \in\left[\frac{1}{n+1}, 1\right]}\left\{\frac{1}{(n+1)^{\beta}}\right.
\end{aligned} \begin{aligned}
2-\beta & \frac{n^{1-\beta}}{(n+1)^{\beta}} \gamma^{2-2 \beta}(1-\gamma)^{\beta} \\
& \left.+\frac{1}{(n+1)^{\beta}} \gamma^{1-\beta}(1-\gamma)+C\left(\frac{n}{n+1}\right)^{\beta}(1-\gamma)^{2-\beta}\right\}
\end{aligned}
$$

To prove the desired inequality we are going to show that the function

$$
g_{n}(\gamma):=\frac{1}{(n+1)^{\beta}} \gamma^{2-\beta}+\frac{n^{1-\beta}}{(n+1)^{\beta}} \gamma^{2-2 \beta}(1-\gamma)^{\beta}+\frac{1}{(n+1)^{\beta}} \gamma^{1-\beta}(1-\gamma)+C\left(\frac{n}{n+1}\right)^{\beta}(1-\gamma)^{2-\beta}
$$

is monotone in the given interval. Its derivative is given by

$$
\begin{aligned}
g_{n}^{\prime}(\gamma) & =\frac{1-\beta}{(n+1)^{\beta}} \gamma^{1-\beta}+2(1-\beta) \frac{n^{1-\beta}}{(n+1)^{\beta}} \frac{(1-\gamma)^{\beta}}{\gamma^{2 \beta-1}}+\frac{1-\beta}{(n+1)^{\beta}} \frac{1-\gamma}{\gamma^{\beta}} \\
& -\beta \frac{n^{1-\beta}}{(n+1)^{\beta}} \frac{\gamma^{2-2 \beta}}{(1-\gamma)^{1-\beta}}-(2-\beta) C \frac{n^{\beta}}{(n+1)^{\beta}}(1-\gamma)^{1-\beta}
\end{aligned}
$$

Comparing negative and positive terms we get the following relationships:

$$
\begin{aligned}
\beta \frac{n^{1-\beta}}{(n+1)^{\beta}} \frac{\gamma^{2-2 \beta}}{(1-\gamma)^{1-\beta}} \geqslant \frac{1-\beta}{(n+1)^{\beta}} \gamma^{1-\beta}, \quad \forall \gamma \geqslant \frac{\left(\frac{1-\beta}{\beta}\right)^{\frac{1}{1-\beta}}}{n+\left(\frac{1-\beta}{\beta}\right)^{\frac{1}{1-\beta}}} \\
\frac{1}{2} \cdot(2-\beta) C \frac{n^{\beta}}{(n+1)^{\beta}}(1-\gamma)^{1-\beta} \geqslant 2(1-\beta) \frac{n^{1-\beta}}{(n+1)^{\beta}} \frac{(1-\gamma)^{\beta}}{\gamma^{2 \beta-1}}, \quad \forall \gamma \geqslant \frac{1}{1+\left(\frac{(2-\beta) C}{(1-\beta) 4}\right)^{\frac{1}{2 \beta-1}} n} \\
\frac{1}{2} \cdot(2-\beta) C \frac{n^{\beta}}{(n+1)^{\beta}}(1-\gamma)^{1-\beta} \geqslant \frac{1-\beta}{(n+1)^{\beta}} \frac{1-\gamma}{\gamma^{\beta}}, \quad \forall \gamma \geqslant \frac{1}{1+\left(\frac{(2-\beta) C}{\left.(1-\beta)^{\frac{1}{2}}\right)^{\frac{1}{\beta}} n}\right.}
\end{aligned}
$$

Our goal is to show that each lower bound on $\gamma$ is always less or equal to $\frac{1}{n+1}$ and hence that the function $g_{n}(\gamma)$ is monotone in the interval of interest. Starting, from the first equation we observe that the function $x /(1+x)$ is increasing in $x \geqslant 0$, so to maximize the lower bound we need to find the maximum of the function:

$$
h_{1}(\beta)=\left(\frac{1-\beta}{\beta}\right)^{\frac{1}{1-\beta}}=e^{\frac{\ln (1-\beta-\ln (\beta)}{1-\beta}}
$$

The function is decreasing for any $\beta \in\left[\frac{1}{2}, 1\right]$ so we have that $\sup _{\beta \in[0.5,1]} h_{1}(\beta)=h_{1}(0.5)=1$. This proves that the first relationship holds for all $\gamma \geqslant \frac{1}{n+1}$. To handle the other two inequalities, we observe that the function $1 /(1+x)$ is decreasing for $x \geqslant 0$ and hence we need to find the minimum of the functions:

$$
\begin{align*}
& h_{2}(\beta)=\left(r_{2} \frac{(2-\beta)}{1-\beta}\right)^{\frac{1}{2 \beta-1}}=e^{\frac{\ln (2-\beta)-\ln (1-\beta)+\ln r_{2}}{2 \beta-1}}  \tag{9.6}\\
& h_{3}(\beta)=\left(r_{3} \frac{(2-\beta)}{1-\beta}\right)^{\frac{1}{\beta}}=e^{\frac{\ln (2-\beta)-\ln (1-\beta)+\ln r_{3}}{\beta}} \tag{9.7}
\end{align*}
$$

where $r_{2}, r_{3}>0$ are constants. The function $h_{2}(\beta)$ is increasing in the interval $[0.5,1]$, hence we get the desired inequality $h_{2}(\beta) \geqslant h_{2}(0.5)=1$. Similarly, the function $h_{3}(\beta)$ satisfies $h_{3}(\beta) \geqslant$ $h_{3}(0.5)=\left(3 r_{3}\right)^{2}=(3 C / 2)^{2}$, which is greater than one for $C=1$. This shows that the function $g_{n}(\gamma)$ is decreasing and consequently that

$$
f_{n+1}^{\beta}(a) \leqslant a^{2-\beta}(n+1)^{\beta} \cdot g_{n}\left(\frac{1}{n+1}\right)=a^{2-\beta}(n+1)^{\beta}
$$

completing the inductive step.

## Acknowledgements

We thank Peter Bailis for valuable conversations that initiated our research on this problem. This research was supported by NSF grants CCF-1617577, CCF-1302518 and a Simons Investigator Award. Paris Siminelakis is partially supported by an Onassis Foundation Scholarship.

## References

[1] J. Acharya, I. Diakonikolas, J. Li, and L. Schmidt. Sample-optimal density estimation in nearlylinear time. In Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 1278-1289. SIAM, 2017.
[2] J. Alman and R. Williams. Probabilistic polynomials and hamming nearest neighbors. In 2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS), pages 136-150. IEEE, 2015.
[3] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. In Proceedings of the twenty-eighth annual ACM Symposium on Theory of Computing, pages 20-29. ACM, 1996.
[4] A. Andoni and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS), pages 459-468. IEEE, 2006.
[5] A. Andoni, T. Laarhoven, I. Razenshteyn, and E. Waingarten. Optimal hashing-based timespace trade-offs for approximate near neighbors. In Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 47-66. SIAM, 2017.
[6] A. Andoni and I. Razenshteyn. Optimal data-dependent hashing for approximate near neighbors. In Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, pages 793-801. ACM, 2015.
[7] E. Arias-Castro, D. Mason, and B. Pelletier. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. Journal of Machine Learning Research, 2015.
[8] H. Avron, K. L. Clarkson, and D. P. Woodruff. Faster kernel ridge regression using sketching and preconditioning. SIAM Journal on Matrix Analysis and Applications, 38(4):1116-1138, 2017.
[9] A. Backurs, P. Indyk, and L. Schmidt. On the fine-grained complexity of empirical risk minimization: Kernel methods and neural networks. In Advances in Neural Information Processing Systems, pages 4308-4318, 2017.
[10] S. Bochner. Monotone funktionen, stieltjessche integrale und harmonische analyse. Mathematische Annalen, 108(1):378-410, 1933.
[11] K. Böröczky Jr and G. Wintsche. Covering the sphere by equal spherical balls. In Discrete and computational geometry, pages 235-251. Springer, 2003.
[12] S.-O. Chan, I. Diakonikolas, R. A. Servedio, and X. Sun. Efficient density estimation via piecewise polynomial approximation. In Proceedings of the 46th Annual ACM Symposium on Theory of Computing, pages 604-613. ACM, 2014.
[13] M. Charikar and P. Siminelakis. Hashing-based-estimators for kernel density in high dimensions. In 2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS), pages 1032-1043. IEEE, 2017.
[14] Y.-C. Chen, C. R. Genovese, and L. Wasserman. Density level sets: Asymptotics, inference, and visualization. Journal of the American Statistical Association, (just-accepted), 2016.
[15] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In Proceedings of the twentieth annual symposium on Computational geometry, pages 253-262. ACM, 2004.
[16] L. Devroye and G. Lugosi. Combinatorial methods in density estimation. Springer Science \& Business Media, 2012.
[17] I. Diakonikolas, G. Kamath, D. M. Kane, J. Li, A. Moitra, and A. Stewart. Robust estimators in high dimensions without the computational intractability. In 2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS), pages 655-664. IEEE, 2016.
[18] J. Dongarra and F. Sullivan. Guest editors' introduction: The top 10 algorithms. Computing in Science E Engineering, 2(1):22-23, 2000.
[19] J. Fan and I. Gijbels. Local polynomial modelling and its applications: monographs on statistics and applied probability 66, volume 66. CRC Press, 1996.
[20] W. Feller. An Introduction to Probability Theory and Its Applications, volume 1. Wiley, New York, third edition, 1968.
[21] A. Goldenshluger and O. Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. The Annals of Statistics, pages 1608-1632, 2011.
[22] L. Greengard and V. Rokhlin. A fast algorithm for particle simulations. Journal of computational physics, 73(2):325-348, 1987.
[23] L. Greengard and J. Strain. The fast gauss transform. SIAM Journal on Scientific and Statistical Computing, 12(1):79-94, 1991.
[24] S. Joshi, R. V. Kommaraji, J. M. Phillips, and S. Venkatasubramanian. Comparing distributions and shapes using the kernel distance. In Proceedings of the twenty-seventh annual symposium on Computational geometry, pages 47-56. ACM, 2011.
[25] P. Kar and H. Karnick. Random feature maps for dot product kernels. In AISTATS, volume 22, pages 583-591, 2012.
[26] M. Kearns, Y. Mansour, D. Ron, R. Rubinfeld, R. E. Schapire, and L. Sellie. On the learnability of discrete distributions. In Proceedings of the twenty-sixth annual ACM symposium on Theory of computing, pages 273-282. ACM, 1994.
[27] K. A. Lai, A. B. Rao, and S. Vempala. Agnostic estimation of mean and covariance. In 2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS), pages 665-674. IEEE, 2016.
[28] D. Lee, A. Gray, and A. Moore. Dual-tree fast gauss transforms. Advances in Neural Information Processing Systems, 18:747, 2006.
[29] W. B. March, B. Xiao, and G. Biros. Askit: Approximate skeletonization kernel-independent treecode in high dimensions. SIAM Journal on Scientific Computing, 37(2):A1089-A1110, 2015.
[30] M. Mitzenmacher and E. Upfal. Probability and Computing: Randomized algorithms and probabilistic analysis. Cambridge University Press, 2005.
[31] C. Musco and C. Musco. Recursive sampling for the nystrom method. In Advances in Neural Information Processing Systems, pages 3833-3845, 2017.
[32] R. Panigrahy, K. Talwar, and U. Wieder. Lower bounds on near neighbor search via metric expansion. In 2010 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS), pages 805-814. IEEE, 2010.
[33] N. Pham and R. Pagh. Fast and scalable polynomial kernels via explicit feature maps. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 239-247. ACM, 2013.
[34] J. M. Phillips. $\varepsilon$-samples for kernels. In Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 1622-1632. Society for Industrial and Applied Mathematics, 2013.
[35] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In Advances in neural information processing systems, pages 1177-1184, 2007.
[36] A. Rinaldo and L. Wasserman. Generalized density clustering. The Annals of Statistics, pages 2678-2722, 2010.
[37] B. Scholkopf and A. J. Smola. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press, 2001.
[38] E. Schubert, A. Zimek, and H.-P. Kriegel. Generalized outlier detection with flexible kernel density estimates. In Proceedings of the 2014 SIAM International Conference on Data Mining, pages 542-550. SIAM, 2014.
[39] R. Spring and A. Shrivastava. A new unbiased and efficient class of lsh-based samplers and estimators for partition function computation in log-linear models. arXiv preprint arXiv:1703.05160, 2017.
[40] G. Valiant. Finding correlations in subquadratic time, with applications to learning parities and juntas. In 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science (FOCS), pages 11-20. IEEE, 2012.
[41] S. Vempala and G. Wang. A spectral algorithm for learning mixture models. Journal of Computer and System Sciences, 68(4):841-860, 2004.
[42] Y. Zheng and J. M. Phillips. Coresets for kernel regression. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17, pages 645-654, New York, NY, USA, 2017. ACM.

## A Chernoff Bounds

For completeness we state here the version of Chernoff bounds that we use (e.g. [30]).
Theorem A.1. Let $X_{1}, \ldots, X_{n}$ be i.i.d Bernouli random variables with mean $\mathbb{E}\left[X_{i}\right]=\bar{x}$, then for $\delta \in(0,1]$

$$
\begin{align*}
& \mathbb{P}\left[\sum_{i=1}^{n} X_{i} \leqslant(1-\delta) n \bar{x}\right] \leqslant e^{-\frac{\delta^{2} n \bar{x}}{2}}  \tag{A.1}\\
& \mathbb{P}\left[\sum_{i=1}^{n} X_{i} \geqslant(1+\delta) n \bar{x}\right] \leqslant e^{-\frac{\delta^{2} n \bar{x}}{3}} \tag{A.2}
\end{align*}
$$

## B Random Fourier Features Estimator

The basis of using random Fourier features is Bochner's theorem [10], that for the gaussian kernel us equivalent to.

Fact B.1. Let $\omega \sim \mathcal{N}\left(0, I_{d}\right), \mathbb{E}\left[\cos \left(\omega^{\top}(x-y)\right)\right]=e^{-\|x-y\|^{2}}$
The random fourier features basic estimator is given by

$$
Z_{R F F}(x):=\frac{2}{|P|} \sum_{y \in P} \cos \left(\omega^{\top} x+b\right) \cos \left(\omega^{\top} y+b\right)
$$

where $\omega \sim \mathcal{N}\left(0, I_{d}\right)$ and $b \sim U[0,2 \pi]$
Proposition B.2. The first two moments of the RFF estimator are given by

$$
\begin{align*}
\mathbb{E}\left[Z_{R F F}(x)\right] & =\frac{1}{|P|} \sum_{y \in P} k(x, y)  \tag{B.1}\\
\mathbb{E}\left[Z_{R F F}^{2}\right] & =\frac{1}{|P|}+\frac{1}{|P|} \sum_{y \in P} k(2 x, 2 y)+\frac{1}{|P|^{2}} \sum_{y \neq z \in P} k(y, z)+\frac{1}{|P|^{2}} \sum_{y \neq z \in P} k(2 x, y+z) \tag{B.2}
\end{align*}
$$

To prove the lemma we require the following trigonometric identities.
Proposition B. 3 (Random angle). Let $B \sim U[0,2 \pi]$ and $a, b, c \in \mathbb{R}$, we have the following identities:

$$
\begin{align*}
\underset{B}{\mathbb{E}}[\cos (a+B) \cos (c+B)] & =\frac{1}{2} \cos (a-c)  \tag{B.3}\\
\underset{B}{\mathbb{E}}\left[\cos ^{2}(a+B) \cos ^{2}(c+B)\right] & =\frac{1}{4}+\frac{1}{8} \cos (2 a-2 c)  \tag{B.4}\\
\underset{B}{\mathbb{E}}\left[\cos ^{2}(a+B) \cos (b+B) \cos (c+B)\right] & =\frac{1}{4} \cos (b-c)+\frac{1}{8} \cos (2 a-(b+c)) \tag{B.5}
\end{align*}
$$

Proof of Proposition B. 2 We have that $\mathbb{E}\left[Z_{R F F}(x)\right]=\operatorname{KDE}_{P}(x)=\frac{1}{|P|} \sum_{y \in P} k(x, y)$ Further

$$
\mathbb{E}\left[Z_{R F F}^{2}\right]=\frac{1}{|P|^{2}} \mathbb{E}\left\{\left|\sum_{y \in P} Z(x, y)\right|^{2}\right\}=\frac{1}{|P|^{2}}\left\{\sum_{y \in P} \mathbb{E}\left[Z(x, y)^{2}\right]+\sum_{y \neq z \in P} \mathbb{E}[Z(x, y) Z(x, z)]\right\}
$$

where $Z(x, y):=2 \cos \left(\omega^{\top} x+b\right) \cos \left(\omega^{\top} y+b\right)$. We start with the first term:

$$
\mathbb{E}\left[Z(x, y)^{2}\right]=4 \mathbb{E}\left[\cos ^{2}\left(\omega^{\top} x+b\right) \cos ^{2}\left(\omega^{\top} y+b\right)\right]=1+\frac{1}{2} k(2 x, 2 y)
$$

Next, we analyze the second term

$$
\begin{equation*}
\mathbb{E}[Z(x, y) Z(x, z)]=4 \mathbb{E}\left[\cos ^{2}\left(\omega^{\top} x+b\right) \cos \left(\omega^{\top} y+b\right) \cos \left(\omega^{\top} z+b\right)\right]=k(y, z)+\frac{1}{2} k(2 x, y+z) \tag{B.6}
\end{equation*}
$$

The second moment thus becomes:

$$
\begin{equation*}
\mathbb{E}\left[Z_{R F F}^{2}\right]=\frac{1}{|P|}+\frac{1}{|P|} \sum_{y \in P} k(2 x, 2 y)+\frac{1}{|P|^{2}} \sum_{y \neq z \in P} k(y, z)+\frac{1}{|P|^{2}} \sum_{y \neq z \in P} k(2 x, y+z) \tag{B.7}
\end{equation*}
$$

Imagine, that $P$ consists of $n$ points all placed at distance $\sqrt{\log (1 / \mu)}$ of $x$. In this case the term $\frac{1}{|P|} \sum_{y \neq z \in P} k(y, z)$ is constant independent of $\mu$. So, when the dataset is highly correlated the variance of the estimator is very large.

Proof of Proposition B. 3 Using partial integration and the fact that sin, cos are periodic functions with period $2 \pi$, we get

$$
\begin{aligned}
\underset{B}{\mathbb{E}}[\sin (a+B) \sin (c+B)] & =-[\cos (a+B) \sin (c+B)]_{0}^{2 \pi}+\underset{B}{\mathbb{E}}[\cos (a+B) \cos (c+B)] \\
& =\underset{B}{\mathbb{E}}[\cos (a+B) \cos (c+B)]
\end{aligned}
$$

The first equation then follows from the identity $\cos (A) \cos (B)=\cos (A-B)-\sin (A) \sin (B)$. To get the second equation we use the identities $\cos ^{2}(a)=\frac{1}{2}(1+\cos (2 a))$ and $\mathbb{E}_{B}[\cos (2 a+2 B)]=0$ to get

$$
\underset{B}{\mathbb{E}}\left[\cos ^{2}(a+B) \cos ^{2}(c+B)\right]=\frac{1}{4}+\frac{1}{4} \underset{B}{\mathbb{E}}[\cos (2 a+2 B) \cos (2 c+2 B)]
$$

Finally, invoking the first part completes the proof.
Cost of Linearity. Looking at the RFF estimator one might think that it requires linear time per query. However, observe that if we spend linear time once to compute $\phi_{P}=\frac{1}{|P|} \sum_{y \in P} \sqrt{2} \cos \left(\omega^{\top} y+\right.$ $\theta$ ), the estimate $Z_{R F F}(x)=\sqrt{2} \cos \left(\omega^{\top} x\right) \cdot \phi_{P}$ can be computed in $O(1)$ time per query. Perhaps counter-intuitively this shows that a simple rescaling gives an unbiased estimate for all queries $x$ ! Thus, it is not surprising that the variance of the estimator is essentially as large as possible. If we want to reduce the variance, we can take the average of $m$ such estimates, that would be equivalent to constructing a randomized embedding $\phi: \mathbb{R}^{d} \rightarrow \mathbb{R}^{m}$ of the points into a euclidean space such that $\operatorname{KDE}_{P}(x) \approx\left\langle\phi(x), \frac{1}{|P|} \sum_{y \in P} \phi(y)\right\rangle$. This approach is quite general and can be extended to a large range of kernels [35, 25, 33]. However, the inherent variance of the randomized embedding suggests that this approach cannot beat even the Random Sampling estimator.


[^0]:    *A preliminary version of this paper appeared in FOCS'2017 [13].

