Multi-Resolution Hashing for Fast Pairwise Summations

Moses Charikar
Department of Computer Science
Stanford University
Stanford, USA
Email: moses@cs.stanford.edu

Paris Siminelakis
Department of Electrical Engineering
Stanford University
Stanford, USA
Email: paris@cs.stanford.edu

***Abstract*—A basic computational primitive in the analysis of massive datasets is summing simple functions over a large number of objects. Modern applications pose an additional challenge in that such functions often depend on a parameter vector y (query) that is unknown a priori. Given a set of points X and a pairwise function w(x,y), we study the problem of designing a data-structure that enables sublinear-time approximation of the summation of w(x,y) for all x in X for any query point y. By combining ideas from Harmonic Analysis (partitions of unity and approximation theory) with *Hashing-Based-Estimators* [Charikar, Siminelakis FOCS'17], we provide a general framework for designing such data structures through hashing that reaches far beyond what previous techniques allowed.**

A key design principle is constructing a collection of hash families, each inducing a different collision probability between points in the dataset, such that the pointwise supremum of the collision probabilities scales as the square root of the function w(x,y). This leads to a data-structure that approximates pairwise summations using a sub-linear number of samples from each hash family. Using this new framework along with *Distance Sensitive Hashing* [Aumuller, Christiani, Pagh, Silvestri PODS'18], we show that such a collection can be constructed and evaluated efficiently for log-convex functions of the inner product between two vectors.

Our method leads to data structures with sub-linear query time that significantly improve upon random sampling and can be used for Kernel Density, Partition Function Estimation and sampling.

***Keywords*-Hashing; Kernel Density; Partition Function; Importance Sampling;Sub-linear algorithms.**

I. Introduction

The analysis of massive datasets very often involves summing simple functions over a very large number of objects [1], [2], [3]. While in all cases one can compute the sum of interest exactly in time and space polynomial or even linear in the number of objects, practical considerations, such as space usage and update/query time, require developing significantly more efficient algorithms that can provably approximate the quantity in question arbitrarily well. For $\alpha \geq 1$, we say that $\hat{\mu}$ is an α-approximation to μ if $\alpha^{-1}\mu \leq \hat{\mu} \leq \alpha\mu$ and an $(1 \pm \epsilon)$-approximation if $(1-\epsilon)\mu \leq \hat{\mu} \leq (1+\epsilon)\mu$.

Modern applications in Machine Learning pose an additional challenge in that such functions depend on a parameter vector $y \in \mathbb{R}^d$ that is unknown a priori or changes with time. Such examples include outlier detection [4], text generation [5], [6], and empirical risk minimization (ERM) [7], [8]. Motivated by such applications, we seek sub-linear time algorithms for summing pairwise functions in high dimensions.

Given a set of points $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, a non-negative function $w : \mathbb{R}^d \times \mathbb{R}^d \to [0,1]$, and a parameter $\epsilon > 0$, we study the problem of designing a data structure that for any query $y \in \mathbb{R}^d$ provides in sub-linear time a $(1 \pm \epsilon)$-approximation to the sum:

$$Z_w(y) = \frac{1}{n}\sum_{i=1}^{n} w(x_i, y) \tag{1}$$

and show how it relates to the problem of sampling from the distribution that assigns probability to points in X proportional to $w(x,y)$. The actual (normalized) value of the sum $Z_w(y) \in [0,1]$ for a given query y, will be denoted by μ and, as we see next, we can use a lower bound $\tau \leq \mu$ to bound the complexity of the problem. Let $w_{\max}$ be a number such that $\max_{x \in X}\{w(x,y)\} \leq w_{\max}$. The assumption that $Z_w(y) \in [0,1]$ is justified as we can equivalently estimate $Z_w(y)/w_{\max}$.

A prominent method to approximate such sums is constructing unbiased estimators of low variance. The simplest and extremely general approach to get such estimators is through uniform random sampling. Letting $\chi \in (0,1)$ be an upper bound on the failure probability, a second moment argument shows that storing and querying a uniform random sample of size $O\left(\frac{1}{\epsilon^2}\frac{1}{\tau}\log(1/\chi)\right)$ is sufficient and necessary in general [9], [10], to approximate the sum $\mu = Z_w(y)$ for any $\mu \geq \tau$. The dependence on ϵ, χ is standard and easily shown to be necessary, so the question is *for which class of functions can we improve the dependence on τ?*

In this paper, we focus on the class of log-convex functions of the inner product between two vectors. For the unit sphere such functions can be written as $w(x,y) = e^{\phi(\langle x,y\rangle)}$ for some convex function $\phi : [-1,1] \to \mathbb{R}$ of the inner product between $x, y \in \mathcal{S}^{d-1}$. Approximate summation of such functions has several fundamental applications in Machine Learning, including:

- **Partition Function Estimation** [11], [12]: a basic

2575-8454/19/$31.00 ©2019 IEEE
DOI 10.1109/FOCS.2019.00051

Table I
EXAMPLES OF LOG-CONVEX FUNCTIONS OF THE INNER PRODUCT $\rho = \langle \frac{x}{\|x\|}, \frac{y}{\|y\|} \rangle$ FOR $x, y \in r\mathcal{S}^{d-1}$. $L(\phi)$ DENOTES THE LIPSCHITZ CONSTANT OF $\phi : [-1, 1] \to \mathbb{R}$.

$w(x,y)$	$\phi(\rho)$	$L(\phi)$
$e^{\langle x,y \rangle}$	$r^2\rho$	r^2
$e^{-\|x-y\|_2^2}$	$2r^2(\rho - 1)$	$2r^2$
$(\|x-y\|_2^2 + 1)^{-1}$	$-\log(1 + (1-\rho)2r^2)$	$2r^2$
$(1 + \exp(-\langle x,y \rangle))^{-1}$	$-\log(1 + e^{-r^2\rho})$	r^2
$(\langle x,y \rangle + cr^2)^{-k}$	$-k\log(r^2(\rho + c))$	$\frac{k}{c-1}$

workhorse in statistics are exponential families where, given a parameter vector $y \in \mathbb{R}^d$, for all $x \in X \subseteq \mathbb{R}^d$ a probability distribution is defined by setting $p_y(x) \propto e^{\langle x,y \rangle}$. Exponential families find many applications in Natural Language Processing (NLP) such as word embeddings and text generation [13], [5], [14], [6]. The normalizing constant $Z(y) = \sum_{x \in X} e^{\langle x,y \rangle}$ is called the *partition function*. Approximating this quantity is important for sampling, hypothesis testing and inference.

- **Kernel Density Estimation:** a non-parametric way [15] to estimate the "density of a set X at y" is through $Z(y) = \frac{1}{n\sigma^d} \sum_{i=1}^{n} \exp(-\frac{\|x_i - y\|^2}{\sigma^2})$. Such an estimate is used in algorithms for outlier detection [16], [17], topological data analysis [18] and clustering [19].
- **Logistic activation and Multi-label Classification:** setting $\phi(\rho) = -\log(1 + e^{-\rho})$, we get the logistic function $e^{\phi(\rho)} = \frac{1}{1+e^{-\rho}}$. A basic building block in classification with n labels, is for each label i to train a separate linear classifier (vector x_i) and then assign a query point y to a label $J \in \{1, \ldots, n\}$ by setting $J = j$ with probability $\propto \frac{1}{1+e^{-\langle x_j, y \rangle}}$. When the number of labels is large, going through all the labels is impractical and faster algorithms are sought [20], [21]. Approximating the sum of the activations is, as we will see, intimately related to sampling.

More examples of log-convex functions are presented in Table I. Obtaining fast algorithms for approximating summations gives speedups to all of the above settings. For such functions we denote $Z_w(y)$ as $Z_\phi(y)$ and normalize by $e^{\phi_{\max}}$ whenever $\phi_{\max} \neq 0$. Let $L(\phi)$ be the lipschitz constant of the function ϕ. For points on the unit sphere, we have that $\mu := \frac{Z_\phi(y)}{e^{\phi_{\max}}} \geq e^{-(\phi_{\max} - \phi_{\min})} \geq e^{-2L(\phi)} := \tau$ and hence random sampling requires $O(\frac{1}{\epsilon^2} e^{2L(\phi)})$ samples. For $L(\phi) \geq \frac{1}{2}\log n$, random sampling offers *no improvement* over the trivial algorithm. In this work we design the first sub-linear algorithms for the problem of summing general log-convex functions of the inner product.

A. Our results

At a high level, we significantly generalize the recent approach of Hashing-Based-Estimators [10] to handle more general functions. This is done by combining classical ideas from *Harmonic analysis* (partitions of unity and approximation theory) with recent results for *similarity search*. We give a general technique for approximating pairwise summations that gives the following result for log-convex functions:

Theorem 1 (Main Result)**.** *Given a convex function $\phi : [-1, 1] \to \mathbb{R}$ with lipschitz constant $L(\phi) < (1-\delta)\log n$ for $\delta > 0$, there exists a data structure that for $\epsilon > 0$ and any set of n vectors $X \subset \mathcal{S}^{d-1}$ can provide a $(1 \pm \epsilon)$-approximation to $Z_\phi(y)$ for any query $y \in \mathcal{S}^{d-1}$ with constant probability and query time $n^{1-\delta+o(1)}/\epsilon^2$ using space/pre-processing time $n^{2-\delta+o(1)}/\epsilon^2$.*

The exact dependence on the lipschitz constant is $e^{(1+o(1))L(\phi)}$. To put our result into context, compared to random sampling that in worst case requires $O(\frac{1}{\epsilon^2} e^{2L(\phi)})$ samples, we offer a square root improvement for a large family of functions.

Although this theorem is phrased for points on the unit sphere, our results are more general. In particular, for $r_X, r_Y > 0$ assume that $X \subset r_X\mathcal{S}^{d-1}$, $y \in r_Y\mathcal{S}^{d-1}$ and that we wish to sum the function $e^{\tilde{\phi}(\langle x,y \rangle)}$. We may map this setting to the unit sphere by setting $\hat{x} = \frac{x}{\|x\|}$ for all $x \neq 0$ and defining $\phi(\langle \hat{x}, \hat{y} \rangle) := \tilde{\phi}(r_X r_Y \langle \hat{x}, \hat{y} \rangle) = \tilde{\phi}(\langle x, y \rangle)$ with the lipschitz constant $L(\phi) = r_X r_Y \cdot L(\tilde{\phi})$ increased by a factor of $r_X r_Y$. For example, in the natural case of binary vectors $X \subset \{\pm 1\}^d$ with $d = c\log n$ we have that $r_X = \sqrt{c \log n}$. Hence, if $L(\tilde{\phi}) = 1$ (e.g. $w(x,y) = e^{\langle x,y \rangle}$) our theorem applies when $r_Y \leq \frac{1-\delta}{\sqrt{c}}\sqrt{\log n}$. Moreover, we can handle the general case (Section VI), where points do not lie exactly on a sphere, by partitioning the space in "thin" annuli $(1+\gamma)^i r_0 \leq \|x\| < (1+\gamma)^{i+1} r_0$ with $\gamma = O(1/r_X r_Y L(\tilde{\phi}))$ and applying Theorem 1 for each possible pair of annuli (points and query).

We show that under popular conjectures a restriction on $L(\phi)$ is necessary in order to obtain sublinear algorithms for the problem even on average over n queries. In fact, it turns out that $L(\phi)$ needs to be $O(\log n)$ even if one allows for polynomially large approximation factors. Our hardness result is based on either of the following two conjectures that have been the base of a flurry of quadratic hardness results in the past years.

Conjecture 1 (Strong Exponential Time Hypothesis (SETH)[22])**.** *For any $\epsilon > 0$, there exists $k = k(\epsilon)$ such that k-SAT on n variables cannot be solved in time $O(2^{(1-\epsilon)n})$.*

Conjecture 2 (Orthogonal Vectors Conjecture (OVC) [23], [24])**.** *For every $\delta > 0$ there exists $c = c(\delta)$ such that given two sets $A, B \subset \{0,1\}^m$ of cardinality N, where $m = c\log N$, deciding if there is a pair $(a, b) \in A \times B$ such that*

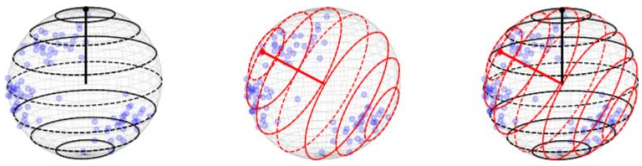


Figure 1. Angular partitions for two different query points (black and red) for a fixed dataset.

$a^\top b = 0$ *cannot be solved in time* $O(N^{2-\delta})$.

Using a recent result of Rubinstein [25], we show the following.

Theorem 2. *Unless SETH and OVC fail, for every* $\delta > 0$ *and* $\alpha \geq 1$ *there exists a constant* $C(\delta, \alpha) > 0$ *such that for two sets* $X, Y \subset \mathcal{S}^{d-1}$ *of size* n *with* $d = O_\delta(\log n)$ *and* $L > C(\delta, \alpha) \cdot \log n$*, there exists no* $n^{2-O(\delta)}$ *algorithm that produces an* α*-approximation to* $\frac{1}{n} \sum_{y \in Y} \left(\frac{1}{n} \sum_{x \in X} e^{L \cdot \langle x, y \rangle} \right)$.

The precise dependence is $C(\delta, \alpha) = O(e^{e^{\frac{\delta}{c(\delta)}}})(1 + \log \alpha / 2 \log n)$ where $c(\delta)$ is a constant. Even if we allow for approximation factor $\alpha = n^s$ with $s > 0$, we see that $C(\delta, n^s)$ is still a constant. The intuition behind this result is that when $L = \Omega(\log n)$ the function $e^{L\langle x,y \rangle}$ varies fast enough so that the presence or absence of a single pair of "relatively close" points can dominate the sum. In applications, though, the Lipschitz constant encountered is often small (e.g. [14, Section 2]).

B. Motivation: Partitions of Unity

Next, we offer some motivating remarks on how might one go about designing algorithms for pairwise summation problems.

Given a natural number T, let $[T] := \{1, \ldots, T\}$. A general way to estimate sums over X is to define a query-dependent partition $\mathcal{P}(y) = \{P_1(y), \ldots, P_T(y)\}$ of X in T parts and express the sum as $\sum_{t \in [T]} \left(\sum_{x \in P_t} w(x, y) \right)$. If for the specific partition there exist $M \geq 1$ such that $\forall t \in [T]$ and $\forall x_1, x_2 \in P_t(y)$:

$$\frac{1}{M} \cdot w(x_2, y) \leq w(x_1, y) \leq M \cdot w(x_2, y). \tag{2}$$

taking $O(M/\epsilon^2)$ random samples would give us an accurate estimate of each term $\sum_{x \in P_t} w(x, y)$ and using at most $O(MT/\epsilon^2)$ samples we would obtain a good estimate of the sum. The problem is that generating and sampling from such a partition efficiently for any query y can be computationally challenging. For example if $w(x, y) = e^{-\|x-y\|^2}$ and points $X \subset r\mathcal{S}^{d-1}$ lie on a sphere of radius $r > 0$, then $L(\phi) = 2r^2$ and such partitions are equivalent to being able to sample from a certain *angular (inner product) range* around the query $y \in r\mathcal{S}^{d-1}$ (Figure 1), as

$$\frac{w(x_1, y)}{w_2(x_2, y)} = e^{2r^2(\langle \frac{x_1}{r}, \frac{y}{r} \rangle - \langle \frac{x_2}{r}, \frac{y}{r} \rangle)} \leq M$$
$$\Rightarrow \left| \left\langle \frac{x_1}{r}, \frac{y}{r} \right\rangle - \left\langle \frac{x_2}{r}, \frac{y}{r} \right\rangle \right| \leq \frac{\log M}{L(\phi)}$$

Setting $M = e^{(c^2-1)L}$, the resulting angular range of length $c^2 - 1$ corresponds to distances differing by a factor of $1 \leq c \leq \sqrt{3}$. Sampling from such partitions in high dimensions can be costly [26], [27].

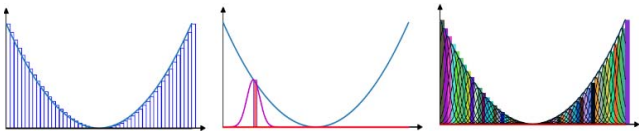


Figure 2. Partitions of unity as a tool of rewriting integrals in terms of of localized functions

Partitions of unity: Instead of a partition $\mathcal{P}$, consider a collection of functions $\tilde{w}_t(x, y)$ such that $\sum_{t \in [T]} \tilde{w}_t(x, y) = 1, \forall x \in X$. Each such function concentrates its mass on a small portion of the space – this can be thought of as a soft partition. Such a collection of functions is called a *partition of unity* (Figure 2) and is widely used in Harmonic analysis. We will use partitions of unity to define estimators for which we can control their first and second moments through linearity of expectation and provide a generic recipe to use them within the framework of Hashing-based-Estimators, e.g. hashing-based important sampling, to bound the overall variance. The intuition is that we can use carefully designed hash functions to sample points according to a soft partition.

C. Our techniques

The main conceptual contribution of this work is a new framework for approximating pairwise summations. Our framework is based on a class of estimators that we introduce, called *Multi-resolution* Hashing-Based-Estimators, that significantly generalizes previous work [10]. The main idea is that, instead of a single hashing scheme, we have a collection of hash families $\mathcal{H}_t$ for $t \in [T]$, where each $\mathcal{H}_t$ is responsible for a different portion of the angular range around the query; $\mathcal{H}_t$ has relatively high collision probability within the range assigned to it and relatively low outside. We divide up the task of estimating the summation of interest amongst these various hash families by assigning data points $x \in X$ to $t \in [T]$ via a soft partition (i.e. a partition of unity). Our end goal is to produce an unbiased estimator and bound its variance by selecting the hashing scheme and partition of unity appropriately. While this overall scheme sounds complicated, we show that a particular choice of weights for the soft partition (as a function of collision probabilities) makes the analysis modular and tractable: *for the purpose of analysis, the collection of hash families behaves like a single*

hash family whose collision probability is the supremum of the collision probabilities for $\mathcal{H}_t, t \in [T]$. We now flesh out this informal description.

Multi-resolution HBE (MR-HBE): Given a collection $\mathcal{H}_1, \ldots, \mathcal{H}_T$ of hashing schemes with collision probabilities $p_1, \ldots, p_T : \mathbb{R}^d \times \mathbb{R}^d \to [0,1]$ and functions $\tilde{w}_t : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ for $t \in [T]$, such that $\sum_{t \in T} \tilde{w}_t(x,y) = 1$ (*partition of unity*) and $w_t(x,y) := \tilde{w}_t(x,y) w(x,y)$, we form an unbiased estimator by:

- **Preprocessing:** for all $t \in [T]$, sample a hash function $h_t \sim \mathcal{H}_t$ and evaluate it on X creating hash table H_t . Let $H_t(z) \subseteq X$ denote the subset of points in X that are mapped to the same hash bucket as $z \in \mathbb{R}^d$ under h_t.
- **Querying:** given a query $y \in \mathbb{R}^d$, for all $t \in [T]$ let $X_t \sim H_t(y)$ be a random element from $H_t(y)$ or $\perp$ if $H_t(y) = \emptyset$. Return $Z_T(y) = \frac{1}{|X|} \sum_{t \in [T]} \frac{w_t(X_t,y)}{p_t(X_t,y)} |H_t(y)|$, where it is understood that if $X_t = \perp$ the corresponding term is 0.

The conditions on $\{\tilde{w}_t\}$ and $\{p_t\}$ ensure that the estimator is unbiased. The motivation behind these estimators is to use the extra freedom in selecting $\{\tilde{w}_t\}$ and $\{p_t\}$ so that we can obtain better bounds on the overall variance. This is quite challenging as the variance of each of the T terms in the sum depends on the whole data set through $|H_t(y)|$. This raises the question whether there exist design principles for $\{\tilde{w}_t\}$ and $\{p_t\}$ that lead to low variance? We introduce two key design principles:

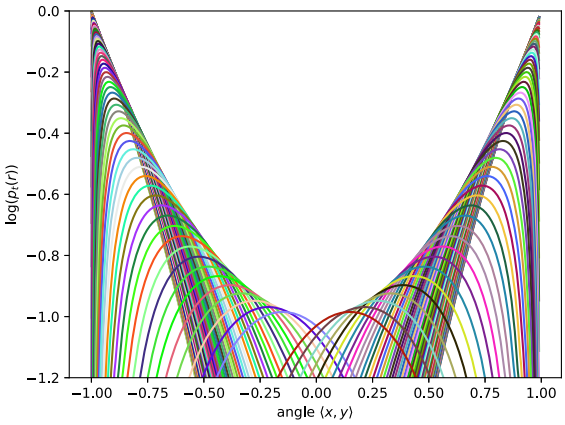
- **Variance bounds and p^2-Weighting:** For a fixed collection of weight functions $\{\tilde{w}_t\}$ and collision probabilities $\{p_t\}$, by utilizing a lemma from [10], we get an explicit bound on the variance of the estimator for a query $y \in \mathbb{R}^d$ only as a function of $\{w_t(\cdot, y)\}, \{p_t(\cdot, y)\}$ and $\mu := Z_\phi(y)$. We then minimize a separable relaxation of our upper bound to obtain the *p^2-weighting scheme* where

$$\tilde{w}_t(x,y) = \frac{p_t^2(x,y)}{\sum_{t'=1}^{T} p_{t'}^2(x,y)} \text{ for all } x, y \in \mathbb{R}^d. \quad (3)$$

- **Approximation by a supremum of functions:** Using the p^2-weighting scheme and after some algebraic manipulations, we are able to get an upper bound on the variance that depends only on $w(x,y)$, $\mu = Z_w(y)$ and on the *pointwise supremum* of the collision probabilities $p_*(x,y) := \sup_{t \in [T]} \{p_t(x,y)\}$. An interesting fact that comes out from the analysis is that the resulting bound is closely related to the variance of a single HBE, i.e. $T = 1$, with collision probability equal to $p_*(x,y)$. Exploiting this connection and by providing a simplified proof for a theorem of [10] that bounds the variance of *scale-free HBE*, we identify the second design principle, namely designing $\{p_t\}$ such that:

$$p_*(x,y) = \sup_{t \in T} \{p_t(x,y)\} = \Theta(\sqrt{w(x,y)}). \quad (4)$$

Figure 3. Approximation of the *squared inner product* $\phi_2(\rho) = \rho^2 - 1$ function by elements of (5).



Observe that so far our discussion has been about the variance, or on how many independent realizations of Multi-resolution HBE we need to efficiently estimate $Z_\phi(y)$, and we have not mentioned the time needed to compute each one. The natural question is then: for which family of functions $w(x,y)$, does there exist a family of hashing schemes $\{(\mathcal{H}_t, p_t)\}$ satisfying (4) that can be efficiently constructed and evaluated?

Approximating Log-convex Functions via Distance Sensitive Hashing: We show that constructing a family of hash functions satisfying (4) is indeed possible for log-convex functions of the inner product by utilizing a family of hashing schemes introduced recently by Aumuller et al. [27], referred to as *Distance Sensitive Hashing* (DSH). This family is defined through two parameters $\gamma \geq 0$ and $s > 0$, with collision probability $p_{\gamma,s}(\rho)$ having the following dependence on the inner product $\rho = \langle x, y \rangle$ between two vectors $x, y \in \mathcal{S}^{d-1}$

$$\log\left(1/p_{\gamma,s}(\rho)\right) = \Theta\left(\left(\frac{1-\rho}{1+\rho} + \gamma^2 \frac{1+\rho}{1-\rho}\right) \frac{s^2}{2}\right). \quad (5)$$

Exploiting the fact that $w(x,y) = \frac{e^{\phi(\langle x,y \rangle)}}{e^{\phi_{\max}}}$ is only a function of the inner product $\langle x, y \rangle$, (4) becomes equivalent to constructing hash families with collision probabilities $p_1, \ldots, p_T$ such that:

$$\left|\log \sup_{t \in [T]} \{p_t(\rho)\} - \frac{1}{2}(\phi(\rho) - \phi_{\max})\right| = O(1). \quad (6)$$

The approximation is achieved by: (a) producing a sequence of explicit "interpolation points" $\rho_1, \ldots, \rho_T \in [-1, 1]$, (b) using a single DSH scheme with parameters $\gamma_t, s_t \geq 0$ and log-collision probabilities $\log p_t$ to approximate the function $\frac{1}{2}(\phi(\rho) - \phi_{\max})$ locally around each ρ_t (value and derivative) *(multi-resolution)*, (c) and then using convexity of ϕ and "concavity" of p_{γ_t, s_t} to bound the error in (6) (Section

IV). An interesting fact is that in order to achieve the above approximation guarantee using DSH, *convexity of the function ϕ is instrumental* (Lemma 3 and Proposition 3). We give an example of the resulting approximation in Figure 3.

The number of hash families T as well as the approximation error in (6) are *sub-linear* in the Lipschitz constant $L(\phi)$ of the function. This sub-linear dependence is the result of achieving a trade-off between evaluation time of the hash functions and fidelity of approximation in (6). The evaluation time of the hash functions roughly scales like $e^{O(\max_{t\in[T]} s_t^2)}$ whereas an error Δ in (6) increases the variance by a factor of $e^{O(\Delta)}$ (Section V). To trade-off the two terms we design the hash families in order to approximate the scaled-down version $\frac{1}{2k}(\phi - \phi_{\max})$ to error Δ/k (decreases $\max_{t\in[T]}\{s_t^2\}$) and then use concatenation of k i.i.d hash functions to get the collision probabilities to be of the correct order (has the effect of increasing the approximation error by a factor of k) and overall error Δ.

D. Applications

To illustrate the above techniques, we give concrete examples for which our data structures have $n^{0.5+o(1)}$ query time, i.e. $L(\phi) \leq \log(n)/2$.

Corollary 1. *Let $\Phi_{r,k,c}$ be the set of functions in Table I with parameters $r \leq \frac{1}{2}\sqrt{\log n}$ and $0 \leq k \leq \frac{c-1}{2}\log n$. Then for any $\phi \in \Phi_{r,k,c}$ and $X \subset r\mathcal{S}^{d-1}$, there exists a data structure using space $n^{1.5+o(1)}/\epsilon^2$ that for any $y \in r\mathcal{S}^{d-1}$ can produce a $(1\pm\epsilon)$-approximation to $Z_\phi(y)$ in time $n^{0.5+o(1)}/\epsilon^2$.*

This corollary highlights the main point of our paper: we provide a *general technique* that enables the design of data structures that solve a variety of pairwise integration problems. For the special case of the Gaussian kernel for points on a sphere, our data structure has the same dependence in ϵ, r (up to poly-logarithmic factors in n) as the currently best known algorithm [10]. At the same time we are able to handle the important case of the logistic function $\frac{1}{1+e^{-\langle x,y\rangle}}$ and non-monotone functions such as $e^{|\langle x,y\rangle|}$ or $e^{\langle x,y\rangle^2} = e^{\sum_{\ell,\ell'=1}^d y_\ell y_{\ell'} x_\ell x_{\ell'}}$ for which no previous algorithms were known.

Sampling: The problem of approximating pairwise sums sometimes referred to as partition function approximation is closely related to the problem of *sampling from discrete distributions*. Given a non-negative function $w(x,y)$ and a set $\{x_1,\ldots,x_n\} \subset \mathbb{R}^d$ we would like to produce a random variable $I \in [n] := \{1,\ldots,n\}$ such that $\mathbb{P}[I=i] = \frac{w(x_i,y)}{\sum_{i=1}^n w(x_i,y)}$ for all $i \in [n]$. This can always be done by spending linear time, the question that we ask here is *how fast can one produce a random variable $\tilde{I}$ whose distribution is close in total variation distance to the distribution of I?*

We show in a generic way that not only our methods can be used to approximate the partition function but also to sample from a distribution that has small total variation distance to the desired distribution. In fact this is true for any method that produces a sample with the following properties.

Definition 1. *Given non-negative weights $w_1,\ldots,w_n$ and parameters $\epsilon,\zeta,\chi \in [0,1)$ a sequence of m pairs of random variables $(\hat{w}_t, I_t) \in \mathbb{R}_+ \times [n]$ for $t \in [m]$ is called (ζ,ϵ,χ)-*sample *iff:*

- *for all $i \in [n]$, $\mathbb{E}[\sum_{t=1}^m \hat{w}_t \mathbb{I}[I_t = i]] \leq (1+\zeta)w_i$,*
- *$\mathbb{P}[|\sum_{t=1}^m \hat{w}_t - \sum_{i=1}^n w_i| \geq \epsilon \sum_{i=1}^n w_i] \leq \chi$.*

For two random variables $I, J \in [n]$ let $\mathrm{TV}(I,J) := \max_{A\subset[n]} |\sum_{i\in A}\mathbb{P}[I=i] - \sum_{j\in A}\mathbb{P}[J=j]|$.

Lemma 1. *Given a (ζ,ϵ,χ)-sample for $w_1,\ldots,w_n$ of size m, one can construct a random variable $\tilde{I} \in [n]$ in time $O(m)$ such that $\mathrm{TV}(\tilde{I}, I) \leq \frac{\zeta+\epsilon}{1-\epsilon} + \chi$ where $\mathbb{P}[I=i] = \frac{w_i}{\sum_{j=1}^n w_j}$.*

Proof: Define $\tilde{I}$ to be the random variable such that $\mathbb{P}[\tilde{I}=j|\{(\hat{w}_t,I_t)\}_{t\leq m}] = \frac{\sum_{t=1}^m \hat{w}_t \mathbb{I}[I_t=j]}{\sum_{t'=1}^m \hat{w}_{t'}}$, i.e. after we get the sample $\{(\hat{w}_t,I_t)\}_{t\in[m]}$ we return I_t with probability proportional to $\hat{w}_t$. Let $n\mu = \sum_{i=1}^n w_i > 0$ and F be the event that $|\sum_{t=1}^m \hat{w}_t - n\mu| \geq \epsilon n\mu$, by our assumption $\mathbb{P}[F] \leq \chi$.

$$\begin{aligned}\mathbb{P}[\tilde{I}=j] &= \mathbb{E}[\mathbb{P}[\tilde{I}=j|\{(\hat{w}_t,I_t)\}_{t\leq m}]\mathbb{I}[F]] \\ &\quad + \mathbb{E}[\mathbb{P}[\tilde{I}=j|\{(w_t,I_t)\}_{t\leq m}]\mathbb{I}[F^c]]\end{aligned}$$

Consider the set $A = \{j : \mathbb{P}[\tilde{I}=j] \geq \frac{w_j}{n\mu}\}$. Then, the total variation distance can be bounded as

$$\begin{aligned}\mathrm{TV}(\tilde{I},I) &= \sum_{j\in A}\left\{\mathbb{P}[\tilde{I}=j] - \frac{w_j}{n\mu}\right\} \\ &= \sum_{j\in A}\Big\{\mathbb{E}[\mathbb{P}[\tilde{I}=j|\{(\hat{w}_t,I_t)\}_{t\leq m}]\mathbb{I}[F^c]] \\ &\qquad + \mathbb{E}[\mathbb{P}[\tilde{I}=j|\{(\hat{w}_t,I_t)\}_{t\leq m}]\mathbb{I}[F]] - \frac{w_j}{n\mu}\Big\} \\ &= \sum_{j\in A}\left\{\mathbb{E}[\mathbb{P}[\tilde{I}=j|\{(\hat{w}_t,I_t)\}_{t\leq m}]\mathbb{I}[F^c]] - \frac{w_j}{n\mu}\right\} \\ &\qquad + \mathbb{E}\left[\sum_{j\in A}\mathbb{P}[\tilde{I}=j|\{(\hat{w}_t,I_t)\}_{t\leq m}]\mathbb{I}[F]\right] \\ &\leq \sum_{j\in A}\left\{\mathbb{E}[\frac{\sum_t \hat{w}_t\mathbb{I}[I_t=j]}{\sum_{t'}\hat{w}_{t'}}\mathbb{I}[F^c]] - \frac{w_j}{n\mu}\right\} + \mathbb{P}[F] \\ &\leq \sum_{j\in A}\left\{\frac{\mathbb{E}[\sum_t \hat{w}_t\mathbb{I}[I_t=j]\mathbb{I}[F^c]]}{(1-\epsilon)n\mu} - \frac{w_j}{n\mu}\right\} + \mathbb{P}[F] \\ &\leq \sum_{j\in A}\left\{\frac{1+\zeta}{1-\epsilon} - 1\right\}\frac{w_j}{n\mu} + \mathbb{P}[F] \\ &\leq \frac{\zeta+\epsilon}{1-\epsilon}\sum_{j\in A}\frac{w_j}{n\mu} + \mathbb{P}[F] \\ &\leq \frac{\zeta+\epsilon}{1-\epsilon} + \chi.\end{aligned}$$

■

To the best of our knowledge this lemma establishes a novel connection between sampling from discrete distributions and partition function approximation ("counting"), not captured by the self-reducible setting of Jerrum, Valiant, Vazirani [28] and Jerrum, Sinclair [29]. As such it might be of independent interest. For the special case of log-convex functions on the unit sphere we have the following result.

Corollary 2. *Given a convex function $\phi : [-1, 1] \to \mathbb{R}$ with lipschitz constant $L(\phi)$, there exists a data structure that for $\epsilon > 0$ and a set of n vectors $X \subset \mathcal{S}^{d-1}$ produces for any query $y \in \mathcal{S}^{d-1}$ an $(0, \epsilon, \epsilon)$-sample for the weights defined by $w(x, y) = e^{\phi(\langle x,y\rangle)}$ of size $O(e^{(1+o(1))L(\phi)}/\epsilon^3)$ in time $O(e^{(1+o(1))L(\phi)}/\epsilon^3)$ using space/pre-processing time $O(n \cdot e^{(1+o(1))L(\phi)}/\epsilon^3)$.*

Summary: Our work provides a general technique that reduces the computational task of summing a pairwise function over a large dataset to the task of constructing a family of hash functions whose square root of the pointwise supremum of collision probabilities approximates the function in question.

E. Related work

Recent approaches on obtaining sub-linear algorithms for pairwise summation are based on two different ideas: *Hashing-based Importance Sampling* and *Well-conditioned Partitions* .

1) Hashing-based Importance Sampling: Importance Sampling aims to reduce the variance of uniform random sampling by sampling points according to some biased distribution that assigns *greater probability* to points with *higher value* $w(x, y)$. The challenge in our setting is that such a distribution needs to be adaptive to the query $y \in \mathbb{R}^d$ and to admit an efficient sampling algorithm at query time. The approach of using hashing to perform importance sampling was introduced independently by Charikar-Siminelakis [10] and Spring-Shrivastava [30]. Since then these ideas have found many applications in machine learning and data analysis [31], [32], [33].

Hashing-Based-Estimators (HBE): In a previous work of the authors [10], the general approach of using hashing to create importance sampling schemes with provable low-variance was introduced under the name of Hashing-Based-Estimators. Given a *single* hashing scheme $\mathcal{H}$ with collision probability $p(x, y) = \mathbb{P}_{h\sim\mathcal{H}}[h(x) = h(y)]$ an *unbiased estimator* for $Z_w(y)$ is constructed through a two-step sampling process that corresponds to the $T = 1$ case of Multi-Resolution HBE.

Limitations of HBE: The approach of HBE hinges upon constructing a *single hashing scheme* that has the property $p(x, y) = \Theta(\sqrt{w(x, y)})$. This can be quite difficult to achieve with hash functions that can be efficiently stored and evaluated. In fact, the authors were able to carry out this approach *for exactly three functions*: the Gaussian $e^{-\|x-y\|_2^2}$, Exponential $e^{-\|x-y\|_2}$, and Generalized t-Student $1/(1 + \|x - y\|_2^p)$ kernels using *Locality Sensitive Hashing* schemes of Andoni-Indyk [34] and Datar et al. [35]. This is due the fact that these LSH schemes exhibited collision probabilities that matched the aforementioned functions. Hence, there are severe restrictions on the classes of functions for which sub-linear algorithms can be obtained through HBE.

Comparison: In this work, we essentially remove the main bottleneck of the Hashing-based approach and make it more broadly applicable. This is done by using the idea of Partitions of Unity via Multi-Resolution HBE, and identifying key design principles (3) and (4) that provably lead to an overall low-variance estimator. In doing so we also provide a more general theorem for the variance of scale-free estimators (Theorem 6).

2) Partition-based approaches and Smoothness: The idea of partition-based approaches, is to efficiently partition points in a small number of parts such that some simple primitive (Random Sampling or Polynomial approximation) can be used to accurately estimate the contribution of each part. This approach in low dimensions, is known under the names of Fast-Multipole Methods [36] or Well Separated Pair Decomposition [37] and the complexity scales typically as $\log(1/\epsilon)^{O(d)}$ [38] for additive error ϵ.

Due to the explosion in Machine learning applications the problem was revisited in the high-dimensional case through works on "Dual-tree Algorithms" [39], [40], [41] that aimed to exploit an underlying low dimensional structure [42] (when it exists). However, no theoretical results were known for the general case.

"Non-smooth" functions: The lower bound presented here, inspired by [43], shows that this is for good reason. In high dimensions $d = \Omega(\log n)$, even for simple functions (e.g. Gaussian kernel), and under no restrictions on the rate that the function changes we do not expect to be able to get sub-linear algorithms barring major progress in complexity theory (e.g. refuting SETH).

"Smooth" functions: In a recent work [44], it was established that indeed in high dimensions quick variation of the pairwise function is the only obstacle in obtaining efficient algorithms. In particular, the authors of [44] introduced the following notion of (C, L)-smoothness that captures functions that vary polynomially fast with distance:

$$\max\left\{\frac{w(x,y)}{w(x',y)}, \frac{w(x',y)}{w(x,y)}\right\} \le C\left\{\frac{\|x-y\|}{\|x'-y\|}, \frac{\|x'-y\|}{\|x-y\|}\right\}^L$$

and showed that one can get $\mathrm{poly}(2^L, \log n, \frac{1}{\epsilon})$-time algorithms giving exponential improvement over the linear time algorithm for small values of $L = o(\log n)$. This was achieved by showing that one can efficiently construct query-dependent partitions (in time roughly $2^{O(L)}$) that are *"good on average"* when random sampling is used to approximate the contribution of each part. Interestingly, ideas related to

hashing were instrumental to both constructing and analyzing the partitions. The authors also provided an intimate connection to the problem of Approximate Near Neighbor Search (ANNS) by showing that for "radial" and smooth functions one can solve the problem given oracle access to an c-ANNS data structure using $\mathrm{poly}(c^L, \log n, \frac{1}{\epsilon})$ calls.

Comparison: The class of log-convex functions studied in this paper *does not* satisfy (in general) this definition of smoothness (exponential vs polynomial). Still, in order to compare the two methods for $x, y \in r\mathcal{S}^{d-1}$, we may use the function $\frac{1}{(\langle x,y\rangle+2r^2)^k} = \frac{1}{(3r^2-\frac{1}{2}\|x-y\|^2)^k}$ that is both $(O(1), 2k)$-smooth and log-convex with Lipschitz constant k. For $k = \frac{1}{2}\log n$ our algorithms run in time $n^{0.5+o(1)}$ (Corollary 1) whereas the approach in [44] offers no improvement over the linear algorithm.

3) Partition Function Estimation: For the special case of *log-linear models*, there is a different approach that relies on LSH to approximate the partition function [45], [46]. In the heart of this approach are two reductions. For $\alpha \geq 1$, the first one is reducing the problem of obtaining a α-approximation to the inverse of the Parition Function to obtaining an $\log(\alpha)$-*additive approximation* for the problem of Maximum Inner Product Search (Gumbel trick). The second one, is reducing the problem of MIPS to the problem of $(1+\gamma(\alpha))$-approximate nearest neighbor search (ANNS). Using the best known data-structure for ANNS [47], this method requires *worst case* time/space $\Omega(n^{1-O(\gamma(\epsilon))})$, which is tight [48]. For vectors in $r\mathcal{S}^{d-1}$, the dependence is $\gamma(\alpha) = O(\frac{\log\alpha}{r^2})$. Hence, at least for adversarial data-sets this approach cannot bring forth significant improvements unless $r = O(\sqrt{\log\alpha})$. Nevertheless, the authors [46] have shown experimentally that their method is still competitive compared to uniform sampling.

F. Outline of the paper

In the next section, we describe the basis of our approach and introduce the main tools we need. In Section III, we derive the key design principles for Multi-resolution HBE and show how they yield provable bounds on the variance. In Section IV, we use an idealized version of the collision probabilities provided by Distance Sensitive Hashing to approximate log-convex functions. In Section V, we finish the construction of our estimators for the unit sphere and prove our main result. In Sections VI and VIII, we show respectively how to extend this construction to Euclidean space and to estimate vector functions, whereas in Section VII we give the proof of the lower bound. Finally, in Section IX, we provide the proofs for some intermediate lemmas and conclude with some future directions in Section X.

II. Preliminaries

We introduce some parameters that capture the complexity of a function for our purposes.

Definition 2. *Let $S \subset \mathbb{R}$, a function $\phi : S \to \mathbb{R}$ is called* Lipschitz with constant $0 \leq L < \infty$ *if for all $\rho_1, \rho_2 \in S$, $|\phi(\rho_1) - \phi(\rho_2)| \leq L|\rho_1 - \rho_2|$. For given ϕ, we denote by $L(\phi)$ the minimum such constant.*

For a function $\phi : S \to \mathbb{R}$, let also $R(\phi) := \phi_{\max} - \phi_{\min}$ denote the range of ϕ.

Proposition 1. *Given $a, b \in \mathbb{R}$, we have $L(a\phi + b) = |a|L(f)$, $R(a\phi + b) = |a|R(\phi)$, and $R(\phi) \leq L(\phi) \cdot \sup_{\rho_1,\rho_2\in S} |\rho_1 - \rho_2|$.*

Proof: If $a > 0$, $R(a\phi+b) = a\phi_{\max}+b-(a\phi_{\min}+b) = aR(\phi)$. If $a < 0$, $R(a\phi + b) = a\phi_{\min} + b - (a\phi_{\max} + b) = -aR(\phi) = |a|R(\phi)$. Finally, $|a\phi(x) + b - (a\phi(y) + b)| \leq |a||\phi(x) - \phi(y)| \leq |a|L|x - y|$. The last inequality follows directly by Definition 2. ∎

Throughout the paper for a query $y \in \mathbb{R}^d$ we use $\mu := \mu(y) = Z_w(y)/w_{\max}$ and make the simplifying assumption that floating point operations and evaluation of functions take constant time. For log-convex functions, we assume that $L(\phi)$ is greater than some small constant. Otherwise $O(1/\epsilon^2 \log(1/\chi))$ uniform random samples are sufficient to estimate any $\mu \in [e^{-R(\phi)}, 1]$.

A. Basis of the approach

The starting point of our work is the *method of unbiased estimators*. Assume that we would like to estimate a quantity $\mu = \mu(y)$ using access to samples from a distribution $\mathcal{D}$, such that for $\hat{Z} \sim \mathcal{D}$, $\mathbb{E}[\hat{Z}] = \mu$ and $\mathrm{Var}[\hat{Z}] \leq \mu^2 V_{\mathcal{D}}(\mu)$. The quantity $V_{\mathcal{D}}(\mu)$ (depending possibly on μ) bounds the *relative variance* $\mathrm{RelVar}[\hat{Z}] := \frac{\mathrm{Var}[\hat{Z}]}{(\mathbb{E}[\hat{Z}])^2}$. For $\epsilon > 0$, we get through Chebyshev's inequality that the average of $O(\epsilon^{-2}V_{\mathcal{D}}(\mu))$ samples are sufficient to get $(1 \pm \epsilon)$-multiplicative approximation to μ with constant probability. Moreover, using the median-of-means technique [49], we can make the failure probability to be less than $\chi > 0$ by only increasing the number of samples by a $O(\log(1/\chi))$ factor.

V-bounded Estimators: The above discussion seems to suggest that as long as one has an unbiased estimator $\hat{Z} \sim \mathcal{D}$ for μ and a bound $V_{\mathcal{D}}(\mu)$ on the relative variance, one can accurately estimate μ. The caveat of course is that in cases where $V_{\mathcal{D}}$ is indeed a function of μ, setting the requisite number of samples requires knowledge of μ. An unbiased estimator for which $\mu^2 V_{\mathcal{D}}(\mu)$ is increasing and $V_{\mathcal{D}}(\mu)$ is decreasing is called V*-bounded* [10]. An estimator has complexity $\mathcal{C}$, if using space $O(\mathcal{C}n)$ we can evaluate it, i.e. sample from $\mathcal{D}$, in $O(\mathcal{C})$ time. A general way to construct data-structures to solve estimation problems using V-bounded estimators was recently proposed.

Theorem 3 ([10])**.** *Given a V-bounded estimator of complexity $\mathcal{C}$ and parameters $\epsilon, \tau, \chi \in (0, 1)$, there exists a data structure that using space $O(\frac{1}{\epsilon^2}V_{\mathcal{D}}(\tau)\mathcal{C}\log(1/\chi) \cdot n)$ can provide a $(1 \pm \epsilon)$ approximation to any $\mu \geq \tau$ in time*

$O(\frac{1}{\epsilon^2}\mathcal{C}V_{\mathcal{D}}(\mu)\log(1/\chi))$ *with probability at least* $1-\chi$. *The data-structure can also detect when* $\mu < \tau$.

Our goal is to construct such estimators through hashing and bound their complexity. The above theorem turns our construction into an efficient data-structure for *estimating pairwise summations*.

B. Analytical Tools

For a positive vector $w \in \mathbb{R}^n_{++}$ of coefficients we define the weighted ℓ_1 norm as $\|f\|_{w,1} := \sum_{i=1}^n w_i|f_i|$. The following variational inequality was first proved in [10] and bounds the *maximum of a quadratic form* over the intersection of two weighted ℓ_1-balls. This is going to be the key lemma that will allow us to obtain worst-case bounds on the variance of our estimators.

Lemma 2 ([10])**.** *Given positive vector* $w \in \mathbb{R}^n$, *number* $\mu > 0$, *define* $f_i^* := \min\{1, \frac{\mu}{w_i}\}$. *For any matrix* $A \in \mathbb{R}^{n\times n}$ *:*

$$\sup_{\|f\|_{w,1}\leq\mu, \|f\|_1\leq 1} \{f^\top A f\} \leq 4 \sup_{ij\in[n]} \left\{f_i^*|A_{ij}|f_j^*\right\}.$$

The following crucial lemma, that upper bounds the value of a convex function away from the natural boundary, lies in the core of our ability to use the family of functions (5) to approximate convex functions of the inner product.

Lemma 3. *Let* $\phi : [-1,1] \to \mathbb{R}$ *be a non-constant, non-positive, convex, differentiable function, then*

$$2\phi(\rho_0) < -(1-\rho_0^2)|\phi^{'}(\rho_0)|, \ \forall \rho_0 \in (-1,1). \tag{7}$$

Proof: Let $g(\rho) = \phi^{'}(\rho_0)(\rho-\rho_0)+\phi(\rho_0)$ be the linear approximation of ϕ around $\rho_0 \in (-1,1)$, by convexity we have that $g(\rho) \leq \phi(\rho) \leq 0$. First let's assume that g is increasing, then:

$$g(1) \leq 0 \Rightarrow \phi'(\rho_0) \leq -\frac{\phi(\rho_0)}{1-\rho_0} \tag{8}$$

$$\Rightarrow 2\phi(\rho_0) + (1-\rho_0^2)|\phi^{'}(\rho_0)| \leq 2\phi(\rho_0)(1-\frac{1+\rho_0}{2}) \tag{9}$$

that is always negative. The last inequality follows from the fact that a non-constant convex function attains its maximum only at the boundary of a convex domain. Similarly, if g is decreasing:

$$g(-1) \leq 0 \Rightarrow \phi^{'}(\rho_0) \geq \frac{\phi(\rho_0)}{1+\rho_0} \tag{10}$$

$$\Rightarrow 2\phi(\rho_0) - (1-\rho_0^2)\phi^{'}(\rho_0) \leq 2\phi(\rho_0)\left[1-\frac{1-\rho_0}{2}\right]. \tag{11}$$

We also utilize a structural result for convex functions. ■

Theorem 4 ([50])**.** *Given* $\epsilon > 0$, *there exists an algorithm that given a univariate convex function* f *on an interval* $[a,b]$ *constructs a piecewise linear convex function* ℓ *such that* $0 \leq f(x) - \ell(x) \leq \epsilon$ *for all* $x \in [a,b]$ *using* $O(\sqrt{\frac{(b-a)\Delta}{\epsilon}})$ *linear segments where* $\Delta = f'(b_-) - f^{'}(a_+)$.

C. Hashing

Definition 3 (Asymmetric Hashing)**.** *Given a set of functions* $\mathcal{H} \subset \{h : \mathcal{X} \to \mathcal{U}\}$ *and a probability distribution* ν *on* $\mathcal{H}\times\mathcal{H}$, *we write* $(h,g) \sim \mathcal{H}_\nu$ *to denote a random element sampled from* ν, *and call* $\mathcal{H}_\nu$ *a hashing scheme on* $\mathcal{X}$.

Definition 4 (Hash Bucket)**.** *Given a finite set* $X \subset \mathcal{X}$ *and an element* $(h,g) \in \mathcal{H}\times\mathcal{H}$, *we define for all* $y \in \mathcal{X}$ *the hash bucket of* X *with respect to* y *as* $H_X(y) := \{x \in X | h(x) = g(y)\}$. *For such a hash bucket we write* $X_0 \sim H_X(y)$ *to denote the random variable* X_0 *that is uniformly distributed in* $H_X(y)$ *when the set is not empty and equal to* $\perp$ *when it is.*

The *collision probability* of a hashing scheme $\mathcal{H}_\nu$ on $\mathcal{X}$ is defined by $p_{\mathcal{H}_\nu}(x,y) := \mathbb{P}_{(h,g)\sim\mathcal{H}_\nu}[h(x) = g(y)]$ for all $x, y \in \mathcal{X}$. Whenever it is clear from the context we will omit ν from $\mathcal{H}_\nu$ and X from $H_X(y)$. We also define $\mathcal{H}^{\otimes k}$ to denote the hashing scheme resulting from stacking k independent hash functions from $\mathcal{H}$. For such hashing schemes we have $p_{\mathcal{H}^{\otimes k}}(x,y) = [p_{\mathcal{H}}(x,y)]^k$ for $x, y \in \mathcal{X}$.

D. Multi-resolution Hashing Based Estimators

We define next the class of estimators that we employ.

Definition 5. *Given hashing schemes* $\mathcal{H}_1, \ldots, \mathcal{H}_T$, *with collision probabilities* $p_1, \ldots, p_T : \mathcal{X}\times\mathcal{X} \to [0,1]$, *and weight functions* $w_1, \ldots, w_T : \mathcal{X}\times\mathcal{X} \to \mathbb{R}_+$, *we define for a given set* $X \subseteq \mathcal{X}$, *the Multi-Resolution Hashing-Based-Estimator for all* $y \in \mathcal{X}$ *as:*

$$Z_T(y) := \frac{1}{|X|}\sum_{t=1}^T \frac{w_t(X_t,y)}{p_t(X_t,y)}|H_t(y)|, \tag{12}$$

where $X_t \sim H_t(y) = (H_t)_X(y)$ *and by setting* $w_t(\perp,\cdot) = p_t(\perp,\cdot) = 1$ *for* $t \in [T]$. *We denote such an estimator by* $Z_T \sim \mathrm{HBE}_X(\{\mathcal{H}_t, p_t, w_t\}_{t\in[T]})$.

Again we drop the dependence on X when it is clear from the context. Manipulating conditional expectations gives us the following basic properties for such estimators.

Lemma 4 (Moments)**.** *For any* $y \in \mathcal{X}$ *and* $x \in X$ *let* $T(x,y) =: \{t \in [T] | p_t(x,y) > 0\}$ *and assume that* $\forall x \in X, \sum_{t\in T(x,y)} w_t(x,y) = w(x,y)$ *for a non-negative function* $w : \mathcal{X}\times\mathcal{X} \to \mathbb{R}_+$. *Then,*

$$\mathbb{E}[Z_T(y)] = \mu := \frac{1}{|X|}\sum_{x\in X} w(x,y), \tag{13}$$

$$\mathbb{E}[Z_T^2(y)] \leq \frac{1}{|X|^2}\sum_{x\in X}\left(\sum_{t\in T(x,y)}\frac{w_t^2(x,y)}{p_t(x,y)} \sum_{z\in X}\frac{\min\{p_t(z,y), p_t(x,y)\}}{p_t(x,y)}\right) + \mu^2. \tag{14}$$

The upper bound on the variance comes from $\mathbb{E}\left[|H_t(y)|\middle|x \in H_t(y)\right] \leq \sum_{z\in X} \frac{\min\{p_t(z,y),p_t(x,y)\}}{p_t(x,y)}$.

E. Distance Sensitive Hashing on the unit Sphere

In this subsection, we describe the hashing scheme of Aumuller et al. [27] (see also [51], [47]) and give slightly different bounds on the collision probability that are more appropriate for our purposes.

LSH for unit sphere: We define the hash family $\mathcal{D}_+ = \mathcal{D}_+(t,\zeta)$ that takes as parameters real numbers $t > 0$, $\zeta \in (0,1)$ and defines a pair of hash functions $h_+ : \mathcal{S}^{d-1} \to [m] \cup \{m+1\}$ and $g_+ : \mathcal{S}^{d-1} \to [m] \cup \{m+2\}$, where m is given by

$$m(t,\zeta) = \left\lceil \sqrt{2\pi}(t+1)\log(\frac{2}{\zeta})e^{\frac{t^2}{2}} \right\rceil. \tag{15}$$

To define the functions h_+, g_+, we sample m normal random vectors $g_1, \ldots, g_m \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d)$ and use them to create $m+2$ hash buckets through the mappings

$$h_+(x) := \min\left(\{i\middle|\langle x, g_i\rangle \geq t\} \cup \{m+1\}\right), \tag{16}$$

$$g_+(x) := \min\left(\{i\middle|\langle x, g_i\rangle \geq t\} \cup \{m+2\}\right). \tag{17}$$

The time and memory required for evaluating the function are both bounded by $O(dm) = O(dt\log(\frac{1}{\zeta})e^{\frac{t^2}{2}})$. We also define the hash family $\mathcal{D}_-(t,\zeta)$ that is identical to $\mathcal{D}_+$ except from the fact that instead of using g_+ we use:

$$g_-(x) := \min\left(\{i\middle|\langle x, g_i\rangle \geq -t\} \cup \{m+2\}\right) \tag{18}$$

The need to use a pair of hash functions arises from the fact that we treat the points in the dataset X and the queries differently. We will write $(h,g) \sim \mathcal{D}_s$ for $s \in \{+,-\}$ to indicate such pairs of hash functions. Due to isotropy of the normal distribution the collision probability only depends on $\langle x, y\rangle$,

$$\mathbb{P}_{(h,g)\sim D_\pm}[h(y) = g(x)] = p_\pm(\langle x, y\rangle), \tag{19}$$

and satisfies $p_+(\rho) = p_-(-\rho)$ for all $\rho \in [-1,1]$. Utilizing results for Gaussian integrals [52], [53], we obtain the following explicit bounds.

Lemma 5 (Pointwise bounds)**.** *The collision probability $p_+(\rho)$ is decreasing and for $\delta > 0$ satisfies:*

- $\forall |\rho| \leq 1-\delta$,

$$\frac{\sqrt{2}(1-\zeta)\delta^2}{148\sqrt{\pi}} e^{-\frac{1-\rho}{1+\rho}\frac{t^2}{2}} \leq p_+(\rho) \leq \frac{2}{\sqrt{\pi}\sqrt{\delta}} e^{-\frac{1-\rho}{1+\rho}\frac{t^2}{2}}, \tag{20}$$

- $\forall 1-\delta < \rho \leq 1$,

$$\frac{1-\zeta}{2\sqrt{2\pi}(1+\sqrt{2})} e^{-\frac{\delta}{2-\delta}\frac{t^2}{2}} \leq p_+(\rho) \leq 1, \tag{21}$$

- $\forall -1 \leq \rho \leq -1+\delta$,

$$0 \leq p_+(\rho) \leq \frac{2}{\sqrt{\pi}\sqrt{\delta}} e^{-\frac{2-\delta}{\delta}\frac{t^2}{2}}. \tag{22}$$

The family D_+ tends to map correlated points to the same bucket, whereas D_- tends to map anti-correlated points together. Combining the two hash families, Aumuller et al. [27] created a *Distance Sensitive Hashing* scheme.

DSH for unit sphere: Given real numbers $t, \gamma > 0$ and $\zeta \in (0, 1/2)$, we define the following hash family $\mathcal{D}_\gamma(t,\zeta)$ by sampling a $(h_+, g_+) \sim \mathcal{D}_+(t,\zeta)$ and $(h_-, g_-) \sim \mathcal{D}_-(\gamma t, \zeta)$. We create the hash functions by $h_\gamma(x) := (h_+(x), h_-(x))$ and $g_\gamma(x) := (g_+(x), g_-(x))$ and write $(h_\gamma, g_\gamma) \sim \mathcal{D}_\gamma(t,\zeta)$. Define the collision probability $p_{\gamma,t}(\rho) := \mathbb{P}_{(h_\gamma,g_\gamma)\sim\mathcal{D}_\gamma(t,\zeta)}[h_\gamma(x) = g_\gamma(y)]$.

Corollary 3. *Given constants $\gamma, t > 0$ and $\zeta \in (0, \frac{1}{2})$ define $t_\gamma = t\max\{\gamma, 1\}$ a pair of hash functions $(h_\gamma, g_\gamma) \sim \mathcal{D}_\gamma(t,\zeta)$ can be evaluated using space and time $O(dt_\gamma \log(\frac{1}{\zeta})e^{t_\gamma^2/2})$. Furthermore, for $\delta > 0$ let $C_1(\delta) := \left(\frac{148\sqrt{\pi}}{\sqrt{2}(1-\zeta)\delta^2}\right)^2$ depending only on ζ, δ such that:*

- $\forall\ |\rho| \leq 1-\delta$,

$$\frac{1}{C_1} \leq \frac{p_{\gamma,t}(\rho)}{e^{-\left(\frac{1-\rho}{1+\rho}+\gamma^2\frac{1+\rho}{1-\rho}\right)\frac{t^2}{2}}} \leq C_1, \tag{23}$$

- $\forall |\rho| > 1-\delta$

$$p_{\gamma,t}(\rho) \leq \sqrt{C_1} e^{-\frac{2-\delta}{\delta}\frac{t_\gamma^2}{2}}. \tag{24}$$

Proof: As we sample hash functions from the families $D_+(t,\zeta)$ and $D_-(\gamma t,\zeta)$ independently, the collision probability $p_{\gamma,t}(\rho) = p_+(\rho)p_-(\rho)$ is the product of the two collision probabilities. Using Lemma 5 we get the required statement with $C_1(\delta) := \max\{(\frac{\sqrt{2}(1-\zeta)\delta^2}{148\sqrt{\pi}})^{-2}, (\frac{2}{\sqrt{\pi}\sqrt{\delta}})^2\}$. ■

III. Variance of Multi-resolution HBE

In this section, we analyze the variance of Multi-resolution HBE and identify two key design principles: the p^2*-weighting scheme*, and the *scale-free property* of HBE, for which we give strong theoretical bounds on the variance. Our first step is to obtain a more tractable bound on (14).

Lemma 6. *Given an n point set X and an unbiased $Z_T \sim \mathrm{HBE}_X(\{\mathcal{H}_t, p_t, w_t\}_{t\in[T]})$, there exists explicit $A \in \mathbb{R}^{n\times n}$ and vector $v \in \mathbb{R}^n_{++}$ such that: $\mathbb{E}[Z_T^2(y)] \leq \sup_{\|f\|_1\leq 1, \|f\|_{v,1}\leq\mu}\{f^\top A f\} + \mu^2$.*

Proof: Fix $x_1, \ldots, x_n$ potential positions for the n points in the dataset and let $f_1, \ldots, f_n \in [0,1]$ be the fraction of points that are assigned to each of this positions. Moreover for any two positions x_i, x_j let L_{ij} be the set of hash functions such that $p_t(x_i, y) < p_t(x_j, y)$ and G_{ij} be the complement. We get:

$$\sum_{j\in[n]} \frac{\min\{p_t(x_j,y), p_t(x_i,y)\}}{p_t(x_i,y)} \leq \sum_{j\in[n]} n f_j (\mathbb{I}[t \in L_{ij}] + \mathbb{I}[t \in G_{ij}]\frac{p_t(x_j,y)}{p_t(x_i,y)}). \tag{25}$$

Using (14), and (25), the lemma follows by setting $\nu_i = w(x_i, y)$ and

$$A_{ij} = \sum_{t \in L_{ij}} \frac{w_t^2(x_i, y)}{p_t(x_i, y)} + \sum_{t \in G_{ij}} \frac{w_t^2(x_i, y)}{p_t^2(x_i, y)} p_t(x_j, y). \quad (26)$$

■

The main question that the above lemma leaves open, is to how select the functions $\{w_t\}$ so that, the estimator is still unbiased, but the variance is minimized.

A. The p^2-weighting scheme for HBE

Our goal is to find a set of weights that are only a function of the query y and any point x. To select such a weights we first obtain the following upper bound on (26)

$$\sum_{t \in L_{ij}} \frac{w_t^2(x_i, y)}{p_t(x_i, y)} + \sum_{t \in G_{ij}} \frac{w_t^2(x_i, y)}{p_t^2(x_i, y)} p_t(x_j, y) \leq \sum_{t \in [T]} \frac{w_t^2(x_i, y)}{p_t^2(x_i, y)}. \quad (27)$$

The set of weights that minimize (27) and for which the HBE is still unbiased are given by: $w_t^*(x, y) = \frac{p_t^2(x,y)}{W(x,y)} w(x, y)$, where $W(x, y) := \sum_{t \in T(x,y)} p_t^2(x, y)$. In what follows we denote any unbiased $\mathrm{HBE}_X(\{\mathcal{H}_t, w_t, p_t\}_{t \in [T]})$ with $w_t \propto p_t^2 w$ as $\mathrm{HBE}_X^2(\{\mathcal{H}_t, p_t\}_{t \in [T]})$. We aim to quantify precisely how well these estimators can perform by choosing $\{p_t\}$ judiciously. To that end, using Lemmas 6 and 2, we obtain the following upper bound on the variance.

Theorem 5. *Given a set $X \subseteq S \subset \mathcal{X}$, and $Z_T \sim \mathrm{HBE}_X^2(\{\mathcal{H}_t, p_t\}_{t \in [T]})$ let $p_*(x, y) := \sup_{t \in [T]} \{p_t(x, y)\}$, then for all $y \in Y$ such that $Z(y) = \mu > 0$ and $f_i = f(x_i) := \min\{1, \frac{\mu}{w(x_i, y)}\}$, we get:*

$$\mathbb{E}[Z_T^2(y)] \leq \mu^2 + 4 \sup_{x_1, x_2 \in S} \{ f_1^2 \frac{w^2(x_1, y)}{p_*(x_1, y)} + f_2^2 \frac{w^2(x_2, y)}{p_*(x_2, y)} + f_1 f_2 (\frac{w^2(x_1, y)}{p_*^2(x_1, y)} + \frac{w^2(x_2, y)}{p_*^2(x_2, y)}) D_T(x_1, x_2) \}, \quad (28)$$

where $D_T(x_1, x_2) := \max_{t \in [T]} \min\{p_t(x_1, y), p_t(x_2, y)\} \leq \min\{p_*(x_1, y), p_*(x_2, y)\}$.

Proof: Using Lemma 2 we get for all $f \geq 0$ such thatt $\|f\|_{w,1}$ and $\|f\|_1 \leq 1$:

$$f^\top A f \leq 4 \sup_{ij} \left\{ \min\{1, \frac{\mu}{w(x_i, y)}\} |A_{ij}| \min\{1, \frac{\mu}{w(x_j, y)}\} \right\},$$

with $A_{ij} = \left(\sum_{t \in L_{ij}} \frac{w_t^2(x_i, y)}{p_t(x_i, y)} + \sum_{t \in G_{ij}} \frac{w_t^2(x_i, y)}{p_t^2(x_i, y)} p_t(x_j, y) \right)$. Setting $f_i = \min\{1, \frac{\mu}{w(x_i, y)}\}$ and $\tilde{A}_{ij} = f_i |A_{ij}| f_j$, we get by the above $\sup_{\|f\|_{w,1} \leq \mu, \|f\|_1 \leq 1} \{f^\top A f\} \leq 4 \sup_{ij} \left\{ \tilde{A}_{ii} + \tilde{A}_{jj} + \tilde{A}_{ij} + \tilde{A}_{ji} \right\}$. Let V_{ij} be the expression in brackets. For the p^2-weighting scheme $w_t(x, y) = \frac{p_t^2(x,y)}{W(x,y)} w(x, y)$ we get

$$\begin{aligned} V_{ij} = f_i^2 \sum_{t \in L_{ij}} \frac{w^2(x_i, y)}{W^2(x_i, y)} p_t^3(x_i, y) + f_j^2 \sum_{t \in G_{ij}} \frac{w^2(x_j, y)}{W^2(x_j, y)} p_t^3(x_j, y) \\ + f_i f_j \left(\sum_{t \in L_{ij}} \frac{w^2(x_i, y)}{W^2(x_i, y)} p_t^3(x_i, y) + \sum_{t \in G_{ij}} \frac{w^2(x_i, y)}{W^2(x_i, y)} p_t^2(x_i, y) p_t(x_j, y) \right) \\ + f_j f_i \left(\sum_{t \in L_{ji}} \frac{w^2(x_j, y)}{W^2(x_j, x)} p_t^3(x_j, y) + \sum_{t \in G_{ji}} \frac{w^2(x_j, y)}{W^2(x_j, y)} p_t^2(x_j, y) p_t(x_i, y) \right). \end{aligned}$$

Using $W(x, y) = \sum_{t \in [T]} p_t^2(x, y) \geq p_*^2(x, y)$ and $p_t \leq p_*(x, y)$ we

$$V_{ij} \leq f_i^2 \frac{w^2(x_i, y)}{p_*(x_i, y)} + f_j^2 \frac{w^2(x_j, y)}{p_*(x_j, y)} + f_i f_j \left(\frac{w^2(x_i, y)}{p_*^2(x_i, y)} + \frac{w^2(x_j, y)}{p_*^2(x_j, y)} \right) \cdot \max \left\{ \max_{t \in L_{ij}} p_t(x_i, y), \max_{t \in G_{ij}} p_t(x_j, y) \right\}.$$

Since $G_{ji} \subseteq L_{ij}$ and vice versa, setting $D_T(x_i, x_j) := \max\left\{\max_{t \in L_{ij}} p_t(x_i, y), \max_{t \in L_{ji}} p_t(x_j, y)\right\}$ we arrive at the following bound on:

$$V_{ij} \leq f_i^2 \frac{w^2(x_i, y)}{p_*(x_i, y)} + f_j^2 \frac{w^2(x_j, y)}{p_*(x_j, y)} + f_i f_j \left(\frac{w^2(x_i, y)}{p_*^2(x_i, y)} + \frac{w^2(x_j, y)}{p_*^2(x_j, y)} \right) D_T(x_i, x_j).$$

To complete the proof we show the following:

$$\begin{aligned} D_T(x_1, x_2) &= \max \left\{ \max_{t \in L_{12}} p_t(x_1, y), \max_{t \in L_{21}} p_t(x_2, y) \right\} \\ &= \max_t \min\{p_t(x_1, y), p_t(x_2, y)\}. \end{aligned}$$

Noticing that $\max_t \min\{p_t(x_1, y), p_t(x_2, y)\} \leq p_*(x_1, y)$ and $\max_t \min\{p_t(x_1, y), p_t(x_2, y)\} \leq p_*(x_2, y)$, we get the statement. ■

B. Scale-free Multi-Resolution Hashing

The development above has revealed that the crucial parameter for consideration of HBE^2 is the *pointwise maximum hashing probability* $p_*(x, y)$. Here, we analyze a specific family of estimators where $p_*(x, y)$ has polynomial dependence with $w(x, y)$.

Definition 6. *Given $M \geq 1$, $\beta \in [0, 1]$ and function w, an estimator $Z_T \sim \mathrm{HBE}_X^2(\{\mathcal{H}_t, p_t\}_{t \in [T]})$ is called (β, M)-*scale free*, if $M^{-1} \cdot w^\beta(x, y) \leq p_*(x, y) \leq M \cdot w^\beta(x, y)$ for all $x \in X$ and $y \in \mathcal{X}$.*

Exploiting the scale-free property we get explicit bounds on the variance.

Theorem 6 (Scale-free)**.** *Let $Z_T \sim \mathrm{HBE}_X^2(\{\mathcal{H}_t, p_t\}_{t\in[T]})$ be a (β, M)-scale free estimator, then:*

$$\mathbb{E}[Z_T^2(y)] \le V_{\beta,M}(\mu) := 8M^3\mu^2\left[\frac{1}{\mu^\beta} + \frac{1}{\mu^{1-\beta}}\right] + \mu^2.$$

Our theorem shows that the optimal worst-case variance is achieved for $\beta^* = 1/2$ and improves over uniform random sampling by a factor of $O(\frac{1}{\sqrt{\mu}})$. A theorem of similar nature but with a more involved proof was given in [10] for $\beta \in [\frac{1}{2}, 1]$.

Proof: For $i \in [1,2]$ let $w_i := w(x_i, y)$ and f_i as in Theorem 5. Using the scale-free property, Theorem 5 and $D_T(x_1, x_2) \le \min\{p_*(x_1, y), p_*(x_2, y)\}$ we arrive at:

$$\mathbb{E}[Z_T^2] \le \mu^2 + 4M^3 \sup_{x_1,x_2\in S}\Big\{f_1^2 w_1^{2-\beta} + f_2^2 w_2^{2-\beta} + f_1 f_2\left(w_1^{2-2\beta} + w_2^{2-2\beta}\right)\min\{w_1, w_2\}^\beta\Big\}$$

Due to the definition of f_i the last expression is only a function of w_1, w_2 and solving the optimization problem boils down to a case analysis. We focus on the case $w_1 \ge \mu$, $w_2 \le \mu$, for which the expression in the parenthesis becomes:

$$\mu^2 w_1^{-\beta} + w_1^{1-2\beta} w_2^{2\beta}\mu + w_2^{2-\beta} w_1^\beta \mu + w_2^{2-\beta} \tag{29}$$

The weights that maximize the expression are $w_1^* = 1$ and $w_2^* = \mu$. $\mu^2 + \mu^{1+2\beta} + \mu^{1+\beta} + \mu^{2-\beta} \le 2\mu^2[\mu^{-\beta} + \mu^{\beta-1}]$. The other cases $w_1, w_2 \le \mu$ and $w_1, w_2 \ge \mu$ follow similarly. ■

IV. Approximation of Convex Functions

In this section, we show how to use the logarithm $h_{\gamma,t}(\rho)$, given below, of the *idealized hashing probability* of the Distance Sensitive Hashing scheme to construct a set of functions whose supremum approximates any non-positive convex Lipschitz function $\phi(\rho)$.

$$h_{\gamma,t}(\rho) := -\left(\frac{1-\rho}{1+\rho} + \gamma^2\frac{1+\rho}{1-\rho}\right)\frac{t^2}{2} \tag{30}$$

Some basic properties of this family of functions are given below.

Proposition 2 (Concavity)**.** *For $\gamma \ge 0$, the function $h_{\gamma,t}$ attains its maximum at $\rho^*(\gamma) = \frac{1-\gamma}{1+\gamma}$ and*

(a) *If $0 \le \gamma \le 1$, the function is concave for all $\rho \in [\rho^*(\gamma^{\frac{2}{3}}), 1]$ and $\rho^*(\gamma) \ge \rho^*(\gamma^{\frac{2}{3}})$ holds.*
(b) *If $\gamma \ge 1$, the function is concave for all $\rho \in [-1, \rho^*(\gamma^{\frac{2}{3}})]$ and $\rho^*(\gamma) \le \rho^*(\gamma^{\frac{2}{3}})$ holds.*

The above properties will be used to show that, by picking parameters γ_0, t_0 appropriately, if we approximate the convex function ϕ *locally* at some point $\rho_0 \in [-1, 1]$ up to *first order* (value and derivative), then $h_{\gamma_0,t_0}(\rho) \le \phi(\rho)$ for all $\rho \in [-1, 1]$. Thus even a single hash function is sufficient to provide a lower bound. Most of the work is devoted to show that we can get a good *upper bound* on ϕ using a small number of functions to approximate ϕ locally at a set of interpolation points $\rho_1, \ldots, \rho_T$. We define the following parametrization. Given $\delta > 0$ for $|\rho_0| \le 1 - \delta$, let

$$\gamma_0^2 := \left(\frac{1-\rho_0}{1+\rho_0}\right)^2 \frac{2\phi(\rho_0) + (1-\rho_0^2)\phi'(\rho_0)}{2\phi(\rho_0) - (1-\rho_0^2)\phi'(\rho_0)} \tag{31}$$

$$t_0^2 := -\frac{1}{2}\frac{1+\rho_0}{1-\rho_0}\left[2\phi(\rho_0) - (1-\rho_0^2)\phi'(\rho_0)\right] \tag{32}$$

and for fixed ϕ and $\rho_0 \in [-1+\delta, 1-\delta]$ define $h_{\rho_0}(\rho) := h_{\gamma_0,t_0}(\rho)$. This parametrization is well defined due to Lemma 3. For $\rho_0 \in \{-1, +1\}$ (boundary) we define $h_{\pm1}(\rho) := -\frac{1\mp\rho}{1\pm\rho}\frac{t_{\pm1}^2}{2} + \phi(\pm1)$, where $t_{\pm1}^2 = 4\max\{\pm\phi'(\pm1), 0\}$. Under our assumptions $\phi \le 0$, hence the constant term above can be implemented by subsampling the data set with probability $e^{\phi(\pm1)}$. The following bounds on the parameters γ_0, t_0 will be useful.

Corollary 4 (Complexity)**.** *Under the conditions of Lemma 3, we have the following bounds: $t_0^2 \le -2\frac{1+\rho_0}{1-\rho_0}\phi(\rho_0)$, $t_0^2\gamma_0^2 \le -2\frac{1-\rho_0}{1+\rho_0}\phi(\rho_0)$, and $t_0^2\max\{\gamma_0^2, 1\} \ge -\frac{1+\rho_0^2}{1-\rho_0^2}\phi(\rho_0)$.*

Using this family of functions we show we can approximate a convex function arbitrarily well.

Theorem 7 (Approximation)**.** *Given $\epsilon > 0$, for every convex function ϕ there exists a set $\mathcal{T}_\epsilon(\phi) \subset [-1, 1]$ of size $O\left(\sqrt{\frac{L(\phi)}{\epsilon}}\log(\frac{L(\phi)}{\epsilon})\right)$ such that $0 \le \phi(\rho) - \sup_{\rho_0\in\mathcal{T}_\epsilon}\{h_{\rho_0}(\rho)\} \le 2\epsilon$ for all $\rho \in [-1, 1]$.*

A. Proof of Approximation Theorem

To prove the above theorem it is sufficient, due to Theorem 4, to only show how to *approximate linear functions*. For ρ away from $\{-1, 1\}$, this is done in Lemma 7, where the *interpolation points are given explicitly*. Lemma 8 treats the case near the boundary. By symmetry of the family of hash functions we only need to show our result for $[-1, 0]$.

Lemma 7. *Let ℓ be a linear function on $[\rho_-, \rho_+] \subseteq [-1+\delta, 0]$. Given $\epsilon > 0$, let $T = \lfloor\frac{\log(\frac{1-|\rho_+|}{1-|\rho_-|})}{\log(1+\sqrt{\frac{\epsilon}{8|\ell_{\min}|}})}\rfloor$ and define $\rho_i := \rho_- + (1-|\rho_-|)\left[\left(1+\sqrt{\frac{\epsilon}{8R(\ell)}}\right)^i - 1\right]$ for $i = 0, \ldots, T$. Then, for all $\rho \in [\rho_-, \rho_+]$ there exists $i(\rho) \in [T] \cup \{0\}$ such that $0 \le \ell(\rho) - h_{\rho_{i(\rho)}}(\rho) \le \epsilon$.*

Lemma 8. *Given $\epsilon > 0$, let $\delta(\epsilon) := \min\{1, \sqrt{\frac{\epsilon}{4L(\phi)}}, \frac{\epsilon}{L(\phi)}\}$. Then $0 \le \phi(\rho) - h_{-1}(\rho) \le \epsilon$ for all ρ in the interval $[-1, -1+\delta(\epsilon)]$.*

Proof: If $\phi'(-1) \ge 0$, then $0 \le \phi(\rho) - h(\rho) = \phi(\rho) - \phi(-1) \le L(\rho+1) \le L\delta$. If $\phi'(-1) < 0$ then by the Taylor

remainder theorem and $0 \leq \delta \leq 1$ we get

$$0 \leq \phi(\rho) - h_{-1}(\rho) \leq \frac{1}{2}\frac{2}{(2-\delta)^3} 4|\phi^{'}(-1)|\delta^2 \leq 4L\delta^2$$

Using the definition of $\delta(\epsilon)$ we get the statement. ∎

The previous lemmas provide only local approximation to the function. Proposition 3 below is used to show that the functions we construct are a lower bound to the piecewise linear approximation on the whole interval $\rho \in [-1,1]$, which in turn implies a lower bound for the function $\phi(\rho)$.

Proposition 3. *Let $\phi : [-1,1] \to \mathbb{R}$ be an non-decreasing (resp non-increasing) convex function and $g : [-1,1] \to \mathbb{R}$ a function that attains a global maximum at ρ^*, is concave in $[-1,\rho^*]$ (resp $[\rho^*,1]$), and $\exists \rho_0 \in [-1,\rho^*]$ (resp. $[\rho^*,1]$) such that $\phi'(\rho_0) = g^{'}(\rho_0)$, then $\inf_{\rho\in[-1,1]}\{\phi(\rho) - g(\rho)\} = \phi(\rho_0) - g(\rho_0)$.*

Proof of Theorem 7: Given $\epsilon > 0$, let $\delta(\epsilon)$ as in Lemma 8. We start by applying Theorem 4 separately on the function ϕ restricted on the interval $[-1+\delta, 0]$ and ϕ restricted on $[0, 1-\delta]$ to get piecewise linear convex approximation ℓ to ϕ such that $0 \leq \phi(\rho) - \ell(\rho) \leq \epsilon$ for all $|\rho| \leq 1-\delta$. Let $I^- = \{[\rho^-_{j-1}, \rho^-_j]\}_{j\in[J^-]}$ and $I^+ = \{[\rho^+_{j-1}, \rho^+_j]\}_{j\in[J^+]}$ with $J^{\pm} = O(\sqrt{\frac{L(\phi)}{\epsilon}})$ be the corresponding decompositions of $[-1+\delta, 0]$ and $[0, 1-\delta]$ in contiguous subintervals where the function ℓ is linear. For each $j \in [J^{\pm}]$, let $\mathcal{T}^{\pm}_j$ be the set of points resulting by applying Lemma 7 to $[\rho^{\pm}_{j-1}, \rho^{\pm}_j]$ and set $T^{\pm}_j = |\mathcal{T}^{\pm}_j|$. We define the following set of points $\mathcal{T}_\epsilon(\phi) := \left(\cup_{j=1}^{J^+}\mathcal{T}^+_j\right) \cup \left(\cup_{j=1}^{J^-}\mathcal{T}^-_j\right) \cup \{1,-1\}$. We have

$$|\cup_{j=1}^{J_\pm}\mathcal{T}^{\pm}_j| \leq \sum_{j=1}^{J^{\pm}}(1+T^{\pm}_j) \leq J^{\pm} + \frac{\log(\frac{1}{\delta})}{\log(1+\sqrt{\frac{\epsilon}{8R(\phi)}})}$$

Using $\log(1+x) \geq \frac{2}{3}x$ for $x \in [0,1]$ and $R(\phi) \leq 2L(\phi)$, we get that $|\mathcal{T}_\epsilon(\phi)| = O\left(\sqrt{\frac{L(\phi)}{\epsilon}}\log(\frac{L(\phi)}{\epsilon})\right)$.

Let $\hat{\phi}(\rho) := \sup_{\rho_0\in\mathcal{T}_\epsilon(\phi)}\{h_{\rho_0}(\rho)\}$. Due to Propositions 2 and 3, we get $\phi(\rho) \geq \ell(\rho) \geq h_{\rho_0}(\rho)$ for all ρ and $\rho_0 \in \mathcal{T}_\epsilon(\phi)$ and consequently $\phi(\rho) - \hat{\phi}(\rho) \geq 0$. Let $T = |\mathcal{T}_\epsilon(\phi)|$ and $\rho_1, \ldots, \rho_T$ an increasing ordering of points in $\mathcal{T}_\epsilon(\phi)$. We have

$$\begin{aligned}\sup_{\rho\in[-1,1]}\{\phi(\rho) - \hat{\phi}(\rho)\} &= \max_{i\in[T-1]}\sup_{\rho\in[\rho_i,\rho_{i+1}]}\{\phi(\rho) - \hat{\phi}(\rho)\} \\ &\leq \max_{i\in[T-1]}\sup_{\rho\in[\rho_i,\rho_{i+1}]}\left\{\phi(\rho) - \max\{h_{\rho_i}(\rho), h_{\rho_{i+1}}(\rho)\}\right\}\end{aligned}$$

which is bounded by 2ϵ due to Theorem 4 and Lemmas 8, 7. ∎

V. Scale-free Multi-Resolution Hashing for Log-convex functions

In the previous section, we have shown that using the idealized hashing probabilities one can approximate a log-convex function up to arbitrary multiplicative accuracy. In this section, we use this fact to construct explicit scale-free Multi-resolution HBE, that constitutes the main ingredient needed to prove our main result.

Theorem 8. *Given a convex function ϕ, $X \subset \mathcal{S}^{d-1}$ and $\beta \in [0,1]$, there exist an explicit constant M_ϕ and (β, M_ϕ)-scale free estimator $Z_T \sim \mathrm{HBE}^2_X(\{\mathcal{H}_t, p_t\}_{t\in[T]})$ for $Z_\phi(y)$ with complexity $O(d\{L(\phi)\}^{5/6}M_\phi)$.*

Proof: The main challenge in proving the result is to trade-off *complexity of evaluating* the hashing scheme versus the *fidelity of the approximation* of $\beta[\phi(\langle x,y\rangle) - \phi_{\max}]$ by $\log p_*(x,y)$ that affects the variance. In order to do that, set $\delta^* = \frac{1}{2\beta L(\phi)}$ and for $C^* = C_1(\delta^*)$ as in Corollary 3, define

$$k^* = \left\lceil \left\{ \frac{2\beta^2}{\log C^*} L(\phi)R(\phi) \right\}^{1/3} \right\rceil \tag{33}$$

We further define a "smoothed" version of ϕ as $\tilde{\phi}(\rho) := \frac{\beta(\phi(\rho)-\phi_{\max})}{k^*}$. If $L(\tilde{\phi}) = \frac{\beta}{k^*}L(\phi) < 2$ then the variation in the function $R(\tilde{\phi}) < 4$ is too small and a constant number of random samples suffice to answer any query. So, we only deal with the interesting case when and $L(\tilde{\phi}) \geq 2$ and $R(\tilde{\phi}) \geq 4$.

1) *Approximation:* let $\mathcal{T}_{1/2} = \mathcal{T}_{1/2}(\tilde{\phi})$ be the set of interpolation points resulting from invoking Theorem 7 for $\tilde{\phi}$ and $\epsilon = \frac{1}{2}$. For this set of points we have $\left|\sup_{\rho_0\in\mathcal{T}_{1/2}}\{h_{\rho_0}(\rho)\} - \tilde{\phi}(\rho)\right| \leq 1$.
2) *Hashing scheme:* let $\rho_1 < \ldots < \rho_T$ be an increasing enumeration of points in $\mathcal{T}_{1/2}$. For each $t \in [T]$, let $\tilde{\mathcal{H}}_t$ be the DSH family with collision probability $\tilde{p}_t$ and parameters given by (31) and (32) (for $\tilde{\phi}$ and ρ_t). We raise each hashing scheme to the k^*-th power to get $\mathcal{H}_t := \tilde{\mathcal{H}}_t^{\otimes k^*}$ with collision probability $p_t := \tilde{p}_t^{k^*}$. Using Lemma 5 and Corollary 3 we show:

   Lemma 9. *For all $\rho \in [-1,1]$.*

   $$\left|\sup_{t\in[T]}\{\log p_t(\rho)\} - k^*\sup_{t\in[T]}\{h_{\rho_t}(\rho)\}\right| \leq k^*\log C_1$$

3) *Scale-free property:* by the previous two steps and noting that $\log w^\beta(x,y) = k^*\tilde{\phi}(\langle x,y\rangle)$

   $$\begin{aligned}\left|\sup_{t\in[T]}\{\log p_t(\rho)\} - \log w(x,y)^\beta\right| &\leq k^* + k^*\log C_1 \\ &\leq 2k^*\log C_1 \quad (34)\end{aligned}$$

   This shows that $Z_T \sim \mathrm{HBE}^2_X(\{\mathcal{H}_t, p_t\}_{t\in[T]})$ is (β, M_ϕ)-scale free with $M_\phi := e^{2k^*\log C_1}$.
4) *Complexity:* To bound the complexity of the estimator $Z_T \sim \mathrm{HBE}^2_X(\{\mathcal{H}_t, p_t\}_{t\in[T]})$, we need by (15), (31), (32) to bound $t^2_{\gamma_0} = t_0^2\max\{\gamma_0^2, 1\}$ for $\rho_0 \in \mathcal{T}_{1/2}(\tilde{\phi})$. Using Corollary 4 we get

   Lemma 10. *If $L(\tilde{\phi}) \geq 2$ and $R(\tilde{\phi}) \geq \frac{1}{2}$, then $\forall \rho_0 \in \mathcal{T}_{1/2}(\tilde{\phi})$, $t^2_{\gamma_0} \leq 8\left(\frac{\beta}{k}\right)^2 L(\phi)R(\phi)$.*

Hence, the complexity of evaluating the estimator is $O\left(|\mathcal{T}_{\frac{1}{2}(\tilde{\phi})}|k^* d\log(\frac{1}{\zeta})e^{4\left(\frac{\beta}{k^*}\right)^2 L(\phi)R(\phi)}\right)$, by Theorem 7 and our choice (33), this is bounded by $O(dL(\phi)^{5/6}M_\phi)$.

■

A. Main Result

Theorem 9. *Given $\epsilon, \tau \in (0,1)$, for every convex function ϕ with Lispchitz constant $L(\phi)$, there exists an explicit constant M_ϕ and a data structure using space $O(dL(\phi)^{5/6}M_\phi^3 \frac{1}{\epsilon^2}\frac{1}{\sqrt{\tau}}\cdot n)$ and query time $O(dL(\phi)^{5/6}M_\phi^4 \frac{1}{\epsilon^2}\frac{1}{\sqrt{\mu}})$ that for any $y \in \mathcal{S}^{d-1}$ with constant probability can either produce an $(1+\epsilon)$ approximation to $\mu = Z_\phi(y) \geq \tau$ or assert that $\mu < \tau$.*

Proof: Follows by invoking Theorems 8, 6 and 3 for $\beta^* = 1/2$. ■

The explicit constant $M_\phi := e^{\{2\log(C^*)\sqrt{L(\phi)R(\phi)}\}^{2/3}}$ (where $R(\phi) \leq 2L(\phi)$ is the range of ϕ and $\log(C^*) = O(\log L(\phi))$) is sub-exponential in $L(\phi)$ and is of similar nature to the evaluation time of the Andoni-Indyk LSH [34] and Spherical LSH [47]. It corresponds to the number of randomly placed spherical caps of certain size that are required to cover most of the unit sphere.

Proof of Theorem 1: The simplified version of our main result follows by setting $L \leq (1-\delta)\log n$. We have that $\mu \geq e^{-2L(\phi)} \geq n^{2(1-\delta)} \Rightarrow \frac{1}{\sqrt{\mu}} \leq n^{1-\delta}$ and $L(\phi)^{5/6}M_\phi^4 = e^{O(\log^{2/3}(n)\log\log n)} = n^{o(1)}$. ■

Proof of Corollary 2: The corollary follows by invoking Theorem 1 with $\epsilon' = \epsilon^{3/2}$ to define a data structure that can be used to return a $(0,\epsilon,\epsilon)$-sample of size $O(e^{(1+o(1))L(\phi)}/\epsilon'^2)$. To see this observe that if for a random variable Z with mean μ and relative variance bounded by V, we average $m = 6V/\epsilon'^2$ independent samples then by Chebyshev's $\mathbb{P}[|\frac{1}{m}\sum_{i=1}^m Z_i - \mu| > \epsilon\mu] \leq \frac{V}{\epsilon^2 m} \leq \frac{\epsilon}{6}$. ■

VI. Reduction from Euclidean Space to Unit Sphere

In order to extend our method from unit sphere to bounded subsets of Euclidean space the main observation is that given $\gamma \in (0,1]$, if for two sets $S_x, S_y \subset \mathbb{R}^d$ we have that $\forall x_1, x_2 \in S_x, \|x_1\|/\|x_2\| \leq (1+\gamma)$ and $\forall y_2, y_1 \in S_y$, $\|y_1\|/\|y_2\| \leq (1+\gamma)$, then $\forall x_1, x_2 \in S_x, \forall y_1, y_2 \in S_y$

$$\langle x_1, y_1\rangle \approx \|x_2\|\|y_2\|\left\langle \frac{x_1}{\|x_1\|}, \frac{y_1}{\|y_1\|}\right\rangle. \tag{35}$$

This fact suggests the following strategy:

1) Partition the data set $X = X_1 \uplus \ldots \uplus X_K$ and the set of possible queries $Y = Y_1 \uplus \ldots \uplus Y_K$ in *spherical annuli* $\{X_i\}_{i\in[K]}$ and $\{Y_j\}_{j\in[K]}$.
2) For each pair (X_i, Y_j) use the approximation (35) and assume that for some r_i and r_j all points in X_i and Y_j approximately lie on $r_i\mathcal{S}^{d-1}$ and $r_j\mathcal{S}^{d-1}$ respectively.
3) For each such pair construct a Multi-resolution HBE to obtain a low-variance unbiased estimator of the contribution of points in X_i for any possible value of $j \in [K]$ (annulus the query might belong to).
4) Sum up the contribution for all $i \in [K]$ to obtain the final estimator and bound its variance.

Our approach applies to the following general class of functions:

$$w(x,y) = p_0(\|x\|)e^{\phi(\langle x,y\rangle)+\mathcal{A}(y)}, \tag{36}$$

where ϕ is convex and Lipschitz, $\mathcal{A}(y)$ arbitrary[1] and $p_0 : \mathbb{R}_{++} \to \mathbb{R}_{++}$ satisfies a notion of smoothness that is related to Lipschitz continuity under the *Hilbert metric* $d_H(x,y) := |\log(\frac{x}{y})|$ for $x, y \in \mathbb{R}_+$.

Definition 7. *For $H, \delta \geq 0$ and $\gamma \in (0,1]$, a function $p_0 : \mathbb{R}_{++} \to \mathbb{R}_{++}$ is called (H,δ,γ)-*log-Lipschitz*, if for all $r_1 \geq r_2 > 0$ such that $r_1 \leq (1+\gamma)r_2$ we have $|\log(p_0(r_1)/p_0(r_2))| \leq H\cdot\gamma + \delta$.*

This notion of smoothness implies that the function *changes multiplicatively within each annulus*.

Proposition 4. *For $\gamma \in (0,1]$ and all $r \in (0,R]$ the function $r^q e^{f(r)}$ is $(|q|,\ L(f)R\gamma,\ \gamma)$-log-Lipschitz.*

Proof: Let $r_1, r_2 \in (0,R]$ such that $r_2 \leq r_1 \leq (1+\gamma)r_2$, then

$$\left|\log\left(r_1^q e^{f(r_1)}/(r_2^q e^{f(r_2)})\right)\right| \leq |q||\log(\frac{r_1}{r_2})| + |f(r_1)-f(r_2)| \leq |q|\gamma + L(f)R\gamma.$$

■

Functions that are of the form (36) include the Gaussian kernel $e^{-\|x-y\|^2}$ or the norm of the derivative of the logistic log-likelihood $\|\nabla_y \log(1+\exp(\langle x,y\rangle))\| = \|x\|(1+e^{-\langle x,y\rangle})^{-1}$. For concreteness we are going to assume that the function p_0 is $(q, HR\gamma, \gamma)$-log-Lipshcitz for some $q, H > 0$, as in Proposition 4, instead of using general δ as in Definition 7. However, our result applies also to the more general case. In the rest of this section, we carry out the strategy outlined above.

A. Partitioning in Spherical Annuli

Given $0 < \gamma \leq 1$, a dataset X and a set of possible queries Y, define

$$r_0 := r_0(X,Y) = \inf\{\|z\| : z \in X \cup Y, z \neq 0\} \tag{37}$$

$$R := R(X,Y) = \sup\{\|z\| : z \in X \cup Y\} \tag{38}$$

$$K := K(R/r_0,\gamma) = \lceil \log(R/r_0)/\log(1+\gamma)\rceil \tag{39}$$

Further for $i \in \mathbb{Z}$ define $r_i := (1+\gamma)^{i-1}r_0$ and $S_i := S_i(\gamma) = [r_i, r_{i+1})$ and the corresponding sets:

$$X_i := \{x \in X \big| \|x\| \in S_i\},\ i \in [K] \tag{40}$$

[1]For any given query y, $e^{\mathcal{A}(y)}$ is a constant factor that can be factored out.

For any point $x \in \mathbb{R}^d$ define $i(x) := \arg\min_{i\in\mathbb{Z}}\{\|x\| \in S_i\}$, and its *norm-truncated version*:

$$\tilde{x} := \tilde{x}_\gamma = \frac{x}{\|x\|} r_{i(x)}. \tag{41}$$

For any point $x \neq 0$ let $\hat{x} := \frac{x}{\|x\|}$. Note that $\hat{x} = \frac{\tilde{x}}{r_{i(x)}}$ is also the normalized version of $\tilde{x}$. The motivation for partitioning the space in such annuli and projecting points on the inner boundary of each spherical annulus is that in doing so the ratio between the function $w(x,y)$ and $w(\tilde{x},\tilde{y})$ does not change too much.

Lemma 11. *For points $x, y \in \mathbb{R}^d$ such that $\|x\|, \|y\| \in [r_0, R]$ and $\gamma \in (0,1]$, let $w(x,y) = p_0(\|x\|)e^{\phi(\langle x,y\rangle)}$ with p_0 being $(q, HR\gamma, \gamma)$-log-Lipshcitz and ϕ being L Lipschitz. Then for $M = \exp\left(q + HR + 3Lr_{i(x)}r_{i(y)}\right)$ we have:*

$$\frac{1}{M} \leq \frac{w(\tilde{x}_\gamma, \tilde{y}_\gamma)}{w(x,y)} \leq M \tag{42}$$

This suggests that if we pick γ appropriately we can use the framework of Multi-resolution HBE to perform importance sampling for each annulus separately and bound the variance of the overall estimator.

Theorem 10. *For a set $X \subset \mathbb{R}^d$ and a set of possible queries Y define r_0, R by* (37), (38) *respectively. For every convex function $\phi : [-R^2, R^2] \to \mathbb{R}$ and a $(q, HR\gamma, \gamma)$-log-Lipschitz function p_0, let $w(x,y) = p_0(\|x\|)e^{\phi(\langle x,y\rangle)}$. There exists constants $\gamma^* \in (0,1]$, K^* and a distribution $\mathcal{D}^*$ such that for every $y \in Y$, the estimator $Z(y) \sim \mathcal{D}_*$ is unbiased $\mathbb{E}[Z(y)] = Z_w(y) = \mu$, V-bounded with $V(\mu) = 2e^{5/2}(8M^3_{\phi_{K^*K^*}} + 1)\mu^{-1/2}$ and has complexity $O\left(d(K^*)^2(L(\phi)R^2)^{5/6}M_{\phi_{K^*K^*}}\right)$ where $M_{\phi_{K^*K^*}} = \exp(O\left(\left\{\log(L(\phi)(K^*)^2)L(\phi)(K^*)^2\right\}^{2/3}\right))$.*

Invoking Theorem 3 with the estimators given by Theorem 10 results in a data structure to approximate $Z_w(y)$ for all $y \in Y$.

B. Proof of Theorem 10

Step 1: Our first concern is to pick a constant $\gamma \in (0,1]$ so that the partitioning scheme in subsection VI-A is fully defined. The constant on one hand affects the space/time (complexity) it takes to evaluate our estimator and on the other hand the variance through the approximation $\langle x, y\rangle \approx \langle \tilde{x}_\gamma, \tilde{y}_\gamma\rangle$. To simplify things we pick γ so that the value of $w(x,y)$ changes at most by a factor of e when projecting points on the inner boundary of the spherical annulus.

$$\gamma^* = 1/\max\left\{1, q + HR + 3LR^2\right\} \tag{43}$$

For this choice by (39) and $\log(1+x) \geq \frac{2x}{2+x}$ we get $K^* = \lceil \frac{3}{2}\log(R/r_0)\max\{1, q + HR + 3LR^2\}\rceil$.

Step 2: For all pairs $i, j \in [K^*]$ we are going to construct an unbiased estimator for:

$$Z^{(ij)}_w(y) = \frac{\mathbb{I}\{\|y\| \in S_j\}}{nw_{\max}} \sum_{x\in X_i} p_0(\|x\|)e^{\phi(\langle x,y\rangle)} \tag{44}$$

It is easy to see that if $\|y\| \in S_j$ then $Z_w(y) = \sum_{i\in[K^*]} Z^{(ij)}_w(y)$. For a given pair $i, j \in [K^*]$, we define a modified version of ϕ. Let $\phi_{ij} : [-1,1] \to \mathbb{R}$ be the function given by $\phi_{ij}(\rho) = \phi(r_i r_j \rho)$ for all $\rho \in [-1,1]$ and set $\phi^*_{ij} = \sup\left\{\phi_{ij}(\rho) \middle| |\rho| \leq 1\right\}$. We are going to use these functions to perform "importance sampling" in each spherical annulus X_i. To that end, we define for every pair $i, j \in [K^*]$:

$$\mu_{ij} := \frac{1}{|X_i|e^{\phi^*_{ij}}} \sum_{x\in X_i} e^{\phi_{ij}(\langle \hat{x},\hat{y}\rangle)} \leq 1 \tag{45}$$

$$A_{ij} := \frac{p_0(r_i)|X_i|e^{\phi^*_{ij}}}{nw_{\max}} \leq 1 \tag{46}$$

Using these two quantities we can upper and lower bound the density $Z_w(y)$.

Lemma 12. *For any $y \in \mathbb{R}^d$ such that $\|y\| \in S_j$ we have for $\mu = Z_w(y)$ that*

$$e^{-1} \cdot \sum_{i\in[K^*]} A_{ij}\mu_{ij} \leq \mu \leq e \cdot \sum_{i\in[K^*]} A_{ij}\mu_{ij} \tag{47}$$

Proof: We only show the lower bound. Using Lemma 11 and the definition of γ^* we get:

$$\begin{aligned}
\mu &= \frac{1}{nw_{\max}} \sum_{i\in[K^*]} \sum_{x\in X_i} w(x,y) \\
&\geq e^{-1}\frac{1}{nw_{\max}} \sum_{i\in[K^*]} \sum_{x\in X_i} w(\tilde{x},\tilde{y}) \\
&= e^{-1} \sum_{i\in[K^*]} \left(\frac{|X_i|p_0(r_i)e^{\phi^*_{ij}}}{nw_{\max}}\right) \frac{1}{|X_i|e^{\phi^*_{ij}}} \sum_{x\in X_i} e^{\phi_{ij}(\langle \hat{x},\hat{y}\rangle)}
\end{aligned}$$

The upper bound follows similarly. ∎

Before constructing the estimators for $Z^{(ij)}_w(y)$, we relate the Lipschitz constants of ϕ and ϕ_{ij}.

Proposition 5 (Rescaling)**.** *Given $\alpha > 0$, and a convex function $\phi : [-a, a] \to \mathbb{R}$ with constant L, the function $\phi(\alpha\rho)$ is convex and αL-Lipschitz.*

Proof: Convexity is trivial, and $|\phi(\alpha\rho_1) - \phi(\alpha\rho_2)| \leq L|\alpha\rho_1 - \alpha\rho_2| \leq L\alpha|\rho_1 - \rho_2|$. ∎

Thus, under our assumption $L(\phi_{ij}) \leq Lr_ir_j$.

Step 3: For each $i, j \in [K^*]$, define $\hat{X}_i := \{\hat{x} : x \in X_i\}$. Let $\{\mathcal{H}^{ij}_t, p^{ij}_t\}_{t\in T_{ij}}$ be the hashing scheme resulting from invoking Theorem 8 for ϕ_{ij}, $\hat{X}_i$ and $\beta = 1/2$.

- *Preprocessing:* for all $t \in [T_{ij}]$, sample a hash function $h^{ij}_t \sim \mathcal{H}^{ij}_t$ and evaluate it on $\hat{X}_i$ creating hash table H^{ij}_t . Let $H^{ij}_t(z) \subseteq \hat{X}_i$ denote the hash bucket where $z \in \mathcal{S}^{d-1}$ maps to under h^{ij}_t.

- *Querying:* given a query y ($\|y\| \in S_j$), for all $t \in [T_{ij}]$ let $\hat{X}_t^{ij} \sim H_t^{ij}(\hat{y})$ be a random element from $H_t^{ij}(\hat{y})$ or $\perp$ if $H_t^{ij}(\hat{y}) = \emptyset$. Return $Z_{ij}(y) = \frac{1}{nw_{\max}} \sum_{t\in[T_{ij}]} \left\{ \frac{p_t^{ij}(\hat{X}_t^{ij},\hat{y})}{W^{ij}(\hat{X}_t^{ij},\hat{y})} |H_t^{ij}(\hat{y})| w(X_t, y) \right\}$.

where $W^{ij}(x,y) = \sum_{t\in[T_{ij}]} (p_t^{ij}(x,y))^2$. For $\|y\| \in S_j$, we denote this estimator as $Z_{ij} \sim \mathcal{D}_{ij}(y)$. The estimator is unbiased and has complexity $\mathcal{C}_{ij}$ bounded by $O(dL(\phi_{ij})^{5/6} M_{\phi_{ij}})$ where $M_{\phi_{ij}} = \exp(O\left(\{\log(L(\phi_{ij}))L(\phi_{ij})\}^{2/3}\right))$ and given explicitly below (34) in the proof of Theorem 8. We next bound its variance. Towards that end, we define a different estimator:

$$\begin{aligned}\tilde{Z}_{ij} &= \frac{1}{nw_{\max}} \sum_{t\in[T_{ij}]} \left\{ \frac{p_t^{ij}\left(\hat{X}_t^{ij},\hat{y}\right)}{W^{ij}\left(\hat{X}_t^{ij},\hat{y}\right)} |H_t^{ij}(\hat{y})| w(\tilde{X}_t, \tilde{y}) \right\} \\ &= \left(\frac{p_0(r_i)|X_i|e^{\phi_{ij}^*}}{nw_{\max}} \right) \frac{1}{|X_i|e^{\phi_{ij}^*}} \\ &\quad \cdot \sum_{t\in[T_{ij}]} \left\{ \frac{p_t^{ij}\left(\hat{X}_t^{ij},\hat{y}\right)}{W^{ij}\left(\hat{X}_t^{ij},\hat{y}\right)} |H_t^{ij}(\hat{y})| e^{\phi_{ij}(\langle \hat{x},\hat{y}\rangle)} \right\}.\end{aligned}$$

For this estimator we get by (45) and (46) that $\mathbb{E}[\tilde{Z}_{ij}] = A_{ij}\mu_{ij}$. Furthermore, by our construction of $\{\mathcal{H}_t^{ij}, p_t^{ij}\}_{t\in[T_{ij}]}$ and Theorem 6 for $\beta = 1/2$ it follows that:

$$\mathbb{E}[\tilde{Z}_{ij}^2] \le A_{ij}^2 \cdot \left(16M_{\phi_{ij}}^3 + 1\right) \mu_{ij}^{3/2} \tag{48}$$

Finally, due to Lemma 11 we have that $Z_{ij} \le e\tilde{Z}_{ij}$.

Step 4: We are now in position to define the final estimator and bound its variance. For $\|y\| \in S_j$ and $i \in [K^*]$, let $Z_{ij} \sim \mathcal{D}_{ij}$ as before, and define:

$$Z_j(y) = \sum_{i\in[K^*]} Z_{ij}(y) \tag{49}$$

The estimator is unbiased $\mathbb{E}[Z_j(y)] = Z_w(y)$ and the variance is bounded by

$$\begin{aligned}\mathbb{E}[Z_j^2] &\le (\mathbb{E}[Z_j(y)])^2 + \sum_{i\in[K^*]} \mathbb{E}[Z_{ij}^2] \\ &\le \mu^2 + e^2 \sum_{i\in[K^*]} (16M_{\phi_{ij}}^3 + 1) A_{ij}^2 \mu_{ij}^{3/2} \\ &\le \mu^2 + e^2 \sum_{i\in[K^*]} (16M_{\phi_{ij}}^3 + 1) A_{ij}^{1/2} e^{3/2} (e^{-1} A_{ij}\mu_{ij})^{3/2} \\ &\le \mu^2 + e^{5/2} \max_{i\in[K^*]} \{16M_{\phi_{ij}}^3 + 1\} (e^{-1} \sum_{i\in[K^*]} A_{ij}\mu_{ij})^{3/2} \\ &\le \mu^2 + e^{5/2}(16M_{\phi_{K^*K^*}}^3 + 1)\mu^{3/2}\end{aligned}$$

where in the penultimate inequality we used $A_{ij} \le 1$, Hölder's inequality and super-additivity of $g(x) := x^{3/2}$. The final steps follows from Lemma 12 and monotonicity of $g(x)$. This shows that our estimator is V-bounded with $V(\mu) = 2e^{5/2}(8M_{\phi_{K^*K^*}}^3 + 1)\mu^{-1/2}$ and complexity $O\left(d(K^*)^2(LR^2)^{5/6} M_{\phi_{K^*K^*}}^4\right)$ with $M_{\phi_{K^*K^*}} = \exp(O\left(\left\{\log(L(K^*)^2)L(K^*)^2\right\}^{2/3}\right))$.

C. Proof of Lemma 11

We first show that for all $x_1, x_2 \in S_i(\gamma)$, $y_1, y_2 \in S_j(\gamma)$, and $\gamma \le 1$ we have:

$$\begin{aligned}\left|\|x_1\| - \|x_2\|\right| &\le r_i\gamma, &(50)\\ \left|\|x_1\|\|y_1\| - \|x_2\|\|y_2\|\right| &\le 3r_ir_j\gamma. &(51)\end{aligned}$$

To see the first part, assume without loss of generality that $\|x_1\| \ge \|x_2\|$ and $\|y_1\| \ge \|y_2\|$. We have for $z \in \{x, y\}$: $\|z_1\| - \|z_2\| \le (1+\gamma)^{i(z_1)} r_0 - (1+\gamma)^{i(z_2)-1} r_0 \le (1+\gamma)^{i(z_1)-1} r_0\gamma$. For the second part, we used that $\gamma \le 1$.

$$\begin{aligned}\|y_1\|\|x_1\| - \|y_2\|\|x_2\| &\le (1+\gamma)^{i(y_1)+i(x_1)} r_0^2 \\ &\quad - (1+\gamma)^{i(y_2)+i(x_2)-2} r_0^2 \\ &\le (1+\gamma)^{i(y_1)+i(x_1)-2} r_0^2 \\ &\quad \cdot \left((1+\gamma)^2 - 1\right) \\ &\le 3r_ir_j\gamma.\end{aligned}$$

Using (50), (51) and the fact that $\langle x, y\rangle = \|x\|\|y\|\langle \hat{x}, \hat{y}\rangle$ we get:

$$\begin{aligned}\phi(\langle \tilde{x}, \tilde{y}\rangle) &\ge \phi(\langle x, y\rangle) - L(\phi)(\|x\|\|y\| - \|\tilde{x}\|\|\tilde{y}\|)|\langle \hat{x}, \hat{y}\rangle| \\ &\ge \phi(\langle x, y\rangle) - 3L(\phi) r_{i(x)} r_{i(y)} \gamma,\end{aligned}$$

and

$$\begin{aligned}\phi(\langle \tilde{x}, \tilde{y}\rangle) &\le \phi(\langle x, y\rangle) + L(\phi)(\|x\|\|y\| - \|\tilde{x}\|\|\tilde{y}\|)|\langle \hat{x}, \hat{y}\rangle| \\ &\le \phi(\langle x, y\rangle) + 3L(\phi) r_{i(x)} r_{i(y)} \gamma.\end{aligned}$$

Putting these two together and by the fact that p_0 is $(q, HR\gamma, \gamma)$-log-Lipschitz the statement follows.

VII. Lower bound under SETH or OVC

Conjecture 3 (Strong Exponential Time Hypothesis (SETH)[22])**.** *For any $\epsilon > 0$, there exists $k = k(\epsilon)$ such that k-SAT on n variables cannot be solved in time $O(2^{(1-\epsilon)n})$.*

A conjecture that is implied by SETH [23], [24], concerns the complexity of finding a pair of orthogonal vectors amongst two set of binary vectors.

Conjecture 4 (Orthogonal Vectors Conjecture (OVC))**.** *For every $\delta > 0$ there exists $c = c(\delta)$ such that given two sets $A, B \subset \{0,1\}^m$ of cardinality N, where $m = c \log N$, deciding if there is a pair $(a, b) \in A \times B$ such that $a \top b = 0$ cannot be solved in time $O(N^{2-\delta})$.*

These popular conjectures have been the base of a flurry of quadratic hardness results in the past years. The basis of our hardness result is the following recent theorem by Aviad Rubinstein [25]. Let $d^2(A, B) := \min_{a\in A} \min_{b\in B} \{\|a - b\|_2^2\}$ be the minimum squared distance between $A, B \subset \mathbb{R}^d$.

Theorem 11 (Theorem 4.1[25])**.** *Unless SETH and OVC are false, the following holds: for every $\delta > 0$ and $\epsilon \in (0, e^{-1})$ there exist constants $c(\delta) > 0$, $T(\epsilon) = O(\frac{\log \frac{1}{\epsilon}}{\log\log \frac{1}{\epsilon}})$ and $T' = 2^{O(T \log T)} = O(\frac{1}{\epsilon})$ such that given two sets $A, B \subset \{0,1\}^d$ of N vectors with*

- *Dimension: $d \geq 2mT'$, with $m = c(\delta)\log N$*
- *Sparisty: for all $x \in A \cup B$, $\|x\|_2^2 = mT'$*

there is no algorithm that decides whether $d^2(A,B) \begin{cases} = m(T'-1) \\ \text{or} \\ \geq mT' \end{cases}$ in time $N^{2-O\left(\delta + c(\delta)\frac{\log^2\log\frac{1}{\epsilon}}{\log\frac{1}{\epsilon}}\right)}$.

Our proof will proceed by translating hardness for the problem of Approximate Bi-chromatic Closest pair to our setting. This connection was first established in [43] to obtain quadratic hardness results for Kernel Methods and Neural Networks.

A. Proof of Theorem 2

Proof: The proof proceeds by showing how to reduce an instance (A, B) of the approximate Bi-chromatic closest pair in Theorem 11 to an instance $(X, Y) \subset \mathcal{S}^{d-1} \times \mathcal{S}^{d-1}$ of producing a α approximation to: $\frac{1}{N^2}\sum_{x\in X}\sum_{y\in Y} e^{L\cdot(\langle x,y\rangle - 1)}$.

Setting $\epsilon = e^{-e^{\delta/c(\delta)}} \in (0, e^{-1})$ in Theorem 11: We start by finding a constant $\epsilon \in (0, e^{-1})$ such that:

$$c(\delta)\frac{\log^2\log\frac{1}{\epsilon}}{\log\frac{1}{\epsilon}} \leq \delta \tag{52}$$

$$\Leftrightarrow \log^2\log\frac{1}{\epsilon} \leq \left(\frac{\delta}{c(\delta)}\right)\log\frac{1}{\epsilon} \tag{53}$$

$$\Leftrightarrow \zeta^2 \leq \left(\frac{\delta}{c(\delta)}\right)e^{\zeta} \tag{54}$$

where $\zeta = \log\log\frac{1}{\epsilon} > 0$. Setting $\zeta = \frac{\delta}{c(\delta)} > 0$ we get $e^{\zeta} \geq 1 + \zeta \geq \zeta > 0$. For this choice we have:

$$\epsilon = e^{-e^{\zeta}} < e^{-1} \Leftrightarrow e^{\zeta} > 1 \tag{55}$$

Hence, we may pick $\epsilon = e^{-e^{\frac{\delta}{c(\delta)}}}$ for which $\tilde{T}(\delta) = O(\frac{e^{\delta/c(\delta)}}{\delta/c(\delta)})$ and $\tilde{T}'(\delta) = O(e^{e^{\delta/c(\delta)}})$. Theorem 11 then shows that there is no $N^{2-O(\delta)}$ algorithm to decide between: $d^2(A,B) \begin{cases} = m(\tilde{T}'(\delta)-1) \\ \text{or} \\ \geq m\tilde{T}'(\delta) \end{cases}$.

Translating distance bounds to Density bounds for Gaussian Kernel: We next show that distinguishing between the two cases for $d^2(A,B)$ distinguishes between two values for the average of the Gaussian kernel between points in the two datasets. In the case where $d^2(A,B) \geq m\tilde{T}'(\delta)$, we have that:

$$\frac{1}{N^2}\sum_{a\in A}\sum_{b\in B} e^{-\beta\|a-b\|^2} \leq e^{-\beta m\tilde{T}'(\delta)} \tag{56}$$

In the other case, where $d^2(A,B) = m(\tilde{T}'(\delta)-1)$ we get:

$$\frac{1}{N^2}\sum_{a\in A}\sum_{b\in B} e^{-\beta\|a-b\|^2} \geq \frac{1}{N^2}e^{-\beta d^2(A,B)} = e^{-\beta m\tilde{T}'(\delta)}\cdot e^{-2\log N+\beta m}$$

So as long as $e^{-2\log N+\beta m} > \alpha \Leftrightarrow \beta > \frac{2\log N+\log\alpha}{m}$ any algorithm that can produce a α-approximation to $\frac{1}{N^2}\sum_{a\in A}\sum_{a\in B} e^{-\beta\|a-b\|^2}$ distinguishes between the two cases as such it cannot run in time $N^{2-O(\delta)}$.

Gaussian Kernel to Log-convex (linear) and Bound on Lipschitz Constant: To complete the proof we observe that:

$$\begin{aligned}\beta\|a-b\|^2 &= -\beta 2m\tilde{T}'(\langle\frac{a}{\sqrt{m\tilde{T}'}},\frac{b}{\sqrt{m\tilde{T}'}}-1\rangle) \\ &= L(\langle\frac{a}{\sqrt{m\tilde{T}'}},\frac{b}{\sqrt{m\tilde{T}'}}\rangle - 1)\end{aligned}$$

with $L := 2\beta m\tilde{T}'$. Setting $Y := \{a/\sqrt{m\tilde{T}'} : a \in A\}$ and $X := \{b/\sqrt{m\tilde{T}'} : b \in B\}$ we have that:

$$e^{-\beta\|a-b\|^2} = e^{L(\langle y,x\rangle - 1)}$$

and $X, Y \subset \mathcal{S}^{d-1}$. Hence, substituting the lower bound on β we get that for:

$$L > 2\tilde{T}'(\delta)(2\log N+\log\alpha) = \left\{C(\delta)\left(1+\frac{\log\alpha}{2\log N}\right)\right\}\cdot\log N$$

where $C(\delta) = O\left(e^{e^{\frac{\delta}{c(\delta)}}}\right)$ there is no algorithm that approximates the sum in time less than $N^{2-O(\delta)}$. ∎

VIII. Importance Sampling for Vector Functions

In this section, we show that for a class of unbiased estimators, that result from *jointly sampling* a random weight function $U : X \cup \{\perp\} \to \mathbb{R}_+$ and a random point $Y \in X \cup \{\perp\}$ according to some *balanced distribution*, the variance of an unbiased estimator for the sum of vectors is bounded by that of the same distribution applied for the vector norms (Corollary 5). The class of such estimators include trivially classical importance sampling as well as Hashing-Based-Estimators (Lemma 14). Using this connection we will show how to estimate sum of gradients when the gradient norms are log-convex functions of the inner product.

A. Randomly weighted estimators via Balanced distributions

We start by defining a class of estimators that work by sampling a point Y from $X \cup \{\perp\}$ and a, possibly random and correlated with Y, function $U : X \cup \{\perp\} \to \mathbb{R}_+$ with support possibly on a subset S of X.

Definition 8 (Balanced distribution)**.** *Given a finite set $S \subset X$, let $\mathcal{D}$ be a distribution of a pair of random variables $(U, Y) \sim \mathcal{D}$ where $Y \in X \cup \{\perp\}$ and $U : X \cup \{\perp\} \to \mathbb{R}_+$.*

A distribution is called S-balanced *if* $U(S^c \cup \{\perp\}) = \{0\}$, *and* $\mathbb{E}[U(x)|Y=x] = \frac{1}{\mathbb{P}[Y=x]} \in (0,\infty)$ *for all* $x \in S$.

Classical importance sampling schemes correspond to the case where $U(x) = \frac{1}{\mathbb{P}[Y=x]}$ is a deterministic function of x. We show next that any such distribution, even with random U, can be used to create unbiased estimators for the sum of a function on S.

Lemma 13 (Moments)**.** *Let* $S \subseteq X$, $f : X \cup \{\perp\} \to \mathbb{R}$ *a bounded function, and* $\mathcal{D}$ *an* S*-balanced distribution. For* $(U,Y) \sim \mathcal{D}$ *it holds that*

- $\mathbb{E}[U(Y)f(Y)] = \sum_{x\in S} f(x)$, *and*
- $\mathbb{E}[\{U(Y)f(Y)\}^2] = \sum_{x\in S} \frac{\mathbb{E}[U^2(x)|Y=x]}{\mathbb{E}[U(x)|Y=x]} f^2(x)$.

Proof: Using the law of total probability we have:

$$\begin{aligned}\mathbb{E}[U(Y)f(Y)] &= \sum_{x\in S} \mathbb{E}[U(Y)f(Y)|Y=x]\mathbb{P}[Y=x] \\ &= \sum_{x\in S} f(x)\mathbb{E}[U(x)|Y=x]\mathbb{P}[Y=x] \\ &= \sum_{x\in S} f(x).\end{aligned}$$

We proceed similarly:

$$\begin{aligned}\mathbb{E}[\{U(Y)f(Y)\}^2] &= \sum_{x\in S} \mathbb{E}[\{U(Y)f(Y)\}^2|Y=x]\mathbb{P}[Y=x] \\ &= \sum_{x\in S} \mathbb{E}[U^2(x)|Y=x]\mathbb{P}[Y=x]f^2(x) \\ &= \sum_{x\in S} \frac{\mathbb{E}[U^2(x)|Y=x]}{\mathbb{E}[U(x)|Y=x]} f^2(x).\end{aligned}$$

■

Finally, we show that for vector functions the variance is controlled by the variance of the corresponding estimator for the sum of the gradient norms.

Corollary 5 (Vectors to Norms)**.** *Let* $g : X \cup \{\perp\} \to \mathbb{R}^d$ *a bounded function, and* $S \subseteq X$. *For any* S*-balanced distribution* $(U,Y) \sim \mathcal{D}$, *we have* $\mathbb{E}[U(Y)g(Y)] = \sum_{x\in S} g(x)$ *and*

$$\mathbb{E}[\|U(Y)g(Y)\|^2] = \sum_{x\in S} \frac{\mathbb{E}[U^2(x)|Y=x]}{\mathbb{E}[U(x)|Y=x]} \|g(x)\|^2. \quad (57)$$

Proof: The first equation follows by applying Lemma 13 for $i \in [d]$, $g_i : X \to \mathbb{R}$ and linearity of expectation, while the second part by applying the lemma for $f(x) = \|g(x)\|$. ■

B. Hashing-Based-Estimators

We next show that Hashing-Based-Estimators induce indeed balanced distributions for the support of the collision probability on X for a given query y.

Lemma 14 (HBE)**.** *Given a set* $X \subset \mathcal{X}$, *and a hashing scheme* $\mathcal{H}$ *with collision probabilities* $p : \mathcal{X} \cup \{\perp\} \times \mathcal{X} \to [0,1]$, *let* $(h,g) \sim \mathcal{H}$. *For any given* $y \in \mathcal{X}$, *let* $Y \sim H_X(y)$ *and* $S(y) := \{x \in X | p(x,y) > 0\}$, *the distribution of* $\left(\frac{|H(y)|}{p(Y,y)}, Y\right)$ *is* $S(y)$*-balanced.*

Proof: For all $x \in S(y)$,

$$\begin{aligned}\mathbb{E}\left[\frac{|H(y)|}{p(Y,y)}\middle| Y=x\right] &= \frac{\mathbb{E}\left[\frac{|H(y)|}{p(x,y)}\mathbb{I}[Y=x]\right]}{\mathbb{P}[Y=x]} \\ &= \frac{\mathbb{E}\big[|H(y)|\mathbb{I}[Y=x]\mathbb{I}[x\in H(y)]\big]}{p(x,y)\mathbb{P}[Y=x]} \\ &= \frac{\mathbb{E}\big[|H(y)|\mathbb{I}[Y=x]\big|x\in H(y)\big]p(x,y)}{p(x,y)\mathbb{P}[Y=x]} \\ &= \frac{1}{\mathbb{P}[Y=x]} \in (0,\infty).\end{aligned}$$

■

C. Multi-resolution HBE

To cover Multi-resolution HBE, or their Multi-scale extension described in Section VI, we show that adding together randomly weighted estimators, resulting from balanced distributions that are pairwise independent, produces the results we expect.

Corollary 6. *Given* $X \subset \mathcal{X}$, $y \in \mathcal{X}$, *let* $(U_t, Y_t) \sim \mathcal{D}_t(y)$ *for* $t \in [T]$ *being pairwise independent and* $D_t(y)$ t *being* $S_t(y)$*-balanced. Let* $T(x,y) = \{t \in [T] | x \in S_t(y)\}$. *For a collection of bounded functions* $\{f_t : \mathcal{X} \cup \{\perp\} \to \mathbb{R}^d\}_{t\in[T]}$, *we have:*

$$\mathbb{E}[\sum_{t\in[T]} U_t(Y_t)f_t(Y_t)] = \sum_{x\in X} \sum_{t\in T(x,y)} f_t(x) \quad (58)$$

and

$$\begin{aligned}\mathbb{E}[\|\sum_{t\in[T]} U_t(Y_t)f_t(Y_t)\|^2] \le \sum_{t\in[T]} \mathbb{E}[\{U_t(Y_t)\|f_t(Y_t)\|\}^2] \\ + \mathbb{E}[\sum_{t\in[T]} U_t(Y_t)\|f_t(Y_t)\|]^2.\end{aligned} \quad (59)$$

Proof: The first part follows easily due to linearity and Lemma 13, while the second one follows from triangle inequality. ■

This shows that if Multi-resolution HBE has small variance in estimating the sum of the vector norms, it can be used to estimate the sum of the vectors with the same variance up to constants.

Corollary 7. *Let* $g : \mathcal{S}^{d-1} \times \mathcal{S}^{d-1} \to \mathbb{R}^m$ *be a vector function such that* $\|g(x,y)\|_2 = e^{\phi(\langle x,y\rangle)}$ *for some convex function* ϕ. *Given* $\epsilon, \tau \in (0,1)$, *there exists an explicit constant* M_ϕ *and a data structure using space* $O\left(dL(\phi)^{5/6}M_\phi^3 \frac{1}{\epsilon^2}\frac{1}{\sqrt{\tau}} \cdot n\right)$ *and query time*

$O(dL(\phi)^{5/6}M_\phi^4\frac{1}{\epsilon^2}\frac{1}{\sqrt{\mu}})$ *that for any* $y \in \mathcal{S}^{d-1}$ *with constant probability can either produce a a vector* G *such that*

$$\left\| G - \frac{1}{ne^{\phi_{\max}}} \sum_{x \in X} g(x,y) \right\|_2 \leq \epsilon\mu \tag{60}$$

if $\mu := \frac{1}{ne^{\phi_{\max}}} \sum_{x\in X} \|g(x,y)\|_2 \geq \tau$ *or assert that* $\mu < \tau$.

Proof: We first call Theorem 9 to construct an Multi-resolution HBE for the problem of approximating $Z_\phi(y)$, where $\phi = \log(\|g\|_2)$. By Corollary 6, this shows that we can turn our MR-HBE estimator to an unbiased estimator for $\sum_{x\in X} g(x,y)$ and that the variance is bounded by that of estimating $Z_\phi(y)$. ∎

IX. Remaining Proofs

This section contains proofs of lemmas and theorems stated in the main paper as well as various auxiliary results.

A. Proof of Corollary 1

Under the condition $r \leq \frac{1}{2}\sqrt{\log n}$ we have that the Lipschitz constants of the first four functions in Table I are bounded by $L(\phi) \leq 2r^2 \leq \frac{1}{2}\log n$. This is also true for the last function under the condition $0 \leq k \leq \frac{c-1}{2}\log n$. The result follows from $\mu \in [e^{-2L(\phi)}, 1] \subseteq [\frac{1}{n}, 1]$.

B. Moments of Multi-resolution HBE

Proof of Lemma 4: We start by computing the first moment:

$$\mathbb{E}[Z_T(y)] = \frac{1}{|X|}\sum_{t\in[T]} \mathbb{E}\left[\frac{w_t(X_t,y)}{p_t(X_t,y)}|H_t(y)|\right] \tag{61}$$

$$= \frac{1}{|X|}\sum_{x\in X}\sum_{t\in T(x,y)} w_t(x,y) \tag{62}$$

$$= \frac{1}{|X|}\sum_{x\in X} w(x,y) \tag{63}$$

The second moment is given by

$$\mathbb{E}[Z_T^2] = \sum_{t\in[T]}\sum_{t'\in[T]} \mathbb{E}[\frac{w_t(X_t,y)|H_t(y)|}{p_t(X_t,x)|X|}\frac{w_{t'}(X_{t'},x)|H_{t'}(y)|}{p_{t'}(X_{t'},x)|X|}]$$
$$\leq \frac{1}{|X|^2}\sum_{t\in[T]}\mathbb{E}\left[\frac{w_t^2(X_t,y)}{p_t^2(X_t,y)}|H_t(y)|^2\right] + \mu^2$$

The proof is concluded by $\mathbb{E}\left[\frac{w_t^2(X_t,y)}{p_t^2(X_t,y)}|H_t(y)|^2\right] = \sum_{x\in X}\frac{w_t^2(x,y)}{p_t(x,y)}\mathbb{E}\left[|H_t(y)||x\in H_t(y)\right] \leq \sum_{x\in X}\frac{w_t^2(x,y)}{p_t(x,y)}\sum_{z\in X}\frac{\min\{p_t(z,y),p_t(x,y)\}}{p_t(x,y)}$. ∎

C. Distance Sensitive Hashing on the unit sphere

To analyze the collision probability of the DSH scheme we closely follow the proof of Aumuller et al. [27] with the difference that we use Proposition 7 to bound bi-variate Gaussian integrals.

Proposition 6 (Proposition 3 [52])**.** *Let* $X_1 \sim \mathcal{N}(0,1)$ *and* $t > 0$*,*

$$\frac{1}{\sqrt{2\pi}}\frac{1}{t+1}e^{-\frac{t^2}{2}} \leq \mathbb{P}[X_1 \geq t] \leq \frac{1}{\sqrt{2\pi}}\frac{1}{t}e^{-\frac{t^2}{2}} \tag{64}$$

Proposition 7 (Propositions 3.1 & 3.2 [53])**.** *Let* $(X_1,X_2) \sim \mathcal{N}(0, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix})$ *be two* ρ*-correlated standard normal random variables. For all* $\rho < 1$ *and* $t > 0$*:*

$$\mathbb{P}[X_1 > t \wedge X_2 > t] \geq \frac{4}{(1+\sqrt{1+4\frac{(1+\rho)^2}{\min(1-\rho,1+\rho)}})^2} \cdot \frac{\min(1-\rho,1+\rho)}{(1+\rho)^2}\frac{1+|\rho|}{2\pi\sqrt{1-\rho^2}}e^{-\frac{2}{1+\rho}\frac{t^2}{2}}, \tag{65}$$

and

$$\mathbb{P}[X_1 > t \wedge X_2 > t] \leq \frac{(1+\rho)^{\frac{3}{2}}}{2\pi\sqrt{1-\rho}}e^{-\frac{2}{1+\rho}\frac{t^2}{2}}. \tag{66}$$

We first simplify the sub-exponential terms appearing on the above inequalities using our assumption that $|\rho| < 1-\delta$. Since the function $\frac{(1+\rho)^{\frac{3}{2}}}{2\pi\sqrt{1-\rho}}$ is increasing in ρ we get $\frac{(1+\rho)^{\frac{3}{2}}}{2\pi\sqrt{1-\rho}} \leq \frac{\sqrt{2}}{\pi\sqrt{\delta}}$. Additionally, we have that $\frac{\min(1-\rho,1+\rho)}{(1+\rho)^2} \geq \frac{\delta}{4}$ and $(a+b)^2 \leq 2(a^2+b^2)$ for all $a,b \in \mathbb{R}$. Using the above bounds we get:

$$\frac{4}{(1+\sqrt{1+4\frac{(1+\rho)^2}{\min(1-\rho,1+\rho)}})^2}\frac{\min(1-\rho,1+\rho)}{(1+\rho)^2}\frac{1+|\rho|}{2\pi\sqrt{1-\rho^2}}$$
$$\geq \frac{2}{2+\frac{16}{\delta}}\frac{\delta}{4}\frac{1}{2\pi} \geq \frac{\delta^2}{8+\delta}\frac{1}{8\pi}.$$

We are now in a position ot bound the collision probability.

Proof of Lemma 5: The collision probability can be written as:

$$\mathbb{P}[h(x) = g(y)] = \mathbb{P}[h(x) \leq m \wedge g(y) \leq m]P_t(x,y) \tag{67}$$

where $P_t(x,y) := \frac{\mathbb{P}[\langle x,g\rangle \geq t \wedge \langle y,g\rangle \geq t]}{\mathbb{P}[\langle x,g\rangle \geq t \vee \langle y,g\rangle \geq t]}$. We are going to obtain upper and lower bounds for both terms. We start first with the second term. An easy calculation shows that the vector $(X_1,X_2) := (\langle x,g\rangle, \langle y,g\rangle) \sim \mathcal{N}(0, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix})$ follows a bivariate normal distribution with unit variances and correlation $\rho = \langle x,y\rangle$. Hence, $\mathbb{P}[\langle x,g \geq t\rangle] = \mathbb{P}[\langle y,g \geq t\rangle] = \mathbb{P}[X_1 \geq t]$ and $\mathbb{P}[\langle x,g\rangle \geq t \wedge \langle y,g\rangle \geq t] = \mathbb{P}[X_1 \geq t \wedge X_2 \geq t]$. Using monotonicity and union bound we get that:

$$\frac{1}{2}\frac{\mathbb{P}[X_1 \geq t \wedge X_2 \geq t]}{\mathbb{P}[X_1 \geq t]} \leq P_t(x,y) \leq \frac{\mathbb{P}[X_1 \geq t \wedge X_2 \geq t]}{\mathbb{P}[X_1 \geq t]}. \tag{68}$$

Using (68) and the estimates from Propositions 6, 7

$$\frac{\sqrt{2}\delta^2}{148\sqrt{\pi}}e^{-\frac{1-\rho}{1+\rho}\frac{t^2}{2}} \leq P_t(x,y) \leq \frac{2}{\sqrt{\pi}\sqrt{\delta}}e^{-\frac{1-\rho}{1+\rho}\frac{t^2}{2}} \quad (69)$$

Next, we bound the remaining term as

$$\begin{aligned}\mathbb{P}[h(x) \leq m \wedge g(y) \leq m] &\geq 1 - \mathbb{P}[h(x) > m \wedge g(y) > m] \\ &\geq 1 - 2(1 - \mathbb{P}[X_1 \geq t])^m \\ &\geq 1 - 2e^{-\mathbb{P}[X_1 \geq t]m} \\ &\geq 1 - \zeta. \end{aligned} \quad (70)$$

where in the last step we used the definition of $m(t,\zeta)$ and the lower bound from (64). Using the last inequality along with (69) and (67), we arrive at:

$$\frac{\sqrt{2}(1-\zeta)\delta^2}{148\sqrt{\pi}}e^{-\frac{1-\rho}{1+\rho}\frac{t^2}{2}} \leq p_+(\rho) \leq \frac{2}{\sqrt{\pi}\sqrt{\delta}}e^{-\frac{1-\rho}{1+\rho}\frac{t^2}{2}}. \quad (71)$$

Next, we treat the case where $1-\delta < \rho \leq 1$, let Z_1, Z_2 be standard normal random variables then:

$$\frac{\mathbb{P}[Z_1 \geq t \wedge \rho Z_1 + \sqrt{1-\rho^2}Z_2 \geq t]}{\mathbb{P}[Z_1 \geq t \vee \rho Z_1 + \sqrt{1-\rho^2}Z_2 \geq t]} \geq \frac{1}{2}\mathbb{P}[Z_2 \geq \sqrt{\frac{1-\rho}{1+\rho}}t],$$

and

$$\begin{aligned}\mathbb{P}[Z_2 \geq \sqrt{\frac{1-\rho}{1+\rho}}t] &\geq \frac{1}{\sqrt{2\pi}}\frac{\sqrt{1+\rho}}{\sqrt{1-\rho}+\sqrt{1+\rho}}e^{-\frac{1-\rho}{1+\rho}\frac{t^2}{2}} \\ &\geq \frac{1}{\sqrt{2\pi}}\frac{1}{1+\sqrt{2}}e^{-\frac{\delta}{2-\delta}\frac{t^2}{2}}.\end{aligned}$$

Lastly, we show an upper bound on $p_+(\rho)$ for $-1 \leq \rho \leq -1+\delta$, we have that:

$$\begin{aligned}p_+(\rho) &\leq \frac{\mathbb{P}[Z_1 \geq t \wedge \rho Z_1 + \sqrt{1-\rho^2}Z_2 \geq t]}{\mathbb{P}[Z_1 \geq t \vee \rho Z_1 + \sqrt{1-\rho^2}Z_2 \geq t]} \\ &\leq \frac{1}{\mathbb{P}[Z_1 \geq t]}\int_t^\infty \mathbb{P}[Z_2 \geq \frac{t-\rho u}{\sqrt{1-\rho^2}}]\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}du \\ &\leq \frac{1}{\mathbb{P}[Z_1 \geq t]}\int_t^\infty \mathbb{P}[Z_2 \geq \frac{t-(-1+\delta)u}{\sqrt{1(-1+\delta)^2}}]\frac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}}du \\ &\leq \frac{2}{\sqrt{\pi}\sqrt{\delta}}e^{-\frac{2}{\delta}\frac{t^2}{2}}.\end{aligned}$$

This concludes the proof. ∎

D. Idealized Hashing

We consider the idealized hashing probability $h_{\gamma,t}(\rho) = -\left(\frac{1-\rho}{1+\rho} + \gamma^2\frac{1+\rho}{1-\rho}\right)\frac{t^2}{2}$. Its first and second derivatives are given by:

$$h'_{\gamma,t}(\rho) = \left(\frac{1}{(1+\rho)^2} - \gamma^2\frac{1}{(1-\rho)^2}\right)t^2, \quad (72)$$

$$h''_{\gamma,t}(\rho) = -2\left(\frac{1}{(1+\rho)^3} - \gamma^2\frac{1}{(1-\rho)^3}\right)t^2. \quad (73)$$

Proof of Proposition 2: Using (72), we see that the derivative becomes zero only at $\rho^*(\gamma) = \frac{1-\gamma}{1+\gamma}$ and that the second derivative becomes zero at $\rho^{**}(\gamma) = \frac{1-\gamma^{\frac{2}{3}}}{1+\gamma^{\frac{2}{3}}}$. Let $g(x) = \frac{1-x}{1+x}$, the function $h_{\gamma,t}$ is concave for all $\rho \geq \rho^{**}(\gamma) = g(\gamma^{\frac{2}{3}})$. Since g is decreasing for all $\rho \geq -1$, we have:

$$\gamma \leq 1 \Rightarrow \gamma \leq \gamma^{\frac{2}{3}} \Rightarrow g(\gamma) \geq g(\gamma^{\frac{2}{3}}) \Leftrightarrow \rho^*(\gamma) \geq \rho^{**}(\gamma),$$
$$\gamma \geq 1 \Rightarrow \gamma \geq \gamma^{\frac{2}{3}} \Rightarrow g(\gamma) \leq g(\gamma^{\frac{2}{3}}) \Leftrightarrow \rho^*(\gamma) \leq \rho^{**}(\gamma).$$

∎

Proof of Proposition 3: We only show the case where ϕ is non-decreasing the other case follows similarly. We have that $g(\rho) \leq g(\rho_*)$ for all $\rho \in [-1,1]$. By concavity, we know that:

$$g(\rho) \leq g(\rho_0) + g'(\rho_0)(\rho - \rho_0), \ \forall \rho \in [-1, \rho^*]$$

Therefore, we have that for all $\rho \in [1, \rho_*]$

$$\begin{aligned}\phi(\rho) - g(\rho) &\geq \phi(\rho) - g(\rho_0) - g'(\rho_0)(\rho - \rho_0) \\ &\geq \phi'(\rho_0) - g(\rho_0) + [\phi'(\rho_0) - g'(\rho_0)](\rho - \rho_0) \\ &= \phi(\rho_0) - g(\rho_0)\end{aligned}$$

Finally, for $\rho \in [\rho^*, 1]$ we have by monotonicity $\phi(\rho) - g(\rho) \geq \phi(\rho_*) - g(\rho_*) \geq \phi(\rho_0) - g(\rho_0)$. ∎

Proof of Corollary 4: Using the fact that $a+b \leq 2\max\{a,b\}$ and estimates from Lemma 3, we get that

$$\begin{aligned}t_0^2 &= -\frac{1}{2}\frac{1+\rho_0}{1-\rho_0}(2\phi(\rho_0) - (1-\rho_0^2)\phi'(\rho_0)) \\ &\leq -2\frac{1+\rho_0}{1-\rho_0}\phi(\rho_0),\end{aligned}$$

and

$$\begin{aligned}\gamma_0^2 t_0^2 &= -\frac{1}{2}\frac{1-\rho_0}{1+\rho_0}(2\phi(\rho_0) + (1-\rho_0^2)\phi'(\rho_0)) \\ &\leq -2\frac{1-\rho_0}{1+\rho_0}\phi(\rho_0).\end{aligned}$$

When $\phi'(\rho_0) \geq 0$, we get by (9) that:

$$\begin{aligned}t_0^2 &\geq -\frac{1+\rho_0}{1-\rho_0}\phi(\rho_0) \\ \gamma_0^2 t_0^2 &\geq -\frac{1}{2}\frac{1-\rho_0}{1+\rho_0}2\phi(\rho_0)\frac{1-\rho_0}{2} \geq -\frac{(1-\rho_0)^2}{2(1+\rho_0)}\phi(\rho_0)\end{aligned}$$

Similarly, when $\phi'(\rho_0) \leq 0$, we get by (11):

$$t_0^2 \geq -\frac{1}{2}\frac{1+\rho_0}{1-\rho_0}2\phi(\rho_0)\frac{1+\rho_0}{2} \geq -\frac{(1+\rho_0)^2}{2(1-\rho_0)}\phi(\rho_0) \quad (74)$$

$$\gamma_0^2 t_0^2 \geq -\frac{1-\rho_0}{1+\rho_0}\phi(\rho_0) \quad (75)$$

Using again $\max\{a,b\} \geq \frac{a+b}{2}$, we get in both cases that $\max\{\gamma_0^2 t_0^2, t_0^2\} \geq -\frac{1+\rho_0^2}{1-\rho_0^2}\phi(\rho_0)$. ∎

E. Approximation

Proof of Lemma 7: The idea is to select a set of points $\rho_1, \ldots, \rho_T$ and break $[\rho_-, \rho_+]$ in intervals $\rho_i \leq \rho \leq \rho_i + \Delta(\rho_i)$ of length $\Delta(\rho_i)$ such that within each interval $\ell(\rho)$ is well approximated by $h_{\rho_i}(\rho)$. For $\rho \geq \rho_0$ using the Taylor Remainder theorem and the fact that $\ell(\rho_0) = h(\rho_0)$ as well as $\ell^{'}(\rho_0) = h^{'}(\rho_0)$, there exists $\xi = \xi(\rho, \rho_0) \in [\rho_0, \rho]$ such that

$$\ell(\rho) - h_{\rho_0}(\rho) = -\frac{1}{2} h^{''}_{\rho_0}(\xi(\rho, \rho_0))(\rho - \rho_0)^2 \geq 0 \qquad (76)$$

Where the inequality follows by concavity of h. To obtain an upper bound, we need an absolute bound on the second derivative. Using (73), we get that

$$|h^{''}_{\gamma_0, t_0}| \leq 2 \max\left\{ \frac{1}{(1+\rho)^3} t_0^2, \gamma_0^2 t_0^2 \frac{1}{(1-\rho)^3} \right\}$$

Substituting the upper bounds from Corollary 4 in turn gives

$$|h^{''}_{\gamma_0, t_0}| \leq 8 \max\left\{ \frac{1+\rho_0}{1-\rho_0} \frac{1}{(1+\rho)^3}, \frac{1-\rho_0}{1+\rho_0} \frac{1}{(1-\rho)^3} \right\} R(\ell)$$

For $\rho_0 \leq \rho \leq \rho_0 + \Delta(\rho_0) \leq 0$ we have $|h^{''}_{\rho_0}| \leq \frac{16}{(1-|\rho_0|)^2} R(\ell)$. Setting $\Delta(\rho_0) = \sqrt{\frac{\epsilon}{8R(\ell)}}(1 - |\rho_0|)$, gives

$$\ell(\rho) - h_{\rho_0}(\rho) \leq \frac{1}{2}|h^{''}_{\rho_0}|\Delta^2(\rho_0) \leq \frac{8}{(1-|\rho_0|)^2}|\ell_{\min}|\Delta^2(\rho_0) \leq \epsilon$$

Hence, we have the following inductive definition of points ρ_i:

$$1 + \rho_i = 1 + \rho_{i-1} + \Delta(\rho_{i-1}) \qquad (77)$$

$$= (1 + \rho_{i-1}) + \sqrt{\frac{\epsilon}{8R(\ell)}}(1 + \rho_{i-1}) \qquad (78)$$

$$= (1 + \sqrt{\frac{\epsilon}{8R(\ell)}})(1 + \rho_{i-1}) \qquad (79)$$

multiplying both sides with $\sqrt{\frac{\epsilon}{8R(\ell)}}$ gives us the updates for $\Delta(\rho_i)$. We are now in a position to write an explicit expression for ρ_i:

$$\begin{aligned}
\rho_i &= \rho_- + \sum_{j=1}^{i} \Delta(\rho_{j-1}) \\
&= \rho_- + \sum_{j=1}^{i} \left(1 + \sqrt{\frac{\epsilon}{8R(\ell)}}\right)^{j-1} \sqrt{\frac{\epsilon}{8R(\ell)}}(1 - |\rho_-|) \\
&= \rho_- + \left[\left(1 + \sqrt{\frac{\epsilon}{8R(\ell)}}\right)^{i} - 1 \right] (1 - |\rho_-|)
\end{aligned}$$

for $i = 0, \ldots, T$ with $T = \lfloor \frac{\log(\frac{1-|\rho_+|}{1-|\rho_-|})}{\log(1+\sqrt{\frac{\epsilon}{8R(\ell)}})} \rfloor$. The floor function is justified by the fact that if $\rho_T < \rho_+$ then $\rho_T + \Delta_T > \rho_+$ and as such ϕ is well approximated between $[\rho_T, \rho_+]$ by h_{ρ_T}. The lemma follows by setting $i(\rho) := \min\{j \in \{0, \ldots, T\} | \rho_i \leq \rho\}$. ∎

F. Scale-free Multi-resolution HBE

Proof of Lemma 9: We bound the difference

$$E_\epsilon(\phi) := \sup_{\rho \in [-1,1]} | \sup_{\rho_0 \in \mathcal{T}_\epsilon(\phi)} \{h_{\rho_0}(\rho)\} - \sup_{\rho_0 \in \mathcal{T}_\epsilon(\phi)} \{\log(p_{\gamma_0, t_0}(\rho))\}|$$

We break the analysis into three parts depending where ρ belongs to. The first case $\rho \in [-1 + \delta, 1 - \delta]$ is the easier one, as due to Lemma 5 and Corollary 3 we have for all $\rho_0 \in \mathcal{T}_\epsilon(\phi)$,

$$-\log(C_1) \leq \log(p_{\gamma_0, t_0}(\rho)) - h_{\rho_0}(\rho) \leq \log C_1. \qquad (80)$$

Hence, for all $\rho \in [-1 + \delta, 1 - \delta]$:

$$| \sup_{\rho_0 \in \mathcal{T}_\epsilon(\phi)} \{h_{\rho_0}(\rho)\} - \sup_{\rho_0 \in \mathcal{T}_\epsilon(\phi)} \{\log(p_{\gamma_0, t_0}(\rho))\}| \leq \log C_1.$$

We next treat the case $\rho \in [-1, -1 + \delta]$. Recall that $h_{\pm 1}(\rho) := -\frac{1 \mp \rho}{1 \pm \rho} \frac{t^2_{\pm 1}}{2} + \phi(\pm 1)$, where $t^2_{\pm 1} = 4 \max\{\pm \phi^{'}(\pm 1), 0\}$. Assuming that ϕ is increasing at -1, by construction $t^2_{-1} = 0$ and hence for all $\rho \in \mathcal{T}_\epsilon(\phi)$,

$$\log(p_{\gamma_0, t_0}(\rho)), h_{\rho_0}(\rho) \geq h_{-1}(\rho) \geq \phi(-1). \qquad (81)$$

Assuming that ϕ is decreasing at -1, we have $t^2_- = 4|\phi^{'}(-1)|$ and for all $\rho \in \mathcal{T}_\epsilon(\phi)$,

$$h_{\rho_0}(\rho) \geq h_{-1}(\rho) = \phi(-1) - 2\frac{1+\rho}{1-\rho}|\phi^{'}(-1)| \geq \phi(-1) \qquad (82)$$

Moreover, by (21) in Lemma 5 applied to $p_+(-\rho)$, we get for all $\rho_0 \in \mathcal{T}_{\frac{1}{2}}(\tilde{\phi})$

$$\begin{aligned}
\log(p_{\gamma_0, t_0}(\rho)) &\geq -\frac{1}{2}\log C_1 - \frac{2}{2-\delta}\delta\phi'(-1) + \phi(-1) \\
&\geq -\frac{1}{2}\log C_1 + \phi(-1). \qquad (83)
\end{aligned}$$

By Proposition 3 and the fact that ϕ can be written as the supremum of linear functions we get that $\sup_{\rho_0 \in \mathcal{T}_\epsilon(\phi)}\{h_{\rho_0}(\rho)\} \leq \phi(\rho)$. Using Corollary 3 and Corollary 4, we obtain at for all $\rho_0 \in \mathcal{T}_\epsilon(\phi) \setminus \{-1, +1\}$,

$$\begin{aligned}
p_{\gamma_0, t_0}(\rho) &\leq -\frac{2-\delta}{\delta} t^2_{\gamma_0} + \frac{1}{2}\log C_1 \\
&\leq -\frac{2-\delta}{\delta}(-\frac{1+\rho_0^2}{1-\rho_0^2}\phi(\rho_0)) + \frac{1}{2}\log C_1
\end{aligned}$$

To bound the above quantity further, distinguish two cases: $\phi(-1) = 0$ or $\phi(1) = 0$. By convexity, in the former case we have $\phi(\rho_0) \leq \frac{1+\rho_0}{2}\phi(1)$ and $\phi(\rho_0) \leq \frac{1-\rho_0}{2}\phi(-1)$ in the latter. Substituting these bounds and solving the optimization problem we find that the minimizer in the first case is $\rho_0 = -\sqrt{2} + 1$ and in the latter case $\rho_0 = \sqrt{2} - 1$. In both cases we may obtain:

$$\begin{aligned}
&\sup_{\rho_0 \in \mathcal{T}_{\frac{1}{2}}(\tilde{\phi}) \setminus \{-1, +1\}} \{p_{\gamma_0, t_0}(\rho)\} \\
&\leq -\frac{2-\delta}{\delta}(\sqrt{2} - 1)\max\{|\tilde{\phi}(1)|, |\tilde{\phi}(-1)|\} + \frac{1}{2}\log C_1.
\end{aligned}$$

Next, we obtain bounds for $\rho_0 \in \{-1, +1\}$:

$$\log(p_{-1}(\rho)) \leq \phi(-1),$$
$$\log(p_{+1}(\rho)) \leq \frac{1}{2}\log C_1 - 2\frac{2-\delta}{\delta}\max\{\phi^{'}(1), 0\}$$

Using the above inequalities we may conclude that:

$$\sup_{\rho\in[-1,-1+\delta]}\sup_{\rho_0\in\mathcal{T}_\epsilon(\phi)}\{\log p_{\gamma_0,t_0}(\rho)\} \leq \max\{\phi(-1), \frac{1}{2}\log C_1\} \leq \phi(-1) + \frac{1}{2}\log C_1 \tag{84}$$

We have for $\rho \in [-1, -1+\delta]$ by (81) and (83)

$$\sup_{\rho_0\in\mathcal{T}_\epsilon(\phi)}\{h_{\rho_0}(\rho)\} - \sup_{\rho_0\in\mathcal{T}_\epsilon(\phi)}\{\log(p_{\gamma_0,t_0}(\rho))\} \leq \phi(\rho) - \phi(-1) + \frac{1}{2}\log C_1 \leq L(\phi)\delta + \frac{1}{2}\log C_1$$

where in the last step we used the fact that ϕ is Lipischitz. In the same vein by (81) and (84)

$$\sup_{\rho_0\in\mathcal{T}_\epsilon(\phi)}\{h_{\rho_0}(\rho)\} - \sup_{\rho_0\in\mathcal{T}_\epsilon(\phi)}\{\log(p_{\gamma_0,t_0}(\rho))\} \geq \phi(-1) - \phi(-1) - \frac{1}{2}\log C_1 = -\frac{1}{2}\log C_1$$

Using $\delta \leq \frac{\epsilon}{L(\phi)} \Rightarrow L(\phi)\delta \leq \epsilon \leq \epsilon \log C_1$. By symmetry the case $\rho \in [1-\delta, 1]$ follows. Overall, for $\epsilon = 1/2$ we obtain the bound $E_{1/2}(\phi) \leq \log C_1$. ∎

Proof of Lemma 10: Let $\delta_{1/2} := \frac{k^*}{2\beta L(\phi)}$ be the constant from Lemma 8 applied for $\tilde{\phi}$, then for all $\rho_0 \in \mathcal{T}_{\frac{1}{2}}(\tilde{\phi}) \setminus \{-1, +1\}$ we have $|\rho_0| \leq 1 - \delta_{1/2}$. Using $\tilde{\phi}(\rho_0) \leq R(\tilde{\phi}) = \frac{\beta}{k}R(\phi)$ and $|\rho_0| \leq 1 - \delta_{1/2} = 1 - \frac{k}{2\beta L(\phi)}$ for $\rho_0 \neq \pm 1$, we get by Corollary 4 that $\sup_{\rho_0\in\mathcal{T}_{\frac{1}{2}}(\tilde{\phi})} t^2_{\gamma_0} \leq 8L(\tilde{\phi})R(\tilde{\phi}) \leq 8\left(\frac{\beta}{k}\right)^2 L(\phi)R(\phi)$. For $\rho_0 \in \{-1, 1\}$ we have $t^2 \leq 4|\phi'(\rho)| \leq 4L(\tilde{\phi}) \leq 8L(\tilde{\phi})R(\tilde{\phi})$ for $R(\tilde{\phi}) \geq 1/2$. ∎

X. Future Directions

Data-dependent LSH: Both the HBE and Multi-Resolution HBE approaches exhibit $1/\sqrt{\mu}$ complexity depending on $\mu = Z_w(y)$. For HBE [10], the instance that realizes the worst-case variance of the estimator is when there are $O(n\mu)$ points very close to the query such that $w(x_1, y) = \Theta(1)$ and $O(n)$ points "away" from the query such that $w(x_2, y) = \Theta(\mu)$. On the other hand for MR-HBE, if one uses the full power of Theorem 5 (see Section III) by analyzing $D_T(x_1, x_2)$ rather than its simplified version Theorem 6, the worst case instance for the variance appears to have $O(n\sqrt{\mu})$ points with $w(x_1, y) = \Theta(\sqrt{\mu})$ and $O(n)$ points with $w(x_2, y) = \Theta(\mu)$. For the Gaussian kernel this essentially means that it involves solving a c-ANN problem with $c = \sqrt{2}$. Utilizing this connection to the ANN problem and subsequently adapting the Data-Dependent LSH approach [48] for this setting is an intriguing direction for future work.

Locality Sensitive Hashing: One disadvantage of many LSH based approaches is that hash functions can often be slow to evaluate at least in the form suggested by the theory. In recent years there has been an effort to design practical hash functions that come close to the performance of the optimal ones. For example the papers [55], [56] study practical functions for the unit sphere, while [57] study functions for the binary hypercube. Combining these novel LSH methods with the method of Hashing Based Estimators introduced in [10] and extended here, is a promising direction to getting practical algorithms for estimation problems.

Variance Reduction: The topic of Variance Reduction for Stochastic Gradient [58], [59], [60] is an important field of current research. There are roughly three almost orthogonal approaches to this problem: re-weighting schemes [61], [62], [63], importance sampling schemes [64], [65] and partition-based schemes [66], [67]. For almost all these approaches, the distribution that gradients are sampled is independent of the current iterate (e.g. uniform or based on Lipschitz constants of gradients), or changes with the current iterate and requires linear time to update the new distributions. The latter approaches are referred to as Adaptive Variance Reduction methods [68], [69], [70]. Our approach sidesteps the issue of recomputing such distributions through the use of Locality Sensitive Hashing. An intriguing direction is to utilize our techniques within an optimization algorithm to obtain faster optimization methods.

Acknowledgments

The authors would like to thank Dimitris Achlioptas, Clement Canonne and anonymous reviewers for valuable feedback on improving the presentation of the paper, as well as Aviad Rubinstein for helpful conversations on conditional lower bounds. We are also grateful to Casper Freksen for pointing out a number of typos on an earlier version of the paper. This research was supported by NSF grant CCF-1617577, a Simons Investigator Award, a Google Faculty Research Award and an Amazon Research Award. The second author is partially supported by a Onassis Foundation Scholarship.

References

[1] S. Muthukrishnan *et al.*, "Data streams: Algorithms and applications," *Foundations and Trends® in Theoretical Computer Science*, vol. 1, no. 2, pp. 117–236, 2005.

[2] D. P. Woodruff *et al.*, "Sketching as a tool for numerical linear algebra," *Foundations and Trends® in Theoretical Computer Science*, vol. 10, no. 1–2, pp. 1–157, 2014.

[3] A. McGregor, "Graph stream algorithms: a survey," *ACM SIGMOD Record*, vol. 43, no. 1, pp. 9–20, 2014.

[4] S. Zou, Y. Liang, H. V. Poor, and X. Shi, "Unsupervised nonparametric anomaly detection: A kernel method," in *Communication, Control, and Computing (Allerton), 2014 52nd Annual Allerton Conference on*. IEEE, 2014, pp. 836–841.

[5] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[6] S. Arora, Y. Li, Y. Liang, T. Ma, and A. Risteski, "Linear algebraic structure of word senses, with applications to polysemy," *Transactions of the Association of Computational Linguistics*, vol. 6, pp. 483–495, 2018.

[7] V. Vapnik, *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

[8] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[9] Y. Li, P. M. Long, and A. Srinivasan, "Improved bounds on the sample complexity of learning," *Journal of Computer and System Sciences*, vol. 62, no. 3, pp. 516–527, 2001.

[10] M. Charikar and P. Siminelakis, "Hashing-based-estimators for kernel density in high dimensions," in *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*. IEEE, 2017, pp. 1032–1043.

[11] M. J. Wainwright, M. I. Jordan *et al.*, "Graphical models, exponential families, and variational inference," *Foundations and Trends® in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2008.

[12] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[14] S. Arora, Y. Li, Y. Liang, T. Ma, and A. Risteski, "A latent variable model approach to pmi-based word embeddings," *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 385–399, 2016.

[15] L. Devroye and G. Lugosi, *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012.

[16] 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*. SIAM, 2014, pp. 542–550.

[17] E. Gan and P. Bailis, "Scalable kernel density classification via threshold-based pruning," in *Proceedings of the 2017 ACM International Conference on Management of Data*. ACM, 2017, pp. 945–959.

[18] 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*. ACM, 2011, pp. 47–56.

[19] 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.

[20] J. Liu, W.-C. Chang, Y. Wu, and Y. Yang, "Deep learning for extreme multi-label text classification," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2017, pp. 115–124.

[21] Y. Prabhu, A. Kag, S. Harsola, R. Agrawal, and M. Varma, "Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2018, pp. 993–1002.

[22] R. Impagliazzo and R. Paturi, "On the complexity of k-sat," *Journal of Computer and System Sciences*, vol. 62, no. 2, pp. 367–375, 2001.

[23] R. Williams, "A new algorithm for optimal 2-constraint satisfaction and its implications," *Theoretical Computer Science*, vol. 348, no. 2-3, pp. 357–365, 2005.

[24] R. Williams and H. Yu, "Finding orthogonal vectors in discrete structures," in *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*. SIAM, 2014, pp. 1867–1877.

[25] A. Rubinstein, "Hardness of approximate nearest neighbor search," in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 2018, pp. 1260–1268.

[26] T. D. Ahle, M. Aumüller, and R. Pagh, "Parameter-free locality sensitive hashing for spherical range reporting," in *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017, Barcelona, Spain, Hotel Porta Fira, January 16-19*, 2017, pp. 239–256. [Online]. Available: https://doi.org/10.1137/1.9781611974782.16

[27] M. Aumüller, T. Christiani, R. Pagh, and F. Silvestri, "Distance-sensitive hashing," in *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*. ACM, 2018, pp. 89–104.

[28] M. R. Jerrum, L. G. Valiant, and V. V. Vazirani, "Random generation of combinatorial structures from a uniform distribution," *Theoretical Computer Science*, vol. 43, pp. 169–188, 1986.

[29] A. Sinclair and M. Jerrum, "Approximate counting, uniform generation and rapidly mixing markov chains," *Information and Computation*, vol. 82, no. 1, pp. 93–133, 1989.

[30] 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.

[31] C. Luo and A. Shrivastava, "Arrays of (locality-sensitive) count estimators (ace): Anomaly detection on the edge," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2018, pp. 1439–1448.

[32] B. Chen, Y. Xu, and A. Shrivastava, "Lsh-sampling breaks the computational chicken-and-egg loop in adaptive stochastic gradient estimation," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*, 2018. [Online]. Available: https://openreview.net/forum?id=Hk8zZjRLf

[33] R. Spring and A. Shrivastava, "Scalable estimation via LSH samplers (LSS)," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*, 2018. [Online]. Available: https://openreview.net/forum?id=BJazbHkPG

[34] A. Andoni and P. Indyk, "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions," in *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*. IEEE, 2006, pp. 459–468.

[35] 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*. ACM, 2004, pp. 253–262.

[36] L. Greengard and V. Rokhlin, "A fast algorithm for particle simulations," *Journal of computational physics*, vol. 73, no. 2, pp. 325–348, 1987.

[37] P. B. Callahan and S. R. Kosaraju, "A decomposition of multidimensional point sets with applications to k-nearest-neighbors and n-body potential fields," *Journal of the ACM (JACM)*, vol. 42, no. 1, pp. 67–90, 1995.

[38] L. Greengard and J. Strain, "The fast gauss transform," *SIAM Journal on Scientific and Statistical Computing*, vol. 12, no. 1, pp. 79–94, 1991.

[39] A. G. Gray and A. W. Moore, "Nonparametric density estimation: Toward computational tractability," in *Proceedings of the 2003 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, 2003, pp. 203–211.

[40] C. Yang, R. Duraiswami, N. A. Gumerov, and L. Davis, "Improved fast gauss transform and efficient kernel density estimation," in *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2003.

[41] D. Lee, A. W. Moore, and A. G. Gray, "Dual-tree fast gauss transforms," in *Advances in Neural Information Processing Systems*, 2006.

[42] P. Ram, D. Lee, W. March, and A. G. Gray, "Linear-time algorithms for pairwise statistical problems," in *Advances in Neural Information Processing Systems*, 2009.

[43] 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*, 2017, pp. 4308–4318.

[44] A. Backurs, M. Charikar, P. Indyk, and P. Siminelakis, "Efficient density evaluation for smooth kernels."

[45] S. Mussmann and S. Ermon, "Learning and inference via maximum inner product search," in *International Conference on Machine Learning*, 2016, pp. 2587–2596.

[46] S. Mussmann, D. Levy, and S. Ermon, "Fast amortized inference and learning in log-linear models with randomly perturbed nearest neighbor search," *arXiv preprint arXiv:1707.03372*, 2017.

[47] 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*. ACM, 2015, pp. 793–801.

[48] A. Andoni, T. Laarhoven, I. Razenshteyn, and E. Waingarten, "Optimal hashing-based time-space trade-offs for approximate near neighbors," in *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2017, pp. 47–66.

[49] 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*. ACM, 1996, pp. 20–29.

[50] G. Rote, "The convergence rate of the sandwich algorithm for approximating convex functions," *Computing*, vol. 48, no. 3, pp. 337–361, 1992.

[51] A. Andoni, P. Indyk, H. L. Nguyen, and I. Razenshteyn, "Beyond locality-sensitive hashing," in *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2014, pp. 1018–1028.

[52] S. J. Szarek and E. Werner, "A nonsymmetric correlation inequality for gaussian measure," *Journal of multivariate analysis*, vol. 68, no. 2, pp. 193–211, 1999.

[53] E. Hashorva and J. Hüsler, "On multivariate gaussian tails," *Annals of the Institute of Statistical Mathematics*, vol. 55, no. 3, pp. 507–522, 2003.

[54] D. Chakraborty, L. Kamma, and K. G. Larsen, "Tight cell probe bounds for succinct boolean matrix-vector multiplication," in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 2018, pp. 1297–1306.

[55] C. Kennedy and R. Ward, "Fast cross-polytope locality-sensitive hashing," in *LIPIcs-Leibniz International Proceedings in Informatics*, vol. 67. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[56] A. Andoni, P. Indyk, T. Laarhoven, I. Razenshteyn, and L. Schmidt, "Practical and optimal lsh for angular distance," in *Advances in Neural Information Processing Systems*, 2015, pp. 1225–1233.

[57] A. Andoni, I. Razenshteyn, and N. S. Nosatzki, "Lsh forest: Practical algorithms made theoretical," in *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2017, pp. 67–78.

[58] N. Le Roux, M. W. Schmidt, F. R. Bach *et al.*, "A stochastic gradient method with an exponential convergence rate for finite training sets." in *NIPS*, 2012, pp. 2672–2680.

[59] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Advances in neural information processing systems*, 2013, pp. 315–323.

[60] S. Shalev-Shwartz and T. Zhang, "Stochastic dual coordinate ascent methods for regularized loss minimization," *Journal of Machine Learning Research*, vol. 14, no. Feb, pp. 567–599, 2013.

[61] G. Lan, "An optimal method for stochastic composite optimization," *Mathematical Programming*, vol. 133, no. 1-2, pp. 365–397, 2012.

[62] Z. Allen-Zhu and E. Hazan, "Variance reduction for faster non-convex optimization," in *International Conference on Machine Learning*, 2016, pp. 699–707.

[63] Z. Allen-Zhu, "Katyusha: The first direct acceleration of stochastic gradient methods," in *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, 2017, pp. 1200–1205.

[64] P. Zhao and T. Zhang, "Stochastic Optimization with Importance Sampling for Regularized Loss Minimization," in *International Conference on Machine Learning*, 2015, pp. 1–9.

[65] Z. Allen-Zhu, Z. Qu, P. Richtárik, and Y. Yuan, "Even faster accelerated coordinate descent using non-uniform sampling," in *International Conference on Machine Learning*, 2016, pp. 1110–1119.

[66] P. Zhao and T. Zhang, "Accelerating minibatch stochastic gradient descent using stratified sampling," *arXiv preprint arXiv:1405.3080*, 2014.

[67] Z. Allen-Zhu, Y. Yuan, and K. Sridharan, "Exploiting the structure: Stochastic gradient methods using raw clusters," in *Advances in Neural Information Processing Systems*, 2016, pp. 1642–1650.

[68] D. Csiba, Z. Qu, and P. Richtárik, "Stochastic dual coordinate ascent with adaptive probabilities," in *International Conference on Machine Learning*, 2015, pp. 674–683.

[69] H. Namkoong, A. Sinha, S. Yadlowsky, and J. C. Duchi, "Adaptive sampling probabilities for non-smooth optimization," in *International Conference on Machine Learning*, 2017, pp. 2574–2583.

[70] F. Salehi, E. Celis, and P. Thiran, "Stochastic optimization with bandit sampling," *arXiv preprint arXiv:1708.02544*, 2017.