Uniform Approximations for Randomized Hadamard Transforms with Applications

Yeshwanth Cherapanamjeri
yeshwanth@berkeley.edu
University of California at Berkeley
Berkeley, California, USA

Jelani Nelson
minilek@berkeley.edu
University of California at Berkeley
Berkeley, California, USA

ABSTRACT

Randomized Hadamard Transforms (RHTs) have emerged as a computationally efficient alternative to the use of dense unstructured random matrices across a range of domains in computer science and machine learning. For several applications such as dimensionality reduction and compressed sensing, the theoretical guarantees for methods based on RHTs are comparable to approaches using dense random matrices with i.i.d. entries. However, several such applications are in the low-dimensional regime where the number of rows sampled from the matrix is rather small. Prior arguments are not applicable to the high-dimensional regime often found in machine learning applications like kernel approximation. Given an ensemble of RHTs with Gaussian diagonals, $\{M^i\}_{i=1}^m$, and any 1-Lipschitz function, $f : \mathbb{R} \to \mathbb{R}$, we prove that the average of f over the entries of $\{M^i v\}_{i=1}^m$ converges to its expectation uniformly over $\|v\| \leq 1$ at a rate comparable to that obtained from using truly Gaussian matrices. We use our inequality to then derive improved guarantees for two applications in the high-dimensional regime: 1) kernel approximation and 2) distance estimation. For kernel approximation, we prove the first *uniform* approximation guarantees for random features constructed through RHTs lending theoretical justification to their empirical success while for distance estimation, our convergence result implies data structures with improved runtime guarantees over previous work by the authors. We believe our general inequality is likely to find use in other applications.

CCS CONCEPTS

• **Theory of computation → Random projections and metric embeddings**; **Sketching and sampling**; **Kernel methods**; *Pseudorandomness and derandomization*; *Computational geometry*.

KEYWORDS

adaptive statistics, random matrix theory, hadamard transforms, kernel approximation, distance estimation, pseudo-randomness

ACM Reference Format:
Yeshwanth Cherapanamjeri and Jelani Nelson. 2022. Uniform Approximations for Randomized Hadamard Transforms with Applications. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing (STOC '22), June 20–24, 2022, Rome, Italy*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3519935.3519961

STOC '22, June 20–24, 2022, Rome, Italy

ACM ISBN 978-1-4503-9264-8/22/06.
https://doi.org/10.1145/3519935.3519961

1 INTRODUCTION

Randomized linear mappings find ubiquitous application in diverse domains across computer science and machine learning. Representing a linear transformation $f : \mathbb{R}^d \to \mathbb{R}^k$ as a matrix $\Pi \in \mathbb{R}^{k \times d}$ such that $f(x) = \Pi x$, a commonly examined randomized linear mapping is one where the entries of Π are drawn i.i.d. from a simple distribution; say, a standard normal. Randomized matrices of the previous form have found use as tools for compressed sensing [7, 13], dimensionality reduction [15], machine learning [21], and differential privacy [5], amongst other areas. However, one downside to the use of such transformations is that they can be slow, as applying the map amounts to dense matrix-vector multiplication.

Randomized Hadamard Transforms (RHTs) have emerged as a versatile alternative to the use of fully random matrices in applications ranging from the construction of fast Johnson-Lindenstrauss transforms [1], to speeding up iterative recovery methods in compressed sensing [8, 20], designing fast algorithms for approximate regression and low-rank approximation [23], and building faster algorithms for deep learning [10]; their special structure allowing for faster computation of the mapping. Assuming $d = 2^\ell$ for some $\ell \in \mathbb{N}$, the RHT is defined as follows:

$$f(x) = H_d D x \text{ where } H_d = \begin{bmatrix} H_{d/2} & H_{d/2} \\ H_{d/2} & -H_{d/2} \end{bmatrix} \text{ with}$$

$$H_1 = [1] \text{ and } D_{i,j} \sim \begin{cases} \mathcal{N}(0,1) & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}.$$

Due to the recursive structure of the Hadamard matrix, the mapping $f(x)$ can be computed in time $O(d \log d)$ as opposed to $O(d^2)$ for a matrix with i.i.d. entries. While each row of the matrix is distributed as a standard normal vector, entires in a column are correlated due to the shared randomness from D. Despite these correlations, in "low-dimensional" applications such as dimensionality reduction and compressed sensing, where a small number of rows are sampled from $f(x)$, prior work has shown that (subsampled) RHTs provide guarantees competitive with the use of Gaussian random matrices.

However, for "high-dimensional" applications frequently found in machine learning where k may be significantly larger than d, known guarantees for RHTs are not comparable to those for Gaussian random matrices. As a concrete example, consider the problem of approximating the RBF kernel, defined as $K_{\text{RBF}}(x, y) := \exp\left\{-\frac{\|x-y\|^2}{2}\right\}$ where $\|\cdot\|$ denotes the Euclidean norm. In their pioneering work, Rahimi and Recht construct an embedding, $h_g(x)$, of dimension (d/ε^2) based on Gaussian random matrices such that

$|\langle h_g(x), h_g(y)\rangle - K_{\mathrm{RBF}}(x,y)| \leq \varepsilon$ for all x, y in a bounded ball. These embeddings have grown to become one of the most widely adopted techniques for scaling up kernel methods and as such have been hugely influential in machine learning, with its impact recognized in NeurIPS 2017 with a Test of Time Award. Due to their widespread use, much effort has been devoted toward improving the computational complexity of these methods with approaches based on RHTs emerging as a popular alternative with comparable empirical performance and significantly faster runtimes [18, 26]. However, in sharp contrast to the situation for Gaussian matrices, there are no known uniform concentration results for methods based on RHTs, despite their superior computational properties and empirical performance [12, 26].

Our main result is a uniform concentration inequality on RHTs with the goal of bridging the gap between RHTs and full Gaussian matrices in the "high-dimensional" setting where we show that for any Lipschitz function, its average over the entries of the output of a RHT converges uniformly (over inputs to the RHT) to its expectation at a rate comparable to that obtained for full Gaussian matrices. We use our result to establish improved theoretical guarantees for two "high-dimensional" problems: kernel approximation and distance estimation, illustrating its broad applicability. We introduce some notation then state our main result as Theorem 1.1. Below and in the rest of the paper, $\mathcal{N}(0, \sigma^2)$ denotes the Gaussian with mean 0 and variance σ^2.

$$\{D^j\}_{j=1}^m \in \mathbb{R}^{d\times d} \text{ s.t } (D^j)_{k,l} \overset{iid}{\sim} \begin{cases} \mathcal{N}(0,1), & \text{if } k = l \\ 0, & \text{otherwise} \end{cases}$$

$$\forall z \in \mathbb{R}^d : \widetilde{h}(z) := \begin{bmatrix} HD^1 z \\ HD^2 z \\ \vdots \\ HD^m z \end{bmatrix}, \ \widetilde{h}_{j,k}(z) = (HD^j z)_k, \ \widetilde{h}^j(z) = HD^j z \tag{RHT}$$

Theorem 1.1. *Let $d \in \mathbb{N}, \delta, \varepsilon \in (0, 1/2)$ and $f : \mathbb{R} \to \mathbb{R}$ be a 1-Lipschitz function. Then we have with probability at least $1 - \delta$:*

$$\forall z, \|z\| \leq 1 : \left| \frac{1}{md} \cdot \sum_{j=1}^{m} \sum_{k=1}^{d} f(\widetilde{h}_{j,k}(z)) - \mathbb{E}_{Z\sim\mathcal{N}(0,\|z\|^2)}[f(z)] \right| \leq \varepsilon$$

as long as $m \geq C\varepsilon^{-2} \log^5(d/\varepsilon) \log 1/\delta$ for some absolute constant $C > 0$.

We pause to make some remarks regarding Theorem 1.1. First, note that the number of rows in the linear transformation is md, and hence the concentration properties obtained in Theorem 1.1 are within a small logarithmic factor of those obtained from the use of full Gaussian matrices where $\approx d/\varepsilon^2$ suffice (see Section 2 for a standard proof). Secondly, similar results are *not* obtainable when an alternative distribution, $\mathcal{D}$, is used in place of Gaussians in the diagonal matrices in the definition of the RHT. To see this, consider the case where the $\mathcal{D}$ is symmetric and observe that the empirical distribution of the entries of $\widetilde{h}(e_1)$ converge to $\mathcal{D}$ while the entries of $\widetilde{h}(\mathbf{1})$ converge to a Gaussian as a consequence of the central limit theorem. We conclude our discussion with two complementary lower bounds establishing the tightness of Theorem 1.1. In Theorem A.1, we show that the constraint on the embedding dimension in terms of m is optimal up to log factors by exhibiting a candidate 1-Lipschitz function requiring $m \geq \varepsilon^{-2} \log d/\delta$ and finally, in Theorem A.2, we show that there exists a 1-Lipschitz function requiring an embedding dimension of at least $\varepsilon^{-2} d$ for uniform concentration when true Gaussian random matrices are used. Taken together, these results imply that RHTs are *optimally* comparable (in the sense of Theorem 1.1) to random Gaussian matrices and that this phenomenon is *not* an artifact of the looseness of either analysis.

To our knowledge, this is the first uniform concentration inequality of this type for RHTs and we anticipate its use beyond the applications illustrated in our work. We now discuss applications of Theorem 1.1 to two tasks featuring high dimensional embeddings: kernel approximation and distance estimation. For both of these applications, our result yields significant runtime improvements over prior work.

1.1 Kernel Approximation

Kernel functions drastically increase the ability of machine learning based methods to learn complex functions of the underlying data. Roughly speaking, these techniques allow the use of a user specified "inner-product" function, $K(x, y)$, which corresponds to the inner product $\langle \phi(x), \phi(y)\rangle$ for some function $\phi : \mathbb{R}^d \to \mathcal{H}$ and Hilbert space $\mathcal{H}$. For instance, consider a simple classification task where the input data consists of pairs, $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \{0, 1\}$, and the goal is learn a classifier predicting the label y on a new input x. Kernel methods represent the classifier as a linear combination $q(x) = \sum_{i=1}^n \alpha_i K(x_i, x)$ where the coefficients α_i are learnt from data. By parameterizing the classifier in this way, Kernel methods can exploit the flexibility offered by the use of high dimensional embeddings without explicitly performing the embedding which may be computationally expensive/infeasible depending on the kernel used.

One major drawback of kernel functions is that naively evaluating the classifier on even a single input point can potentially incur a runtime of nd. One approach to improve this runtime is the use of Random Fourier Features [21], in which one embeds the data points into a Euclidean space such that inner products of the embeddings roughly correspond to the evaluation of the kernel. For the popular RBF kernel, their embedding is defined as follows where the function $\cos(\cdot)$ is applied elementwise:

$$h(x) = \sqrt{\frac{2}{md}} \cdot \cos(\Pi x + b) \text{ where}$$

$$\Pi \in \mathbb{R}^{k\times d}, b \in \mathbb{R}^k \text{ with } \Pi_{i,j} \overset{iid}{\sim} \mathcal{N}(0,1), \ b_i \overset{iid}{\sim} \mathrm{Unif}([0, 2\pi])$$

For fixed $B > 0$ and $k \approx d/\varepsilon^2$, Rahimi and Recht establish the following claim with high probability:

$$\forall x, y \in \mathbb{B}(0, B) : |\langle h(x), h(y)\rangle - K_{\mathrm{RBF}}(x, y)| \leq \varepsilon.$$

However, there are no proven universal approximation results when the random transformation Πx is replaced by a RHT despite their empirical success across a range of machine learning applications [10–12, 24, 26]. With this context in mind, we present our theorem for the approximation of the RBF kernel:

Theorem 1.2. *Let $d \in \mathbb{N}, \delta, \varepsilon \in (0, 1/2)$ and $\mathcal{W} \subset \mathbb{R}^d$ be arbitrary. Then, defining:*

$$h(x) = \sqrt{\frac{2}{md}} \cdot \cos(\widetilde{h}(x) + b)$$

where $\widetilde{h}(\cdot)$ is defined in RHT and $b_i \overset{iid}{\sim} \mathrm{Unif}([0, 2\pi])$, we have:

$$\forall x, y \in \mathcal{W} : |\langle h(x), h(y)\rangle - K_{\mathrm{RBF}}(x, y)| \leq \varepsilon$$

with probability at least $1 - \delta$ when $m \geq \widetilde{\Omega}\left(\varepsilon^{-2}\mathrm{Diam}(\mathcal{W})^2 \log 1/\delta\right)$.

While previous approaches have shown approximation results for *fixed* (x, y) in expectation [26, Theorem 1], Theorem 1.2 is the first *uniform* approximation guarantee for RHTs thus providing theoretical justification for their empirical success. While in-expectation guarantees suffice if a classifier has already been trained and test vectors are chosen independently of the classifier, these approximations are often used in tandem with an iterative optimization procedure during training and in deployment, may face test points which are potentially correlated with predictions on previous inputs. In both these scenarios featuring potentially adaptive inputs, in-expectation guarantees break down while uniform guarantees continue to hold. Note that standard approaches such as generating a new random embedding for each step of an optimization procedure or each input query fail as the coefficients of a linear method utilizing these embeddings are *specific* to the embedding and are unlikely to transfer to a new randomly chosen one. While the dependence of the embedding dimension on $\mathrm{Diam}(\mathcal{W})$ are weaker than those obtained for full Gaussian matrices which have logarithmic dependence on $\mathrm{Diam}(\mathcal{W})$, note that the most interesting regime is when $\|x - y\| \approx \widetilde{O}(1)$ as the RBF kernel decays rapidly in $\|x - y\|$.

1.2 Distance Estimation

The second application of our result is in the construction of adaptive algorithms for distance estimation. Formally, the distance estimation problem is defined as follows:

Problem 1.3 (Distance Estimation)**.** For a known metric, $d(\cdot, \cdot)$, we are given $X = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and $\varepsilon \in (0, 1)$ and we are required to construct a data structure $\mathcal{D}$, which when given input query q, outputs distance estimates $\{d_i\}_{i=1}^n$ satisfying:

$$(1 - \varepsilon)d(q, x_i) \leq d_i \leq (1 + \varepsilon)d(q, x_i).$$

Our goal will be to build a data structure for distance estimation in the adaptive setting where the sequence of queries seen by the data structure are potentially adversarially chosen with knowledge of answers to previous queries and even potentially the instantiation of the data structure. Note however, that the query cannot depend on future randomness that the algorithm may draw in the process of answering it. The construction of adaptive data structures has received much attention in the recent literature [2–4, 9, 14, 16]. For the particular problem of distance estimation, the approach devised in [9] achieves nearly optimal space complexity and query time for all ℓ_p "norms" for $p \in (0, 2)$.

We now focus solely on the Euclidean setting where our results apply and briefly recall the construction from [9]. The approach first draws l i.i.d random Gaussian matrices $\{\Pi_i \in \mathbb{R}^{k \times d}\}_{i=1}^l$ with $l \approx d$ and $k \approx \varepsilon^{-2}$. For each $x_j \in X$, its embedding, $\Pi_i x_j$ is computed for each Π_i and stored. When given query, q, the data structure samples $p = O(\log n)$ random matrices, $\{\Pi_{i_r}\}_{r=1}^p$ and outputs $d_i = \mathrm{Median}(\{\|\Pi_{i_r}(q - x_i)\|\}_{r=1}^p)$. This approach yields nearly optimal querytimes of $\widetilde{O}(n/\varepsilon^2)$. Unfortunately, the update and construction times of the data structure are slow. As the matrices, $\{\Pi_i\}_{i=1}^l$, have no special structure, adding a new point to the data structure takes time $\widetilde{O}(d^2)$ and despite the existence of fast methods for matrix multiplication, construction the data structure is slow ($O(d^\omega)$ for $n = d$ where ω is the matrix multiplication constant).

Our result for distance estimation constructs an algorithm for distance estimation in Euclidean space:

Theorem 1.4. *Let $\varepsilon, \delta \in (0, 1/2)$. Then, there is a data structure for Distance Estimation in Euclidean space which is initialized correctly with probability at least $1 - \delta$ and supports the following operations:*

1. *Output a correct answer to a possibly adaptively chosen distance estimation query with probability at least $1 - \delta$*
2. *Add input $x \in \mathbb{R}^d$ to the database, X.*

Furthermore, the query and update times of the data structure are $\widetilde{O}(\varepsilon^{-2}(n + d)\log 1/\delta)$ and $\widetilde{O}(\varepsilon^{-2}d \log 1/\delta)$ respectively while the data structure is constructed in time $\widetilde{O}(\varepsilon^{-2}(nd)\log 1/\delta)$.

In comparison to [9, Theorem 4.1], Theorem 1.4 implies a factor d improvement in update and construction times which is significant in high-dimensional applications. Furthermore, Theorem 1.4 yields nearly optimal guarantees as the time taken to construct the data structure is near linear in the size of the data structure and the space complexity was also shown to be optimal in [9].

Organization: The rest of the paper is organized as follows. We give a brief overview of the proof of Theorem 1.1 in Section 2. We then present the formal proof in Section 3 and describe applications to kernel approximation in Section 4 where we prove Theorem 1.2 and distance estimation in Section 5 which proves Theorem 1.4. Finally, Appendix A contains proofs of our lower bounds establishing the optimality of Theorem 1.1 while Appendix B contains standard inequalities and basic technical results used in our proofs.

Notation: Throughout the paper, $\widetilde{h}(\cdot)$ denotes the RHT defined in RHT. For $x \in \mathbb{R}^d$, we use $\mathbb{B}(x, r)$ to denote the ball of radius r around x, $\|x\|$ and $\|x\|_\infty$ to denote its Euclidean and infinity norm and $\|x\|_0$ and $\mathrm{Supp}(x)$ will denote the size of its support and its support respectively. When used with a matrix $M \in \mathbb{R}^{p \times q}$, $\|M\|$ will denote the spectral norm of M. For $x \in \mathbb{R}^d$ and a diagonal matrix $B \in R^{d \times d}$, we use $\mathrm{Diag}(x)$ to denote the diagonal matrix, D, with the entries of x along the diagonal while $\mathrm{diag}(B)$ denotes the vector consisting of the diagonal entries of B in order. For two sets of identically indexed subsets of $\mathbb{R}^d$, $V_S = \{v_s\}_{s \in S}$ and $U_S = \{u_s\}_{s \in S}$, we abuse notation and let $\|V_S - U_S\| = \sqrt{\sum_{s \in S}\|u_s - v_s\|^2}$. For a set $\mathcal{W} \subset \mathbb{R}^d$, $\mathrm{Diam}(\mathcal{W})$ will denote its diameter. A scalar function, f, when applied to a vector is applied elementwise. For $x \in \mathbb{R}^d$ and $S \subset [d]$, we let x_S to denote the vector obtained by setting the entries of x not in S to 0. For $\alpha \in (0, 1)$ and finite $S \subset \mathbb{R}$, we use $\mathrm{Quant}_\alpha(S)$ to denote the αth quantile of the set and ϕ and Φ will denote the pdf and cdf of a standard Gaussian random variable.

2 PROOF OVERVIEW

We now briefly describe the main ideas behind the proof of Theorem 1.1. Before we begin, it is instructive to inspect standard methods of establishing similar results for the Gaussian setting and their shortcomings in our scenario. Specifically, for n i.i.d standard Gaussian vectors $\{g_i\}_{i=1}^n$ (the rows of a Gaussian matrix), our goal will be to establish the following inequality for some absolute constant $C > 0$:

$$Z\left(\{g_i\}_{i=1}^n\right) := \max_{\|v\|\le 1}\left|\frac{1}{n}\cdot\sum_{i=1}^n f\left(\langle g_i, v\rangle\right) - \mathbb{E}_{g\sim\mathcal{N}(0,\|v\|^2)}\left[f(g)\right]\right| \le C\left(\sqrt{\frac{d}{n}} + \sqrt{\frac{\log 1/\delta}{n}}\right) \tag{1}$$

with probability at least $1-\delta$. Note that n here is analogous to md for our RHTs.

We now bound the expectation and concentration terms of Z separately. We start with the mildly more involved concentration term. In particular, we will show that Z is a Lipschitz function of the g_i. Let $\{g_i'\}_{i=1}^n \in \mathbb{R}^d$ be an alternative choice of vectors. We have:

$$\begin{aligned}
&\left|Z(\{g_i\}_{i=1}^n) - Z(\{g_i'\}_{i=1}^n)\right| \\
&= \left|\max_{\|v\|\le 1}\left|\frac{1}{n}\cdot\sum_{i=1}^n f\left(\langle g_i, v\rangle\right) - \mathbb{E}_{g\sim\mathcal{N}(0,\|v\|^2)}\left[f(g)\right]\right| - \right. \\
&\qquad \left. \max_{\|v\|\le 1}\left|\frac{1}{n}\cdot\sum_{i=1}^n f\left(\langle g_i', v\rangle\right) - \mathbb{E}_{g\sim\mathcal{N}(0,\|v\|^2)}\left[f(g)\right]\right|\right| \\
&= \max_{\|v\|\le 1}\left|\frac{1}{n}\cdot\sum_{i=1}^n \left(f\left(\langle g_i, v\rangle\right) - f\left(\langle g_i', v\rangle\right)\right)\right| \\
&\le \max_{\|v\|\le 1}\frac{1}{n}\cdot\sum_{i=1}^n\left|f\left(\langle g_i, v\rangle\right) - f\left(\langle g_i', v\rangle\right)\right| \\
&\le \max_{\|v\|\le 1}\frac{1}{n}\cdot\sum_{i=1}^n\left|\langle g_i, v\rangle - \langle g_i', v\rangle\right| \le \frac{1}{\sqrt{n}}\cdot\|\{g_i\}_{i=1}^n - \{g_i'\}_{i=1}^n\|.
\end{aligned}$$

The above display establishes that Z is a $n^{-1/2}$-Lipschitz function of $\{g_i\}_{i=1}^n$. Hence, we have:

$$\left|Z(\{g_i\}_{i=1}^n) - \mathbb{E}[Z(\{g_i\}_{i=1}^n)]\right| \le \sqrt{\frac{2\log 2/\delta}{n}}$$

with probability at least $1-\delta$ by concentration of Lipschitz functions of Gaussians (Theorem B.1).

Letting $g_i' \sim \mathcal{N}(0, I)$ and $\sigma_i \sim \{\pm 1\}$ be mutually independent standard normal vectors and Rademacher random variables respectively, we bound the expectation of Z as follows:

$$\begin{aligned}
\mathbb{E}[Z] &= \mathbb{E}\left[\max_{\|v\|\le 1}\left|\frac{1}{n}\cdot\sum_{i=1}^n f\left(\langle g_i, v\rangle\right) - \mathbb{E}_{g\sim\mathcal{N}(0,\|v\|^2)}\left[f(g)\right]\right|\right] \\
&\le \mathbb{E}_{g_i, g_i'}\left[\max_{\|v\|\le 1}\left|\frac{1}{n}\cdot\sum_{i=1}^n f\left(\langle g_i, v\rangle\right) - f\left(\langle g_i', v\rangle\right)\right|\right] \\
&= \mathbb{E}_{g_i, g_i', \sigma_i}\left[\max_{\|v\|\le 1}\left|\frac{1}{n}\cdot\sum_{i=1}^n \sigma_i\left(f\left(\langle g_i, v\rangle\right) - f\left(\langle g_i', v\rangle\right)\right)\right|\right] \\
&\le 2\mathbb{E}_{g_i, \sigma_i}\left[\max_{\|v\|\le 1}\left|\frac{1}{n}\cdot\sum_{i=1}^n \sigma_i f\left(\langle g_i, v\rangle\right)\right|\right] \\
&\le 4\mathbb{E}_{g_i, \sigma_i}\left[\max_{\|v\|\le 1}\left|\frac{1}{n}\cdot\sum_{i=1}^n \sigma_i\langle g_i, v\rangle\right|\right] = 4\mathbb{E}_{g_i}\left[\left\|\frac{1}{n}\cdot\sum_{i=1}^n g_i\right\|\right] \\
&\le 4\sqrt{\frac{d}{n}}
\end{aligned}$$

where the second to last inequality follows from Ledoux-Talagrand contraction [19, Theorem 4.12]. The previous two displays now yield Eq. (1). This succinct argument, unfortunately, breaks down when working with RHTs in the place of Gaussians. While the concentration term can be modified to yield a weaker bound with m in the denominator, the expectation term crucially relies on the mutual independence of all the g_i and g_i' which does not hold true for RHTs.

An alternative approach is to use a standard gridding argument. Consider a γ-net, $\mathcal{G}$, of $\mathbb{B}(0,1)$ for some small γ (Definition B.3). For each $v \in \mathcal{G}$, we have by noting that $f(\langle v, g_i\rangle)$ is $\|v\|$-subGaussian:

$$\left|\frac{1}{n}\sum_{i=1}^n f\left(\langle v, g_i\rangle\right) - \mathbb{E}_{g\sim\mathcal{N}(0,\|v\|^2)}\left[f\left(g\right)\right]\right| \le \sqrt{\frac{2\log 2/\delta'}{n}}$$

with probability at least $1-\delta'$. Setting $\delta' = \delta/|\mathcal{G}|$ and an application of the union bound yield the desired conclusion on the net as $\mathcal{G}$ can be chosen to satisfy $|\mathcal{G}| \le (C\gamma^{-1})^d$ (Corollary B.5). Unfortunately, this simple argument also fails when working with RHTs. To see this, consider the case $v = e_1$. Here, the previous application of Hoeffding's Inequality yields the weaker inequality:

$$\left|\frac{1}{md}\sum_{i=1}^m\sum_{j=1}^d f\left(\widetilde{h}_{i,j}(v)\right) - \mathbb{E}_{g\sim\mathcal{N}(0,\|v\|^2)}\left[f\left(g\right)\right]\right| \le \sqrt{\frac{2\log 2/\delta'}{m}}.$$

A naive union bound would then require $md = \Omega(d^2)$ (whereas our aim is to have md nearly linear in d). This is reminiscent of the situation in compressed sensing, in which a naive union bound provides a similarly weak result [7, 22]. In the next subsection, we present our approach to circumvent this issue in our context (which is not related to the chaining technique used in the compressed sensing context).

Our Approach. The first key observation underlying our approach is that while standard basis vectors lead to sub-optimal tail bounds, a typical vector behaves quite differently. For example, consider the vector $v = \mathbf{1}/\sqrt{d}$. In this case, it is not hard to show that each entry of $\widetilde{h}(v)$ is independent due to the orthogonality of the rows of H_d. Hence, for this particular vector, we obtain with probability at least $1-\delta'$:

$$\left|\frac{1}{md}\sum_{i=1}^m\sum_{j=1}^d f(\widetilde{h}_{i,j}(v)) - \mathbb{E}_{g\sim\mathcal{N}(0,\|v\|^2)}\left[f(g)\right]\right| \le \sqrt{\frac{2\log 2/\delta'}{md}}.$$

Since "most" vectors on the unit sphere are typically closer to v than a standard basis vector, one could hope to apply the stronger inequality for most vectors while treating sparse vectors like those in the standard basis separately. Intuitively, our proof establishes

the following concentration inequality:

$$\left|\frac{1}{md}\sum_{i=1}^{m}\sum_{j=1}^{d} f(\widetilde{h}_{i,j}(v)) - \mathbb{E}_{g\sim\mathcal{N}(0,\|v\|^2)}[f(g)]\right| \le C\sqrt{\frac{\|v\|_\infty^2 \log 1/\delta'}{m}}. \tag{2}$$

Observe that the above inequality interpolates between the settings $v = e_1$ and $v = \mathbf{1}/\sqrt{d}$ depending on how well-spread the input vector is. We now use the inequality to establish our result for a simpler set of vectors.

Consider the following sets

$$\forall S \subseteq [d] : \mathcal{G}_S := \left\{ \|v\| \le 1 : \begin{array}{c} \forall i,j \in S\ \frac{|v_i|}{2} \le |v_j| \le 2|v_i| \text{ and} \\ \forall i \notin S\ v_i = 0 \end{array} \right\}$$

$$\forall k \in [d] : \mathcal{G}_k := \cup_{\substack{S\subset[d]\\|S|=k}} \mathcal{G}_S$$

$$\mathcal{G} := \cup_{k\in[d]} \mathcal{G}_k$$

Hence, $\mathcal{G}_S$ consists of vectors uniformly spread over S and for any $v \in \mathcal{G}_S$, we have $\|v\|_\infty \le 2/\sqrt{|S|}$. We will use Eq. (2) to perform a union bound over $\mathcal{G}$. First, notice that a γ-net of $\mathcal{G}_S$ has size $(C/\gamma)^{|S|}$ and hence, there exists a γ-net of $\mathcal{G}_k$, $\widetilde{\mathcal{G}}_k$ of size at most $(Cd/\gamma)^k$. A union bound over only the elements in $\widetilde{\mathcal{G}}_k$ yields:

$$\forall v \in \widetilde{\mathcal{G}}_k : \left|\frac{1}{md}\sum_{i=1}^{m}\sum_{j=1}^{d} f(\widetilde{h}_{i,j}(v)) - \mathbb{E}_{g\sim\mathcal{N}(0,\|v\|^2)}[f(g)]\right| \le C\sqrt{\frac{\|v\|_\infty^2 \log 1/\delta + \log(d/\gamma)}{m}}$$

with probability at least $1 - \delta/d$. Ignoring discretization errors, this establishes our concentration result for the restricted set $\mathcal{G}$. While $\mathcal{G}$ is quite restricted and this inequality is not strong enough to prove Theorem 1.1, the ideas used in establishing it will play a key part in proving the general result.

Our next key observation is that any $v \in \mathbb{B}(0,1)$ can be well approximated by a linear combination of a small number of vectors from $\mathcal{G}$; that is, $v \approx \sum_{i=1}^{r} v_i$ for $r \approx \log(d/\varepsilon)$ and $v_i \in \mathcal{G}$ with $\|v_i\|_0 \le \|v_{i+1}\|_0$. While the previously established inequalities are strong enough to ensure the conclusion of Theorem 1.1 for the individual components, v_i, this does not ensure that their combination enjoys similar concentration properties and it is not clear how these vectors behave when their embeddings are combined.

The final ingredient in our argument is the stronger conditional inequality for $u = u_1 + u_2$ where $\mathrm{Supp}(u_1) \cap \mathrm{Supp}(u_2) = \phi$ and $U = \{(D^j)_{k,k}\}_{j\in[m],k\in\mathrm{Supp}(u_2)}$:

$$\left|\frac{1}{md}\sum_{i=1}^{m}\sum_{j=1}^{d} f(\widetilde{h}_{i,j}(u)) - \mathbb{E}\left[\frac{1}{md}\sum_{i=1}^{m}\sum_{j=1}^{d} f(\widetilde{h}_{i,j}(u)) \mid U\right]\right| \le C\sqrt{\frac{\|u_1\|_\infty^2 \log 1/\delta}{m}}. \tag{3}$$

The above inequality shows that once we fix the variables in U, the concentration properties of the entries of $\widetilde{h}(u)$ are solely determined by the properties of u_1. This inequality allows us to establish uniformly over v, for all $k \in [r]$:

$$\left|\frac{1}{md}\sum_{i=1}^{m}\sum_{j=1}^{d} f(\widetilde{h}_{i,j}(v^k)) - \mathbb{E}\left[\frac{1}{md}\sum_{i=1}^{m}\sum_{j=1}^{d} f(\widetilde{h}_{i,j}(v^k)) \mid V_{k-1}\right]\right| \le \frac{\varepsilon}{r} \tag{4}$$

where $v^k = \sum_{i=1}^{k} v_i$ and $V_i = \{(D^j)_{k,k}\}_{j\in[m],k\in\mathrm{Supp}(v^i)}$. The final step of our argument involves showing through a careful recursive argument that the conditional expectation is close to its unconditional expectation where we additionally require that Eq. (4) holds not just for the original function, f, but also for offset versions of the function, f_t, defined as $f_t(x) = f(x - t)$ for a large range of t. We prove Eq. (3) and carry out this argument in full detail in Section 3.

3 PROOF OF UNIFORM LIPSCHITZ CONCENTRATION

In this section, we formally prove Theorem 1.1 by expanding on the outline presented in Section 2. We begin by defining the class of functions for which our concentration properties will hold:

$$\forall S \subseteq [d] : V_S := \bigcup_{j=1}^{m} \left\{D^j_{i,i}\right\}_{i\in S} \tag{LIP-NOT}$$

$$\forall S \subset [d], t \in \mathbb{R}, z \in \mathbb{R}^d : F_{S,t}(z) := \frac{1}{md}\cdot\sum_{j=1}^{m}\sum_{k=1}^{d} f\left(\widetilde{h}_{j,k}(z_S) - t\right)$$

$$\forall S,T \subseteq [d], t \in \mathbb{R}, z \in \mathbb{R}^d : \widetilde{F}_{S,T,t}(z) := \mathbb{E}\left[F_{S\cup T,t}(z) \mid V_T\right]$$

To help define the nets used in our argument, we introduce the following notation and define:

$$\rho := \left(\frac{\varepsilon}{10d}\right)^3,\ \lambda := \left(10\sqrt{\log(d/\varepsilon)}\right)^6,$$

$$\gamma := \left(10\sqrt{\log(d/\varepsilon)}\right)^3,\ \nu := \left(\frac{\varepsilon}{256\log(d/\varepsilon)}\right)$$

$$\forall S,T \subseteq [d] : \mathcal{G}_{S,T} := \left\{ \|z\| \le 1 : \begin{array}{c} \forall i \notin S\cup T, z_i = 0 \text{ and} \\ \forall i,j \in S, \frac{1}{2}|z_j| \le |z_i| \le 2|z_j| \end{array} \right\}$$

$$r := 32\log(d/\varepsilon),\quad \zeta := \left(\frac{\varepsilon}{10d}\right)^3,\quad \mathcal{T} := \{\pm i\cdot\zeta\}_{i=0}^{\lambda/\zeta} \tag{LIP-NET}$$

For all disjoint $S, T \subseteq [d]$ such that $|T| \le r\cdot|S|$, let $\widetilde{\mathcal{G}}_{S,T}$ be an ρ-net of $\mathcal{G}_{S,T}$ (Definition B.3). Note, we may assume that $|\widetilde{\mathcal{G}}_{S,T}| \le \left(\frac{10}{\rho}\right)^{(r+1)\cdot|S|}$ (Corollary B.5). We now have the following claim:

Claim 3.1. We have $\forall t \in \mathcal{T}, S, T$ s.t $|T| \le r\cdot|S|, T\cap S = \phi, z \in \widetilde{\mathcal{G}}_{S,T}$:

$$|F_{S\cup T,t}(z) - \widetilde{F}_{S,T,t}(z)| \le \nu$$

with probability at least $1 - \delta/4$.

Proof. For the proof, we start by conditioning on V_T. Note that for all $i \in S$, we must have:

$$|z_i| \le \frac{2}{\sqrt{|S|}}.$$

Let V_S^1 and V_S^2 be two distinct settings of the random variables V_S and let $F_{S,t}^1(x), F_{S,t}^2(x)$ be the values of $F_{S,t}$ computed by setting the variables in V_S to V_S^1 and V_S^2 respectively fixing all the rest to be the

same. Similarly, let $\widetilde{h}^1(\cdot), \widetilde{h}^2(\cdot)$ denote the vectors $\widetilde{h}$ computed with the corresponding settings of V_S and $D^{j,1}, D^{j,2}$ be the corresponding diagonal matrices for $j \in [m]$. Recalling that $f(\cdot)$ is a 1-Lipschitz function, we have:

$$
\begin{aligned}
&|F^1_{S\cup T,t}(z) - F^2_{S\cup T,t}(z)| \\
&\le \frac{1}{md} \cdot \sum_{j=1}^{m}\sum_{k=1}^{d} |\widetilde{h}^1_{j,k}(z) - \widetilde{h}^2_{j,k}(z)| \\
&\le \frac{1}{\sqrt{md}} \cdot \sqrt{\sum_{j=1}^{m}\sum_{k=1}^{d} (\widetilde{h}^1_{j,k}(z) - \widetilde{h}^2_{j,k}(z))^2} \\
&= \frac{1}{\sqrt{md}} \cdot \left\| \begin{bmatrix} H(D^{1,1} - D^{1,2})z \\ H(D^{2,1} - D^{2,2})z \\ \vdots \\ H(D^{m,1} - D^{m,2})z \end{bmatrix} \right\| \\
&= \frac{1}{\sqrt{md}} \cdot \left\| \begin{bmatrix} H & 0 & \dots & 0 \\ 0 & H & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & H \end{bmatrix} \cdot \begin{bmatrix} (D^{1,1} - D^{1,2})z \\ (D^{2,1} - D^{2,2})z \\ \vdots \\ (D^{m,1} - D^{m,2})z \end{bmatrix} \right\| \\
&= \frac{1}{\sqrt{m}} \cdot \left\| \begin{bmatrix} (D^{1,1} - D^{1,2})z \\ (D^{2,1} - D^{2,2})z \\ \vdots \\ (D^{m,1} - D^{m,2})z \end{bmatrix} \right\| \\
&= \frac{1}{\sqrt{m}} \cdot \left\| \begin{bmatrix} \mathrm{Diag}(z)\,\mathrm{diag}(D^{1,1} - D^{1,2}) \\ \mathrm{Diag}(z)\,\mathrm{diag}(D^{2,1} - D^{2,2}) \\ \vdots \\ \mathrm{Diag}(z)\,\mathrm{diag}(D^{m,1} - D^{m,2}) \end{bmatrix} \right\| \le \frac{2}{\sqrt{m|S|}} \cdot \|V^1_S - V^2_S\|
\end{aligned}
$$

Therefore, $F_{S\cup T,t}(z)$ is a $\frac{2}{\sqrt{m|S|}}$-Lipschitz function of V_S conditioned on V_T. Hence, we have by Theorem B.1:

$$
\begin{aligned}
&\mathbb{P}\left\{|F_{S\cup T,t}(z) - \widetilde{F}_{S,T,t}(z)| \le \nu\right\} \\
&\qquad \ge 1 - \frac{\delta}{32 \cdot (10/\rho)^{(r+1)\cdot|S|} \cdot (d+1)^{(r+1)\cdot|S|} \cdot |\mathcal{T}| \cdot d^2}.
\end{aligned}
$$

The claim now follows from a union bound over all possible S, T satisfying the constraints. □

From this point on, we condition on the conclusions of Lemma B.6 and Claim 3.1; i.e we condition on the following event which occurs with probability at least $1 - \delta/2$ via Claim 3.1 and Lemma B.6:

$$
\begin{aligned}
&\forall t \in \mathcal{T}, S, T \text{ s.t } |T| \le r \cdot |S|, T \cap S = \phi, z \in \widetilde{\mathcal{G}}_{S,T} : \\
&\qquad |F_{S\cup T,t}(z) - \widetilde{F}_{S,T,t}(z)| \le \nu \\
&\forall x, y \in \mathbb{R}^d : \frac{\left\|\widetilde{h}(x) - \widetilde{h}(y)\right\|}{\sqrt{md}} \le 2\|x - y\|
\end{aligned}
$$

We now extend the conclusion of Claim 3.1 to all $z \in \mathcal{G}_{S,T}$.

Claim 3.2. We have $\forall t \in \mathcal{T}, |T| \le r \cdot |S|, z \in \mathcal{G}_{S,T},\ T \cap S = \phi$:

$$
|F_{S\cup T,t}(z) - \widetilde{F}_{S,T,t}(z)| \le 2\nu.
$$

Proof. Let $z \in \mathcal{G}_{S,T}, t \in \mathcal{T}$ for S, T satisfying the constraints and $\widetilde{z} = \arg\min_{u \in \widetilde{\mathcal{G}}_{S,T}} \|z - u\|$. Note that $\|z - \widetilde{z}\| \le \rho$. We simply show that $F_{S\cup T,t}, \widetilde{F}_{S,T,t}$ are close to their corresponding values for $\widetilde{z}$. For the first term (i.e for $F_{S\cup T,t}$), we have:

$$
\begin{aligned}
&|F_{S\cup T,t}(z) - F_{S\cup T,t}(\widetilde{z})| \\
&= \frac{1}{md} \cdot \left| \sum_{j=1}^{m}\sum_{k=1}^{d} f(\widetilde{h}_{j,k}(z_{S\cup T}) - t) - f(\widetilde{h}_{j,k}(\widetilde{z}_{S\cup T}) - t) \right| \\
&\le \frac{1}{md} \cdot \sum_{j=1}^{m}\sum_{k=1}^{d} \left| f(\widetilde{h}_{j,k}(z_{S\cup T}) - t) - f(\widetilde{h}_{j,k}(\widetilde{z}_{S\cup T}) - t) \right| \\
&\le \frac{1}{md} \cdot \sum_{j=1}^{m}\sum_{k=1}^{d} \left| \widetilde{h}_{j,k}(z_{S\cup T}) - \widetilde{h}_{j,k}(\widetilde{z}_{S\cup T}) \right| \\
&\le \frac{1}{md} \cdot \sqrt{md} \cdot \sqrt{\sum_{j=1}^{m}\sum_{k=1}^{d} \left(\widetilde{h}_{j,k}(z_{S\cup T}) - \widetilde{h}_{j,k}(\widetilde{z}_{S\cup T}) \right)^2} \\
&\le \frac{1}{\sqrt{md}} \cdot 2\sqrt{md} \cdot \rho \le \frac{\nu}{4}
\end{aligned}
$$

concluding the proof for the first term. For the second term (i.e $\widetilde{F}_{S,T,t}$), we proceed as follows:

$$
\begin{aligned}
&\left|\widetilde{F}_{S,T,t}(z) - \widetilde{F}_{S,T,t}(\widetilde{z})\right| \\
&= \left|\mathbb{E}\left[F_{S\cup T,t}(z) - F_{S\cup T,t}(\widetilde{z}) \mid V_T\right]\right| \\
&\le \frac{1}{md} \cdot \sum_{j=1}^{m}\sum_{k=1}^{d} \mathbb{E}\left[|f(\widetilde{h}_{j,k}(z_{S\cup T}) - t) - f(\widetilde{h}_{j,k}(\widetilde{z}_{S\cup T}) - t)| \mid V_T \right] \\
&\le \frac{1}{md} \cdot \sum_{j=1}^{m}\sum_{k=1}^{d} \mathbb{E}\left[|\widetilde{h}_{j,k}(z_{S\cup T}) - \widetilde{h}_{j,k}(\widetilde{z}_{S\cup T})| \mid V_T \right] \\
&\le \frac{1}{md} \cdot \sum_{j=1}^{m}\sum_{k=1}^{d} \mathbb{E}\left[|\widetilde{h}_{j,k}(z_S) - \widetilde{h}_{j,k}(\widetilde{z}_S)| \mid V_T \right] + \\
&\qquad \frac{1}{md} \cdot \sum_{j=1}^{m}\sum_{k=1}^{d} \mathbb{E}\left[|\widetilde{h}_{j,k}(z_T) - \widetilde{h}_{j,k}(\widetilde{z}_T)| \mid V_T \right] \\
&\le \frac{1}{md} \cdot \left(\sum_{j=1}^{m}\sum_{k=1}^{d} \rho \cdot \mathbb{E}_{g\sim\mathcal{N}(0,1)}\left[|g|\right] + |\widetilde{h}_{j,k}(z_T) - \widetilde{h}_{j,k}(\widetilde{z}_T)| \right) \\
&\le \rho\mathbb{E}_{g\sim\mathcal{N}(0,1)}\left[|g|\right] + \frac{1}{md} \cdot \sqrt{md} \cdot \sqrt{\sum_{j=1}^{m}\sum_{k=1}^{d} (\widetilde{h}_{j,k}(z_T) - \widetilde{h}_{j,k}(\widetilde{z}_T))^2} \\
&\le \frac{\nu}{4} + \frac{1}{\sqrt{md}} \cdot 2\sqrt{md} \cdot \rho \le \frac{\nu}{2}.
\end{aligned}
$$

The previous two bounds along with the conclusion of Claim 3.1 yield the claim. □

Let $z \in \mathbb{R}^d$ with $\|z\| \le 1$. We decompose the coordinates of z into disjoint sets defined as follows:

$$
\tau^* := \max_{k\in[d]} |z_k|, \quad \forall l \in [r] : S_l := \left\{ k : \frac{1}{2^l} \cdot \tau^* < |z_k| \le \frac{1}{2^{l-1}} \cdot \tau^* \right\}.
$$

Let $S_{(1)}, \ldots, S_{(r)}$ be an ordering of the S_l such that $|S_{(i)}| \leq |S_{(i+1)}|$ for all $i \in [r-1]$ and define the sets S^l:

$$\forall l \in [r] : S^l \coloneqq \cup_{q=1}^{l} S_{(q)} \text{ and } T \coloneqq [d] \setminus S^r.$$

We now show that as far as the functions F are concerned, z is well approximated by z_{S^r}.

Claim 3.3. We have for all $t \in \mathbb{R}$:

$$\left|\mathbb{E}\left[F_{[d],t}(z) - F_{S^r,t}(z_{S^r})\right]\right| \leq \frac{\nu}{4}$$
$$\left|F_{[d],t}(z) - F_{S^r,t}(z_{S^r})\right| \leq \frac{\nu}{4}.$$

Proof. We make the following simple observation:

$$\|z_T\| \leq \sqrt{d} \cdot \frac{1}{2^r} \leq \rho. \tag{5}$$

From Eq. (5), we have:

$$\begin{aligned}
&\left|\mathbb{E}\left[F_{[d],t}(z) - F_{S^r,t}(z_{S^r})\right]\right| \\
&= \left|\mathbb{E}\left[F_{[d],t}(z) - F_{[d],t}(z_{S^r})\right]\right| \\
&\leq \frac{1}{md} \cdot \sum_{j=1}^{m}\sum_{k=1}^{d} \mathbb{E}\left[\left|f(\widetilde{h}_{j,k}(z) - t) - f(\widetilde{h}_{j,k}(z_{S^r}) - t)\right|\right] \\
&\leq \frac{1}{md} \cdot \sum_{j=1}^{m}\sum_{k=1}^{d} \mathbb{E}\left[\left|\widetilde{h}_{j,k}(z) - \widetilde{h}_{j,k}(z_{S^r})\right|\right] \leq \rho \cdot \mathbb{E}_{g\sim\mathcal{N}(0,1)}[|g|] \leq \frac{\nu}{4}.
\end{aligned}$$

Similarly, we have:

$$\begin{aligned}
&\left|F_{[d],t}(z) - F_{S^r,t}(z_{S^r})\right| \\
&= \left|F_{[d],t}(z) - F_{[d],t}(z_{S^r})\right| \\
&\leq \frac{1}{md} \cdot \sum_{j=1}^{m}\sum_{k=1}^{d}\left|f(\widetilde{h}_{j,k}(z) - t) - f(\widetilde{h}_{j,k}(z_{S^r}) - t)\right| \\
&\leq \frac{1}{md} \cdot \sum_{j=1}^{m}\sum_{k=1}^{d}\left|\widetilde{h}_{j,k}(z) - \widetilde{h}_{j,k}(z_{S^r})\right| \\
&\leq \frac{1}{md} \cdot \sqrt{md} \cdot \sqrt{\sum_{j=1}^{m}\sum_{k=1}^{d}(\widetilde{h}_{j,k}(z) - \widetilde{h}_{j,k}(z_{S^r}))^2} \\
&\leq \frac{1}{\sqrt{md}} \cdot 2\sqrt{md} \cdot \rho \leq \frac{\nu}{4}
\end{aligned}$$

concluding the proof of the claim. □

In our final claim, we show that $F_{[d],t}(z_{S^r})$ is close to its expectation.

Claim 3.4. We have:

$$\forall l \in [r], t \in \mathbb{R} \text{ s.t } |t| \leq \lambda - (l-1)\gamma : \left|F_{S^l,t}(z_{S^l}) - \mathbb{E}\left[F_{S^l,t}(z_{S^l})\right]\right| \leq 3l\nu.$$

and consequently, from Claim 3.3:

$$\forall t \in \mathbb{R} \text{ s.t } |t| \leq \lambda - r\gamma : \left|F_{[d],t}(z) - \mathbb{E}\left[F_{[d],t}(z)\right]\right| \leq 3(r+1)\nu.$$

Proof. We prove the claim inductively and start with the base case of the induction.

Base case: Claim 3.2 establishes the base case for all $t \in \mathcal{T}$. Now, for any $|t| \leq \lambda$, there exists $t' \in \mathcal{T}$ with $|t' - t| \leq \zeta$. For this t', we have by the triangle inequality and the fact that F is 1-Lipschitz in t:

$$\begin{aligned}
&\left|F_{S^1,t}(z_{S^1}) - \mathbb{E}\left[F_{S^1,t}(z_{S^1})\right]\right| \\
&\leq \left|F_{S^1,t'}(z_{S^1}) - \mathbb{E}\left[F_{S^1,t'}(z_{S^1})\right]\right| + \\
&\quad \left|F_{S^1,t}(z_{S^1}) - F_{S^1,t'}(z_{S^1}) - \mathbb{E}\left[F_{S^1,t}(z_{S^1}) - F_{S^1,t'}(z_{S^1})\right]\right| \\
&\leq \left|F_{S^1,t'}(z_{S^1}) - \mathbb{E}\left[F_{S^1,t'}(z_{S^1})\right]\right| + \\
&\quad \left|F_{S^1,t}(z_{S^1}) - F_{S^1,t'}(z_{S^1})\right| + \left|\mathbb{E}\left[F_{S^1,t}(z_{S^1}) - F_{S^1,t'}(z_{S^1})\right]\right| \\
&\leq \left|F_{S^1,t'}(z_{S^1}) - \mathbb{E}\left[F_{S^1,t'}(z_{S^1})\right]\right| + \\
&\quad \left|F_{S^1,t}(z_{S^1}) - F_{S^1,t'}(z_{S^1})\right| + \mathbb{E}\left[\left|F_{S^1,t}(z_{S^1}) - F_{S^1,t'}(z_{S^1})\right|\right] \\
&\leq 2\nu + 2\zeta \leq 3\nu
\end{aligned}$$

establishing the base case of the induction.

Inductive case: For the induction step, we proceed similarly to the base case with one additional step. Assuming the induction up to $l = q$, we establish it for $l = q+1$. First, we show that $\widetilde{F}_{S_{(q+1)},S^q,t}(z_{S^{q+1}})$ is close to $\mathbb{E}\left[F_{S^{q+1},t}(z_{S^{q+1}})\right]$ for all $|t| \leq \lambda - q\gamma$. We proceed as follows:

$$\begin{aligned}
&\left|\widetilde{F}_{S_{(q+1)},S^q,t}(z_{S^{q+1}}) - \mathbb{E}\left[F_{S^{q+1},t}(z_{S^{q+1}})\right]\right| \\
&= \left|\frac{1}{md} \cdot \sum_{j=1}^{m}\sum_{k=1}^{d} \mathbb{E}\left[f(\widetilde{h}_{j,k}(z_{S^{q+1}}) - t) \mid V_{S^q}\right] - \mathbb{E}\left[F_{S^{q+1},t}(z_{S^{q+1}})\right]\right| \\
&= \left|\frac{1}{md} \cdot \sum_{j=1}^{m}\sum_{k=1}^{d} \mathbb{E}_{g\sim\mathcal{N}(0,1)}\left[f(\|z_{S_{(q+1)}}\| \cdot g + \widetilde{h}_{j,k}(z_{S^q}) - t) \mid V_{S^q}\right]\right. \\
&\qquad \left. - \mathbb{E}\left[F_{S^{q+1},t}(z_{S^{q+1}})\right]\right| \\
&= \left|\frac{1}{md} \cdot \sum_{j=1}^{m}\sum_{k=1}^{d} \int_{-\infty}^{\infty} f(\|z_{S_{(q+1)}}\| \cdot y + \widetilde{h}_{j,k}(z_{S^q}) - t)\phi(y)dy\right. \\
&\qquad \left. - \mathbb{E}\left[F_{S^{q+1},t}(z_{S^{q+1}})\right]\right| \\
&= \left|\int_{-\infty}^{\infty} F_{S^q,t-\left\|z_{S_{(q+1)}}\right\|\cdot y}(z_{S^q})\phi(y)dy - \right. \\
&\qquad \left. \mathbb{E}_{y_1,y_2\sim\mathcal{N}(0,1)}\left[f(\|z_{S_{(q+1)}}\| \cdot y_1 + \|z_{S^q}\| \cdot y_2 - t)\right]\right| \\
&= \left|\int_{-\infty}^{\infty} F_{S^q,t-\left\|z_{S_{(q+1)}}\right\|\cdot y}(z_{S^q})\phi(y)dy - \right. \\
&\qquad \left. \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\|z_{S^q}\| \cdot y_1 + \|z_{S_{(q+1)}}\| \cdot y_2 - t)\phi(y_1)\phi(y_2)dy_1dy_2\right| \\
&\leq \int_{-\infty}^{\infty}\left|F_{S^q,t-\left\|z_{S_{(q+1)}}\right\|\cdot y}(z_{S^q}) - \mathbb{E}\left[F_{S^q,t-\left\|z_{S_{(q+1)}}\right\|\cdot y}(z_{S^q})\right]\right|\phi(y)dy \\
&= \underbrace{\int_{-\gamma}^{\gamma}\left|F_{S^q,t-\left\|z_{S_{(q+1)}}\right\|\cdot y}(z_{S^q}) - \mathbb{E}\left[F_{S^q,t-\left\|z_{S_{(q+1)}}\right\|\cdot y}(z_{S^q})\right]\right|\phi(y)dy}_{\alpha} +
\end{aligned}$$

$$\underbrace{\int_{\mathbb{R}\setminus[-\gamma,\gamma]}\left|F_{S^q,t-\left\|z_{S_{(q+1)}}\right\|y}(z_{S^q})-\mathbb{E}\left[F_{S^q,t-\left\|z_{S_{(q+1)}}\right\|y}(z_{S^q})\right]\right|\phi(y)dy}_{\beta} \tag{6}$$

For the first term, α, we have by the inductive hypothesis as $|t - \|z_{S_{(q+1)}}\|y| \le \lambda - (q-1)\gamma$ when $|y| \le \gamma$:

$$\alpha \le 3q\nu.$$

For the second term, β, we proceed as follows noting $F_{S,t}(z)$ is 1-Lipschitz in t for all $z \in \mathbb{R}^d$:

$$\begin{aligned}
&\int_{\mathbb{R}\setminus[-\gamma,\gamma]}\left|F_{S^q,t-\left\|z_{S_{(q+1)}}\right\|y}(z_{S^q})-\mathbb{E}\left[F_{S^q,t-\left\|z_{S_{(q+1)}}\right\|y}(z_{S^q})\right]\right|\phi(y)dy\\
&\le \int_{\mathbb{R}\setminus[-\gamma,\gamma]}\left|F_{S^q,0}(z_{S^q})-\mathbb{E}\left[F_{S^q,0}(z_{S^q})\right]\right|\phi(y)dy+\\
&\quad\int_{\mathbb{R}\setminus[-\gamma,\gamma]}\left|F_{S^q,t-\left\|z_{S_{(q+1)}}\right\|\cdot y}(z_{S^q})-F_{S^q,0}(z_{S^q})-\right.\phi(y)dy\\
&\qquad\left.\mathbb{E}\left[F_{S^q,t-\left\|z_{S_{(q+1)}}\right\|\cdot y}(z_{S^q})-F_{S^q,0}(z_{S^q})\right]\right|\phi(y)dy\\
&\le \int_{\mathbb{R}\setminus[-\gamma,\gamma]}\left(3q\nu+2\left|t-\|z_{S_{(q+1)}}\|\cdot y\right|\right)\phi(y)dy\\
&\le 3\int_{\mathbb{R}\setminus[-\gamma,\gamma]}\left(q\nu+|t|+\|z_{S_{(q+1)}}\|\cdot|y|\right)\phi(y)dy\\
&\le 6\int_{\mathbb{R}\setminus[-\gamma,\gamma]}(\lambda+|y|)\,\phi(y)dy \le \frac{\nu}{2}
\end{aligned}$$

where the last inequality follows the setting of γ, λ, ν (LIP-NET). Putting the previous two bounds together:

$$\left|\widetilde{F}_{S_{(q+1)},S^q,t}(z_{S^{q+1}})-\mathbb{E}\left[F_{S^{q+1},t}(z_{S^{q+1}})\right]\right| \le 3q\nu+\frac{\nu}{2}.$$

Now, as in the base case, we simply bound the deviations of F from its expectation. Claim 3.2 now yields:

$$\forall t\in\mathcal{T}:\left|F_{S^{q+1},t}(z_{S^{q+1}})-\mathbb{E}\left[F_{S^{q+1},t}(z_{S^{q+1}})\right]\right| \le \frac{5\nu}{2}+3q\nu.$$

Similarly to the base case, the previous display establishes the inductive hypothesis for all $t \in \mathcal{T}$. For any t such that $|t| \le \lambda - q\gamma$, there exists $t' \in \mathcal{T}$ with $|t - t'| \le \zeta$. Then, we have:

$$\begin{aligned}
&\left|F_{S^{q+1},t}(z_{S^{q+1}})-\mathbb{E}\left[F_{S^{q+1},t}(z_{S^{q+1}})\right]\right|\\
&\le\left|F_{S^{q+1},t'}(z_{S^{q+1}})-\mathbb{E}\left[F_{S^{q+1},t'}(z_{S^{q+1}})\right]\right|+\\
&\quad\left|F_{S^{q+1},t}(z_{S^{q+1}})-F_{S^{q+1},t'}(z_{S^{q+1}})-\right.\\
&\qquad\left.\mathbb{E}\left[F_{S^{q+1},t}(z_{S^{q+1}})-F_{S^{q+1},t'}(z_{S^{q+1}})\right]\right|\\
&\le 3q\nu+\frac{5\nu}{2}+\left|F_{S^{q+1},t}(z_{S^{q+1}})-F_{S^{q+1},t'}(z_{S^{q+1}})\right|+\\
&\left|\mathbb{E}\left[F_{S^{q+1},t}(z_{S^{q+1}})-F_{S^{q+1},t'}(z_{S^{q+1}})\right]\right|\\
&\le 3q\nu+\frac{5\nu}{2}+\left|F_{S^{q+1},t}(z_{S^{q+1}})-F_{S^{q+1},t'}(z_{S^{q+1}})\right|+\\
&\mathbb{E}\left[\left|F_{S^{q+1},t}(z_{S^{q+1}})-F_{S^{q+1},t'}(z_{S^{q+1}})\right|\right]\\
&\le 3q\nu+\frac{5\nu}{2}+2\zeta \le 3(q+1)\nu
\end{aligned}$$

establishing the hypothesis for all $|t| \le \lambda - q\gamma$. The final statement of the claim now follows by an application of Claim 3.3 along with the above inductive hypothesis. □

Theorem 1.1 now follows from Claim 3.4 along with a union bound over Claim 3.1 and Lemma B.6. □

4 KERNEL APPROXIMATION PROOF

In this section, we prove Theorem 1.2 leveraging Theorem 1.1. We start with a simple algebraic manipulation. For all $x, y \in \mathbb{R}^d$, we have by standard trigonometric identities:

$$\begin{aligned}
&\langle h(x), h(y)\rangle\\
&=\frac{2}{md}\cdot\sum_{j=1}^{m}\sum_{k=1}^{d}\cos\left((HD^jx)_k+b_k^j\right)\cos\left((HD^jy)_k+b_k^j\right)\\
&=\frac{1}{md}\sum_{j=1}^{m}\sum_{k=1}^{d}\cos\left((HD^j(x+y))_k+2b_k^j\right)+\cos\left((HD^j(x-y))_k\right).
\end{aligned} \tag{KER-DEC}$$

We will show that the first term is uniformly close to 0 for all $x, y \in \mathcal{W}$. This fact follows straightforwardly by using the fact that the b_k^js are independent of the D^js. Consequently, our efforts will primarily be focussed on the second term. The following simple lemma shows that the first term in Eq. (KER-DEC) is uniformly close to 0 for all $x, y \in \mathcal{W}$.

Lemma 4.1. *For $m \ge \widetilde{\Omega}\left(\varepsilon^{-2}\mathrm{Diam}(\mathcal{W})^2\log 1/\delta\right)$, we have that:*

$$\forall x,y\in\mathcal{W}:\frac{1}{md}\cdot\left|\sum_{j=1}^{m}\sum_{k=1}^{d}\cos\left(\widetilde{h}_{j,k}(x+y)+2b_k^j\right)\right| \le \frac{\varepsilon}{8}$$

with probability at least $1-\delta/2$.

Proof. We have from Lemma B.6 that with probability at least $1-\delta/8$:

$$\forall x,y\in\mathbb{R}^d:\left\|\widetilde{h}(x)-\widetilde{h}(y)\right\| \le 2\sqrt{md}\cdot\|x-y\|.$$

Let $\mathcal{G}$ be a ρ-net of $\mathcal{H} = \{x + y : x, y \in \mathcal{W}\}$, with $\rho = \left(\frac{\varepsilon}{10\cdot d}\right)^{10}$. Note we may assume $|\mathcal{G}| \le \left(\frac{20\cdot\mathrm{Diam}(\mathcal{W})}{\rho}\right)^d$. For any $z \in \mathcal{G}$, we have from the independence of the b_k^j from the D^j and Hoeffding's Inequality:

$$\mathbb{P}\left\{\left|\frac{1}{md}\cdot\sum_{j=1}^{m}\sum_{k=1}^{d}\cos\left(\widetilde{h}_{j,k}(z)+2b_k^j\right)\right| \ge t\right\} \le 2\exp\left\{-\frac{mdt^2}{2}\right\}.$$

Setting $t = \varepsilon/16$ and a union bound over all $z \in \mathcal{G}$ yields that with probability at least $1-\delta/8$:

$$\forall z\in\mathcal{G}:\left|\frac{1}{md}\cdot\sum_{j=1}^{m}\sum_{k=1}^{d}\cos\left(\widetilde{h}_{j,k}(z)+2b_k^j\right)\right| \le \frac{\varepsilon}{16}.$$

Now let $z \in \mathcal{H}$ with $\widetilde{z} = \arg\min_{w\in\mathcal{G}}\|z-w\|$. Now, we have from the fact that $\cos(\cdot)$ is 1-Lipschitz:

$$\left|\frac{1}{md}\cdot\sum_{j=1}^{m}\sum_{k=1}^{d}\cos\left(\widetilde{h}_{j,k}(z)+2b_k^j\right)-\cos\left(\widetilde{h}_{j,k}(\widetilde{z})+2b_k^j\right)\right|$$

$$\leq \frac{1}{md}\cdot\sum_{j=1}^{m}\sum_{k=1}^{d}\left|\cos\left(\widetilde{h}_{j,k}(z)+2b_k^j\right)-\cos\left(\widetilde{h}_{j,k}(\widetilde{z})+2b_k^j\right)\right|$$
$$\leq \frac{1}{md}\cdot\sum_{j=1}^{m}\sum_{k=1}^{d}\left|\widetilde{h}_{j,k}(z)-\widetilde{h}_{j,k}(\widetilde{z})\right|$$
$$\leq \frac{1}{\sqrt{md}}\cdot\left(\sum_{j=1}^{m}\sum_{k=1}^{d}\left(\widetilde{h}_{j,k}(z)-\widetilde{h}_{j,k}(\widetilde{z})\right)^2\right)^{1/2}$$
$$\leq \frac{2}{\sqrt{md}}\cdot 2\sqrt{md}\cdot\|z-\widetilde{z}\|\leq\frac{\varepsilon}{16}.$$

The previous two displays yield the conclusion of the lemma by a triangle inequality. □

We now prove a lemma which shows that the second term in Eq. (KER-DEC) is close to its expectation.

Lemma 4.2. *We have* $\forall x,y\in\mathcal{W}$*:*

$$\left|\frac{1}{md}\cdot\sum_{j=1}^{m}\sum_{k=1}^{d}\cos\left(\widetilde{h}_{j,k}(x-y)\right)-\mathbb{E}_{Z\sim\mathcal{N}(0,\|x-y\|^2)}\left[\cos(Z)\right]\right|\leq\frac{\varepsilon}{2}$$

with probability at least $1-\delta/2$ *when* $m\geq\widetilde{\Omega}\left(\varepsilon^{-2}\mathrm{Diam}(\mathcal{W})^2\log 1/\delta\right)$.

Proof. Let $f(x)=\cos(2\mathrm{Diam}(\mathcal{W})\cdot x)$. Note that f is a $2\mathrm{Diam}(\mathcal{W})$-Lipschitz function and hence, we get from Theorem 1.1 and our setting of m that with probability at least $1-\delta/2$:

$$\forall\|z\|\leq 1:\left|\frac{1}{md}\cdot\sum_{j=1}^{m}\sum_{k=1}^{d}f(\widetilde{h}_{j,k}(z))-\mathbb{E}_{Z\sim\mathcal{N}(0,\|z\|^2)}\left[f(Z)\right]\right|\leq\frac{\varepsilon}{2}.$$

The lemma follows from the fact that for all $x,y\in\mathcal{W}$, $\|x-y\|\leq 2\mathrm{Diam}(\mathcal{W})$. □

Theorem 1.2 follows from Lemmas 4.1 and 4.2 and noting:

$$\mathbb{E}_{Z\sim\mathcal{N}(0,\sigma^2)}[\cos Z]=1+\sum_{k=1}^{\infty}\frac{(2k-1)!!}{(2k)!}\cdot\sigma^{2k}$$
$$=1+\sum_{k=1}^{\infty}\frac{1}{k!}\cdot\left(\frac{\sigma^2}{2}\right)^k=\exp\left\{-\frac{\sigma^2}{2}\right\}.$$

□

5 DISTANCE ESTIMATION

In this section, we prove Theorem 1.4. First, in Subsection 5.1, we describe the data structure achieving the guarantees of Theorem 1.4 and then prove its correctness in Subsection 5.2.

5.1 Algorithm

The pseudocode for our algorithm for distance estimation is defined in Algorithms 1 to 3 with Algorithm 1 instantiating the data structure with 0 points in the dataset by only initializing the RHT. Algorithms 2 and 3 then outline the query and update procedures where

$$\forall r>0:\psi_r(x):=\min(|x|,r).$$

Our query procedure is quite simple: we simply draw a small $\widetilde{O}(1)$ many random coordinates from $[md]$, $\{l_j\}_{j=1}^{k}$ and output the α-quantile corresponding the entries $\{(y_{l_j}-(y_i)_{l_j})\}_{j\in[k]}$. If the distribution of the entries of $y-y_i$ were exactly Gaussian, the returned value would be exactly $\|q-x_i\|$. In Subsection 5.2, we simply invoke Theorem 1.1 and bound the incurred errors.

Algorithm 1 Produce Distance Estimation Data Structure

Input: Accuracy ε, Failure probability δ
$m\leftarrow\widetilde{\Omega}(\varepsilon^{-2}\log 1/\delta)$
Let $\widetilde{h}$ be RHTs with m blocks
Return: $\left(\widetilde{h},\phi\right)$

Algorithm 2 Produce Distance Estimates

Input: Data structure with point set $\left(\widetilde{h},\{y_i\}_{i=1}^{n}\right)$, Data point $x\in\mathbb{R}^d$, Failure probability δ
$y\leftarrow\widetilde{h}(x)$
Let $l_1,\ldots,l_k\overset{iid}{\sim}\mathrm{Unif}([md])$ for $k=\widetilde{\Omega}(\varepsilon^{-2}\log 1/\delta)$
$\alpha\leftarrow\Phi(3)$
$r_i\leftarrow 2\sqrt{\log 1/\varepsilon}\cdot\mathrm{Quant}_\alpha\left(\left\{y_{l_p}-(y_i)_{l_p}\right\}_{p\in[k]}\right)$
$d_i\leftarrow\frac{1}{k}\cdot\sqrt{\frac{\pi}{2}}\cdot\sum_{p=1}^{k}\psi_{r_i}(y_{l_p}-(y_i)_{l_p})$
Return: $\{d_i\}_{i=1}^{n}$

Algorithm 3 Update Distance Estimation Data Structure

Input: Data structure $\left(\widetilde{h},\{y_i\}_{i=1}^{n}\right)$, Data point $x_{n+1}\in\mathbb{R}^d$
$y_{n+1}=\widetilde{h}(x_{n+1})$
Return: $\left\{\widetilde{h},\{y_i\}_{i=1}^{n+1}\right\}$

5.2 Proof of Theorem 1.4

The proof of Theorem 1.4 will rely a sequence of applications of Theorem 1.1 applied to appropriately chosen Lipschitz functions outlined in the following claims.

Claim 5.1. Letting $\beta=\phi(4)$, we have:

$$\forall z\text{ s.t }\|z\|=1:2\leq\mathrm{Quant}_{\alpha-\beta/4}\left(\left\{\widetilde{h}_{p,q}(z)\right\}_{p\in[m],q\in[d]}\right)$$
$$\leq\mathrm{Quant}_{\alpha+\beta/4}\left(\left\{\widetilde{h}_{p,q}(z)\right\}_{p\in[m],q\in[d]}\right)\leq 4$$

with probability at least $1-\delta/4$.

Proof. We first define the functions:

$$f(u):=\begin{cases}1 & u\leq 2\\ (3-u) & 2\leq u\leq 3\\ 0 & \text{otherwise}\end{cases}\text{ and }g(x):=\begin{cases}1 & u\leq 3\\ (4-u) & 3\leq u\leq 4\\ 0 & \text{otherwise}\end{cases}.$$

Note that both f and g are 1-Lipschitz functions of u and therefore, an application of Theorem 1.1 yields:

$$\forall \|z\| \le 1 : \left| \frac{1}{md} \cdot \sum_{p=1}^{m} \sum_{q=1}^{d} f(\widetilde{h}_{p,q}(z)) - \mathbb{E}_{Z\sim\mathcal{N}(0,\|z\|^2)}[f(Z)] \right| \le \frac{\beta}{8}$$

$$\forall \|z\| \le 1 : \left| \frac{1}{md} \cdot \sum_{p=1}^{m} \sum_{q=1}^{d} g(\widetilde{h}_{p,q}(z)) - \mathbb{E}_{Z\sim\mathcal{N}(0,\|z\|^2)}[g(Z)] \right| \le \frac{\beta}{8}$$

with probability at least $1 - \delta/4$. Now, let $z \in \mathbb{R}^d$ with $\|z\| = 1$. We now have:

$$\begin{aligned}
&\frac{1}{md} \cdot \sum_{p=1}^{m} \sum_{q=1}^{d} \mathbf{1}\left\{\widetilde{h}_{p,q}(z) \le 2\right\} \\
&\le \frac{1}{md} \cdot \sum_{p=1}^{m} \sum_{q=1}^{d} f(\widetilde{h}_{p,q}(z)) \le \mathbb{E}_{Z\sim\mathcal{N}(0,1)}[f(Z)] + \frac{\beta}{8} \\
&= \alpha + \frac{\beta}{8} - \int_2^3 (u-2)\phi(u)du \le \alpha + \frac{\beta}{8} - \frac{\beta}{2} < \alpha - \frac{\beta}{4}
\end{aligned}$$

yielding the first inequality in the conclusion of the claim. For the second, we have:

$$\begin{aligned}
&\frac{1}{md} \cdot \sum_{p=1}^{m} \sum_{q=1}^{d} \mathbf{1}\left\{\widetilde{h}_{p,q}(z) \le 4\right\} \\
&\ge \frac{1}{md} \cdot \sum_{p=1}^{m} \sum_{q=1}^{d} g(\widetilde{h}_{p,q}(z)) \ge \mathbb{E}_{Z\sim\mathcal{N}(0,1)}[g(Z)] - \frac{\beta}{8} \\
&= \alpha - \frac{\beta}{8} + \int_3^4 (4-u)\phi(u)du \ge \alpha - \frac{\beta}{8} + \frac{\beta}{2} > \alpha + \frac{\beta}{4}
\end{aligned}$$

yielding the second and concluding the proof of the claim. □

Claim 5.2. We have for all $\|z\| = 1, r \ge 4\sqrt{\log 1/\varepsilon}$:

$$\left(1 - \frac{\varepsilon}{2}\right) \le \frac{1}{md} \cdot \sqrt{\frac{\pi}{2}} \cdot \sum_{p=1}^{m} \sum_{q=1}^{d} \psi_r(\widetilde{h}_{p,q}(z)) \le \left(1 + \frac{\varepsilon}{2}\right)$$

with probability at least $1 - \delta/4$.

Proof. Note that the function $\sqrt{\pi/2} \cdot \psi_r(u)$ is $\sqrt{\pi/2}$-Lipschitz in u for any r. Defining $r^* := 4\sqrt{\log 1/\varepsilon}$, two applications of Theorem 1.1 yield for all $\|z\| \le 1$:

$$\left| \frac{1}{md}\sqrt{\frac{\pi}{2}} \sum_{p=1}^{m} \sum_{q=1}^{d} \psi_{r^*}(\widetilde{h}_{p,q}(z)) - \sqrt{\frac{\pi}{2}}\mathbb{E}_{Z\sim\mathcal{N}(0,\|z\|^2)}[\psi_{r^*}(Z)] \right| \le \frac{\varepsilon}{8}$$

$$\left| \frac{1}{md} \cdot \sqrt{\frac{\pi}{2}} \cdot \sum_{p=1}^{m} \sum_{q=1}^{d} |\widetilde{h}_{p,q}(z)| - \sqrt{\frac{\pi}{2}} \cdot \mathbb{E}_{Z\sim\mathcal{N}(0,\|z\|^2)}[|Z|] \right| \le \frac{\varepsilon}{8}$$

with probability at least $1 - \delta/4$. Now, fix $z \in \mathbb{R}^d$ with $\|z\| = 1$. We now have:

$$\sqrt{\frac{\pi}{2}} \cdot \mathbb{E}_{Z\sim\mathcal{N}(0,1)}[|Z|] = 1$$

$$\sqrt{\frac{\pi}{2}} \cdot \mathbb{E}_{Z\sim\mathcal{N}(0,1)}[\psi_{r^*}(Z)] \le \sqrt{\frac{\pi}{2}} \cdot \mathbb{E}_{Z\sim\mathcal{N}(0,1)}[|Z|] = 1.$$

We now derive a lower bound on $\mathbb{E}_{Z\sim\mathcal{N}(0,1)}[\psi_{r^*}(Z)]$ below:

$$\begin{aligned}
&\sqrt{\frac{\pi}{2}} \cdot \mathbb{E}_{Z\sim\mathcal{N}(0,1)}[\psi_{r^*}(Z)] \\
&\ge \sqrt{\frac{\pi}{2}} \cdot \mathbb{E}_{Z\sim\mathcal{N}(0,1)}[|Z|] - \sqrt{\frac{\pi}{2}} \int_{(-\infty,-r^*]\cup[r^*,\infty)} |u|\phi(u)du \\
&= 1 - \sqrt{2\pi} \int_{r^*}^{\infty} u\phi(u)du = 1 - \exp\left\{-\frac{(r^*)^2}{2}\right\} \ge 1 - \frac{\varepsilon}{8}.
\end{aligned}$$

The previous three displays now yield for all $\|z\| = 1$:

$$\left(1 - \frac{\varepsilon}{4}\right) \le \frac{1}{md} \cdot \sqrt{\frac{\pi}{2}} \cdot \sum_{p=1}^{m} \sum_{q=1}^{d} \psi_{r^*}(\widetilde{h}_{p,q}(z)) \le \left(1 + \frac{\varepsilon}{4}\right)$$

$$\left(1 - \frac{\varepsilon}{4}\right) \le \frac{1}{md} \cdot \sqrt{\frac{\pi}{2}} \cdot \sum_{p=1}^{m} \sum_{q=1}^{d} |\widetilde{h}_{p,q}(z)| \le \left(1 + \frac{\varepsilon}{4}\right)$$

which imply the claim by noting that for any $r \ge r^*$, $\psi_{r^*}(x) \le \psi_r(x) \le |x|$. □

For the rest of the proof, we will condition the conclusions of Claims 5.1 and 5.2. We will now analyze the correctness of the query procedure. We first prove a correctness guarantee on estimated truncation levels r_i.

Claim 5.3. Conditioned on Claims 5.1 and 5.2, we have:

$$\forall i \in [n] : 2\|x - x_i\| \le \mathrm{Quant}_\alpha\left(\left\{y_{l_p} - (y_i)_{l_p}\right\}_{p\in[k]}\right) \le 4\|x - x_i\|$$

with probability at least $1 - \delta/4$ over the random indices $\{l_j\}_{j\in[k]}$.

Proof. Fix $i \in [n]$ and for all $j \in [k]$, define the random variables W_j, V_j:

$$W_j := \mathbf{1}\left\{(y_{l_j} - (y_i)_{l_j}) \le 2\|x - x_i\|\right\}$$

$$V_j := \mathbf{1}\left\{(y_{l_j} - (y_i)_{l_j}) \le 4\|x - x_i\|\right\}.$$

From the linearity of $\widetilde{h}$ and Claim 5.1, we get:

$$\mathbb{E}[W_j] \le \alpha - \frac{\beta}{4} \text{ and } \mathbb{E}[V_j] \ge \alpha + \frac{\beta}{4}.$$

An application of Hoeffding's inequaity to the random variables W_j, V_j with our setting of k yields:

$$\frac{1}{k} \cdot \sum_{j=1}^{k} W_j \le \alpha - \frac{\beta}{8} \text{ and } \frac{1}{k} \cdot \sum_{j=1}^{k} V_j \ge \alpha + \frac{\beta}{8}$$

with probability at least $1 - \delta/(4n)$. On the above event, we have:

$$2\|x - x_i\| \le \mathrm{Quant}_\alpha\left(\left\{y_{l_p} - (y_i)_{l_p}\right\}_{p\in[k]}\right) \le 4\|x - x_i\|.$$

A union bound over $i \in [n]$, concludes the proof of the claim. □

To conclude the proof, fix $i \in [n]$ and let $\widetilde{r} := 4\sqrt{\log 1/\varepsilon} \cdot \|x - x_i\|$ and $\widehat{r} := 8\sqrt{\log 1/\varepsilon} \cdot \|x - x_i\|$. Noting that $\psi_r(u) \le r$ for all $u \in \mathbb{R}, r > 0$, we have from Hoeffding's inequality, Claim 5.2 and our setting of k that:

$$\frac{1}{k} \cdot \sqrt{\frac{\pi}{2}} \cdot \sum_{p=1}^{k} \psi_{\widetilde{r}}((y_{l_p} - (y_i)_{l_p})) \ge (1 - \varepsilon) \cdot \|x - x_i\|$$

$$\frac{1}{k}\cdot\sqrt{\frac{\pi}{2}}\cdot\sum_{p=1}^{k}\psi_{\widehat{r}}((y_{l_p}-(y_i)_{l_p}))\le(1+\varepsilon)\cdot\|x-x_i\|$$

with probability at least $1-\delta/4n$. A union bound yields the above condition for all $i\in[n]$ with probability at least $1-\delta/4$. Conditioning on the above display and Claims 5.1 to 5.3 and noting that on this event $\psi_{\widehat{r}}(u)\le\psi_{r_i}(u)\le\psi_{\widehat{r}}(u)$ for all $u\in\mathbb{R}$, we get via a union bound:

$$\forall i\in[n]:(1-\varepsilon)\cdot\|x-x_i\|\le d_i\le(1+\varepsilon)\cdot\|x-x_i\|.$$

with probability at least $1-\delta$.

The runtime guarantee follow from the fact that for each $j\in[m]$, $\widetilde{h}^j(z)$ is computable in time $O(d\log d)$ for all $z\in\mathbb{R}^d$. This concludes the proof of the theorem up to routine runtime analyses. □

ACKNOWLEDGMENTS

J.N is supported by NSF award CCF-1951384, ONR grant N00014-18-1-2562, ONR DORECG award N00014-17-1-2127, and a Google Faculty Research Award. Y.C gratefully acknowledges support from a Microsoft Research-BAIR Commons open research grant.

A OPTIMALITY OF THEOREM 1.1

In this section, we show that the conclusions of Theorem 1.1 are tight up to log factors and furthermore, that the embedding dimension required is within a logarithmic factor of that required by the use of full Gaussian matrices. We start by establishing the near optimality of conclusion of Theorem 1.1:

Theorem A.1. *Let $\widetilde{h},\widetilde{h}^j,\widetilde{h}_{j,k}$ be as in RHT, $\varepsilon\in(0,1)$ and $\delta\in(0,0.01)$. Then, there exists a 1-Lipschitz function, $f:\mathbb{R}\to\mathbb{R}$, such that there exists $z\in\mathbb{R}^d$ s.t $\|z\|\le 1$ ::*

$$\left|\frac{1}{md}\cdot\sum_{j=1}^{m}\sum_{k=1}^{d}f(\widetilde{h}_{j,k}(z))-\mathbb{E}_{Z\sim\mathcal{N}(0,\|z\|^2)}\left[f(Z)\right]\right|\ge\varepsilon$$

with probability at least δ if $m\le\varepsilon^{-2}\log d/\delta$.

Proof. In our construction, we simply let $f(x):=x$ and we pick $z:=e_i$, the basis vectors. Now, we see:

$$\frac{1}{md}\cdot\sum_{j=1}^{m}\sum_{k=1}^{d}f(\widetilde{h}_{j,k}(e_i))=\frac{1}{m}\cdot\sum_{j=1}^{m}D^j_{i,i}=:W_i\overset{iid}{\sim}\mathcal{N}(0,1/m).$$

By noting that for $Z\sim\mathcal{N}(0,1)$, $\mathbb{P}\{|Z|\ge x\}\ge\frac{1}{2}\cdot\left(\frac{1}{x}-\frac{1}{x^3}\right)\cdot\exp\left(-\frac{x^2}{2}\right)$, we get:

$$\mathbb{P}\{|W_i|\ge\varepsilon\}\ge 1.5\cdot\frac{\delta}{d}\implies$$

$$\mathbb{P}\{\exists i\in[d]:|W_i|\ge\varepsilon\}\ge 1-\left(1-1.5\cdot\frac{\delta}{d}\right)^d\ge\delta.$$

The theorem now follows as $\mathbb{E}_{Z\sim\mathcal{N}(0,\sigma^2)}[f(Z)]=0$ for any σ^2. □

We now show that the embedding dimension guaranteed by Theorem 1.1 are within a logarithmic factor of the best obtainable dimension even in the setting where true Gaussian matrices are used.

Theorem A.2. *Let $\{g_i\}_{i=1}^n$ be n i.i.d random vectors such that $g_i\sim\mathcal{N}(0,I)$, $\varepsilon\in(0,1)$ and $d\ge 40$. Then, there exists a 1-Lipschitz function, $f:\mathbb{R}\to\mathbb{R}$, such that:*

$$\exists z\in\mathbb{R}^d\ s.t\ \|z\|\le 1:\left|\frac{1}{n}\sum_{i=1}^{n}f(\langle z,g_i\rangle)-\mathbb{E}_{Z\sim\mathcal{N}(0,\|z\|^2)}[f(Z)]\right|\ge\varepsilon$$

with probability at least $9/10$ as long as $n\le(2\varepsilon)^{-2}d$.

Proof. As in Theorem A.1, we let $f(x):=x$. Observe that we now have:

$$\max_{z\text{ s.t }\|z\|\le 1}\frac{1}{n}\sum_{i=1}^{n}f(\langle z,g_i\rangle)=\max_{z\text{ s.t }\|z\|\le 1}\left\langle z,\frac{1}{n}\cdot\sum_{i=1}^{n}g_i\right\rangle=\left\|\frac{1}{n}\cdot\sum_{i=1}^{n}g_i\right\|.$$

Letting $g:=\frac{1}{n}\cdot\sum_{i=1}^{n}g_i$, we see that $g\sim\mathcal{N}(0,I/n)$ and hence, we get from Theorem B.1 that:

$$\mathbb{P}\left\{\left\|\frac{1}{n}\cdot\sum_{i=1}^{n}g_i\right\|\ge\varepsilon\right\}\ge\frac{9}{10}.$$

The theorem now follows from the fact that $\mathbb{E}_{Z\sim\mathcal{N}(0,\sigma^2)}[f(Z)]=0$ for all $\sigma^2>0$. □

B MISCELLANEOUS RESULTS AND SUPPORTING LEMMAS

In this section, we recall some standard facts from probability theory and some lemmas used in the proofs our main results. The first is the Tsirelson-Ibragimov-Sudakov inequality (see [6, Theorem 5.6], for example):

Theorem B.1. *[6, Theorem 5.6] Let $X=(X_1,\ldots,X_n)$ be a vector of n independent standard normal random variables. Let $f:\mathbb{R}^n\to\mathbb{R}$ be an L-Lipschitz function; that is, f satisfies:*

$$\forall x,y\in\mathbb{R}^d:|f(x)-f(y)|\le L\cdot\|x-y\|.$$

Then, we have:

$$\mathbb{P}\{f(X)-\mathbb{E}f(X)\ge t\}\le e^{-t^2/(2L^2)}.$$

We have the following simple corollary.

Corollary B.2. *Let $X=(X_1,\ldots,X_n)$ be a vector of n independent standard normal random variables. Let $f:\mathbb{R}^n\to\mathbb{R}$ be an L-Lipschitz function; that is, f satisfies:*

$$\forall x,y\in\mathbb{R}^d:|f(x)-f(y)|\le L\cdot\|x-y\|.$$

Then, we have:

$$\mathbb{P}\{|f(X)-\mathbb{E}f(X)|\ge t\}\le 2e^{-t^2/(2L^2)}.$$

Proof. The proof follows by applying Theorem B.1 to $-f$ and a union bound. □

We now recall the definition of an ε-net and a Covering number from [25].

Definition B.3. [25, Definition 4.2.1] Let (T,d) be a metric space. Consider a subset $K\subset T$ and let $\varepsilon>0$. A subset $\mathcal{N}\subseteq K$ is called an *ε-net* of K if every point in K is within a distance ε of some point in $\mathcal{N}$, i.e

$$\forall x\in K\ \exists x_0\in\mathcal{N}:d(x,x_0)\le\varepsilon.$$

Additionally, the smallest possible cardinality of an ε-net of K is called a *covering number* of K and is denoted $\mathcal{N}(K,d,\varepsilon)$.

We now introduce a standard proposition bounding covering numbers of subsets of Euclidean space.

Proposition B.4. *[25, Proposition 4.2.12] Let K be a subset of $\mathbb{R}^d$ and let $\varepsilon > 0$. Then:*

$$\frac{|K|}{|\varepsilon B_2^n|} \leq \mathcal{N}(K, \|\cdot\|_2, \varepsilon) \leq \frac{|K + (\varepsilon/2)B_2^n|}{|(\varepsilon/2)B_2^n|}.$$

Here, $|\cdot|$ denotes the volume in $\mathbb{R}^n$, B_2^n denotes the unit Euclidean ball in $\mathbb{R}^n$; so, εB_2^n is an Euclidean ball with radius ε.

The proposition yields the following simple corollary similar to [25, Corollary 4.2.13].

Corollary B.5. *Assume the setting of Proposition B.4. Additionally, let K be a subset of B_2^n. Then,*

$$\mathcal{N}(K, \|\cdot\|_2, \varepsilon) \leq \left(\frac{2}{\varepsilon} + 1\right)^n.$$

Proof. We have from Proposition B.4:

$$\mathcal{N}(K, \|\cdot\|_2, \varepsilon) \leq \frac{\left|K + (\varepsilon/2)B_2^n\right|}{\left|(\varepsilon/2)B_2^n\right|} \leq \frac{\left|(1+\varepsilon/2)B_2^n\right|}{\left|(\varepsilon/2)B_2^n\right|} = \frac{(1+\varepsilon/2)^n}{(\varepsilon/2)^n} = \left(\frac{2}{\varepsilon} + 1\right)^n.$$

□

We will also use a simple lemma concerning the spectral norm of the linear transformations considered throughout this paper.

Lemma B.6. *Let $\varepsilon, \delta \in (0, 1/2)$ and $d \in \mathbb{N}$. Then, we have:*

$$\forall x, y \in \mathbb{R}^d : (1-\varepsilon)\cdot\|x-y\| \leq \frac{\left\|\widetilde{h}(x) - \widetilde{h}(y)\right\|}{\sqrt{md}} \leq (1+\varepsilon)\cdot\|x-y\|$$

with probability at least $1-\delta$ if $m \geq 4 \cdot \frac{\log d + \log 2/\delta}{\varepsilon^2}$.

Proof. First, note that we may assume $y = 0$ due to the linearity of $\widetilde{h}$ and that it now suffices to show the conclusion for $x \in \mathbb{S}^{d-1}$. We have from the orthogonality of H:

$$\left\|\widetilde{h}(x)\right\|^2 = (\widetilde{h}(x))^\top \widetilde{h}(x) = dx^\top \sum_{j=1}^m (D^j)^2 x$$

$$= dx^\top \begin{bmatrix} \sum_{j=1}^m (D^j_{1,1})^2 & 0 & \cdots & 0 \\ 0 & \sum_{j=1}^m (D^j_{2,2})^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sum_{j=1}^m (D^j_{d,d})^2 \end{bmatrix} x.$$

Therefore, we have:

$$\forall x \in \mathbb{R}^d : \min_k \sqrt{\frac{\sum_{j=1}^m (D^j)^2_{k,k}}{m}} \leq \frac{\|\widetilde{h}(x)\|}{\sqrt{md}} \leq \max_k \sqrt{\frac{\sum_{j=1}^m (D^j)^2_{k,k}}{m}}.$$

Note that we have from Corollary B.2 that:

$$\left|\sqrt{\frac{\sum_{j=1}^m (D^j)^2_{k,k}}{m}} - 1\right| \leq \sqrt{\frac{2\cdot(\log d + \log 2/\delta)}{m}}$$

with probability at least $1 - \delta/d$. A union bound over $k \in [d]$ and our condition on m concludes the proof. □

REFERENCES

[1] Nir Ailon and Bernard Chazelle. 2009. The Fast Johnson–Lindenstrauss Transform and Approximate Nearest Neighbors. *SIAM J. Comput.* 39, 1 (2009), 302–322. https://doi.org/10.1137/060673096

[2] Noga Alon, Omri Ben-Eliezer, Yuval Dagan, Shay Moran, Moni Naor, and Eylon Yogev. 2021. Adversarial Laws of Large Numbers and Optimal Regret in Online Classification. *CoRR* abs/2101.09054 (2021). arXiv:2101.09054 https://arxiv.org/abs/2101.09054

[3] Omri Ben-Eliezer, Rajesh Jayaram, David P. Woodruff, and Eylon Yogev. 2020. A Framework for Adversarially Robust Streaming Algorithms. In *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*.

[4] Omri Ben-Eliezer and Eylon Yogev. 2020. The Adversarial Robustness of Sampling. In *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, PODS 2020, Portland, OR, USA, June 14-19, 2020*, Dan Suciu, Yufei Tao, and Zhewei Wei (Eds.). ACM, 49–62. https://doi.org/10.1145/3375395.3387643

[5] Jeremiah Blocki, Avrim Blum, Anupam Datta, and Or Sheffet. 2012. The Johnson-Lindenstrauss Transform Itself Preserves Differential Privacy. In *53rd Annual IEEE Symposium on Foundations of Computer Science, FOCS 2012, New Brunswick, NJ, USA, October 20-23, 2012*. 410–419.

[6] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. 2013. *Concentration inequalities*. Oxford University Press, Oxford. x+481 pages. https://doi.org/10.1093/acprof:oso/9780199535255.001.0001 A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[7] Emmanuel Candès and Terence Tao. 2005. Decoding by Linear Programming. *IEEE Trans. Inf. Theory* 51, 12 (2005), 4203–4215.

[8] Emmanuel Candès and Terence Tao. 2006. Near-optimal signal recovery from random projections: universal encod- ing strategies? *IEEE Trans. Inf. Theory* 52, 12 (2006), 5406–5425.

[9] Yeshwanth Cherapanamjeri and Jelani Nelson. 2020. On Adaptive Distance Estimation, See [17]. https://proceedings.neurips.cc/paper/2020/hash/803ef56843860e4a48fc4cdb3065e8ce-Abstract.html

[10] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamás Sarlós, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. 2020. Rethinking Attention with Performers. *CoRR* abs/2009.14794 (2020). arXiv:2009.14794 https://arxiv.org/abs/2009.14794

[11] Krzysztof Choromanski, Mark Rowland, Wenyu Chen, and Adrian Weller. 2019. Unifying Orthogonal Monte Carlo Methods. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA (Proceedings of Machine Learning Research)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.), Vol. 97. PMLR, 1203–1212. http://proceedings.mlr.press/v97/choromanski19a.html

[12] Krzysztof Marcin Choromanski, Mark Rowland, and Adrian Weller. 2017. The Unreasonable Effectiveness of Structured Random Orthogonal Embeddings. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 219–228. https://proceedings.neurips.cc/paper/2017/hash/bf8229696f7a3bb4700cfddef19fa23f-Abstract.html

[13] David Donoho. 2006. Compressed sensing. *IEEE Trans. Inf. Theory* 52, 4 (2006), 1289–1306.

[14] Avinatan Hassidim, Haim Kaplan, Yishay Mansour, Yossi Matias, and Uri Stemmer. 2020. Adversarially Robust Streaming Algorithms via Differential Privacy, See [17]. https://proceedings.neurips.cc/paper/2020/hash/0172d289da48c48de8c5ebf3de9f7ee1-Abstract.html

[15] Piotr Indyk and Rajeev Motwani. 1998. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on the Theory of Computing, Dallas, Texas, USA, May 23-26, 1998*, Jeffrey Scott Vitter (Ed.). ACM, 604–613. https://doi.org/10.1145/276698.276876

[16] Haim Kaplan, Yishay Mansour, Kobbi Nissim, and Uri Stemmer. 2021. Separating Adaptive Streaming from Oblivious Streaming. *CoRR* abs/2101.10836 (2021). arXiv:2101.10836 https://arxiv.org/abs/2101.10836

[17] Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). 2020. *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. https://proceedings.neurips.cc/paper/2020

[18] Quoc Le, Tamas Sarlos, and Alexander Smola. 2013. Fastfood - Computing Hilbert Space Expansions in loglinear time. In *Proceedings of the 30th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Sanjoy Dasgupta and David McAllester (Eds.), Vol. 28. PMLR, Atlanta, Georgia, USA, 244–252. http://proceedings.mlr.press/v28/le13.html

[19] Michel Ledoux and Michel Talagrand. 2011. *Probability in Banach spaces*. Springer-Verlag, Berlin. xii+480 pages. Isoperimetry and processes, Reprint of the 1991 edition.

[20] Deanna Needell and Joel A. Tropp. 2009. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comput. Harmon. Anal.* 26 (2009), 301–332.

[21] Ali Rahimi and Benjamin Recht. 2007. Random Features for Large-Scale Kernel Machines. In *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis (Eds.). Curran Associates, Inc., 1177–1184. http://papers.nips.cc/paper/3182-random-features-for-large-scale-kernel-machines

[22] Mark Rudelson and Roman Vershynin. 2008. On sparse reconstruction from Fourier and Gaussian measurements. *Comm. Pure Appl. Math.* 61, 8 (2008), 1025–1045.

[23] Tamás Sarlós. 2006. Improved Approximation Algorithms for Large Matrices via Random Projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*. 143–152.

[24] Yanning Shen, Tianyi Chen, and Georgios B. Giannakis. 2018. Online Multi-Kernel Learning with Orthogonal Random Features. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*. IEEE, 6289–6293. https://doi.org/10.1109/ICASSP.2018.8461509

[25] Roman Vershynin. 2018. *High-dimensional probability*. Cambridge Series in Statistical and Probabilistic Mathematics, Vol. 47. Cambridge University Press, Cambridge. xiv+284 pages. https://doi.org/10.1017/9781108231596 An introduction with applications in data science, With a foreword by Sara van de Geer.

[26] Felix X. Yu, Ananda Theertha Suresh, Krzysztof Marcin Choromanski, Daniel N. Holtmann-Rice, and Sanjiv Kumar. 2016. Orthogonal Random Features. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.). 1975–1983. https://proceedings.neurips.cc/paper/2016/hash/53adaf494dc89ef7196d73636eb2451b-Abstract.html