MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Representation and Coding of Signal Geometry

Boufounos, Petros T.; Rane, Shantanu D.; Mansour, Hassan

TR2017-036 March 15, 2017

Abstract

Approaches to signal representation and coding theory have traditionally focused on how to best represent signals using parsimonious representations that incur the lowest possible distortion. Classical examples include linear and non-linear approximations, sparse representations, and rate-distortion theory. Very often, however, the goal of processing is to extract specific information from the signal, and the distortion should be measured on the extracted information. The corresponding representation should, therefore, represent that information as parsimoniously as possible, without necessarily accurately representing the signal itself. In this paper, we examine the problem of encoding signals such that sufficient information is preserved about their pairwise distances and their inner products. For that goal, we consider randomized embeddings as an encoding mechanism and provide a framework to analyze their performance. We also demonstrate that it is possible to design the embedding such that it represents different ranges of distances with different precision. These embeddings also allow the computation of kernel inner products with control on their inner product-preserving properties. Our results provide a broad framework to design and analyze embeddings, and generalize existing results in this area, such as random Fourier kernels and universal embeddings

Information and Inference: a Journal of the IMA

REPRESENTATION AND CODING OF SIGNAL GEOMETRY

PETROS T. BOUFOUNOS, SHANTANU RANE, AND HASSAN MANSOUR

ABSTRACT. Approaches to signal representation and coding theory have traditionally focused on how to best represent signals using parsimonious representations that incur the lowest possible distortion. Classical examples include linear and non-linear approximations, sparse representations, and rate-distortion theory. Very often, however, the goal of processing is to extract specific information from the signal, and the distortion should be measured on the extracted information. The corresponding representation should, therefore, represent that information as parsimoniously as possible, without necessarily accurately representing the signal itself.

In this paper, we examine the problem of encoding signals such that sufficient information is preserved about their pairwise distances and their inner products. For that goal, we consider randomized embeddings as an encoding mechanism and provide a framework to analyze their performance. We also demonstrate that it is possible to design the embedding such that it represents different ranges of distances with different precision. These embeddings also allow the computation of kernel inner products with control on their inner product-preserving properties. Our results provide a broad framework to design and analyze embeddings, and generalize existing results in this area, such as random Fourier kernels and universal embeddings.

1. INTRODUCTION

Signal representation theory and practice have primarily focused on how to best represent or encode a signal while incurring the smallest possible distortion. For example, image or video representations typically aim to minimize the distortion in the signal so that the visual quality of the signal is maintained when displayed to a user. Quite often, however, the user of a signal is not a human observer, but an algorithm extracting some information about the signal. In this case, the goal is different: the representation should not destroy the information that the algorithm requires, even if the signal itself cannot be completely recovered.

In this paper, we examine signal representations that preserve aspects of the signal's geometry but not necessarily the signal itself. Our approach exploits the geometry-preserving properties of randomized embeddings. Specifically, we develop a framework that generalizes well-known embeddings in a manner that enables the design and control of the distance distortion and the resulting inner product kernel. The results in this paper extend and generalize recently developed theory for efficient universal quantization and universal quantized embeddings [12, 16–18], and random Fourier kernels [64, 65]. We demonstrate and analyze such representations using a very general approach, that can encompass continuous and quantized embeddings.

As we first reported in [16, 18], representations based on universal embeddings—which are special cases of our development—can be used as a geometry-preserving coding mechanism in image retrieval and classification applications. We demonstrated that we are able to improve compression performance up to 25% over previous embedding-based approaches [51, 75], including our own earlier work [48]. The main advantage of our approach is the ability to control the range of distances best preserved by the embeddings, so that we do not represent distance ranges that are not important to the application at hand. In most inference applications it is only necessary to represent distances up to a certain radius, as required by the algorithm, and not farther. Thus, bits are not wasted in coding distances larger than necessary.

1.1. **Motivation.**

Our work is partly motivated by cloud-based image retrieval and classification applications, such as augmented reality. As we discuss in [16, 18, 48], augmented reality and other image retrieval and classification applications can benefit significantly by efficient coding of the geometry of the signal space and of the geometric relationships between signals.

In typical cloud-based image retrieval applications, a client transmits to a cloud server a query image acquired by the user, or features extracted from that image, requesting more information on the objects in the image. The

Key words and phrases. Randomized Embeddings, Dimensionality Reduction, Distance representations, Coding for inference, Kernel methods, Quantization .

S. Rane performed this work while he was a researcher at MERL.

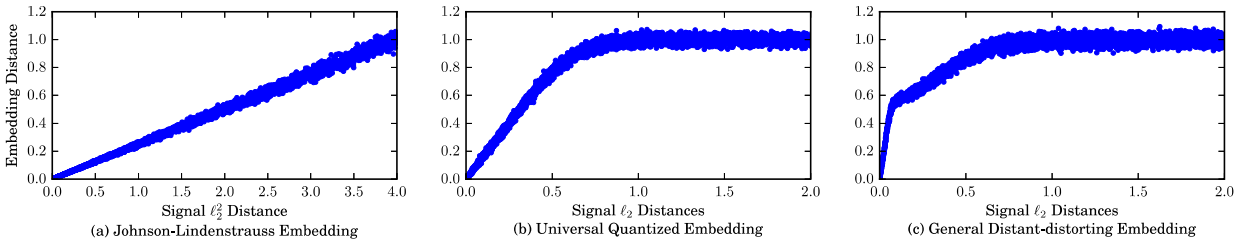

FIGURE 1. Example embeddings designed and analyzed in this paper; each example provides different distance distortion properties, according to application requirements.

server extracts features, if necessary, and uses those to execute the query. The query typically drives a machine learning algorithm, such as nearest neighbor search or support vector machines (SVMs), searching the database for metadata associated with those features. The query typically returns information such as the object class, the object identity, or associated information about the object. This search should be computationally efficient at the client and the database server, and the transmission should be bandwidth-efficient.

While our example application uses nearest-neighbor search and SVMs for classification and inference, our goal is to develop a framework that is agnostic to the underlying mechanism, as long as this mechanism relies on signal geometry for its functioning. Thus, most classical machine learning techniques, such as regressions, mixture models, spectral clustering and deep learning, can immediately exploit our representations. All that is necessary in order to appropriately design the embedding is an understanding of which signal distances should be preserved for the particular application.

Our hope is that this framework will become a useful addition to a system designer's tool-belt, providing methods that can be layered with other machine learning primitives. In this paper, we concentrate on one particular method for designing geometry-preserving embeddings and develop a general method to analyze them. However, some of the theoretical tools developed here can be used in more general settings and applications.

Examples of the embeddings we can design and analyze using our tools are illustrated in Fig. 1. Conventional Johnson-Lindenstrauss embeddings, first introduced in [44], preserve all ranges of distances undistorted, as shown in Fig. 1(a). Instead, universal quantized embeddings, introduced in [16, 18] and shown in Fig. 1(b), distort distances such that all distances beyond a range appear as the same in the embedding. Finally, as shown in Fig. 1(c), we can design more general distortions of the signal geometry, mapping different ranges of signal distances to different ranges of embedding distances. Such selective distortions allow the embedding to preserve different ranges of distances with different accuracy, according to application demands.

1.2. **Contributions.** Our paper contributes several results toward establishing embeddings as general representations of signal geometry:

- We introduce a generalized definition and characterization of geometry-preserving embeddings which allows for selective distortions in the signal geometry. These distortions are captured using a distance map that describes how distances in the original signal space are distorted in the embedding space. This definition covers a large number of existing embeddings, as well as more general designs, and enables analysis of the embedding characteristics given the distance map.
- We develop a very general framework to extend embeddings to infinite sets, such as sparse signals or manifolds, even if the mapping function is discontinuous. Our approach is fundamentally very similar to established approaches using set coverings. However, these methods fail if the embedding function is not continuous, e.g., due to quantization. The tools we introduce extend the notion of Lipschitz continuity to a large variety of discontinuous functions in a way that enables proofs using covering arguments.
- We demonstrate a method to design randomized embeddings such that they achieve the desired distortions in the geometry of the space. The design we describe generalizes existing embedding constructions, such as the random Fourier features [65], and universal quantized embeddings [12, 16–18].
- We present an analysis of the embedding ambiguity in the context of the distance map. We characterize this ambiguity from a new perspective: we assume the embedding is used as a representation of the original signal set. Current embedding guarantees describe the ambiguity in the embedding space, as opposed

to the signal set. While the two are equivalent in many well-known cases, they differ quite often, especially if the embedding distorts the signal geometry.
- We establish a connection between distance embeddings and kernel methods, demonstrating that the distortion of the distance map performed by the embedding is equivalent to the distortion performed by a kernel inner product.
- We provide an analysis of multibit universal embeddings and a generalization of binary universal embeddings to infinite sets, such as sparse signals or manifolds. These generalizations establish new results in this area and serve as examples of how our developments can be used in practice.

1.3. **Related Work.** The best known embeddings are due to Johnson and Lindenstrauss (J-L) [44], which preserve ℓ_2 distances of point clouds. A significant body of work has been devoted to developing such embeddings using a variety of randomizations and for a variety of applications [1, 25]. Their importance was re-established recently thanks to the emergence of compressive sensing (CS) [21, 23, 29]. The Restricted Isometry Property (RIP), which plays a central role in CS theory, is essentially a restatement of the JL property, but applied to unions of signal subspaces instead of point clouds [8, 22]. Consequently, several connections between the two have been established. In addition to the RIP, extensions of the J-L lemma to other infinite sets, such as manifolds [9, 19, 28, 31, 63] and unions of subspaces [7, 10, 19, 28, 32, 63], have also been established.

Significant literature has also studied variations of J-L embeddings. For example, in a number of acquisition systems and coding applications, it is necessary to quantize the representations. Quantized J-L embeddings have been well studied [39, 40, 48], especially down to 1-bit per representation coefficient [43, 60–62]. Furthermore, while J-L embeddings and the RIP preserve ℓ_2 distances, there is a large body of work in preserving other similarity measurements, such as ℓ_p distances for various p's [37, 41, 42, 59], edit distance [3, 6, 47, 57], and angle, i.e., correlation, between signals [13–15].

A common thread in the aforementioned body of work is that distances or other similarity measures are preserved indiscriminately. This is in sharp contrast to our work, which allows the design of embeddings that represent some distances better than others, with control on that design. For example, in our motivating applications in the area of image retrieval, we design embeddings that only encode a short range of distances, as necessary for nearest-neighbor computation and classification. A very narrow notion of locality was discussed in very recent work, fit for the development in that paper [59]. That definition, however, does not capture the richer set of locality properties presented in our line of work.

Recent work has also provided classification guarantees for J-L embeddings [5] on very particular signal models. In particular, it is shown that separated convex ellipsoids remain separated when randomly projected to a space with sufficient dimensions. Our work significantly enhances the available design space compared to J-L embeddings. It should, thus, be possible to establish similar results. However, it is not clear that the techniques in [5] can be used with our designs. Thus, establishing results of similar type remains an interesting problem.

Many of our proof techniques rely on well-established concentration of measure arguments and methods common in the embedding literature, e.g., see [1, 8, 25, 43]. However, we provide a new approach to handle quantization or other discontinuous distortions, which can significantly expand the applicability of established approaches. Our main novelty in computing the embedding is the introduction of a *non-linear, periodic distortion* that enables notable control over the behavior of the embedding. We also develop a framework to analyze the performance of the embedding in preserving distances which, in contrast to the existing literature, takes into consideration the distortion as manifested in the original signal distance, as opposed to the embedded distance.

Our work is also related to locality-sensitive hashing (LSH) methods, which significantly reduce the computational complexity of near-neighbor computation [4, 26, 38]. The LSH literature shares a lot of the tools with the embeddings literature, such as randomized projections, dithering and quantization, but the goal is different: given a query point, LSH will return its near neighbors very efficiently, with $O(1)$ computation. This efficiency comes at a cost: no attempt is made to represent the distances of neighbors. When used to compare signals it only provides a binary decision, whether the distance of the signals is smaller than a threshold or not. This makes it unsuitable for applications that require more accurate distance information.

Our design does not explicitly take retrieval complexity into account; we expect the underlying retrieval machinery to consider complexity issues. Nevertheless, our methods provide dimensionality and bit-rate reduction, which are tightly coupled to complexity. Furthermore, some of our embedding techniques could be used in the context of an LSH-based scheme; some of the LSH techniques in [4, 26, 38] are reminiscent of our approach. It

should also be possible to design mechanisms that reduce complexity which explicitly exploit our methods, for example extending the hierarchical approach in [11]. However, such designs, although quite interesting, are beyond the scope of this paper.

Our work is of similar flavor to [64, 65], which use randomized embeddings to efficiently approximate specific kernel computations. The results we present generalize these approaches, by allowing control over the distance map in the kernel and the ambiguity of the distance preservation. We further provide a general approach to understand the approximation properties of the embedding and its behavior under quantization.

There is also a large body of work focused on learning embeddings from available data, e.g., see [36, 67, 69, 74]. Such approaches exploit a computationally expensive training stage to improve embedding performance with respect to its distance-preserving properties. Still, the embedding guarantees are only applicable to data similar to the training data; the embedding might not perform well on different sets. Instead, our approach relies on a randomization independent of the data. Our designs are universal in the sense that they work on any data set with overwhelming probability, as long as the embedding parameters are drawn independently of the dataset. Of course, using data for training is a promising avenue and a potentially useful extension of our work. However, we do not attempt this in this paper.

1.4. **Notation.** In the remainder of the paper we use regular typeface, e.g., x and y, to denote scalar quantities. Lowercase boldface such as $\mathbf{x}$ denotes vectors and uppercase boldface such as $\mathbf{A}$ denotes matrices. Functions are denoted using regular lowercase typefaces, e.g., $g(\cdot)$. Unless explicitly noted, all functions are scalar functions of one variable. In abuse of notation, a vector input to such functions, e.g., $g(\mathbf{x})$ means that the function is applied element-wise to all the elements of $\mathbf{x}$. Sets and vector spaces are denoted using calligraphic fonts, e.g., $\mathcal{W}$, $\mathcal{S}$.

The Fourier transform of a function $h(x)$ is defined as $H(\xi) = \int_{-\infty}^{+\infty} h(x)e^{-2\pi\imath x\xi}dx$, where $\imath = \sqrt{-1}$ is the imaginary unit. Similarly, the characteristic function of a probability density function $f_x(x)$ is defined as $\phi_x(t) = E\left[e^{\imath tx}\right] = \int_{-\infty}^{+\infty} f_x(x)e^{\imath x\xi}dx$. Thus, the Fourier transform of the density is related to its characteristic function: $F_x(\xi) = \phi_x(-2\pi\xi) = \phi_x^*(2\pi\xi)$, where $(\cdot)^*$ denotes complex conjugation. The Fourier series of a periodic function with period 1 is defined as $H_k = \int_0^1 h(x)e^{-2\pi\imath xk}dx$, where $h(x) = h(x+1)$. Conditional distributions and corresponding characteristic functions are denoted using $\cdot|\cdot$ in their argument, e.g., $f_x(x|y)$ and $\phi_x(x|y)$.

1.5. **Outline.** The next section contains a brief background on embeddings and universal scalar quantization, establishing notation and definitions. It also reviews Lipschitz continuity and introduces a generalization that will prove very useful in our subsequent development, especially for quantized embeddings. Section 3 provides an overview of how embedding results are typically established on point clouds using concentration of measure arguments, and introduces our framework to generalize such embeddings—both quantized and unquantized ones—to continuous sets. Section 4 demonstrates that embeddings and embedding maps can be designed to preserve different ranges of distances with different accuracy by establishing such a design, as well as the tools to analyze its properties.

Examples of embeddings established using our tools are provided in Sec. 5. These include quantized Johnson-Lindenstrauss embeddings, and binary and multibit universal embeddings. These examples demonstrate how our tools can be applied. They also establish some new useful results for universal embeddings. In addition, Sec. 6 provides simulations and application examples that validate our theory and demonstrate how it can be used in practice. Finally, Sec. 7 discusses our results and concludes. For most of our results, in order to improve the flow and readability of our paper, we have relegated the proofs to appendices.

2. Definitions and Background

2.1. **Randomized Linear Embeddings.** An embedding is a transformation of a set of signals in a high-dimensional space to a (typically) lower-dimensional one such that some aspects of the geometry of the set are preserved, as depicted in Fig. 2(a). Since the set geometry is preserved, distance computations can be performed directly on the low-dimensional—and often low bitrate—embeddings, rather than the underlying signals. For the purposes of this paper, we define an embedding as follows.

Definition 2.1. A function $f : \mathcal{S} \to \mathcal{W}$ is a (g, δ, ϵ) embedding of $\mathcal{S}$ into $\mathcal{W}$ if, for all $\mathbf{x}, \mathbf{y} \in \mathcal{S}$, it satisfies

$$(1-\delta)g\left(d_{\mathcal{S}}(\mathbf{x},\mathbf{y})\right) - \epsilon \le d_{\mathcal{W}}\left(f(\mathbf{x}), f(\mathbf{y})\right) \le (1+\delta)g\left(d_{\mathcal{S}}(\mathbf{x},\mathbf{y})\right) + \epsilon. \tag{2.1}$$

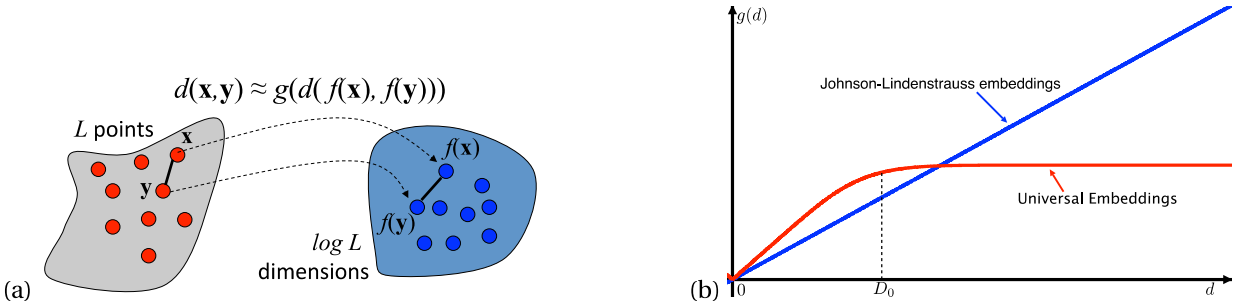

FIGURE 2. (a) Distance-preserving embeddings approximately preserve a function $g(\cdot)$ of the distance, allowing distances to be computed in a space that (typically) has fewer dimensions and produce signals that often require lower transmission rate. (b) For most embeddings, such as JL Embeddings, this function is linear, as shown in blue. For the universal quantized embeddings discussed in this paper, the function is approximately linear initially and quickly flattens out after a certain distance D_0, as shown in red.

In this definition, $g : \mathbb{R} \to \mathbb{R}$ is an invertible function mapping distances $d_{\mathscr{S}}(\cdot,\cdot)$ in $\mathscr{S}$ to distances $d_{\mathscr{W}}(\cdot,\cdot)$ in $\mathscr{W}$ and δ and ϵ quantify, respectively, the multiplicative and the additive ambiguity of the map[1]. We will often refer to $g(\cdot)$ as the distance map and to $f(\cdot)$ as the embedding map. In most known embeddings, such as the ones discussed in this section, the distance map is the identity $g(d) = d$ or a simple scaling.

For most of the development we only require that the distances $d_{\mathscr{S}}(\cdot,\cdot)$ and $d_{\mathscr{W}}(\cdot,\cdot)$ satisfy the triangle inequality. Specifically, the distances we explore are the commonly used ℓ_1, ℓ_2 and Hamming distance, although most of the results are more general.

The best known embeddings are the Johnson-Lindenstrauss embeddings [44]. These are functions $f : \mathscr{S} \to \mathbb{R}^M$ from a finite set of signals $\mathscr{S} \subset \mathbb{R}^N$ to a M-dimensional vector space such that, given two signals $\mathbf{x}$ and $\mathbf{y}$ in $\mathscr{S}$, their images satisfy:

$$(1-\delta)\|\mathbf{x}-\mathbf{y}\|_2^2 \le \|f(\mathbf{x})-f(\mathbf{y})\|_2^2 \le (1+\delta)\|\mathbf{x}-\mathbf{y}\|_2^2. \tag{2.2}$$

In other words, these embeddings preserve Euclidean, i.e., ℓ_2, distances of point clouds within a small factor, measured by δ, and using the identity as a distance map.

Johnson and Lindenstrauss demonstrated that a distance-preserving embedding, as described above, exists in a space of dimension $M = O(\frac{1}{\delta^2}\log L)$, where L is the number of signals in $\mathscr{S}$ (its cardinality) and δ the desired tolerance in the embedding. Remarkably, M is independent of N, the dimensionality of the signal set $\mathscr{S}$. Subsequent work showed that it is straightforward to compute such embeddings using a linear mapping. In particular, the function $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a $M \times N$ matrix whose entries are drawn randomly from specific distributions, satisfies (2.2) for all $\mathbf{x},\mathbf{y} \in \mathscr{S}$ with probability $1 - c_1 e^{\log L - c_2\delta^2 M}$, for some universal constants c_1, c_2, where the probability is with respect to the measure of $\mathbf{A}$. Commonly used distributions for the entries of $\mathbf{A}$ are i.i.d. Gaussian, i.i.d. Rademacher, or i.i.d. uniform [1, 25].

More recently, in the context of compressive sensing, such linear embeddings have been shown to embed infinite sets of signals. For example, the restricted isometry property (RIP) is an embedding of K-sparse signals and has been shown to be achievable with $M = O(K\log(N/K))$ [8, 22]. Similar properties have been shown for other signal set models, such as more general unions of subspaces [7, 10, 19, 28, 32, 63] and manifolds [9, 19, 28, 31, 63]. Typically, these generalizations are established by first proving that the embedding holds in a sufficiently dense point cloud on the signal set and exploiting linearity and smoothness to extend it to all the points of the set.

Such embeddings result in a significant dimensionality reduction. However, dimensionality reduction does not immediately produce rate reduction; the embeddings must be quantized for transmission and, if the quantization is not well designed, performance suffers [48]. In particular, when combined with scalar quantization, the embeddings satisfy

$$(1-\delta)\|\mathbf{x}-\mathbf{y}\|_2 - \epsilon \le \|f(\mathbf{x})-f(\mathbf{y})\|_2 \le (1+\delta)\|\mathbf{x}-\mathbf{y}\|_2 + \epsilon, \tag{2.3}$$

[1]Often it makes more sense to separate the upper and lower multiplicative bounds $(1 \pm \delta)$ to different constants A, B. This does not affect the subsequent development but encumbers the notation, so we avoid it in this paper.

where $\epsilon \propto 2^{-B}$ is the quantizer step size, decreasing exponentially with the number of bits used per dimension, B. On the other hand, δ is a function of M, the projection's dimensionality, and scales approximately as $1/\sqrt{M}$, as is the case for the J-L embedding. Recent work has refined these bounds, demonstrating that ϵ and δ decrease together as the number of measurements decreases when considering an ℓ_2 embedding into ℓ_1 [39, 40]. In addition, [39] establishes a non-linear distance map g for quantized linear embeddings of ℓ_2 into ℓ_2, that becomes linear for large distances. In the extreme case of 1-bit scalar quantization, the quantizer only keeps the sign of each measurement. Thus, a binary embedding does not preserve signal amplitudes, and therefore, ℓ_2 distances. Still, it does preserve angles, or equivalently, correlation coefficients [43, 60–62].

An intermediate case is examined in [27, 35, 46] in the context of recovering sparse signals from saturated measurements. In [46] it is shown that limited saturation and removal of the measurements preserves the embedding properties. In addition, [35] demonstrates that, with respect to recovery, saturated measurements behave similarly to 1-bit scalar quantization when the saturation is significant. However, the result does not establish an embedding; it only describes recovery properties. Still, an embedding result of similar flavor would be desirable and should be possible.

When designing a quantized embedding, the total rate is determined by the dimensionality of the projection and the number of bits used per dimension: $R = MB$. For a fixed bit budget R, as the dimensionality M increases, the accuracy of the embedding before quantization, as reflected in δ, is increased. But to keep the rate fixed, the number of bits per dimension should also decrease, which decreases the accuracy due to quantization, reflected in ϵ. This non-trivial trade-off is explored in detail in [48]; at a constant rate a multibit quantizer outperforms the 1-bit quantizers examined in earlier literature [51, 75].

2.2. **Universal Quantization and Embeddings.** Universal scalar quantization, first introduced in [12], fundamentally revisits scalar quantization and redesigns the quantizer to have non-contiguous quantization regions. This approach also relies on a Johnson-Lindenstrauss style projection, followed by scaling, dithering and scalar quantization:

$$f(\mathbf{x}) = Q(\Delta^{-1}(\mathbf{A}\mathbf{x} + \mathbf{w})), \tag{2.4}$$

where $\mathbf{A}$ is a $M \times N$ random matrix with $\mathcal{N}(0, \sigma^2)$-distributed, i.i.d. elements, Δ^{-1} is an element-wise scaling factor by a scalar Δ, $\mathbf{w}$ a length-M dither vector with i.i.d. elements, uniformly distributed in $[0, 2^B\Delta]$, and $Q(\cdot)$ a B-bit scalar quantizer operating element-wise on its input.

The breakthrough feature in this method is the modified B-bit scalar quantizer, designed to be a periodic binary function with non-contiguous quantization intervals, as shown in Fig. 3(a) for $B = 1$ (top) and $B = 2$ (bottom). The quantizer can be thought of as a regular uniform quantizer, computing a multi-bit representation of a signal and preserving only the least significant bits (LSB) of the representation. For example, for a 1-bit quantizer, scalar values in $[2l, 2l+1)$ quantize to 1 and scalar values in $[2l+1, 2(l+1))$, for any integer l, quantize to 0. If $Q(\cdot)$ is a 1-bit quantizer, this method encodes using as many bits as the rows of $\mathbf{A}$, i.e., M bits, and does not require subsequent entropy coding.

A large part of the development in this paper is inspired by (and generalizes) the periodicity of the quantization function in binary (1-bit) universal quantization. A multi-bit generalization of the universal quantizer is shown in the bottom of Fig. 3(a) but has not, to-date, been analyzed or explored.

As discussed in [12], the modified binary quantizer enables efficient universal encoding of signals. Furthermore, this quantization method is also an embedding of finite signal sets [17]. Specifically, on a set $\mathcal{S}$ with L points, for all $\mathbf{x}, \mathbf{y} \in \mathcal{S}$, the embedding satisfies

$$g\left(\|\mathbf{x}-\mathbf{y}\|_2\right) - \epsilon \le d_H\left(f(\mathbf{x}), f(\mathbf{y})\right) \le g\left(\|\mathbf{x}-\mathbf{y}\|_2\right) + \epsilon, \tag{2.5}$$

with probability $1 - 2e^{2\log L - 2\epsilon^2 M}$ with respect to the measure of $\mathbf{A}$ and $\mathbf{w}$. In (2.5), $d_H(\cdot,\cdot)$ is the Hamming distance of the embedded signals, the function $f(\cdot)$ is as specified in (2.4), and $g(d)$ is the map

$$g(d) = \frac{1}{2} - \sum_{i=0}^{+\infty} \frac{e^{-\left(\frac{\pi(2i+1)\sigma d}{\sqrt{2}\Delta}\right)^2}}{(\pi(i+1/2))^2}, \tag{2.6}$$

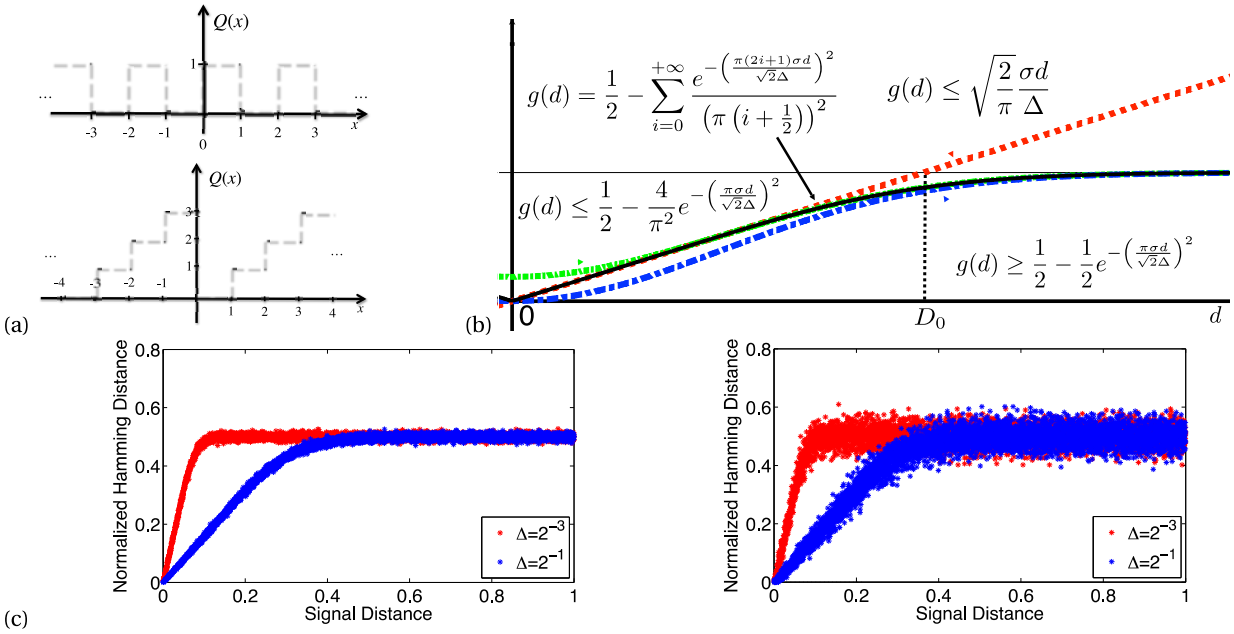

FIGURE 3. (a) This non-monotonic quantization function $Q(\cdot)$ allows for universal rate-efficient scalar quantization. This function is equivalent to using a classical multibit scalar quantizer, and preserving only the least significant bits while discarding all other bits. 1-bit shown on top, multi-bit shown on bottom (b) The embedding map $g(d)$ and its bounds produced by the 1-bit quantization function in (a). (c) Experimental verification of the embedding for small and large Δ in high (left) and low (right) bitrates.

which can be bounded using,

$$g(d) \ge \frac{1}{2} - \frac{1}{2} e^{-\left(\frac{\pi\sigma d}{\sqrt{2}\Delta}\right)^2}, \tag{2.7}$$

$$g(d) \le \frac{1}{2} - \frac{4}{\pi^2} e^{-\left(\frac{\pi\sigma d}{\sqrt{2}\Delta}\right)^2}, \tag{2.8}$$

$$g(d) \le \sqrt{\frac{2}{\pi}}\frac{\sigma d}{\Delta}, \tag{2.9}$$

as shown in Fig. 3(b). The map is approximately linear for small d and becomes a constant equal to $1/2$ exponentially fast for large d, greater than a distance threshold D_0. The slope of the linear section and the distance threshold D_0 is determined by the parameter ratio σ/Δ. In other words, the embedding ensures that the Hamming distance of the embedded signals is approximately proportional to the ℓ_2 distance between the original signals, as long as that ℓ_2 distance was smaller than D_0. Note that a piecewise linear function with slope $\sqrt{\frac{2}{\pi}}\frac{\sigma}{\Delta}$ for $d \le D_0$ and slope equal to zero for $d > D_0$ is a very good approximation to (2.6), in addition to being an upper bound.

To obtain a fixed bound on the probability of failure, the additive ambiguity ϵ in (2.5) scales as $\epsilon \propto 1/\sqrt{M}$, similar to the constant δ in the multiplicative $(1 \pm \delta)$ factor in J-L embeddings. It should be noted, however, that universal embeddings use 1 bit per projection dimension, for a total rate of $R = M$. The trade-off between B and M under constant R exhibited by quantized J-L embeddings does not exist under 1-bit universal embeddings. Still, there is a performance trade-off, controlled by the choice of Δ in (2.4), which is explored in [18] and discussed in subsequent sections.

Figure 3(c) demonstrates experimentally and provides intuition on how the embedding behaves for smaller (red) and larger (blue) Δ and for higher (left) and lower (right) bitrates. The figure plots the embedding (Hamming) distance as a function of the signal distance for randomly generated pairs of signals. The thickness of the curve is quantified by ϵ, whereas the slope of the upward sloping part is quantified by Δ.

Although universal embeddings perform very well in practice, a few theoretical issues have not been resolved to date. One of the most interesting questions is the extension to a multi-bit quantizer. While such an extension, using the quantizer in the bottom of Fig. 3(a) is described in [12], it has not been analyzed and guarantees have not been provided. The techniques used to provide the universal embeddings guarantee can fail or become very tedious for a multi-bit analysis.

Furthermore, the embedding guarantee has been shown in [17] to hold for finite point clouds and not for infinite sets, such as sparse signals or manifolds. In [12] it was shown that a different guarantee, on the distance of signals that map to exactly the same binary vector, can be provided for such sets. However, a general embedding guarantee, similar to the extensions of the Johnson-Lindenstrauss lemma to the RIP [8] and to manifolds [9], does not yet exist. For a number of real-world signals—including, for example, images lying on articulation manifolds—such a guarantee is often desirable, if not necessary.

The development we present in the remainder of this paper addresses both of these issues. Specifically, Sec. 3 provides a general description of how concentration of measure inequalities are typically used to establish the embedding guarantees and how these guarantees can be extended to hold for fairly general embedding designs in infinite sets. It also examines the effect of quantization on the embedding guarantee. Section 4 describes a general embedding design approach which, combined with the development in Sec. 3, provides the desired guarantees for multi-bit universal embeddings, of which binary universal embeddings become a special case. The details are described in Sec. 5, which uses universal embeddings as an example application of the general theory.

Although not immediately relevant to this work, an information-theoretic argument guarantees that using binary universal embeddings while querying a database can preserve the query's privacy [17]. This aspect is not explored in this work for the more general case. Although we are confident that such guarantees can be provided, at least for multi-bit universal embeddings, we defer such a development to a separate publication.

2.3. **Lipschitz Continuity.** A very useful tool in the subsequent development is Lipschitz Continuity. Lipschitz continuity enables us to bound how abruptly a function may vary as its input varies.

Definition 2.2. A function $f : \mathcal{S} \to \mathcal{W}$ is Lipschitz-continuous with Lipschitz constant K, if, for all $\mathbf{x}, \mathbf{y} \in \mathcal{S}$:

$$d_{\mathcal{W}}(f(\mathbf{x}), f(\mathbf{y})) \leq K d_{\mathcal{S}}(\mathbf{x}, \mathbf{y}) \tag{2.10}$$

If a function is Lipschitz continuous with constant K, then it is also Lipschitz continuous for any constant $K' \geq K$. Often, the smallest K that satisfies the Lipscitz property is referred to as the "best" Lipschitz constant for this function.

While a number of functions we consider in the remainder of this paper are Lipschitz continuous, some very interesting ones are not. For example, quantization functions such as the ones in Fig. 3(a), are not Lipschitz continuous. Still, they do exhibit piece-wise continuity properties that we should be able to exploit and characterize. To that end, we introduce a generalization of Lipschitz continuity, which we term *T-part Lipschitz continuity*.

To characterize T-part Lipschitz continuity, we assume that the function $f(\cdot)$ operates in a compact set $\mathcal{S}$ that can be partitioned to T subsets, such that $f(\cdot)$ is Lipschitz continuous when its domain is restricted to each of the T subsets.

Definition 2.3. A function $f : \mathcal{S} \to \mathcal{W}$ is T-part Lipschitz continuous over $\mathcal{S}$ with Lipschitz constant K if there exists a finite partition of $\mathcal{S}$ to T sets $\mathcal{S}_t$, $t = 1, \ldots, T$ such that for all t and for all pairs $\mathbf{x}, \mathbf{y}$ in $\mathcal{S}_t$ the Lipschitz property holds: $d(f(\mathbf{x}), (\mathbf{y})) \leq K d(\mathbf{x}, \mathbf{y})$.

Definition 2.4. A function $f : \mathcal{S} \to \mathcal{W}$ is T-part constant if it is T-part Lipschitz continuous with $K = 0$.

Note that T-part Lipschitz continuity is a much more permissive condition compared to piece-wise continuity, as typically understood and defined. For example, the indicator function of rational numbers

$$I_{\mathbb{Q}}(x) = \begin{cases} 1, & \text{if } x \text{ is rational} \\ 0, & \text{if } x \text{ is irrational} \end{cases} \tag{2.11}$$

is nowhere continuous and definitely not Lipschitz continuous. However, it is 2-part constant, as we can split its domain, $\mathbb{R}$ to two sets $S_1 = \mathbb{Q}$ and $S_2 = \mathbb{R} \setminus \mathbb{Q}$ over which $I_{\mathbb{Q}}(x)$ is constant, i.e., $K = 0$. Of course, the universal quantization functions in Fig. 3(a) are 2^B-part Lipschitz constant, where B is the number of bits available to represent the quantization bins.

A function that is piece-wise continuous with a finite number of pieces, T, is also T-part Lipschitz continuous. However, piece-wise continuous functions with an infinite number of pieces may not fit our definition. A conventional infinite uniform quantizer, for example, is not T-part Lipschitz continuous. If, instead, the quantizer has positive and negative saturation points, then there is a finite number of quantization intervals, and the function becomes T-part constant, with T equal to the number of quantization levels. Typically $T = 2^B$, where B is the number of quantization bits.

We should note that 1-part Lipschitz continuity coincides with the traditional definition of Lipschitz continuity. Furthermore, a T-part Lipschitz continuous function is also $T+1$-part Lipschitz with the same constant K. We use the term "*exactly* T-part Lipschitz", when T represents the minimum number of partitions that can be used to satisfy the definition of T-part Lipschitz continuity.

Definition 2.5. A function $f : \mathscr{S} \to \mathscr{W}$, where $\mathscr{S}$ is a compact set, is *exactly* T-part Lipschitz continuous over $\mathscr{S}$ with Lipschitz constant K if it is T-part Lipschitz continuous with that constant but is not $(T-1)$-part Lipschitz continuous with the same constant.

3. Probabilistic Embedding Construction

As mentioned above, most embedding literature relies on randomized constructions. Typically, such embeddings are demonstrated on finite sets and often extended to cover infinite sets. However the extension is often not trivial. Furthermore, the tools developed in the literature, e.g., [8, 9, 43, 64], are typically specific to each embedding design. Departing from this embedding-specific approach, we now develop general methods that will allow us to extend a wide variety of embedding designs from point clouds to infinite signal sets.

3.1. Embeddings and Concentration of Measure.

Our goal in this paper is to present a fairly general framework to design and analyze embeddings that approximately preserve distances between signals. Typically, such embeddings are designed in a probabilistic manner by drawing the embedding function f randomly from a family of functions. For example, in various constructions of Johnson-Lindenstrauss embeddings or the Restricted Isometry Property (RIP) the function is a linear map $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$, with the elements of $\mathbf{A}$ drawn randomly from a variety of possible distributions [1, 8, 22, 24, 25]. More recently, in quantized and phase embeddings, the linear map is followed by a quantization, phase-extraction, or some other non-linear operation [13–17, 39, 40, 43, 48, 53, 60–62].

Since the embedding function is randomized, we can only prove that the embedding is a (g,δ,ϵ) embedding with high probability. Most proofs rely on concentration of measure arguments to show that (2.1) holds on a pair of points $\mathbf{x},\mathbf{y} \in \mathscr{S}$ with high probability. Typically, the failure probability decays exponentially with the number of measurements, i.e., with the dimensionality of the embedding space $M = \dim(\mathscr{W})$. In other words, the embedding fails on a pair of points with probability bounded by $\Omega(e^{-Mw(\delta,\epsilon)})$, where $w(\delta,\epsilon)$ is an increasing function of ϵ and δ that quantifies the concentration of measure exhibited by the randomized construction.

Once the embedding guarantee is established for a pair of signals, a union bound can be used to extend it to a finite set of signals. If the set $\mathscr{S}$ is finite, containing Q points, then the probability that the embedding fails is upper bounded by $\Omega(Q^2 e^{-Mw(\delta,\epsilon)}) = \Omega(e^{2\log Q - Mw(\delta,\epsilon)})$, which decreases exponentially with M, as long as $M = O(\log(Q))$.

Unfortunately, this union bounding approach does not work for infinite sets, such as signal spaces or sparse signals. Instead, to establish (2.1) for such sets, a covering of the set is constructed using a finite number of ϵ-balls. The concentration of measure is established for the centers of balls and is then extended to all points of the balls using the continuity properties of the embedding map. For example, [8] exploits the properties of the distance metric to establish the RIP for all K-sparse signals. This approach can be used if the distance map is the identity, $g(d) = d$, but does not generalize very well. In the next section, we describe a more general approach that can be used for arbitrary Lipschitz-continuous distance and embedding maps. More recent work exploits a chaining argument [70] to tighten the bounds for linear embeddings [19, 28, 31, 63]. Whether a chaining approach can be used to improve the bounds for more general embeddings, such as the ones described here, is an interesting and open question.

3.2. Embedding of Infinite Sets Using Continuous Maps.

To exploit Lipschitz continuity, we start with the randomized embeddings as described in the previous section, i.e., for which we can show that given a pair of points $\mathbf{x},\mathbf{y} \in \mathscr{S}$, the embedding guarantee (2.1) holds with probability greater than $1 - ce^{-Mw(\delta,\epsilon)}$. We also assume the embedding map $g(d)$ is Lipschitz-continuous with constant K_g and that the embedding map $f(\cdot)$ is Lipschitz-continuous with constant K_f. Next, we use $C_\epsilon^{\mathscr{S}}$ to denote the covering number of the signal set $\mathscr{S}$, i.e., the smallest

number of points $\mathbf{q} \in \mathscr{S}$ s.t. for all $\mathbf{x} \in \mathscr{S}$, $\inf_{\mathbf{q}} \|\mathbf{x}-\mathbf{q}\| \leq \varepsilon$. Its logarithm, $E_\varepsilon^{\mathscr{S}} = \log C_\varepsilon^{\mathscr{S}}$ is the Kolmogorov, or metric, ε-entropy of the set. Appendix A proves the following.

Theorem 3.1. *Consider a signal set $\mathscr{S}$ with r-covering number $C_r^{\mathscr{S}}$ and a (g,δ,ϵ) probabilistic embedding to an M-dimensional space that fails with probability smaller than $ce^{-Mw(\delta,\epsilon)}$ on a pair of points. If the embedding map $f(\cdot)$ is K_f-Lipschitz continuous and the distance map $g(\cdot)$ is K_g-Lipschitz continuous, then for some $\alpha > 0$ the embedding is a $(g,\delta,\epsilon+\alpha)$ embedding that holds with probability greater than $1-ce^{2E_r^{\mathscr{S}}-Mw(\delta,\epsilon)}$ on all pairs of signals $\mathbf{x},\mathbf{y}$ in $\mathscr{S}$, where $r = \frac{\alpha}{(1+\delta)2K_g+2K_f}$.*

As typical with such proofs, the constants are difficult to pin down accurately. However, the main takeaway is that the embeddings preserve distances of an infinite set as long as the number of measurements M is on the order of the Kolmogorov entropy of the set for a radius that depends on the desired accuracy. For example, if the set is bounded-norm signals in $\mathbb{R}^N$, then $E_r^{\mathscr{S}} = O(N\log(1+2/r))$. For bounded-norm K-sparse signals in $\mathbb{R}^N$, $E_r^{\mathscr{S}} = O(K\log(N/rK))$. An extensive discussion on Kolmogorov entropy and other set complexity measures can be found, for example, in [73].

We should note that this theorem introduces an additive ambiguity α, even if the original embedding has $\epsilon = 0$. For general embedding maps, we do not believe this additive ambiguity can be eliminated. In the special case of linear embedding functions $f(\cdot)$ and linear distance maps $g(\cdot)$ the additive constant can be eliminated using proof techniques such as the ones in, e.g., [8, 10].

3.3. **Embedding of Infinite Sets Using Discontinuous Maps.** To extend the mapping to discontinuous embeddings we separate the randomized embedding $f(\cdot)$ into its components $f_m(\cdot)$, $m = 1,\ldots,M$, and examine how the embedding behaves on balls $\mathscr{B}_{r/2}(\mathbf{x})$ of diameter r, i.e., radius $r/2$. In particular, we first examine the behavior of each function component $f_m(\cdot)$ with domain restricted to a given ball $\mathscr{B}_{r/2}(\mathbf{x})$ with respect to its T-part Lipschitz continuity.

Assuming that each function $f_m : \mathscr{B}_{r/2}(\mathbf{x}) \to \mathbb{R}$ is exactly T_m-part Lipschitz over the ball, then we can partition the ball into T_m sets S_{t_m}, $t_m = 1,\ldots,T_m$, over which the $f_m(\cdot)$ is Lipschitz continuous. We can then define $S_{t_1,\ldots,t_M} = S_{t_1} \cap \ldots \cap S_{t_M}$, which is a partition of $\mathscr{B}_{r/2}(\mathbf{x})$ of $T_1 \times \ldots \times T_M$ sets such that all f_m are Lipschitz continuous over each $S_{t_1,\ldots,t_M}$.

Next, we need to quantify T_m for each m. Since the embedding is randomized, we use P_T to denote the probability that $f_m(\cdot)$ is exactly T-part Lipschitz over a ball $\mathscr{B}_{r/2}(\mathbf{x})$. We assume that the probability that $f_m(\cdot)$ is exactly T-part Lipschitz over $\mathscr{B}_{r/2}(\mathbf{x})$ is independent from the probability that $f_{m'}(\cdot), m \neq m'$ is T-part Lipschitz over the same $\mathscr{B}_{r/2}(\mathbf{x})$. This, for example, is ensured if $f_m(\cdot)$ is drawn independently for each m, as is typically the case. We assume that the probability P_T is a decreasing function of the radius r of the ball, and does not depend on the selection of the ball center $\mathbf{x}$, or on any other ball parameter or property. We also select a maximum $T_{\max}$ beyond which the probability that a function $f_m(\cdot)$ is exactly T-part Lipschitz continuous, is negligible or zero.

Theorem 3.2. *Consider a signal set $\mathscr{S}$ with r-covering number $C_r^{\mathscr{S}}$ and a (g,δ,ϵ) probabilistic embedding to an M-dimensional space that fails with probability smaller than $ce^{-Mw(\delta,\epsilon)}$ on a pair of points. If each coordinate of the embedding map $f_m(\cdot)$ is exactly T-part Lipschitz continuous with probability less than P_T, as described above, and the distance map $g(\cdot)$ is K_g-Lipschitz continuous, then the embedding is a $(g,\delta,\epsilon+\alpha)$ embedding that with probability greater than $1-(ce^{2E_{r/2}^{\mathscr{S}}+c_1M-Mw(\delta,\epsilon)} + T_{\max}e^{-2c_0^2M} + P_F)$ holds on all pairs of signals $\mathbf{x},\mathbf{y}$ in $\mathscr{S}$, for any $T_{\max}$, where $r = \frac{\alpha}{(1+\delta)2K_g+2K_f}$, $P_F = \sum_{T=T_{\max}+1}^{\infty} P_T$, $c_1 = \sum_{T=2}^{T_{\max}} P_T(1+c_0)\log T$, and any $\alpha < 1$.*

The proof details can be found in Appendix B.

Assuming P_F is negligible, the embedding fails with probability at most $ce^{2E_{r/2}^{\mathscr{S}}+c_1M-Mw(\delta,\tilde{\epsilon}-\alpha)} + T_{\max}e^{-2c_0^2M} + P_F$, which decays exponentially with M as long as $c_1 < w(\delta,\epsilon)$ and $M = O(E_{r/2}^{\mathscr{S}})$. Note that c_1 depends on P_T and, therefore, on r. The inequality holds for sufficiently small r since c_1 decreases with r for most randomized embedding constructions.

While the theorem provides the guarantee with significant generality, in many practical embedding designs the theorem implies that guarantees established on point clouds and discontinuous maps can be extended to continuous sets with only a small constant oversampling penalty, compared to the continuous maps analyzed in the previous section. Section 5.2.2 and Appendix E analyze such an example.

3.4. **Quantized Embeddings.** Although quantization of the embedding can be analyzed using the framework we describe above, it is often more convenient, especially in the case of high-rate quantization to consider it separately, as an additional step after the projection.

We examine a (g,δ,ϵ) embedding which is subsequently quantized using an M-dimensional vector quantizer $Q(\cdot)$. We assume the quantization error is bounded, i.e., $d(Q(\mathbf{x}),\mathbf{x}) \le E_Q$. The triangle inequality, $|d_{\mathscr{W}}(f(\mathbf{x}),f(\mathbf{w})) - d_{\mathscr{W}}(Q(f(\mathbf{x})),Q(f(\mathbf{w})))| \le 2E_Q$, implies that the quantized embedding guarantee becomes a $(g,\delta,\epsilon+2E_Q)$ embedding, with guarantee

$$(1-\delta)g\left(d_{\mathscr{S}}(\mathbf{x},\mathbf{y})\right) - \epsilon - 2E_Q \le d_{\mathscr{W}}\left(Q(f(\mathbf{x})),Q(f(\mathbf{y}))\right) \le (1+\delta)g\left(d_{\mathscr{S}}(\mathbf{x},\mathbf{y})\right) + \epsilon + 2E_Q. \tag{3.1}$$

Theorem 3.3. *Consider a (g,δ,ϵ) embedding $f(\cdot)$ and a quantizer $Q(\cdot)$ with worst case quantization error E_Q, then the quantized embedding, $Q(f(\cdot))$, is a $(g,\delta,\epsilon+2E_Q)$ embedding.*

In the specific case of a uniform scalar quantizer with quantization interval Δ, the M-dimensional quantization ℓ_2 error is bounded by $E_Q \le \sqrt{M}\Delta/2$, assuming the quantizer is designed such that it does not saturate or such that the saturation error is negligible. The interval of the quantizer is a function of the number of bits B used per coefficient $\Delta = 2^{-B+1}S$, where S is the saturation level of the quantizer. Given a fixed rate to be used by the embedding, $R = MB$, the guarantee becomes

$$\begin{aligned}(1-\delta)g\left(d_{\mathscr{S}}(\mathbf{x},\mathbf{y})\right) - \epsilon - 2^{-\frac{R}{M}+1}\sqrt{M}S \\ \le \|Q(f(\mathbf{x})) - Q(f(\mathbf{y}))\|_2 \le \\ (1+\delta)g\left(d_{\mathscr{S}}(\mathbf{x},\mathbf{y})\right) + \epsilon + 2^{-\frac{R}{M}+1}\sqrt{M}S.\end{aligned} \tag{3.2}$$

Note that the $\sqrt{M}$ factor can often be removed, depending on the normalization of the embedding.

Of course, ℓ_2 is not always the appropriate fidelity metric. If the $d_{\mathscr{S}}(\cdot,\cdot)$ corresponds to the ℓ_1 distance, the quantization error is bounded by $E_Q \le M\Delta/2$. Again, with care in the normalization the M factor can be removed. If, instead, the ℓ_∞ norm is desired, the quantization error is bounded by $E_Q \le \Delta/2$.

One of the issues to consider in designing quantized embeddings using a uniform scalar quantizer is the trade-off between the number of bits per dimension and the total number of dimensions used. Since $R = MB$, increasing the number of bits per dimension B under a fixed bit budget R, requires decreasing the number of dimensions M. While the former reduces the error due to quantization, the latter will typically increase the uncertainty in the embedding by increasing δ and ϵ.

In the case of randomized embeddings, this trade-off can be quantified through the function $w(\epsilon,\delta)$. Given a fixed probability lower bound to guarantee the embedding, then $M = \Omega(1/w(\epsilon,\delta))$. Since $w(\cdot,\cdot)$ is an increasing function of ϵ and δ, which quantify the ambiguity of the embedding, reducing M increases this ambiguity. On the other hand, the quantization ambiguity, quantified in $2^{-R/M+2}S\sqrt{M}$ decreases with M. This trade-off is explored, for example, in the context of quantized J-L embeddings in [48, 66]. Although we describe the trade-off for uniform scalar quantizers, the same issue exists for non-uniform quantizers and for vector quantizers, manifested with different constants but with the same order of magnitude effects (e.g., see [42]).

Sec. 5 provides examples of quantized embeddings, examining cases where the quantization is analyzed as described here, i.e., as an additional step after the embedding is performed, as well as cases in which quantization is analyzed as part of the embedding through the mechanism of T-part Lipschitz functions.

4. Embedding Design and Performance Analysis

4.1. **Embedding Design Using a Periodic Map.** Having provided all the necessary tools to establish the embedding properties over a signal set, we next consider a fairly general embedding design. Specifically, we consider the mapping $\mathbf{y} = h(\mathbf{Ax}+\mathbf{w})$, where the rows $\mathbf{a}_i$ of $\mathbf{A}$ are randomly chosen from some i.i.d. vector distribution and the elements of $\mathbf{w}$ are chosen from an i.i.d. distribution uniform in $[0,1)$. We denote the projection through $\mathbf{A}$ using $\mathbf{u} = \mathbf{Ax}$.

We restrict our attention to functions $h_g(t)$ with finite support, restricted in $[0,1)$ without loss of generality, and their periodic extension $h(t)$ with period 1, i.e., such that $h(t) = h_g(t)$ for $t \in [0,1)$ and $h(t) = h(t+1)$. We use $H_g(\xi)$ and $H(\xi)$ to denote their respective Fourier transform, $R_{h_g}(\tau)$ and $R_h(\tau)$ to denote their deterministic

autocorrelations, and $P_{h_g}(\xi)$ and $P_h(\xi)$ to denote their power spectrum. Their relationship is summarized below.

$$h_g(t) = 0 \text{ if } t \notin [0,1) \overset{\mathscr{F}}{\longleftrightarrow} H_g(\xi) \tag{4.1}$$

$$h(t) = \sum_{k=-\infty}^{+\infty} h_g(t-k) \overset{\mathscr{F}}{\longleftrightarrow} H(\xi) = H_g(\xi) \sum_{k=-\infty}^{+\infty} \delta(\xi - k) = \sum_{k=-\infty}^{+\infty} H_k \delta(\xi - k) \tag{4.2}$$

$$R_{h_g}(\tau) = \int_{-\infty}^{+\infty} h_g(t) h_g(t-\tau)\, dt \overset{\mathscr{F}}{\longleftrightarrow} P_{h_g}(\xi) = |H(\xi)|^2 \tag{4.3}$$

$$R_h(\tau) = \int_0^{+1} h(t) h(t-\tau)\, dt \overset{\mathscr{F}}{\longleftrightarrow} P_h(\xi) = |H_g(\xi)|^2 \sum_{k=-\infty}^{+\infty} \delta(\xi - k) = \sum_{k=-\infty}^{+\infty} |H_k|^2 \delta(\xi - k), \tag{4.4}$$

where $H_k = H_g(k)$ denotes the Fourier series coefficients of $h(t)$. Note that, to avoid convergence issues, the autocorrelation of periodic functions is defined as the integral over a single period, in contrast to the finite-support autocorrelation defined as the integral over all $\mathbb{R}$. Although we use the same notation for simplicity, the appropriate use should be clear from the context. We further assume that $h(t)$ is bounded and denote its range using $\bar{h} = \sup_t h(t) - \inf_t h(t)$.

We first examine the behavior of a single coefficient of $\mathbf{y}$, i.e., $y = h(\langle \mathbf{a}, \mathbf{x} \rangle + w)$, where $\mathbf{a}$ is the corresponding row of $\mathbf{A}$ and w the corresponding coefficient of $\mathbf{w}$. If we measure a pair of signals $\mathbf{x}$ and $\mathbf{x}'$ at distance $d = d_{\mathscr{S}}(\mathbf{x} - \mathbf{x}')$ apart, their (signed) projected distance, denoted $l = \langle \mathbf{a}, \mathbf{x} - \mathbf{x}' \rangle$, is a random variable with density conditioned on d denoted using $f_l(\cdot|d)$ and characteristic function denoted using $\phi_l(\xi|d)$. Then, we can prove the following theorem.

Theorem 4.1. *Consider a set $\mathscr{S}$ of Q points in $\mathbb{R}^N$, measured using $\mathbf{y} = h(\mathbf{A}\mathbf{x} + \mathbf{w})$, with $\mathbf{A}$, $\mathbf{w}$, and $h(t)$ as above. Given $\epsilon > 0$, with probability greater than $1 - e^{2\log Q - 2M\frac{\epsilon^2}{\bar{h}^4}}$ the following holds*

$$g(d) - \epsilon \leq \frac{1}{M} \left\| \mathbf{y} - \mathbf{y}' \right\|_2^2 \leq g(d) + \epsilon \tag{4.5}$$

for all pairs $\mathbf{x}, \mathbf{x}' \in \mathscr{S}$ and their corresponding measurements $\mathbf{y}, \mathbf{y}'$, where

$$g(d) = 2 \sum_k |H_k|^2 (1 - \phi_l(2\pi k | d)). \tag{4.6}$$

defines the distance map of the embedding.

The proof is provided in Appendix C. Note that $\phi_l(0) = 1$ for any distribution. Thus, the DC component H_0 of the embedding map does not affect the distance map in (4.6).

To show that this is a $(g, 0, \epsilon)$ embedding, we derive a bound on the ℓ_2 distance, instead of its square, which follows easily from the fact that $\sqrt{x \pm \epsilon} \lessgtr \sqrt{x} \pm \sqrt{\epsilon}$:

Corollary 4.1. *Consider the signal set $\mathscr{S}$, defined and measured as in Thm. 4.1. Given $\epsilon > 0$, with probability greater than $1 - e^{2\log Q - 2M\left(\frac{\epsilon}{\bar{h}}\right)^4}$ the following holds*

$$\widetilde{g}(d) - \epsilon \leq \frac{1}{\sqrt{M}} \left\| \mathbf{y} - \mathbf{y}' \right\|_2 \leq \widetilde{g}(d) + \epsilon \tag{4.7}$$

for all pairs $\mathbf{x}, \mathbf{x}' \in \mathscr{S}$ and their corresponding measurements $\mathbf{y}, \mathbf{y}'$, where $\widetilde{g}(d) = \sqrt{g(d)}$.

In a similar manner, we can establish that $E\{\|y\|_2^2\} = \sum_k |H_k|^2$ and, therefore,

$$\sum_k |H_k|^2 - \epsilon \leq \frac{1}{M} \|y\|_2^2 \leq \sum_k |H_k|^2 + \epsilon, \tag{4.8}$$

with probability greater than $1 - 2e^{\log Q - 2M\frac{\epsilon^2}{\bar{h}}}$. The proof is similar to the proof of Thm. 4.1, with the differences discussed in App. C. We omit it here for brevity. However, we should note that, to ensure both (4.5) and (4.8) hold, the union bound should be taken over both the point pairs and the points in $\mathscr{S}$. Thus, the probability that both hold is bounded by $1 - 2e^{2\log Q - 2M\frac{\epsilon^2}{\bar{h}}}$. We should also note that in certain cases, such as binary embeddings, the value of $\|y\|_2^2$ can be exactly computed [16].

Of course, using the results of the Section 3 with $w(\epsilon) = 2(\epsilon/\bar{h})^4$ or $w(\epsilon) = 2\epsilon^2/\bar{h}^4$ we can trivially establish the embedding over infinite sets.

4.2. **Projection Randomization and the Distance Map.** There are several choices for the randomization of the embedding projection matrix **A**, which result in interesting properties of the embedding map and the distances preserved. Specifically, the distance $d = d_{\mathscr{S}}(\mathbf{x}-\mathbf{x}')$ of two signals affects l according to the characteristic function $\phi_l(\xi|d)$. By selecting the appropriate distribution on **A**, the map can be designed to preserve a variety of distances. We examine two particularly useful examples, preserving maps of ℓ_2 and ℓ_1 distances. The subsequent discussion, as well as the discussion in Sec. 4.5, exploits and generalizes results in [12, 65].

4.2.1. *Mapping ℓ_2 distances.* If elements of **A** are chosen from an i.i.d. Normal distribution with variance σ^2 then l is a normally distributed random variable with variance $(d_{\ell_2}\sigma)^2$, where $d_{\ell_2} = \|\mathbf{x}-\mathbf{x}'\|_2$. In other words, it is natural to set the distance $d_{\mathscr{S}}(\cdot,\cdot)$ above as the ℓ_2 distance. In that case, the characteristic function is that of a normal distribution $\phi_l(\xi|d) = \phi_{\mathcal{N}(0,\sigma^2 d^2)}(\xi) = e^{-\frac{1}{2}(\sigma d\xi)^2}$, and the distance map becomes

$$g(d) = 2\sum_k |H_k|^2(1-e^{-2(\pi\sigma dk)^2}), \tag{4.9}$$

with d measuring the ℓ_2 distance.

4.2.2. *Mapping ℓ_1 distances.* If, instead, elements of **A** are drawn from an i.i.d. Cauchy distribution with zero location parameter and scale parameter γ, i.e., density

$$f_a(x) = \frac{1}{\pi\gamma}\left[\frac{\gamma^2}{x^2+\gamma^2}\right], \tag{4.10}$$

with corresponding characteristic function $\phi_a(\xi) = e^{-\gamma|\xi|}$, then l is a sum of independent Cauchy-distributed random variables. It is straightforward to show that the resulting characteristic function is a function of the ℓ_1 distance of the two signals, $d_{\ell_1} = \|\mathbf{x}-\mathbf{x}'\|_1$. In particular,

$$\phi_l(\xi|d_{\ell_1}) = e^{-\gamma d_{\ell_1}|\xi|}, \tag{4.11}$$

and the corresponding distance map becomes

$$g(d) = 2\sum_k |H_k|^2(1-e^{-2\pi\gamma dk}), \tag{4.12}$$

with d in this case measuring the ℓ_1 distance. This enables direct embedding of an ℓ_1 space to an ℓ_2 space, in contrast to solutions that first map ℓ_1 to a much larger ℓ_2 space through a "unary" expansion and then embed the mapping, as done, for example, in [49, 66].

It is important to note that JL-type ℓ_1 embeddings have been shown not to be possible [20]. A key reason is that concentration of measure results cannot be established for heavy-tailed distributions, such as the Cauchy distribution we use here. However, our result can be established because of the bounded distortion function $h(\cdot)$, which concentrates the distribution of the measurements. We should also note that the results above can be extended to matrix elements drawn from any α-stable distribution, thus implementing embeddings of ℓ_p distances other than ℓ_1 and ℓ_2, in a manner similar to [58].

4.3. **Properties of the distance map.** Although it would be desirable to be able to generate any distance map desired, the properties of distance computation impose constraints on the distance maps that are possible. In particular, within the ϵ and δ error bounds of the embedding, the distance map should be subadditive.

Definition 4.2. A function $g(x)$ is (ϵ,δ)-subadditive for all a, b in its domain: $(1-\epsilon)g(a+b)-\delta \le g(a)+g(b)$

For the case of the distance map $g(d)$ in a (g,δ,ϵ) embedding, the following proposition is proven in Appendix D:

Proposition 4.3. *Any distance map $g(\cdot)$ satisfying Def. 2.1 for a convex set $\mathscr{S}$, is $(2\epsilon, 3\delta)$-subadditive.*

The subadditivity of the distance map imposes constraints on the distance maps that are achievable with such a scheme. For example, a distance map that is small for a range of distances cannot become large immediately after; it is easy to show that for some positive a, if $g(d) < a$ for $d < d_0$, then $g(d) < (2a+\delta)/(1-\epsilon)$ for $d \le 2d_0$.

In addition to possessing the subadditivity property, distance maps $g(d)$ designed in this section are comprised of linear combinations of increasing functions of d bounded by 1. Thus, they are bounded by

$$g(d) \le \lim_{d\to\infty} g(d) = 2\sum_k |H_k|^2 = 2\int_0^1 |h(t)|^2\,dt = 2R_h(0) \tag{4.13}$$

Since the square root is also a monotonic function, $\sqrt{g(d)}$, used in Cor. 4.1, is also increasing and bounded by $\sqrt{2R_h(0)}$. However, it remains an open question whether all distance maps satisfying (2.1) should satisfy some monotonicity constraint.

4.4. **Error Analysis.** To understand the performance of an embedding in distance computation and to guide our design we want to understand how well the embedding captures the distance. The main question is: given a distance $d_{\mathscr{W}}$ between two embedded signals in the embedding space $\mathscr{W}$, how confident are we about the corresponding distance between the original signals in the signal space $\mathscr{S}$? The function $g(\cdot)$ captures how distance is mapped and can be inverted to approximately determine the distance $d_{\mathscr{S}}$ in the signal space. On the other hand, the constants δ and ϵ capture the ambiguity in the opposite direction, i.e., the ambiguity in the embedding space given the distance in the signal space. Pictorially, taking Fig. 3(c) as an example, (2.1) characterizes the thickness of the curves taking a vertical slice of the plots, while we are now interested in the thickness revealed by taking a horizontal slice instead.

To capture the desired ambiguity, we can reformulate the embedding guarantees as

$$g^{-1}\left(\frac{d_{\mathscr{W}}\left(f(\mathbf{x}),f(\mathbf{y})\right)-\epsilon}{(1+\delta)}\right)\leq d_{\mathscr{S}}(\mathbf{x},\mathbf{y})\leq g^{-1}\left(\frac{d_{\mathscr{W}}\left(f(\mathbf{x}),f(\mathbf{y})\right)+\epsilon}{(1-\delta)}\right), \tag{4.14}$$

which for small δ and ϵ can be approximated using the Taylor expansion of $1/(1\pm\delta)$:

$$g^{-1}\left(\left(d_{\mathscr{W}}\left(f(\mathbf{x}),f(\mathbf{y})\right)-\epsilon\right)(1-\delta)\right)\lesssim d_{\mathscr{S}}(\mathbf{x},\mathbf{y})\lesssim g^{-1}\left(\left(d_{\mathscr{W}}\left(f(\mathbf{x}),f(\mathbf{y})\right)+\epsilon\right)(1+\delta)\right), \tag{4.15}$$

Assuming that $g(\cdot)$ is differentiable, we can approximate the inequality using the Taylor expansion of $g^{-1}(\cdot)$ around $d_{\mathscr{W}}\left(f(\mathbf{x}),f(\mathbf{y})\right)$ and the fact that $(g^{-1})'(x)=1/g'(g^{-1}(x))$. Ignoring the second order term involving $\epsilon\cdot\delta$, and defining the signal distance estimate $\widetilde{d}_{\mathscr{S}}=g^{-1}\left(d_{\mathscr{W}}\left(f(\mathbf{x}),f(\mathbf{y})\right)\right)$ we obtain

$$\widetilde{d}_{\mathscr{S}}-\frac{\epsilon+\delta\, d_{\mathscr{W}}\left(f(\mathbf{x}),f(\mathbf{y})\right)}{g'\left(\widetilde{d}_{\mathscr{S}}\right)}\lesssim d_{\mathscr{S}}(\mathbf{x},\mathbf{y})\lesssim\widetilde{d}_{\mathscr{S}}+\frac{\epsilon+\delta\, d_{\mathscr{W}}\left(f(\mathbf{x}),f(\mathbf{y})\right)}{g'\left(\widetilde{d}_{\mathscr{S}}\right)}. \tag{4.16}$$

In other words, given the distance $d_{\mathscr{S}}$ between two signals in the signal space and using $\widetilde{d}_{\mathscr{S}}$ to denote the estimate of this distance, the ambiguity is less than

$$\left|d_{\mathscr{S}}(\mathbf{x},\mathbf{y})-\widetilde{d}_{\mathscr{S}}\right|\lesssim\frac{\epsilon+\delta\, d_{\mathscr{W}}\left(f(\mathbf{x}),f(\mathbf{y})\right)}{g'\left(\widetilde{d}_{\mathscr{S}}\right)}. \tag{4.17}$$

Thus, ambiguity decreases by decreasing δ or ϵ, or by increasing the slope of the mapping.

This ambiguity locally characterizes the performance of the embedding for any particular pairs of points $\mathbf{x}$, and $\mathbf{y}$, and their corresponding distance. The smaller the ambiguity is, the better the embedding for these two points. Increasing the embedding dimension or the bitrate of the embedding, other things being equal, decreases δ and ϵ, and, thus, improves the embedding quality.

Similarly, changing the distance map to locally increase its gradient, also locally improves the embedding, although it might make the embedding guarantee deteriorate for different distances. Note that simply scaling the embedding map, which similarly scales the distance map and its gradient, will not decrease the ambiguity; the scaling will commensurately scale the distances on which this ambiguity applies. For example, an ambiguity of ϵ at distance d is equivalent to ambiguity $\epsilon/2$ at distance $d/2$.

4.5. **Embeddings of Kernel Inner Products.** Randomized projections, as first demonstrated in [65], can be used to approximate kernel inner products for some shift-invariant kernels. In addition to preserving distances, the embeddings we describe also provide a more general kernel approximation approach, generalizing the results in [65].

The inner product of the measurements $\langle\mathbf{y},\mathbf{y}'\rangle$ can be derived from the ℓ_2^2 difference of the measurements, $\|\mathbf{y}-\mathbf{y}'\|_2^2$. Specifically,

$$\|\mathbf{y}-\mathbf{y}'\|_2^2=\|\mathbf{y}\|_2^2+\|\mathbf{y}'\|_2^2-2\langle\mathbf{y},\mathbf{y}'\rangle\Longrightarrow\langle\mathbf{y},\mathbf{y}'\rangle=\frac{\|\mathbf{y}\|_2^2+\|\mathbf{y}'\|_2^2-\|\mathbf{y}-\mathbf{y}'\|_2^2}{2}. \tag{4.18}$$

Thus, if $d_{\mathscr{W}}(\mathbf{y},\mathbf{y}')=\|\mathbf{y}-\mathbf{y}'\|_2^2$ in Def. 2.1, and substituting (4.5) and (4.8) in (4.18), we can show that the embedding can be designed to approximate a kernel with accuracy $3\epsilon/2$ and with probability of failure bounded by $1-2e^{2\log Q-2M\frac{\epsilon^2}{h}}$, as noted in Sec. 4.1. Using a simple substitution $\epsilon'=3\epsilon/2$, the following theorem follows.

Theorem 4.4. *Consider a set $\mathscr{S}$ of Q points in $\mathbb{R}^N$, measured using $\mathbf{y} = h(\mathbf{A}\mathbf{x}+\mathbf{w})$, with $\mathbf{A}$, $\mathbf{w}$, and $h(t)$ as above. With probability greater than $1-2e^{2\log Q - \frac{\delta}{9}M\frac{\epsilon^2}{h^4}}$ the following holds*

$$K(d)-\epsilon \le \frac{1}{M}\langle \mathbf{y},\mathbf{y}'\rangle \le K(d)+\epsilon \tag{4.19}$$

for all pairs $\mathbf{x},\mathbf{x}' \in \mathscr{S}$ and their corresponding measurements $\mathbf{y},\mathbf{y}'$, where

$$K(d) = \sum_k |H_k|^2 \phi_l(k|d). \tag{4.20}$$

defines the kernel of the embedding.

5. Embedding Examples

5.1. **Quantized J-L Embeddings.** As a first example, we analyze quantized J-L embeddings. In classical J-L embeddings, the distance map is the identity, i.e., $g(d) = d$, and $\epsilon = 0$. Starting with classical J-L embeddings, using the development in Sec. 3.4, we can derive the quantized J-L embedding guarantees described in [48]. In this case, the scaling of the embedding allows the removal of the $\sqrt{M}$ term from (3.2):

$$(1-\delta)\|\mathbf{x}-\mathbf{y}\|_2 - 2^{-\frac{R}{M}+1}S \le \|Q(f(\mathbf{x}))-Q(f(\mathbf{y}))\|_2 \le (1+\delta)\|\mathbf{x}-\mathbf{y}\|_2 + 2^{-\frac{R}{M}+1}S. \tag{5.1}$$

Since $g(d) = d$, which has constant slope equal to 1, the denominator in (4.17) is constant. To reduce the ambiguity, a system designer should reduce the numerator as much as possible. To do so, as discussed in [48], the designer confronts the trade-off between the size of δ and ϵ. The former is controlled by the dimensionality of the projection, M, while the latter by the bit-rate per dimension, B. The greater M is, the smaller δ is. Similarly, the greater B is, the smaller ϵ is.

As we mention above, the total bit-rate of the embedding is equal to $R = MB$. In order to best use a given rate, the system designer should explore the trade-off between fewer projection dimensions at more bits per dimension and more projection dimensions at fewer bits per dimension. This trade-off is explored in detail in [48], where it is shown that, in the image retrieval application considered, the best performance is achieved using $B = 3$ or 4 bits per dimension and $M = R/3$ or $R/4$ dimensions, respectively. The performance of the two choices is virtually indistinguishable and significantly better than previous 1-bit approaches [51, 75], which use $B = 1$, $R = M$.

5.2. **Binary Universal Embeddings.** Universal Embeddings provide a more comprehensive example of the development and analysis above. These embeddings are computed using (2.4) with the periodic quantizer shown in Fig. 3(a). With appropriate scaling on the period of the quantizer, these embeddings satisfy exactly the conditions of Thm. 4.1.

5.2.1. *Embedding Map.* The special case of binary universal embeddings has been extensively studied in [12, 16–18], providing distance embedding results and kernel approximation guarantees. These results, using the development in Sec. 4, become a special case of Thms. 4.1 and 4.4. Specifically, the quantizer $Q(x)$ in Fig. 3(a) has period 1, i.e., $\widetilde{Q}(x) = Q(2x)$ has the correct period. The scaling can be incorporated in the generation of A, i.e., (2.4) becomes

$$Q\left(2\left(\frac{1}{2}\Delta^{-1}\mathbf{A}\mathbf{x}+\frac{1}{2}\Delta^{-1}\mathbf{w}\right)\right) = \widetilde{Q}\left(\widetilde{\mathbf{A}}\mathbf{x}+\widetilde{\mathbf{w}}\right), \tag{5.2}$$

where $\widetilde{\mathbf{A}}$ is drawn i.i.d. Gaussian with variance $(\sigma/2\Delta)^2$ and $\widetilde{\mathbf{w}}$ drawn i.i.d. uniform in $[0,1)$. The distance mapping in (2.6) follows from (4.9) and the Fourier series of the periodic square wave $H_k = \frac{\sin(\pi k/2)}{\pi k}$, exploiting the fact that $|\sin(\pi k/2)|^2$ is 0 for k even, and 1 for k odd. Note, that we are using the Fourier series of a shifted quantizer compared to the one in Fig. 3(a)., i.e., $Q'(x) = Q(x+1/2)$, because it is symmetric and has a real Fourier series. However, the shift is inconsequential in the result because of the dither $\mathbf{w}$.

Using the reformulation above, the distance embedding in [17, Thm. 3.2] and the kernel approximation in [16, Prop. 3.1] follow trivially from Thms. 4.1 and 4.4, respectively. We should also note that the kernel guarantee is slightly tighter in [16, Prop. 3.1], exploiting the fact that a binary embedding with $\mathbf{y}$ taking values in $\{-1,1\}$ has deterministic norm $\|\mathbf{y}\|_2^2 = M$ and not random as is the general case for Thm. 4.4.

Moreover, in addition to verifying existing results, the development above provides several generalizations. For example, embeddings using a matrix with elements drawn from an i.i.d. Cauchy distribution, as described in Sec. 4.2, map the ℓ_1 distance onto the hamming space. Again, starting with (5.2) and $\mathbf{A}$ drawn from a Cauchy

distribution with scale parameter γ, $\widetilde{\mathbf{A}}$ is Cauchy distributed with scale parameter $\gamma/2\Delta$. Thus, the embedding map becomes

$$g(d) = \frac{1}{2} - \sum_{i=0}^{+\infty} \frac{e^{-\frac{(2i+1)\pi\gamma d}{\Delta}}}{(\pi(i+1/2))^2}, \tag{5.3}$$

where d is the ℓ_1 distance.

Of course, other mappings are possible by appropriately constructing the projection matrix $\mathbf{A}$, but we omit them here for brevity.

5.2.2. *Extension to Infinite Sets.* More importantly, using the development in Sec. 3.3 and Thm. 3.2, it is possible to guarantee binary universal embeddings on infinite sets, extending the results on finite point clouds established to-date. In the context of Thm. 3.2, the embedding is satisfied with $c = 1$, $\delta = 0$ and $w(\epsilon) = 2\epsilon^2$. However, the embedding map $Q(\mathbf{Ax}+\mathbf{w})$ is discontinuous and, therefore, we need to examine its T-part Lipschitz continuity property.

In particular, since the scalar binary quantizer $Q(\cdot)$ only takes values in $\{0,1\}$, the embedding function is, at most, exactly 2-part Lipschitz continuous with constant K_f equal to 0. In the context of Thm. 3.2, we can bound $P_2 \le 1$, i.e., $c_1 \le (1+c_0)\log 2$, and set $P_F = 0$. Thus the probability that the embedding does not hold is upper bounded by

$$ce^{2E^{\mathscr{S}}_{r/2}+c_1M-Mw(\delta,\epsilon)} + T_{\max}e^{-2c_0^2M} + P_F = e^{2E^{\mathscr{S}}_{r/2}+M(1+c_0)\log 2-2M\epsilon^2} + 2e^{-2c_0^2M} \tag{5.4}$$

which decreases exponentially in M as long as $(1+c_0)\log 2 < 2\epsilon^2$, which only holds if $\epsilon > \sqrt{0.5\log 2} \approx 0.6$. In other words, this simple bounding approach can only guarantee an embedding with a large error ϵ.

A tighter bound can be found if we better understand and bound P_2. This is the probability that a ball of radius $r/2$ will cross a quantization boundary when projected through a random projection. If the projected ball diameter is Δ or greater, then a quantization boundary will be crossed with probability 1. On the other hand, if the projected ball diameter is $l \le \Delta$, then a boundary crossing only happens with probability l/Δ. Thus, a bound on P_2 can be developed which enables the embedding error ϵ to go to zero. In the interest of brevity in the core of our development, we relegate the details on developing the bound to App. E.

5.2.3. *Error Analysis.* In contrast to quantized J-L embeddings, binary universal embeddings use 1 bit per embedding dimension. Thus, the rate R also determines the dimensionality of the projection, $K = R$, as well as the constant ϵ in the embedding guarantees (2.5). Furthermore, there is no multiplicative term in the guarantees, i.e., $\delta = 0$. Thus, in the ambiguity analysis (4.17), the numerator is fully determined; the system designer can only control the denominator.

This does not mean that there are no design choices and trade-offs: the trade-off in these embeddings is in the choice of the parameter Δ in (2.4). As discussed in the Sec. 2.2 and shown in Fig. 3(b), $g(\cdot)$ exhibits an approximately linear region, followed by a rapid flattening and an approximately flat region. The choice of Δ controls the slope of the linear region and, therefore, how soon the function reaches the flat region.

As mentioned earlier, the linear bound in (2.9) is a very good approximation of the upwards sloping linear region of $g(\cdot)$, which has slope $g'(d) \approx \sqrt{2/\pi}/\Delta$. By decreasing Δ, we can make that slope arbitrarily high, with a corresponding decrease of the ambiguity $\epsilon/g'(\widetilde{d}_{\mathscr{S}})$. However, this linear region does not extend for all d, but only until it reaches the point $d = D_0$ where $g(D_0) \approx 1/2$ and the flat region of $g(d)$ begins. As Δ becomes smaller and the slope of the linear region increases, it reaches the flat region much faster, approximately when $D_0\sqrt{2/\pi}/\Delta = 1/2$, i.e., when $D_0 \approx \Delta\sqrt{\pi/8} \approx 0.6\Delta$.

Unfortunately, beyond that linear region, the slope $g'(d)$ becomes 0 exponentially fast. This implies that the ambiguity in (4.17) approaches infinity. Thus, if the embedding distance $d_{\mathscr{W}}$ is within $0.5 \pm \epsilon$, then it is impossible to know anything about $d_{\mathscr{S}}$ by inverting the mapping, other than $d_{\mathscr{S}} \gtrsim D_0$. This makes the trade-off in designing Δ clear. A smaller Δ reduces the ambiguity in the range of distances it preserves, but also reduces the range of distances it preserves. The system designer should design Δ such that the distances required in the application of the embedding are sufficiently preserved.

As an example, consider the motivating application in Sec. 1.1: retrieval of nearest-neighbors from a database. When a query is executed, its embedding distance is computed with respect to all the entries in the database, embedded using the same parameters. For the query to be successful, there should be at least a few entries in the database with small embedding distance from the query. These entries are selected and returned. For the query to produce meaningful results, the embedding distance of those entries should represent quite accurately the signal

distance between the query signal and the signals from the entries in the database. Furthermore, if the signals are all very distant from the query, the embedding distance should accurately reflect that fact, so that no signal is selected; in this case the embedding does not need to represent how distant each entry is.

In other words, the embedding only needs to represent distances up to a radius D, determined by the system designer, and to only *identify* distances further than D, without necessarily representing those distances. Thus, Δ should be designed to be as small as possible so the ambiguity in representing distances in the linear region is small, but not smaller than necessary to ensure that all distances of interest are contained in the linear region of the embedding and do not spill over into the flat region with high ambiguity.

Note that this notion of locality is much richer that the notion defined in [59]. The latter only ensures that the distances between embeddings of close signals is small and between embeddings of distant signals is large. Instead, our development, further guarantees a linear distant map up to a radius, thus preserving distances up to this radius.

5.3. **Multibit Universal Embeddings.** Another benefit of the development and analysis in this paper is that it facilitates analysis of multibit universal embeddings, i.e., embeddings using the quantizer at the bottom of Fig. 3(a).

In principle, it is possible to consider multi-bit universal quantizers as sums of scaled one-bit quantizers, both in amplitude and the argument. In particular, given the 1-bit quantizer $Q(\cdot)$ defined at the top of Fig. 3(a), the B-bit generalization equals

$$Q_B(x) = \sum_{b=0}^{B-1} 2^b Q\left(\frac{x}{2^b}\right), \tag{5.5}$$

and has period 2^B. Thus, with quantization interval Δ, the embedding can be expressed as

$$\mathbf{y} = \widetilde{Q}_B\left(\widetilde{\mathbf{A}}\mathbf{x} + \widetilde{w}\right), \tag{5.6}$$

where $\widetilde{Q}_B(x) = Q_B(2^B x)$, $\widetilde{\mathbf{A}}$ has elements drawn from an i.i.d. Normal distribution with variance $(\sigma/2^B\Delta)^2$, to map ℓ_2 distances, or an i.i.d. Cauchy distribution with scale parameter $\gamma/2^B\Delta$ to map ℓ_1 distances, as described above. Using the Fourier series of $Q(\cdot)$, appropriately scaled and summed for each bit $b = 0,\ldots,B-1$, and a similar development as in Sec. 5.2.1 it is possible to derive the embedding map. It is important to note, however, that appropriate shifting and scaling of the functions significantly complicates the resulting expressions.

Instead, a simpler expression can be derived by observing that the multibit universal quantization function is, in fact, the result of uniform scalar quantization applied to the sawtooth map $f(x) = x - 1/2$ for $x \in [0,1)$, $f(x) = f(x-1)$ otherwise. Thus, we can use the well-established Fourier series coefficients of the sawtooth function to obtain $|H_k|^2 = (1/2\pi k)^2$ for $k > 0$. The resulting map for a general $\mathbf{A}$ without scaling and quantization equals to

$$g(d) = 2\sum_{k>0} \frac{1}{(2\pi k)^2}(1 - \phi_l(2\pi k|d)) = \frac{1}{12} - \sum_{k>0} \frac{1}{2\pi^2 k^2}\phi_l(2\pi k|d). \tag{5.7}$$

Thus, if the scaling by Δ and 2^B is considered and the elements of $\mathbf{A}$ are drawn from an i.i.d. $\mathcal{N}(0,\sigma^2)$ distribution, the embedding map becomes

$$g(d) = \frac{1}{12} - \sum_{k>0} \frac{1}{2\pi^2 k^2} e^{-2\left(\frac{\pi\sigma dk}{2^B\Delta}\right)^2}, \tag{5.8}$$

where d is the ℓ_2 distance of the signals. Similarly, if elements of $\mathbf{A}$ are drawn from an i.i.d. Cauchy distribution with scale parameter γ, then the embedding map becomes

$$g(d) = \frac{1}{12} - \sum_{k>0} \frac{1}{2\pi^2 k^2} e^{-\frac{2\pi\gamma dk}{2^B\Delta}}, \tag{5.9}$$

where d is now the ℓ_1 distance of the signals.

The sawtooth, which takes values in $[-1/2, 1/2]$ is subsequently quantized by a B-bit uniform scalar quantizer, i.e., one having 2^B levels and interval $\Delta = 2^{-B}$. In the context of Thm. 3.3, this implies a worst-case quantization error of $E_Q = \Delta/2 = 2^{-B-1}$. This error is in the ℓ_2 distance of the embedding, not in the ℓ_2^2 guarantee of Thm. 4.1. The guarantee applies, therefore, to the embedding map in Cor. 4.1. The corresponding guarantee for the ℓ_2^2 distance should use $E_Q = (\Delta/2)^2 = 2^{-2B-2}$.

Using a sawtooth function as a map, followed by scalar quantization to derive multibit universal embedding guarantees should provide looser bounds than explicitly estimating the guarantee using sum of square wave maps.

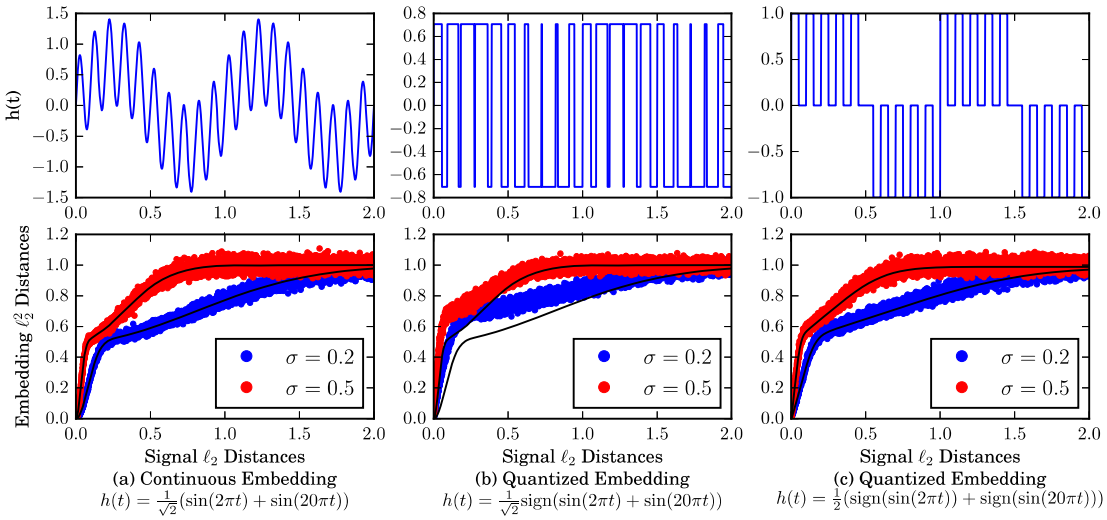

FIGURE 4. Simulation results in embedding design: (a) Unquantized embedding preserving two different distance intervals with different accuracy. (b) Effect of 1-bit quantization on (a). (c) A 3-level quantized embedding with similar performance as (a).

However, as the rate B increases, the two guarantees should converge to the same one. In fact, as B increases and quantization becomes finer, both approaches converge to the guarantees derived using an unquantized sawtooth function as an embedding map.

6. Simulations and Application Examples

To verify and demonstrate the theoretical developments above, we present simulation results verifying our designs. We also demonstrate how our approach can be applied toward encoding features for image retrieval and for image classification over the network.

6.1. **Simulations: Embedding Design.** Existing simulation results on quantized embeddings, such as the ones shown in Fig. 3, demonstrate the validity of our analysis. To further verify the results in Sec. 4.1 we consider a slightly more complex distance map, in which shorter distances should be represented with greater precision, intermediate distances should be represented with less precision and larger distances do not need to be represented with any precision, similar to universal embeddings.

Specifically, we first consider an embedding with $H_1 = H_{10} = \sqrt{2}/2$ and $H_k = 0$ for all other k. In other words, the embedding map $h(t)$ is equal to

$$h(t) = \frac{\sqrt{2}}{2}(\sin(2\pi t) + \sin(20\pi t)), \tag{6.1}$$

as plotted at the top of Fig. 4(a).

The bottom of Figure 4(a) demonstrates the performance of the embedding on randomly generated signals in $N = 10000$ dimensions with different distances in the range $d = 0, \ldots, 2$. The signals are embedded in $M = 2000$ dimensions using matrices with variance $\sigma = 0.2$ and 0.4 for the blue and red dots, respectively. The ℓ_2^2 distance of the embedded signals is plotted against the ℓ_2 distance of the signals. The black line in the figure plots the theoretical embedding map $g(d)$ according to Thm. 4.1. As evident in the figure, the embedding performs as predicted by the theory. The embedding map increases rapidly for a small radius, not as rapidly until a larger radius, and then becomes flat beyond that. The radii are greater for smaller σ, as expected.

The figure also shows the ambiguity, as measured by the horizontal width of the plots and analyzed in Sec. 4.4. As expected, the ambiguity is lower given the vertical ambiguity if the slope of the embedding map is higher. Thus, short distances, up to a first radius, are better represented than longer distances, up to the point where the embedding flattens. Beyond that, the embedding only conveys information that signals are far apart.

We should also note, that it seems that the vertical ambiguity of the embedding seems to be smaller for smaller embedding distances, suggesting a multiplicative ambiguity instead of an additive one.

Figure 4(b) demonstrates a 1-bit quantized version of the embedding in (a), simply taking the sign of the continuous embedding:

$$h(t) = \frac{\sqrt{2}}{2}\text{sign}(\sin(2\pi t) + \sin(20\pi t)), \tag{6.2}$$

with the scaling chosen such that $g(d)$ saturates to 1 asymptotically. The embedding map is shown on the top of the figure, while the embedding performance is shown at the bottom. The black line plots the theoretical embedding map for the unquantized embedding, i.e., the same map as in (a).

As evident by the figure, the actual performance of the embedding concentrates around a curve that is different than the theoretical prediction for the continuous version. However, the embedding is within the bounds of Thm. 3.3 because E_Q is quite large for a 1-bit quantizer. Moreover, the experimental results demonstrate quite good concentration around a curve, even though the embedding uses only one bit per coefficient. Still, a better understanding of the quantized embedding map and its Fourier series would yield a more accurate prediction. Such an understanding is not straightforward and we do not attempt it here.

Instead, in Figure 4(c) we demonstrate a similar quantized embedding that is easier to characterize. Specifically, the embedding map $h(t)$ is equal to

$$h(t) = \frac{1}{2}\text{sign}(\sin(2\pi t)) + \frac{1}{2}\text{sign}(\sin(20\pi t)), \tag{6.3}$$

where

$$\text{sign}(x) = \begin{cases} -1, & \text{if } x < 0 \\ +1, & \text{otherwise,} \end{cases} \tag{6.4}$$

as plotted at the top of the figure. In this case,

$$|H_k| = \begin{cases} 2/\pi k, & \text{if } k \text{ is odd} \\ 20/\pi k, & \text{if } k \text{ is divisible by 10} \\ 0, & \text{otherwise.} \end{cases} \tag{6.5}$$

This embedding, as shown in the figure, has very similar characteristics and distance-preservation properties with the continuous embeddings of Fig. 4(a). The quantized embedding has slightly wider error bounds in the experiments, but this is expected given that it is quantized at only 3 levels per coefficient. Still, despite the effect of quantization, the performance is very close to the performance of the continuous embedding. In contrast to the quantized embedding in Fig. 4(b), the embedding in (c) can be easily analyzed. As expected, it is also a bit tighter than the one in (b) because it is a 3-level per coefficient embedding, i.e., uses approximately 1.6 times the rate.

To further demonstrate the effect of quantization, Fig. 5(a) plots simulation results on the embedding map (6.1) when quantized with a 1-, 2-, and 4-bit scalar quantizer, i.e., for embedding maps of the form

$$h(t) = Q_B(\sin(2\pi t) + \sin(20\pi t)), \tag{6.6}$$

where $Q_B(\cdot)$ is a B-bit scalar quantizer, appropriately scaled. As above, the black line plots the distance map predicted assuming an unquantized embedding, i.e., the embedding in Fig. 4(a). Of course, the case $B = 1$ coincides with the example in Fig. 4(b).

As evident in Fig. 5(a), the higher the rate, the closer the result is to the theoretical prediction for the continuous embedding. Of course, the embedding satisfies the bounds of Thm. 3.3. However, the figure suggests that the bounds are loose. An analysis along the lines of Thm. 4.1, should provide a tighter bound. Such an analysis is not as straightforward for this particular embedding, and we do not attempt it here.

For completeness, Fig. 5(b) plots the performance of quantized universal embeddings for the same choice of bitrates. The figure also plots the theoretical distance maps (2.5) for $B = 1$ and (5.7) for the unquantized sawtooth wave using black lines. Even with $B = 4$, the quantization is sufficiently fine, so that the experimental distance map of the quantized embedding and the theoretical distance map of unquantized embedding are observed to coincide.

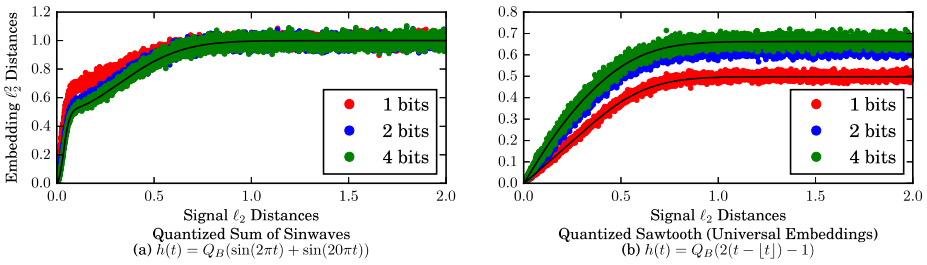

FIGURE 5. Effect of quantization on embeddings. (a) The embedding of Fig. 4(a) with increasingly refined quantization. (b) Multibit universal embeddings with increasingly refined quantization, approaching the performance of a continuous embedding using a sawtooth embedding map.

6.2. **Application Example: Image Retrieval Using Universal Embeddings.** As an example application we consider image retrieval over the cloud. A user wants to retrieve information about a query object by capturing its photograph and transmitting information extracted from the photograph to a database server. The server locates the object in the database that most closely matches the query image, according to a predetermined similarity criterion, and transmits meta-data about that object back to the user. The goal is to reduce, as much as possible, transmission bit-rate given a certain desired performance.

Since the server does not require to exactly reconstruct the image to retrieve similar images, it should be possible to significantly reduce the bit-rate compared to naively transmitting the actual images using lossy compression. As we describe below, this is indeed possible by computing quantized universal embeddings of features extracted from the query and database images.

6.2.1. *Protocol Architecture.* In preparation for the query, server and client agree on the embedding parameters—specifically, $\mathbf{A}$, $\mathbf{w}$, and Δ in the case of universal embeddings—according to the embedding specifications. For example, the server might draw universal parameters for all clients that will access the database.

Next, the server builds a database using features extracted from previously labeled images. In our experiments we use the well-established SIFT features [50], which provide significant invariance properties that facilitate image retrieval. Typically, with such feature extraction methods, a single image might generate a variable number of features. However, each of the generated feature vectors is associated to an image in the database and its associated metadata. The server builds and indexes the database by embedding the features using the predetermined embedding parameters and associating the image and the metadata with the correct embedded feature.

To execute a query, the client first acquires an image that serves as the query image. The client extracts the features from that image, embeds them using the predetermined embedding parameters, and transmits their embedding to the server. The server receives the embedded features, and retrieves from the database the nearest neighbor to each feature using the embedding distance, i.e., a single match for every embedded features. From those matches, the server selects the J closest candidates, with $J = 20$ in our experiments. The metadata are selected using majority voting among those matches.

6.2.2. *Experimental Results.* To validate our approach, we conducted retrieval experiments using the ZuBuD database [68]. This public database contains 1005 images of 201 buildings in the city of Zurich. There are 5 images of each building taken from different viewpoints, all of size 640×480 pixels and compressed in PNG format. Our experimental setup is identical to [48]: One out of the 5 viewpoints of each building was randomly selected as a query image, forming a test set of $s = 201$ images. The server's database comprises of the remaining 4 images of each building, for a total of $t = 804$ images. The query aims to identify which of the 201 possible buildings is depicted in each query image.

Our goal is to examine the performance of embeddings in preserving distances, not the performance of various feature selection methods or retrieval protocols. Thus, we extracted the widely adopted SIFT features [50] from each image and embedded them using quantized J-L embeddings or universal embeddings. Using the protocols described in Sec. 6.2.1 we measured how many of the 201 query images produced the correct result, i.e., correctly identified the building depicted. We conducted our experiments in bitrates ranging from 0 to 80 bits per descriptor.

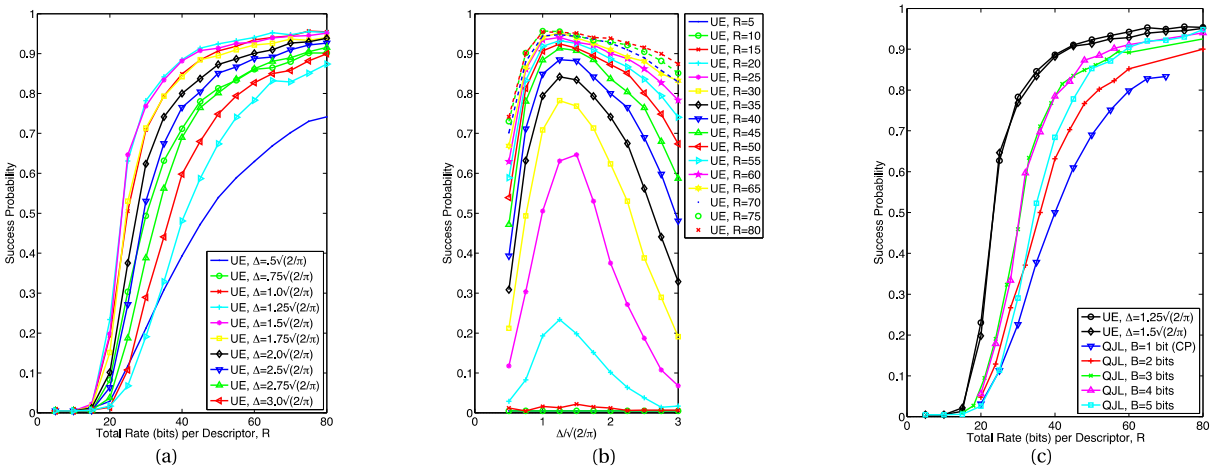

FIGURE 6. Performance of universal embeddings (UE) in metadata retrieval. (a) Probability of correct retrieval as a function of the bitrate for a variety of Δ values. (b) Probability of correct retrieval as a function of Δ for a variety of bitrates. (c) Comparison of universal embeddings using $\Delta = 1.25\sqrt{2/\pi}$ and $1.5\sqrt{2.\pi}$ with quantized J-L methods (QJL). Universal embeddings significantly outperform the alternatives.

Our results are averaged over 100 experiments with different realizations of $\mathbf{A}$ and $\mathbf{w}$, although the variability among individual runs was very small.

The first experiment tested the effect of Δ in the design of the embedding. In particular, we examined the range $\Delta = 0.5\sqrt{2/\pi}, 0.75\sqrt{2/\pi}, \ldots, 3\sqrt{2/\pi}$. The results are shown in Figs. 6(a) and (b). In Fig. 6(a) each curve plots the probability of correct metadata retrieval as a function of the bitrate used per descriptor, given a fixed Δ. The higher the probability of success, the better. Figure 6(b) presents another view on the same data: each curve plots the probability of correct retrieval given a fixed bitrate per descriptor as Δ varies.

The plots in Fig. 6(a) and (b) verify our expectations. As the bitrate increases, the performance improves. With respect to Δ, the behavior is more nuanced. For small Δ, the slope of $g(d)$ is high and the ambiguity in the linear region of $g(d)$ is low, as discussed in Sec. 5.2 and shown in Fig. 3. Thus, the distances represented by the embedding are represented very well. However, D_0 is small, i.e. it can only represent accurately a very small range of distances. Thus, for a large number of queries for which the closest matches are farther than D_0 the results returned are not meaningful. This type of error dominates the results when Δ is low. As Δ increases, more and more queries produce meaningful results and the error performance improves, even though the accuracy of the linear region of the embedding decreases. For larger Δ the reduced accuracy of the embedding starts dominating the error and the performance decreases again. The best performance is obtained for $\Delta = 1.25\sqrt{2/\pi}$, which corresponds to corresponding $D_0 = .625$.

We also compared the performance of our approach using quantized J-L embeddings. Figure 6(c) compares the performance of the two types of embeddings. The figure plots the probability of correct retrieval as a function of the bitrate per descriptor for each of the methods examined. As expected [48], multibit quantized J-L embeddings outperform 1-bit quantized J-L embeddings—known as "compact projections" (CP) [51, 75] and motivated by LSH approaches [4] in earlier literature. More important, universal embeddings—plotted in black circles and black diamonds, for $\Delta = 1.25\sqrt{2/\pi}$ and $1.5\sqrt{2/\pi}$ respectively—significantly outperform quantized J-L embeddings. For example, to achieve a probability of correct retrieval of 80%, universal embeddings require approximately 8 fewer bits per descriptor, a 20% rate reduction. For 90% probability of correct retrieval, universal embeddings require 15 fewer bits per descriptor, a 25% rate reduction. Similarly, using only 40 bits per descriptor, universal embeddings achieve almost 90% success rate, versus almost 80% for the best alternative. The results are robust with respect to Δ: for $\Delta \in [\sqrt{2/\pi}, 2\sqrt{2/\pi}]$, universal embeddings outperform all quantized J-L embeddings.

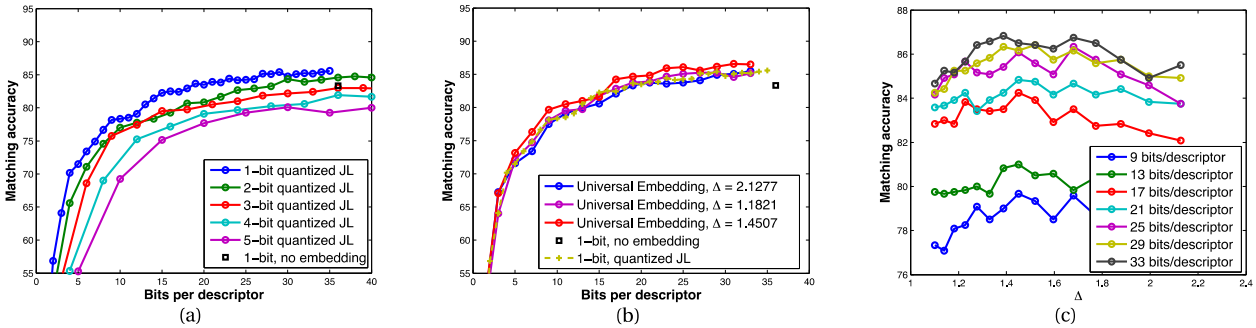

FIGURE 7. Classification accuracy as a function of the bit-rate achieved using (a) quantized JL (QJL) embeddings; and (b) universal embeddings. (c) Classification accuracy as a function of the quantization step size Δ used in computing the universal embeddings.

6.3. **Application Example: Kernel-based Image classification.** We also tested the performance of universal embeddings as a feature representation on a multiclass classification problem. The goal is to identify the class membership of query images belonging to one of 8 different classes. Our experiments demonstrate how the embeddings function as an SVM kernel, and enable a rate-inference trade-off analogous to the rate-distortion trade-off in conventional coding: they allow trading inference performance for reduction in the rate.

6.3.1. *Protocol Architecture.* The protocol for this application is very similar to the one in Sec. 6.2.1. Specifically, to set up this problem, the server extracts a Dalal-Triggs Histogram of Oriented Gradients (HOG) features [54] from the training images. The HOG algorithm extracts a 36 element feature vector (descriptor) for every 8×8 pixel block in an image. The descriptors encode local 1-D histograms of gradient directions in small spatial regions in an image. Every HOG feature is compressed using either quantized JL embeddings or universal quantized embeddings. The compressed features are then stacked to produce a single compressed feature vector for each image. The compressed features of the training images are used together with the image labels to train a number of binary linear SVM classifiers, one for each class.

To classify a query image, the client extracts HOG features, encodes them using the predetermined embedding parameters, and transmits it to the server. The server executes the SVMs directly on the embeddings and decides on the embedding class.

6.3.2. *Experimental Results.* In our simulations, we used tools from the VLFeat library [71] to extract HOG features and train the SVM classifier. We consider eight image classes. One class consists of persons from the INRIA person dataset [54, 55]. The other seven classes—car, wheelchair, stop sign, ball, tree, motorcycle, and face— are extracted from the Caltech 101 dataset [33, 34]. All images are standardized to 128×128 pixels centered around the target object in each class. We use 15 training and 15 test images from each class.

Fig. 7(a) shows the classification accuracy obtained by quantized JL embeddings of HOG descriptors using the trained SVM classifier. The black square corresponds to 1-bit scalar quantization of raw non-embedded HOG descriptors, using a bit-rate of 36 bits—one bit for each element of the descriptor.

The figure shows that 1-bit quantized JL embeddings allow us to achieve a 50% bit-rate reduction, compared to non-embedded quantized descriptors, without reduction in performance (classification accuracy). This is obtained using an 18-dimensional JL embedding of every HOG descriptor, followed by 1-bit scalar quantization. Furthermore, increasing the embedding dimension, and, therefore, the bit-rate, above 18 improves the inference performance beyond that of the 1-bit quantized non-embedded HOG features. Note that, among all quantized JL embeddings, 1-bit quantization achieves the best rate-inference performance.

Fig. 7(b) compares the classification accuracy of universal embeddings for varying values of the step size parameter Δ with that of the 1-bit quantized JL embeddings and the 1-bit quantized non-embedded HOG descriptors. With the choice of $\Delta = 1.4507$, the universal embedded descriptors further improve the rate-inference performance over the quantized JL embeddings by 1% in inference improvement. They also achieve the same classification accuracy as any choice of quantization for non-embedded HOG descriptors, or, even, unquantized ones, at significantly lower bit-rate—points not shown in the figure, as they are out of the interesting part of the bit-rate scale.

Figure 7(c) illustrates the effect of the parameter Δ by plotting the classification accuracy as a function of Δ for different embedding rates. The figure shows that, similar to the findings in Sec. 6.2, if Δ is too small or too large, the performance suffers.

As evident, an embedding-based system design can be tuned to operate at any point on the rate vs. classification performance frontier, not possible just by quantizing the raw HOG features. Furthermore, with the appropriate choice of Δ, universal embeddings improve the classification accuracy given the fixed bit-rate, compared with quantized JL embeddings, or reduce the bit-rate required to deliver a certain inference performance.

7. Discussion and Conclusions

A key contribution of our paper is the notion that embeddings can be designed to have different distance preservation accuracy for different distance ranges. In doing so, we developed a framework to understand how the properties of the embeddings are preserved, exploiting the embedding distance map. However, our work raises more interesting questions than it solves.

A key question in our approach is whether any arbitrary distance preservation design is possible. Our results on the subadditivity of the distance map in Sec. 4.3 place some constraints on what distance maps can be realized. However, it is not clear that this is the only constraint that is realizable. For example, we conjecture that the distance map should be monotonic—within ϵ and δ ambiguities, similar to the subadditivity constraint—but we have not provided such a proof.

It is also not clear that our design method in Sec. 4.1 can achieve any arbitrary design within those constraints. For example, our design is monotonic. If our monotonicity conjecture is false, then our design approach cannot achieve all the possible distance maps. Ideally, we prefer to be able to determine the embedding starting from the distance map instead of the other way around that Thm. 4.1 establishes. In the context of our design approach, we would like to determine $h(t)$ and the corresponding H_k starting from $g(d)$. Of course this could be possible only through a very different design approach.

While our development has demonstrated achievable bounds on the embedding design, fundamental lower bounds would also be desirable. For example, [2] demonstrates a lower bound on the number of measurements required to satisfy the J-L lemma, which is improved in [45] for linear embeddings. Similarly, [76] has demonstrated some lower bounds on the rate of binary embeddings. However, neither of these results account for a distance map and for preserving different ranges with different accuracy.

Of course, part of our goal is to design representations that accurately represent signal geometry between pairs of signals, i.e., distances and inner products, without requiring both signals to be present while the representation is computed. These representations should be easy to compute and easy to compute with, i.e., should provide straightforward mechanisms to compute the necessary geometric quantities. While we are using embeddings as our mechanism, it is not necessarily the only or the optimal approach to the problem.

As a parallel path, consider basis and frame expansions of signals, which provide straightforward mechanisms to represent signals. Using those representations in classical signal processing applications, we can control, for example, the approximation error in the representation, the rate-distortion performance of the quantized representation, and the representation accuracy in certain subspaces, according to the application requirements. We desire to have the same control on the signal geometry and not on the signals themselves. In that sense, the distortion reflects the accuracy of representing distances and different ranges of distances. Thus, for example, we would like to control the rate-distortion performance of a quantized distance representation, given the range of distances we are interested in. General representation and coding methods, as well as representation complexity and rate-distorion bounds are still open for this problem.

Fast computation is also desirable in our methods. While dimensionality reduction does reduce computation in practice, it does not improve the asymptotic behavior of methods such as nearest neighbors. However, there are obvious connections between our work and LSH [4, 26, 38], which are definitely worth exploring. LSH could be trivially used as a layer on top of the embedding, as a separate process to speed up computation. A more interesting approach would be a combination of the two, since many of the fundamental concepts and the resulting algorithmic steps are similar. In some sense, the ideal LSH is also a distance embedding into a discrete space, in which with very high probability $g(d) = 0$ if $d < r$ and $g(d) \neq 0$ if $d > R$ for some $r < R$.

More generally, we are interested in information representation and coding when the end goal is a function computation $g(\mathbf{x}, \mathbf{y})$ and when one of the inputs is not available at the encoder. The information and the corresponding distortion is measured at the function output. The embeddings considered in this work represent a

special case given by $g(\mathbf{x},\mathbf{y}) = g(\|\mathbf{x}-\mathbf{y}\|)$. Of interest are more general functions common in machine learning, such as classifiers and estimators. Some fundamental bounds and coding schemes have been developed using the chromatic entropy of the function, e.g., see [30, 56]. However, these techniques operate on variables $\mathbf{x}$ and $\mathbf{y}$ drawn from a discrete alphabet. Even with discrete sources, they require the construction of a large graph that grows with the size of the alphabet. Such an approach is prohibitive even in simple problems, such as the distance representation discussed in this paper.

Distributed functional coding [52], is a more practical and promising alternative for continuous sources. Unfortunately, distance functions do not satisfy the conditions for optimality in the proposed solution—specifically, an "equivalence-free" property. This will most definitely also be the case for a number of machine learning algorithms. Furthermore, for such algorithms an explicit differentiable functional form, as required for the encoding, might not be easily available. For such functions, an embedding-based coding might be more appropriate. Still, it remains to be seen whether embeddings that preserve function computation, other than geometry, are even possible.

Acknowledgment

The authors would like to thank Laurent Jacques for his constructive feedback and helpful comments in improving the manuscript and some of the proofs.

Appendix A. Proof of Theorem 3.1

Proof. First, we consider two balls of radius r with centers $\mathbf{x}$ and $\mathbf{y}$, denoted $\mathscr{B}_r(\mathbf{x})$ and $\mathscr{B}_r(\mathbf{y})$, respectively. For any vector pair $\mathbf{x}' \in \mathscr{B}_r(\mathbf{x})$, $\mathbf{y}' \in \mathscr{B}_r(\mathbf{y})$ we have

$$|d(\mathbf{x}',\mathbf{y}') - d(\mathbf{x},\mathbf{y})| \le 2r \tag{A.1}$$

$$\Rightarrow |d(f(\mathbf{x}'),f(\mathbf{y}')) - d(f(\mathbf{x}),f(\mathbf{y}))| \le 2K_f r \tag{A.2}$$

$$\text{and } |g(d(\mathbf{x}',\mathbf{y}')) - g(d(\mathbf{x},\mathbf{y}))| \le 2K_g r \tag{A.3}$$

This follows by the triangle inequality and the properties of Lipschitz continuity.

Starting with (A.2) and using (2.1) and (A.3) we can derive

$$\begin{aligned} d(f(\mathbf{x}'),f(\mathbf{y}')) &\le d(f(\mathbf{x}),f(\mathbf{y})) + 2K_f r && \text{(A.4)}\\ &\le (1+\delta)g(d(\mathbf{x},\mathbf{y})) + 2K_f r + \epsilon && \text{(A.5)}\\ &\le (1+\delta)g(d(\mathbf{x}',\mathbf{y}')) \\ &\quad + (1+\delta)2K_g r + 2K_f r + \epsilon && \text{(A.6)} \end{aligned}$$

and

$$\begin{aligned} d(f(\mathbf{x}'),f(\mathbf{y}')) &\ge d(f(\mathbf{x}),f(\mathbf{y})) - 2K_f r && \text{(A.7)}\\ &\ge (1-\delta)g(d(\mathbf{x},\mathbf{y})) - 2K_f r - \epsilon && \text{(A.8)}\\ &\ge (1-\delta)g(d(\mathbf{x}',\mathbf{y}')) \\ &\quad - (1-\delta)2K_g r - 2K_f r - \epsilon && \text{(A.9)}\\ &\ge (1-\delta)g(d(\mathbf{x}',\mathbf{y}')) \\ &\quad - (1+\delta)2K_g r - 2K_f r - \epsilon && \text{(A.10)} \end{aligned}$$

i.e.,

$$\begin{aligned} (1-\delta)g(d(\mathbf{x}',\mathbf{y}')) - (1+\delta)2K_g r - 2K_f r - \epsilon \\ \le d(f(\mathbf{x}'),f(\mathbf{y}')) \le \\ (1+\delta)g(d(\mathbf{x}',\mathbf{y}')) + (1+\delta)2K_g r + 2K_f r + \epsilon \end{aligned} \tag{A.11}$$

Setting $r = \frac{\alpha}{(1+\delta)2K_g + 2K_f}$ and $\widetilde{\epsilon} = \epsilon + \alpha$ for some α, we obtain that the final embedding bound

$$(1-\delta)g(d(\mathbf{x}',\mathbf{y}')) - \widetilde{\epsilon} \le d(f(\mathbf{x}'),f(\mathbf{y}')) \le (1+\delta)g(d(\mathbf{x}',\mathbf{y}')) + \widetilde{\epsilon} \tag{A.12}$$

holds with probability $1 - ce^{-Mw(\delta,\widetilde{\epsilon}-\alpha)}$.

Using the union bound on the $C_\varepsilon^{\mathscr{S}}$ balls that cover the signal set with Kolmogorov entropy $E_r^{\mathscr{S}}$, it follows that (A.12) holds with probability greater than $1-ce^{2E_r^{\mathscr{S}}-Mw(\delta,\widetilde{\epsilon}-\alpha)}$, which decays exponentially with M, as long as $M = O(E_r^{\mathscr{S}})$. □

Appendix B. Proof of Theorem 3.2

Proof. For a single value of T, given M measurements, we can use Hoeffding's inequality to upper bound the probability that more than $P_T(1+c_0)M$ measurements will be exactly T-part Lipschitz over a single ball $\mathscr{B}_{r/2}(\mathbf{x})$:

$$P(\text{more than } P_T(1+c_0)M \text{ measurements are exactly } T\text{-part Lipschitz}) \le e^{-2c_0^2 M}. \tag{B.1}$$

For each T, each of those measurements will partition the ball to T sets. We set a $T_{\max}$, denoting the level beyond which the probability that a function $f_m(\cdot)$ is T-part Lipschitz is negligible. Thus, using the union bound, a lower bound on the probability that for all T, at most $P_T(1+c_0)M$ measurements are exactly T-part Lipschitz continuous is equal to $1-T_{\max}e^{-2c_0^2M}-P_F$, where $P_F = (\sum_{T=T_{\max}+1}^{\infty} P_T)$ is considered negligible.

Therefore, the embedding partitions the ball into at most

$$\#\text{ of Sets} \le \prod_{T=1}^{T_{\max}} T^{P_T(1+c_0)M} = e^{\sum_{T=1}^{T_{\max}} P_T(1+c_0)M\log T} = e^{c_1 M} \tag{B.2}$$

sets, with probability greater than $1-T_{\max}e^{-2c_0^2M}-P_F$, where $c_1 = \sum_{T=1}^{T_{\max}} P_T(1+c_0)\log T = \sum_{T=2}^{T_{\max}} P_T(1+c_0)\log T$, since $\log 1 = 0$. Note that P_T concentrates to lower T's as r decreases and, therefore, we expect c_1 to decrease as r decreases.

The assumption that P_T is independent of the ball center $\mathbf{x}$ can be relaxed if, instead, an upper bound on P_T is used that is independent of $\mathbf{x}$. Moreover, in many practical applications in which the discontinuity arises due to quantization, the assumption can be made to hold using dithering. Dithering also helps P_T concentrate to 0 for $T>1$ as r decreases.

Since the ball has radius $r/2$, i.e., diameter r, each set of its partition also has the same diameter. In other words, if we pick any point in each set of the partition and call it the "center" of the set, all other points of the set are within r of the center. Thus we can repeat the argument of the previous section but on the e^{c_1M} set centers produced by each of the $C_{r/2}^{\mathscr{S}}$ balls that constitute the $r/2$-covering of the set; a total of $e^{2E_{r/2}^{\mathscr{S}}+c_1M}$ centers. Thus, with probability $1-(ce^{2E_{r/2}^{\mathscr{S}}+c_1M-Mw(\delta,\widetilde{\epsilon}-\alpha)}-T_{\max}e^{-2c_0^2M}-P_F)$, the embedding satisfies (A.12):

$$(1-\delta)g(d(\mathbf{x},\mathbf{y}))-\widetilde{\epsilon} \le d(f(\mathbf{x}),f(\mathbf{y})) \le (1+\delta)g(d(\mathbf{x},\mathbf{y}))+\widetilde{\epsilon} \tag{B.3}$$

for all $\mathbf{x}$ and $\mathbf{y}$ in $\mathscr{S}$, where $r = \frac{\alpha}{(1+\delta)2K_g+2K_f}$ □

Of course, an analysis along this line can be extended to multidimensional discontinuous functions for which the continuity in each dimension cannot be considered independently of the other dimension. In this case, we need to considering the T-part Lipschitz continuity property in multiple dimensions and perform the same analysis. In the interest of space, we do not describe this extension.

Appendix C. Proof of Theorem 4.1

Proof. We consider the single coefficient $y = h(\langle \mathbf{a},\mathbf{x}\rangle + w)$, the pair of signals $\mathbf{x}$ and $\mathbf{x}'$ at distance $d = d_{\mathscr{S}}(\mathbf{x}-\mathbf{x}')$ apart, and their (signed) projected distance $l = \langle \mathbf{a},\mathbf{x}-\mathbf{x}'\rangle$. Studying how the mapping operates on this pair provides the basis for how the mapping operates on the whole set of signals, in a manner similar to [1, 8, 12, 25].

Conditioned on l, the squared difference of the signals' mapping has expected value over w equal to

$$E\{(y-y')^2|l\} = \int_0^1 (h(u+w)-h(u+l+w))^2 f_w(w)\,dw \tag{C.1}$$

$$= \int_0^1 h^2(u+w)+h^2(u+l+w)-2h(u+w)h(u+l+w)\,dw \tag{C.2}$$

$$= 2\left(R_h(0)-R_h(l)\right), \tag{C.3}$$

with the last equality following from the shift-invariance of the autocorrelation.

Thus, as a function of d, the expected value of the squared difference is

$$E\{(y-y')^2\} = \int_{-\infty}^{+\infty} E\{(y-y')^2|l\} f_l(l|d)\,dl \tag{C.4}$$

$$= \int_{-\infty}^{+\infty} 2\left(R_h(0) - R_h(l)\right) f_l(l|d)\,dl \tag{C.5}$$

$$= 2\left(R_h(0) - \int_{-\infty}^{+\infty} R_h(l) f_l(l|d)\,dl\right) \tag{C.6}$$

Using Parseval's theorem and the characteristic function of $f_l(l|d)$, denoted using $\phi_l(\xi|d)$ we obtain

$$E\{(y-y')^2\} = 2\left(\sum_k (|H_k|^2 - |H_k|^2 \phi_l(2\pi k|d))\right) \tag{C.7}$$

$$= 2\sum_k |H_k|^2 (1-\phi_l(k|d)) = g(d), \tag{C.8}$$

where $g(d)$ is the distance map (4.6).

Since $h(t)$ is bounded, the squared difference of any measurement is bounded, i.e., $(y-y')^2 \in [0, \bar{h}^2]$. Using Hoeffding's inequality, it follows that, for M measurements,

$$P\left(\left|\frac{1}{M}\sum_m \left(y_m - y'_m\right)^2 - g(d)\right| \geq \epsilon\right) = P\left(\left|\frac{1}{M}\left\|\mathbf{y}-\mathbf{y}'\right\|_2^2 - g(d)\right| \geq \epsilon\right) \leq 2e^{-2M\frac{\epsilon^2}{\bar{h}^4}}. \tag{C.9}$$

As we describe in Section 3.1, using the union bound on a set $\mathscr{S}$ of Q points, i.e., at most $Q^2/2$ point pairs, the embedding guarantee in the theorem follows.

We should also note that if the distribution of $h(\langle \mathbf{a}, \mathbf{x}\rangle + w) - h(\langle \mathbf{a}, \mathbf{x}'\rangle + w)$ can be shown to be sub-Gaussian, even if $h(\cdot)$ is not bounded, a similar result can be shown using concentration of measure results on sub-Gaussian random variables, e.g., see [72]. Restricting $h(\cdot)$ to be bounded is then a special case. Note that the distribution of $h(\langle \mathbf{a}, \mathbf{x}\rangle + w) - h(\langle \mathbf{a}, \mathbf{x}'\rangle + w)$ may be sub-Gaussian, even if the distribution of $\langle \mathbf{a}, \mathbf{x}\rangle$ is not.

The same proof steps can be used to bound the deviation of the measurements from their norm, i.e., to derive (4.8). The main difference is in computing $E\{y^2\}$ instead of $E\{(y-y')^2\}$, i.e.,

$$E\{y^2\} = \int_0^1 h^2(u+w) f_w(w)\,dw \tag{C.10}$$

$$= \int_0^1 h^2(u+w)\,dw = R_h(0), \tag{C.11}$$

and then using Hoeffding's inequality and the union bound on Q point as above. To establish both the norm bound and the embedding bound, the inequality should be established over Q points for the norm and $Q(Q-1)/2$ pairwise distances, i.e., a union of $Q^2/2 + Q/2 \leq Q^2$ events. □

Appendix D. Proof of Proposition 4.3

Proof. By definition, the distance functions on both spaces satisfy the triangle inequality:

$$d(\mathbf{x},\mathbf{z}) \leq d(\mathbf{x},\mathbf{y}) + d(\mathbf{y},\mathbf{z}) \tag{D.1}$$

$$d(f(\mathbf{x}), f(\mathbf{z})) \leq d(f(\mathbf{x}), f(\mathbf{y})) + d(f(\mathbf{y}), f(\mathbf{z})) \tag{D.2}$$

To show that $g(d)$ is subadditive, we pick $\mathbf{y}$ in the line between $\mathbf{x}$ and $\mathbf{z}$, such that $d(\mathbf{x},\mathbf{z}) = d$, $d(\mathbf{x},\mathbf{y}) = \lambda d$, and $d(\mathbf{y},\mathbf{z}) = (1-\lambda)d$. This can be trivially done if $\mathscr{S}$ is a convex set. From (D.2) and (2.1) it follows that

$$(1-\epsilon)g(d) - \delta \leq (1+\epsilon)g(\lambda d) + \delta + (1+\epsilon)g((1-\lambda)d) + \delta \tag{D.3}$$

$$\Rightarrow \frac{1-\epsilon}{1+\epsilon} g(d) - 3\delta \leq g(\lambda d) + g((1-\lambda)d) \tag{D.4}$$

$$\Rightarrow (1-2\epsilon)g(d) - 3\delta \leq g(\lambda d) + g((1-\lambda)d) \tag{D.5}$$

Selecting a, b and d such that $a+b=d$ and setting $\lambda = a/(a+b)$ proves that $g(d)$ is $(2\epsilon, 3\delta)$-subadditive. □

Appendix E. Probability of Crossing a Universal Quantization Threshold

We develop a bound for P_2 assuming the measurement matrix $\mathbf{A}$ has entries drawn from an i.i.d. $\mathcal{N}(0,\sigma^2)$ distribution, followed by scaling with Δ^{-1}. In that case, the m^{th} projection of a ball of radius $r/2$ will have diameter at most $\|\mathbf{a}\|_2 r/\Delta$ where $\|\mathbf{a}\|_2$ is the norm of an N-dimensional Gaussian vector with variance σ^2. If the projection of this ball, with dither added, includes an integer, then a quantization threshold has been crossed. Using I to denote the number of integers in this projection, then $P_2 = P(I=1)$, which is upper bounded by $P_2 \le E\{I\}$.

For example, if the diameter $\|\mathbf{a}\|_2 r/\Delta$ of the projection is less than or equal to 1, then, thanks to the dither, $P_2 = E\{I\} = \|\mathbf{a}\|_2 r/\Delta$ and I can only take the values 0 or 1. Similarly, if $\|\mathbf{a}\|_2 r/\Delta > 1$, then the projection will include at least $\lfloor \|\mathbf{a}\|_2 r/\Delta \rfloor$ integer points and at most $\lfloor \|\mathbf{a}\|_2 r/\Delta \rfloor + 1$, depending on the value of the dither, i.e., with probability $1 - \|\mathbf{a}\|_2 r/\Delta + \lfloor \|\mathbf{a}\|_2 r/\Delta \rfloor$ and $\|\mathbf{a}\|_2 r/\Delta - \lfloor \|\mathbf{a}\|_2 r/\Delta \rfloor$, respectively. In other words, given $\|\mathbf{a}\|_2 r/\Delta$

$$E\{I|\|\mathbf{a}\|_2 r/\Delta\} = \lfloor \|\mathbf{a}\|_2 r/\Delta \rfloor (1 - \|\mathbf{a}\|_2 r/\Delta + \lfloor \|\mathbf{a}\|_2 r/\Delta \rfloor) + \tag{E.1}$$

$$(\lfloor \|\mathbf{a}\|_2 r/\Delta \rfloor + 1)(\|\mathbf{a}\|_2 r/\Delta - \lfloor \|\mathbf{a}\|_2 r/\Delta \rfloor) \tag{E.2}$$

$$= \|\mathbf{a}\|_2 \frac{r}{\Delta} \tag{E.3}$$

Since $P_2 \le E\{I\}$, it follows that

$$P_2 \le E_{\mathbf{a}}\{E\{I|\|\mathbf{a}\|_2 r/\Delta\}\} = E_{\mathbf{a}}\{\|\mathbf{a}\|_2 r/\Delta\} \tag{E.4}$$

$$\le \frac{r}{\Delta}\sqrt{E_{\mathbf{a}}\left\{\|\mathbf{a}\|^2\right\}} = \frac{\sigma r}{\Delta}\sqrt{N} := \overline{P}_2, \tag{E.5}$$

where the second step follows due to Jensen's inequality. The bound becomes meaningful when $r < \frac{\Delta}{\sigma\sqrt{N}}$ and approaches 0 as r decreases.

Thus, $c_1 \le \overline{P}_2(1+c_0)\log 2$ and the probability that the embedding does not hold is upper bounded by

$$ce^{2E^{\mathscr{S}}_{r/2} + c_1 M - M w(\delta,\epsilon)} + T_{\max} e^{-2c_0^2 M} + P_F \le e^{2E^{\mathscr{S}}_{r/2} + M\overline{P}_2(1+c_0)\log 2 - 2M\epsilon^2} + 2e^{-2c_0^2 M} \tag{E.6}$$

which decreases exponentially with M, as long as $2\epsilon^2 > \overline{P}_2(1+c_0)\log 2$, allowing the embedding error ϵ to approach 0 with appropriate choice of r. Note that as r decreases, P_2 decreases approximately linearly whereas $E^{\mathscr{S}}_{r/2}$ increases as $\dim^{\mathscr{S}} \cdot \log(1/r)$, where $\dim^{\mathscr{S}}$ is the Kolmogorov dimension of the set $\mathscr{S}$:

$$\dim^{\mathscr{S}} = \lim_{r\to 0} \frac{\log C^{\mathscr{S}}_r}{\log(1/r)} = \lim_{r\to 0} \frac{\log E^{\mathscr{S}}_r}{\log(1/r)}. \tag{E.7}$$

References

[1] D. Achlioptas, *Database-friendly Random Projections: Johnson-lindenstrauss With Binary Coins*, Journal of Computer and System Sciences, 66 (2003), pp. 671–687.

[2] N. Alon, *Problems and results in extremal combinatorics–i*, Discrete Mathematics, 273 (2003), pp. 31–53.

[3] A. Andoni, M. Deza, A. Gupta, P. Indyk, and S. Raskhodnikova, *Lower bounds for embedding edit distance into normed spaces*, in Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms, 2003, pp. 523–526.

[4] A. Andoni and P. Indyk, *Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions*, Commun. ACM, 51 (2008), pp. 117–122.

[5] A. S. Bandeira, D. G. Mixon, and B. Recht, *Compressive classification and the rare eclipse problem*, arXiv preprint arXiv:1404.3203, (2014).

[6] Z. Bar-Yossef, T. Jayram, R. Krauthgamer, and R. Kumar, *Approximating edit distance efficiently*, in Foundations of Computer Science, 2004. Proceedings. 45th Annual IEEE Symposium on, IEEE, 2004, pp. 550–559.

[7] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde, *Model-based compressive sensing*, IEEE Trans. Info. Theory, 56 (2010), pp. 1982–2001.

[8] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, *A simple proof of the restricted isometry property for random matrices*, Const. Approx., 28 (2008), pp. 253–263.

[9] R. Baraniuk and M. Wakin, *Random projections of smooth manifolds*, Foundations of Computational Mathematics, 9 (2009), pp. 51–77.

[10] T. BLUMENSATH AND M. DAVIES, *Sampling theorems for signals from the union of finite-dimensional linear subspaces*, IEEE Trans. Info. Theory, 55 (2009), pp. 1872–1882.
[11] P. T. BOUFOUNOS, *Hierarchical distributed scalar quantization*, in Proc. Int. Conf. Sampling Theory and Applications (SampTA), Singapore, May 2-6 2011.
[12] ———, *Universal rate-efficient scalar quantization*, IEEE Trans. Info. Theory, 58 (2012), pp. 1861–1872.
[13] ———, *Angle-preserving quantized phase embeddings*, in Proc. SPIE Wavelets and Sparsity XV, San Diego, CA, August 25-29 2013.
[14] ———, *On embedding the angles between signals*, in Proc. Signal Processing with Adaptive Sparse Structured Representations (SPARS), Lausanne, Switzerland, July 8-11 2013.
[15] ———, *Sparse signal reconstruction from phase-only measurements*, in Proc. Int. Conf. Sampling Theory and Applications (SampTA), Bremen, Germany, July 1-5 2013.
[16] P. T. BOUFOUNOS AND H. MANSOUR, *Universal embeddings for kernel machine classification*, in Proc. Sampling Theory and Applications, Washington, DC, May 25-29 2015.
[17] P. T. BOUFOUNOS AND S. RANE, *Secure binary embeddings for privacy preserving nearest neighbors*, in Proc. Workshop on Information Forensics and Security (WIFS), Foz do IguaÃğu, Brazil, November 29–December 2 2011.
[18] ———, *Efficient coding of signal distances using universal quantized embeddings*, in Proc. Data Compression Conference (DCC), Snowbird, UT, March 20-22 2013.
[19] J. BOURGAIN, S. DIRKSEN, AND J. NELSON, *Toward a unified theory of sparse dimensionality reduction in euclidean space*, Geometric and Functional Analysis, 25 (2015), pp. 1009–1088.
[20] B. BRINKMAN AND M. CHARIKAR, *On the impossibility of dimension reduction in ℓ_1*, Journal of the ACM (JACM), 52 (2005), pp. 766–788.
[21] E. CANDÈS, *Compressive sampling*, in Proc. Int. Congress Math., Madrid, Spain, Aug. 2006.
[22] ———, *The restricted isometry property and its implications for compressed sensing*, Comptes rendus de l'Académie des Sciences, Série I, 346 (2008), pp. 589–592.
[23] E. CANDÈS, J. ROMBERG, AND T. TAO, *Stable signal recovery from incomplete and inaccurate measurements*, Comm. Pure and Appl. Math., 59 (2006), pp. 1207–1223.
[24] E. J. CANDÈS AND M. B. WAKIN, *An introduction to compressive sampling*, Signal Processing Magazine, IEEE, 25 (2008), pp. 21–30.
[25] S. DASGUPTA AND A. GUPTA, *An elementary proof of a theorem of Johnson and Lindenstrauss*, Random Structures & Algorithms, 22 (2003), pp. 60–65.
[26] 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, SCG '04, New York, NY, USA, 2004, ACM, pp. 253–262.
[27] M. A. DAVENPORT, J. N. LASKA, P. T. BOUFOUNOS, AND R. G. BARANIUK, *A simple proof that random matrices are democratic*, tech. rep., Rice University ECE Department Technical Report TREE-0906, Houston, TX, November 2009.
[28] S. DIRKSEN, *Dimensionality reduction with subgaussian matrices: a unified theory*, Foundations of Computational Mathematics, (2015), pp. 1–30.
[29] D. DONOHO, *Compressed sensing*, IEEE Trans. Inf. Theory, 6 (2006), pp. 1289–1306.
[30] V. DOSHI, D. SHAH, M. MÉDARD, AND M. EFFROS, *Functional compression through graph coloring*, Information Theory, IEEE Transactions on, 56 (2010), pp. 3901–3917.
[31] A. EFTEKHARI AND M. B. WAKIN, *New analysis of manifold embeddings and signal recovery from compressive measurements*, Applied and Computational Harmonic Analysis, 39 (2015), pp. 67–109.
[32] Y. ELDAR AND M. MISHALI, *Robust recovery of signals from a structured union of subspaces*, IEEE Trans. Info. Theory, 55 (2009), pp. 5302–5316.
[33] L. FEI-FEI, R. FERGUS, AND P. PERONA, *Caltech 101 dataset.* `http://www.vision.caltech.edu/Image_Datasets/Caltech101/`, 2004.
[34] ———, *Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories*, in Proc. IEEE Conf. on Comp. Vision and Pattern Recognition (CVPR), Workshop on Generative-Model Based Vision., June 2004, pp. 178–178.
[35] S. FOUCART AND T. NEEDHAM, *Sparse recovery from saturated measurements*, Information and Inference, (2016). to appear.

[36] C. HEGDE, A. SANKARANARAYANAN, W. YIN, AND R. BARANIUK, *NuMax: A convex approach for learning near-isometric linear embeddings*, IEEE Trans. Signal Processing, 63 (2015), pp. 6109–6121.
[37] P. INDYK, *Stable distributions, pseudorandom generators, embeddings, and data stream computation*, Journal of the ACM (JACM), 53 (2006), pp. 307–323.
[38] P. INDYK AND R. MOTWANI, *Approximate nearest neighbors: towards removing the curse of dimensionality*, in ACM Symposium on Theory of computing, 1998, pp. 604–613.
[39] L. JACQUES, *A quantized johnson-lindenstrauss lemma: The finding of buffon's needle*, IEEE Trans. Info. Theory, 61 (2015), pp. 5012–5027.
[40] ———, *Small width, low distortions: quasi-isometric embeddings with quantized sub-gaussian random projections*, arXiv preprint arXiv:1504.06170, (2015).
[41] L. JACQUES, D. K. HAMMOND, AND J. M. FADILI, *Dequantizing compressed sensing: When oversampling and non-gaussian constraints combine*, Information Theory, IEEE Transactions on, 57 (2011), pp. 559–571.
[42] ———, *Stabilizing nonuniformly quantized compressed sensing with scalar companders*, IEEE Trans. Info. Theory, 59 (2013), pp. 7969–7984.
[43] L. JACQUES, J. N. LASKA, P. T. BOUFOUNOS, AND R. G. BARANIUK, *Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors*, IEEE Trans. Info. Theory, 59 (2013).
[44] W. JOHNSON AND J. LINDENSTRAUSS, *Extensions of Lipschitz mappings into a Hilbert space*, Contemporary Mathematics, 26 (1984), pp. 189 –206.
[45] K. G. LARSEN AND J. NELSON, *The johnson-lindenstrauss lemma is optimal for linear dimensionality reduction*, arXiv preprint arXiv:1411.2404, (2014).
[46] J. N. LASKA, P. T. BOUFOUNOS, M. A. DAVENPORT, AND R. G. BARANIUK, *Democracy in action: Quantization, saturation, and compressive sensing*, Applied and Computational Harmonic Analysis, 31 (2011), pp. 429–443.
[47] V. I. LEVENSHTEIN, *Binary codes capable of correcting deletions, insertions, and reversals*, Soviet Physics Doklady, 10 (1966), pp. 707–710.
[48] M. LI, S. RANE, AND P. T. BOUFOUNOS, *Quantized embeddings of scale-invariant image features for mobile augmented reality*, in Proc. IEEE International Workshop on Multimedia Signal Processing (MMSP), Banff, Canada, Sept. 17–19 2012.
[49] N. LINIAL, E. LONDON, AND Y. RABINOVICH, *The geometry of graphs and some of its algorithmic applications*, Combinatorica, 15 (1995), pp. 215–245.
[50] D. G. LOWE, *Distinctive image features from scale-invariant keypoints*, International Journal of Computer Vision, 60 (2004), pp. 91–110.
[51] K. MIN, L. YANG, J. WRIGHT, L. WU, X.-S. HUA, AND Y. MA, *Compact projection: Simple and efficient near neighbor search with practical memory requirements*, in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, June 13–18 2010.
[52] V. MISRA, V. K. GOYAL, AND L. R. VARSHNEY, *Distributed scalar quantization for computing: High-resolution analysis and extensions*, Information Theory, IEEE Transactions on, 57 (2011), pp. 5298–5325.
[53] Y. MROUEH AND L. ROSASCO, *q-ary compressive sensing*, in Proc. 10th Int. Conf. Sampling Theory and Applications (SampTA), 2013.
[54] D. NAVNEET AND B. TRIGGS, *Histograms of oriented gradients for human detection*, in International Conference on Computer Vision & Pattern Recognition, vol. 2, June 2005, pp. 886–893.
[55] ———, *INRIA Person Dataset*. `http://pascal.inrialpes.fr/data/human/`, 2005.
[56] A. ORLITSKY AND J. R. ROCHE, *Coding for computing*, IEEE Transactions on Information Theory, 47 (2001), pp. 903–917.
[57] R. OSTROVSKY AND Y. RABANI, *Low distortion embeddings for edit distance*, Journal of the ACM (JACM), 54 (2007), p. 23.
[58] D. OTERO AND G. R. ARCE, *Generalized restricted isometry property for alpha-stable random projections*, in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 3676–3679.
[59] S. OYMAK AND B. RECHT, *Near-optimal bounds for binary embeddings of arbitrary sets*, arXiv preprint arXiv:1512.04433, (2015).
[60] Y. PLAN AND R. VERSHYNIN, *One-bit compressed sensing by linear programming*, Communications on Pure and Applied Mathematics, 66 (2013), pp. 1275–1297.
[61] ———, *Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach*, Information Theory, IEEE Transactions on, 59 (2013), pp. 482–494.

[62] ———, *Dimension reduction by random hyperplane tessellations*, Discrete & Computational Geometry, 51 (2014), pp. 438–461.
[63] G. PUY, M. DAVIES, AND R. GRIBONVAL, *Recipes for stable linear embeddings from hilbert spaces to* $\mathbb{R}^m$, arXiv preprint arXiv:1509.06947, (2015).
[64] M. RAGINSKY AND S. LAZEBNIK, *Locality-sensitive binary codes from shift-invariant kernels*, The Neural Information Processing Systems, 22 (2009).
[65] A. RAHIMI AND B. RECHT, *Random features for large-scale kernel machines*, in Advances in neural information processing systems, 2007, pp. 1177–1184.
[66] S. RANE, P. T. BOUFOUNOS, AND A. VETRO, *Quantized embeddings: An efficient and universal nearest neighbor method for cloud-based image retrieval*, in Proc. SPIE Applications of Digital Image Processing XXXVI, San Diego, CA, August 25-29 2013.
[67] A. SADEGHIAN, B. BAH, AND V. CEVHER, *Energy-aware adaptive bi-Lipschitz embeddings*, in Proc. Int. Conf. Sampling Theory and Applications (SampTA), Bremen, Germany, July 1-5 2013.
[68] H. SHAO, T. SVOBODA, AND L. V. GOOL, *ZuBuD : Zurich Buildings database for image based recognition*, Tech. Rep. 260, Computer Vision Lab, Swiss Federal Institute of Technology, Switzerland, Apr. 2003.
[69] C. STRECHA, A. BRONSTEIN, M. BRONSTEIN, AND P. FUA, *LDAHash: Improved matching with smaller descriptors*, IEEE Trans. Pattern Analysis and Machine Intelligence, 34 (2012), pp. 66 –78.
[70] M. TALAGRAND, *The generic chaining: upper and lower bounds of stochastic processes*, Springer Science & Business Media, 2006.
[71] A. VEDALDI AND B. FULKERSON, *VLFeat: An open and portable library of computer vision algorithms.* `http://www.vlfeat.org/`, 2008.
[72] R. VERSHYNIN, *Introduction to the non-asymptotic analysis of random matrices*, in Compressed Sensing: Theory and Applications, Y. C. Eldar and G. Kutyniok, eds., Cambridge University Press, 005 2012, pp. 210–268.
[73] R. VERSHYNIN, *High Dimensional Probability for Mathematicians and Data Scientists*, draft, 2016. available `http://www-personal.umich.edu/~romanv/papers/HDP-book/HDP-book.pdf`.
[74] Y. WEISS, A. TORRALBA, AND R. FERGUS, *Spectral hashing*, in Advances in Neural Information Processing Systems 21, 2009, pp. 1753–1760.
[75] C. YEO, P. AHAMMAD, AND K. RAMCHANDRAN, *Rate-efficient visual correspondences using random projections*, in Proc. IEEE International Conference on Image Processing (ICIP), San Diego, CA, October 12-15 2008.
[76] X. YI, C. CARAMANIS, AND E. PRICE, *Binary embedding: Fundamental limits and fast algorithm*, in Proc. 32nd International Conference on Machine Learning, vol. 37, Lille, France, July 6-11 2015, pp. 2162–2170.

MITSUBISHI ELECTRIC RESEARCH LABORATORIES, 201 BROADWAY, CAMBRIDGE MA 02139,, USA
E-mail address: `petrosb@merl.com`

PALO ALTO RESEARCH CENTER, 3333 COYOTE HILL ROAD, PALO ALTO, CA 94304, USA
E-mail address: `srane@parc.com`

MITSUBISHI ELECTRIC RESEARCH LABORATORIES, 201 BROADWAY, CAMBRIDGE, MA 02139, USA
E-mail address: `mansour@merl.com`