Journal of Artificial Intelligence Research 72 (2021) 215-249 Submitted 01/2021; published 10/2021

A Theoretical Perspective on Hyperdimensional Computing

Anthony Thomas AHTHOMAS@ENG.UCSD.EDU
Sanjoy Dasgupta DASGUPTA@ENG.UCSD.EDU
Tajana Rosing TAJANA@ENG.UCSD.EDU
Department of Computer Science
University of California, San Diego
San Diego, CA 92093, USA

Abstract

Hyperdimensional (HD) computing is a set of neurally inspired methods for obtaining high-dimensional, low-precision, distributed representations of data. These representations can be combined with simple, neurally plausible algorithms to effect a variety of information processing tasks. HD computing has recently garnered significant interest from the computer hardware community as an energy-efficient, low-latency, and noise-robust tool for solving learning problems. In this review, we present a unified treatment of the theoretical foundations of HD computing with a focus on the suitability of representations for learning.

1. Introduction

Hyperdimensional (HD) computing is an emerging area at the intersection of computer architecture and theoretical neuroscience (Kanerva, 2009). It is based on the observation that brains are able to perform complex tasks using circuitry that: (1) uses low power, (2) requires low precision, and (3) is highly robust to data corruption. HD computing aims to carry over similar design principles to a new generation of digital devices that are highly energy-efficient, fault tolerant, and well-suited to natural information processing (Rahimi et al., 2018).

The wealth of recent work on neural networks also draws its inspiration from the brain, but modern instantiations of these methods have diverged from the desiderata above. The success of these networks has rested upon choices that are not neurally plausible, most notably significant depth and training via backpropagation. Moreover, from a practical perspective, training these models often requires high precision and substantial amounts of energy. While a large body of literature has sought to ameliorate these issues with neural networks, these efforts have largely been designed to address specific performance limitations. By contrast, the properties above emerge naturally from the basic architecture of HD computing.

Hyperdimensional computing focuses on the very simplest neural architectures. Typically, there is a single, static, mapping from inputs x to much higher-dimensional "neural" representations $\phi(x)$ living in some space $\mathcal{H}$. All computational tasks are performed in $\mathcal{H}$-space, using simple, operations like element-wise additions and dot products. The mapping ϕ is often taken to be random, and the embeddings have coordinates that have low precision; for instance, they might take values -1 and $+1$. The entire setup is elementary and lends itself to fast, low-power hardware realizations.

©2021 AI Access Foundation. All rights reserved.

Indeed, a cottage industry has emerged around developing optimized implementations of HD computing based algorithms on hardware accelerators (Imani et al., 2017; Rahimi et al., 2018; Gupta et al., 2018; Schmuck et al., 2019; Salamat et al., 2019; Imani et al., 2019). Broadly speaking, this line of work touts HD computing as an energy efficient, low-latency, and noise-resilient alternative to conventional realizations of general purpose ML algorithms like support vector machines, multilayer perceptrons, and nearest-neighbor classifiers. While this work has reported impressive performance benefits, there has been relatively little formal treatment of HD computing as a tool for general purpose learning.

This review has two broad aims. The first, more modest, goal is to introduce the area of hyperdimensional computing to a machine learning audience. The second is to develop a particular mathematical framework for understanding and analyzing these models. The recent literature has suggested a variety of different HD architectures that conform to the overall blueprint given above, but differ in many important details. We present a unified treatment of many such architectures that enables their properties to be compared. The most basic types of questions we wish to answer are:

1. How can individual items, sets of items, and sequences of items, be represented and stored in $\mathcal{H}$-space, in a manner that permits reliable decoding?
2. What kinds of noise can be tolerated in $\mathcal{H}$-space?
3. What kinds of structure in the input x-space are preserved by the mapping to $\mathcal{H}$-space?
4. What is the power of linear separators on the ϕ-representation?

Some of these questions have been introduced in the HD computing literature and studied in isolation (Plate, 2003; Gallant & Okaywe, 2013; Kleyko et al., 2018; Frady et al., 2018). In this work we address these questions formally and in greater generality.

2. Introduction to HD Computing

In the following section we provide an introduction to the fundamentals of HD computing and provide some brief discussion of its antecedents in the neuroscience literature.

2.1 High-Dimensional Representations in Neuroscience

Neuroscience has proven to be a rich source of inspiration for the machine learning community: from the perceptron (Rosenblatt, 1958), which introduced a simple and general-purpose learning algorithm for linear classifiers, to neural networks (Rumelhart et al., 1986), to convolutional architectures inspired by visual cortex (Fukushima, 1980), to sparse coding (Olshausen & Field, 1996) and independent component analysis (Bell & Sejnowski, 1995). One of the most consequential discoveries from the neuroscience community, underlying much research at the intersection of neuroscience and machine learning, has been the notion of *high-dimensional distributed representations* as the fundamental data structure for diverse types of information. In the neuroscience context, these representations are also typically *sparse*.

To give a concrete example, the sensory systems of many organisms have a critical component consisting of a transformation from relatively low dimensional sensory inputs

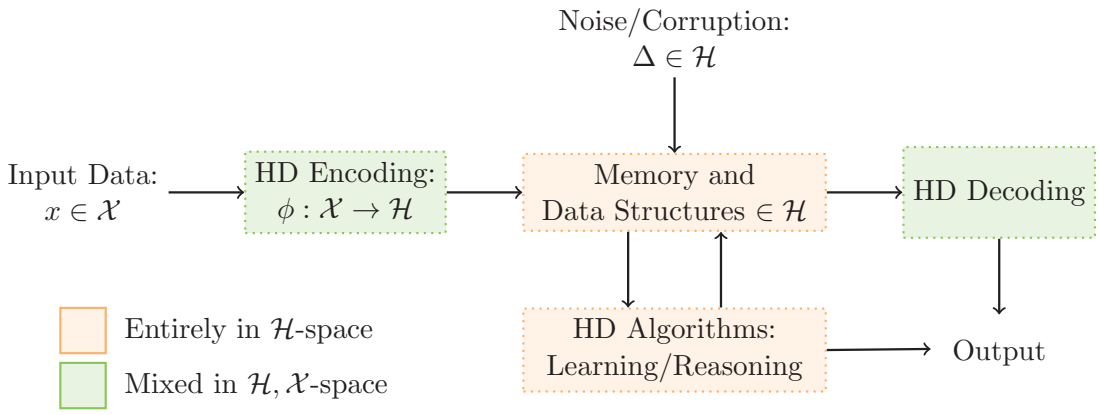

Figure 1: The flow of data in HD computing. Data is mapped from the input space to HD-space under an encoding function $\phi : \mathcal{X} \to \mathcal{H}$. HD representations of data are stored in data structures and may be corrupted by noise or hardware failures. HD representations can be used as input for learning algorithms or other information processing tasks and may be decoded to recover the input data.

to much higher-dimensional *sparse* representations. These latter representations are then used for subsequent tasks such as recall and learning. In the olfactory system of the fruit fly (Masse et al., 2009; Turner et al., 2008; Wilson, 2013; Caron et al., 2013), the mapping consists of two steps that can be roughly captured as follows:

1. An input $\mathbf{x} \in \mathbb{R}^n$ is collected via a sensory organ and mapped under a *random linear transformation* to a point $\phi(\mathbf{x}) \in \mathbb{R}^d$ $(d \gg n)$ in a high-dimensional space.

2. The coordinates of $\phi(\mathbf{x})$ are "sparsified" by a thresholding operation which just retains the locations of the largest k coordinates.

In the fly, the olfactory input is a roughly 50-dimensional vector $(n = 50)$ corresponding to different types of odor receptor neurons while the sparse representation to which it is mapped is roughly 2,000-dimensional $(d = 2000)$. A similar "expand-and-sparsify" template is also found in other species, suggesting that this process somehow exposes the information present in the input signal in a way that is amenable to learning by the brain (Stettler & Axel, 2009; Olshausen & Field, 2004; Chacron et al., 2011). The precise mechanisms by which this occurs are still not fully understood, but may have close connections to some of the literature on the theory of neural networks and kernel methods (Cybenko, 1989; Barron, 1993; Rahimi & Recht, 2008).

2.2 HD Computing

The notion of high-dimensional, distributed, data representations has engendered a number of computational models that have collectively come to be known as *vector symbolic architectures* (VSA) (Levy & Gayler, 2008). In general, VSAs provide a systematic way to generate and manipulate high-dimensional representations of symbols so as to implement

cognitive operations like association between related concepts. Notable examples of VSAs include "holographic reduced representations" (Plate, 1995, 2003), "binary spatter codes" (Kanerva, 1994, 1995), and "matrix binding of additive terms" (Gallant & Okaywe, 2013). HD computing can be seen as a successor to these early VSA models, with a strong additional slant towards hardware efficiency. While our treatment focuses primarily on recent work on HD computing, many of our results apply to these earlier VSA models as well.

An overview of data-flow in HD computing is given in Figure 1. The first step in HD computing is encoding, which maps a piece of input data to its high-dimensional representation under some function $\phi : \mathcal{X} \to \mathcal{H}$. The nature of ϕ depends on the type of input and the choice of $\mathcal{H}$. In this review, we consider inputs consisting of sets, sequences, and structures composed from a finite alphabet as well as vectors in a Euclidean space. The space $\mathcal{H}$ is some d-dimensional inner-product space defined over the real numbers or a subset thereof. Work in the literature on both HD computing and traditional neural networks has also explored complex-valued embeddings (Weiss et al., 2016; Parcollet et al., 2019; Zhang et al., 2016). However, we here focus on the more common case of real-valued embeddings. For computational reasons, it is common to restrict $\mathcal{H}$ to be defined over integers in a limited range $[-b, b]$. We emphasize that the dimension of $\mathcal{H}$ need not, in general, be greater than that of $\mathcal{X}$. Indeed, in several cases the encoding methods discussed can be used to reduce the dimension of the data.

The HD representations of data can be manipulated using simple element-wise operators. Two common and important such operations are "bundling" and "binding." The bundling operator is used to compile a set of elements in $\mathcal{H}$ and takes the form of a function $\oplus : \mathcal{H} \times \mathcal{H} \to \mathcal{H}$. The function takes two points in $\mathcal{H}$ and returns a third point that is similar to both operands. The binding operator is used to create ordered tuples of points in $\mathcal{H}$ and is likewise a function $\otimes : \mathcal{H} \times \mathcal{H} \to \mathcal{H}$. The function takes a pair of points in $\mathcal{H}$ as input, and produces a third point dissimilar to both operands. We make these notions more precise in our subsequent discussion of encoding.

Given the HD representation $\phi(\mathcal{S})$ of a set of items $\mathcal{S} \subset \mathcal{X}$ (produced by bundling the items), we may be interested to query the representation to determine if it contains the encoding of some $x \in \mathcal{X}$. To do so, we compute a metric of similarity $\rho(\phi(x), \phi(\mathcal{S}))$ and declare that the item is present in $\mathcal{S}$ if the similarity is greater than some critical value. This process can be used to decode the HD representation so as to recover the original points in $\mathcal{X}$ (Plate, 2003; Frady et al., 2018). We may additionally wish to assert that we can decode reliably even if $\phi(\mathcal{S})$ has been corrupted by some noise process. One of our chief aims in this paper is to mathematically characterize sufficient conditions for robust decoding under different noise models and input data types.

Beyond simply storing and recalling specific patterns, HD representations may also be used for learning. HD computing is most naturally applicable to classification problems. Suppose we are given some collection of labeled examples $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathcal{X}$ and $y_i \in \{c_i\}_{i=1}^{K}$ is a categorical variable indicating the class label of a particular x_i. One simple form of HD classification bundles together the data corresponding to a particular class to generate a "prototypical" example for the class (Kanerva, 2009; Kleyko et al., 2018;

Rahimi et al., 2018):

$$\phi(c_k) = \bigoplus_{i\,:\,y_i = c_k} \phi(x_i) \tag{1}$$

The resulting $\phi(c_k)$ are sometimes quantized to lower precision or sparsified via a thresholding operation. A nice feature of this scheme is that it is extremely simple to implement in an on-line fashion: that is, on streaming data arriving continuously over time (Rahimi et al., 2018). It is common to fine-tune the class prototypes using a few rounds of perceptron training (Imani et al., 2017, 2019b). Given some subsequent piece of query data $x_q \in \mathcal{X}$ for which we do not know the correct label, we simply return the label of the most similar prototype:

$$k^\star = \operatorname*{argmax}_{k \in 1,\ldots,K} \rho(\phi(x_q), \phi(c_k)).$$

The similarity metric ρ is typically taken to be the dot-product, with the operands normalized if necessary. Thus, on the whole, the scheme is quite similar to classical statistical methods like naive Bayes and Fisher's linear discriminant. In Section 5.2.1, we consider properties of the HD encoding that can make linear models more powerful in HD space than in the original space.

HD computing and closely related techniques have been applied to a wide variety of practical problems in fields ranging from bio-signal processing (Rahimi et al., 2016; Imani et al., 2017), to natural language processing (Sahlgren, 2005), and robotics (Mitrokhin et al., 2019; Neubert et al., 2019). We are here concerned with a more abstract treatment that focuses on the basic properties of HD computing and will not attempt to survey this literature. The interested reader is referred to (Rahimi et al., 2016; Kleyko et al., 2018) for discussions related to practical aspects of HD computing.

3. Encoding and Decoding Discrete Data

A central object in HD computing is the mapping from inputs to their high-dimensional representations. The design of this mapping, typically referred to as "encoding" in the literature on HD computing, has been the subject of considerable research. There is a wide range of possible encoding methods. Some of these have been introduced in the HD computing literature and studied in isolation (Plate, 2003; Gallant & Okaywe, 2013; Kleyko et al., 2018). In this review, we present a novel unifying framework in which to study these mappings and to characterize their key properties in a non-asymptotic setting. We first discuss the encoding and decoding of *sets* in some detail. Many HD encoding procedures for more complex data types such as sequences essentially amount to transforming the data into a set and then applying the standard set-encoding method.

3.1 Finite Sets

Let $\mathcal{A} = \{a_i\}_{i=1}^m$ be some finite alphabet of m symbols. Symbols $a \in \mathcal{A}$ are mapped to $\mathcal{H}$ under an encoding function $\phi : \mathcal{A} \to \mathcal{H}$. Our goal in this section is to consider the encoding of sets $\mathcal{S}$ whose elements are drawn from $\mathcal{A}$. The HD representation of $\mathcal{S}$ is constructed by superimposing the embeddings of the constituent elements using the bundling operator

$\oplus : \mathcal{H} \times \mathcal{H} \to \mathcal{H}$. The encoding of $\mathcal{S}$ is defined to be $\phi(\mathcal{S}) = \oplus_{a \in \mathcal{S}} \phi(a)$. We first focus on the intuitive setting in which $\oplus$ is the element-wise sum and then address other forms of bundling.

To determine if some $a \in \mathcal{A}$ is contained in $\mathcal{S}$, we check if the dot product $\langle \phi(a), \phi(\mathcal{S}) \rangle$ exceeds some fixed threshold. If the codewords $\{\phi(a) : a \in \mathcal{A}\}$ are orthogonal and have a constant length L, then we have $\langle \phi(a), \phi(\mathcal{S}) \rangle = L^2 \, \mathbb{1}(a \in \mathcal{S})$, where $\mathbb{1}$ is the indicator function which evaluates to one if its argument is true and zero otherwise. However, when the codewords are not perfectly orthogonal, we have $\langle \phi(a), \phi(\mathcal{S}) \rangle = L\mathbb{1}(a \in \mathcal{S}) + \Delta$, where Δ is the "cross-talk" caused by interference between the codewords. In order to decode reliably, we must ensure the contribution of the cross-talk is small and bounded. We formalize this using the notion of incoherence popularized in the sparse coding literature. We define incoherence formally as (Donoho et al., 2005):

Definition 1 ***Incoherence.*** *For $\mu \geq 0$, we say $\phi : \mathcal{A} \to \mathcal{H}$ is μ-incoherent if for all distinct $a, a' \in \mathcal{A}$, we have*

$$|\langle \phi(a), \phi(a') \rangle| \leq \mu L^2$$

where $L = \min_{a \in \mathcal{A}} \|\phi(a)\|$.

When $d \geq m$, it is possible to have codewords that are mutually orthogonal, whereupon $\mu = 0$. In general, we will be interested in results that do not require $d \geq m$.

3.1.1 Exact Decoding of Sets

In the following section, we show how the cross-talk can be bounded in terms of the incoherence of ϕ, and use this to derive a simple threshold rule for exact decoding.

Theorem 2 *Let $L = \min_{a \in \mathcal{A}} \|\phi(a)\|$ and let the bundling operator be the element wise sum. To decode whether an element a lies in set S, we use the rule*

$$\langle \phi(a), \phi(S) \rangle \geq \frac{1}{2} L^2.$$

This gives perfect decoding for sets of size $\leq s$ if ϕ is $1/(2s)$-incoherent.

Proof Consider some symbol a. Then:

$$\langle \phi(a), \phi(\mathcal{S}) \rangle = \mathbb{1}(a \in \mathcal{S}) \langle \phi(a), \phi(a) \rangle + \sum_{a' \in \mathcal{S} \setminus \{a\}} \langle \phi(a), \phi(a') \rangle$$

If $a \in \mathcal{S}$, then the above is lower bounded by $L^2 - sL^2\mu$, where μ is the incoherence of ϕ. Otherwise, it is upper bounded by $sL^2\mu$. So we decode perfectly if $sL^2\mu < L^2/2$, or $\mu < 1/(2s)$. ■

3.1.2 Random Codebooks

In practice, each $\phi(a)$ is usually generated by sampling from some distribution over $\mathcal{H}$ or a subset of $\mathcal{H}$ (Kanerva, 2009; Kleyko et al., 2018; Rahimi et al., 2018). We typically

require that this distribution is factorized so that coordinates of $\phi(a)$ are i.i.d.. Intuitively, the incoherence condition stipulated in Theorem 2 will hold if dot products between two different codewords are concentrated around zero. Furthermore, we would like it to be the case that this concentration occurs quickly as the encoding dimension is increased. It turns out that a fairly broad family of simple distributions satisfies these properties.

As an example, suppose $\phi(a)$ is sampled from the uniform distribution over $\{\pm 1\}^d$, which we denote $\phi(a) \sim \{\pm 1\}^d$. In this case, $L = \sqrt{d}$ exactly, and a direct application of Hoeffding's inequality and the union bound yields:

$$\mathbb{P}(\exists \text{ distinct } a, a' \in \mathcal{A} \text{ s.t. } |\langle \phi(a), \phi(a') \rangle| \geq \mu d) \leq m^2 \exp\left(-\frac{\mu^2 d}{2}\right).$$

(Recall that $m = |\mathcal{A}|$.) Stated another way, with high probability $\mu = O(\sqrt{(\ln m)/d})$, meaning that we can make μ as small as desired by increasing d.

In fact, the same basic approach holds for the much broader class of *sub-Gaussian* distributions, which can be characterized as follows (Wainwright, 2019):

Definition 3 ***Sub-Gaussian Random Variable.*** *A random variable $X \sim P_X$ is said to be sub-Gaussian if there exists $\sigma \in \mathbb{R}^+$, referred to as the sub-Gaussian parameter, such that:*

$$\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp\left(\frac{\sigma^2 \lambda^2}{2}\right) \text{ for all } \lambda \in \mathbb{R}.$$

Intuitively, the tails of a sub-Gaussian random variable decay at least as fast those of a Gaussian. We say the encoding ϕ is σ-sub-Gaussian if $\phi(a)$ is generated by sampling its d coordinates independently from the same sub-Gaussian distribution with parameter σ. We say ϕ is "centered" if the distribution from which it is sampled is of mean zero. In general, we assume ϕ is centered unless stated otherwise.

Codewords drawn from a sub-Gaussian distribution have the useful property that their lengths concentrate fairly rapidly around their expected value. This concentration is, in general, worse than sub-Gaussian but well behaved nonetheless. The following result is well known but we reiterate it here as it is useful for our subsequent discussion. A proof is available in the appendix.

Theorem 4 *Let ϕ be centered and σ-sub-Gaussian. Then:*

$$\mathbb{P}(\exists\, a \in \mathcal{A} \text{ s.t. } |\|\phi(a)\|_2^2 - \mathbb{E}[\|\phi(a)\|_2^2]| \geq t) \leq 2m \exp\left(-c \min\left\{\frac{t^2}{d\sigma^4}, \frac{t}{\sigma^2}\right\}\right)$$

for some positive absolute constant c.

Like the conventional Gaussian distribution, sub-Gaussianity is preserved under linear transformations. That is, if $\mathbf{x} = \{x_i\}_{i=1}^n$ is a sequence of i.i.d. sub-Gaussian random variables and $\mathbf{a}$ is an arbitrary vector in $\mathbb{R}^n$, then $\langle \mathbf{a}, \mathbf{x} \rangle$ is sub-Gaussian with parameter $\sigma \|a\|_2$ (Wainwright, 2019). We can obtain a more general version of the previous result about $\phi \sim \{\pm 1\}^d$ which applies to $\phi(a)$ sampled from any sub-Gaussian distribution.

Theorem 5 *Let ϕ be σ-sub-Gaussian. Then, for $\mu > 0$,*

$$\mathbb{P}(\exists \,\text{distinct}\; a, a' \in \mathcal{A} \;\; s.t. \;\; |\langle \phi(a), \phi(a') \rangle| \geq \mu L^2) \leq m^2 \exp\left(-\frac{\mu^2 \kappa L^2}{2\sigma^2}\right)$$

where $\kappa = (\min_a \|\phi(a)\|^2)/(\max_a \|\phi(a)\|^2)$.

Proof Fix some a and a'. Treating $\phi(a)$ as a fixed vector in $\mathbb{R}^d$ and using the fact that sub-Gaussianity is preserved under linear transformations, we may apply a Chernoff bound for sub-Gaussian random variables (e.g. Prop 2.1 of (Wainwright, 2019)) to obtain:

$$\mathbb{P}(|\langle \phi(a), \phi(a') \rangle| \geq \mu L^2) \leq 2 \exp\left(-\frac{\mu^2 L^4}{2\sigma^2 \|\phi(a)\|_2^2}\right) \leq 2 \exp\left(-\frac{\mu^2 L^4}{2\sigma^2 L_{\max}^2}\right)$$

where $L_{\max} = \max_{a \in \mathcal{A}} \|\phi(a)\|_2$. Therefore, taking $\kappa = L^2/L_{\max}^2$, we have:

$$\mathbb{P}(|\langle \phi(a), \phi(a') \rangle| \geq \mu L^2) \leq 2 \exp\left(-\frac{\mu^2 \kappa L^2}{2\sigma^2}\right)$$

and the claim follows by applying the union bound over all $\binom{m}{2} < m^2/2$ pairs of codewords. We note that, per Theorem 4, $\kappa \to 1$ as d becomes large. ■

To be concrete and provide useful practical guidance, we here introduce three running examples of codeword distributions.

Dense Binary Codewords. In our first example, the most common in practice in our impression, $\phi(a)$ is sampled from the uniform distribution over the d-dimensional unit cube $\{-1, +1\}^d$. This approach is advantageous because it leads to efficient hardware implementations (Imani et al., 2017; Rahimi et al., 2018) and is simple to analyze.

Gaussian Codewords. Our second example consists of codewords sampled from the d-dimensional Gaussian distribution (Plate, 2003). That is, $\phi(a) \sim \mathcal{N}(\mathbf{0}_d, \sigma^2 \mathbf{I}_d)$, where $\mathbf{0}_d$ is the d-dimensional zero vector. Here, the codewords will not be of exactly the same length. However, Theorem 4 ensures that squared codeword lengths are concentrated around their expected value of $\sigma^2 d$. More formally, for some $\tau > 0$:

$$\mathbb{P}(\exists\, a \in \mathcal{A} \text{ s.t. } |\|\phi(a)\|_2^2 - \sigma^2 d| \geq \tau \sigma^2 d) \leq 2m \exp\left(-c \min\left\{\tau^2 d, \tau d\right\}\right).$$

In both cases, we can see that to obtain a μ-incoherent codebook with probability $1 - \delta$, is it sufficient to choose:

$$d = O\left(\frac{2}{\mu^2} \ln \frac{m}{\delta}\right)$$

Or, stated another way, we have $\mu = O(\sqrt{(\ln m)/d})$ with high probability. The key point in the two examples above is that the encoding dimension is inversely proportional to μ^2. Per Theorem 2, to decode correctly it is sufficient to have $\mu = 1/(2s)$, meaning that the encoding dimension scales quadratically with the number of elements in the set, but only logarithmically in the alphabet size and probability of error.

We will also consider a third example in which the codewords are *sparse* and binary. However, we defer this for the time being as slightly different encoding methods and analysis techniques are appropriate.

3.1.3 Decoding with Small Probability of Error

The analysis above gives strong *uniform* bounds showing that, with probability at least $1-\delta$ over random choice of the codebook, *every* subset of size at most s will be correctly decoded. However, this guarantee requires us to impose the unappealing restriction that $s \ll \sqrt{d}$ which is a significant practical limitation. We here show that we can obtain $s = O(d)$ but with a weaker *pointwise* guarantee: any arbitrarily chosen set of size at most s will be correctly decoded with probability $1-\delta$ over the random choice of codewords. Rather than insist on a hard upper bound on the incoherence of the codebook, we can instead require the milder condition that random sums over dot-products between $\leq s$ codewords are small with high-probability. We define this property more formally as follows:

Definition 6 ***Subset Incoherence****. For $\tau > 0$, we say a random mapping $\phi : \mathcal{A} \to \mathcal{H}$ satisfies (s, τ, δ)-subset incoherence if, for any $\mathcal{S} \subset \mathcal{A}$ of size at most s, with probability at least $1-\delta$ over the choice of ϕ:*

$$\max_{a \notin \mathcal{S}} \left| \sum_{a' \in \mathcal{S}} \langle \phi(a), \phi(a') \rangle \right| \leq \tau L^2$$

where $L = \min_{a \in \mathcal{A}} ||\phi(a)||$.

Once again, it turns out that sampling the codewords from a sub-Gaussian distribution can readily be seen to satisfy a subset-incoherence condition with high-probability:

Theorem 7 *Let ϕ be σ-sub-Gaussian and fix some $\mathcal{S} \subset \mathcal{A}$ of size s. Then*

$$\mathbb{P}\left(\max_{a \notin \mathcal{S}} \left| \sum_{a' \in \mathcal{S}} \langle \phi(a), \phi(a') \rangle \right| \geq \tau L^2 \right) \leq 2m \exp\left(-\frac{\kappa \tau^2 L^2}{2s\sigma^2} \right)$$

where κ and L are as in Theorem 5.

The proof is similar to Theorem 5 and is available in the appendix. As a concrete example, in the practically relevant case that $\phi \sim \{\pm 1\}^d$ the above boils down to:

$$\mathbb{P}\left(\exists\, a \notin \mathcal{S} \text{ s.t. } \left| \sum_{a' \in \mathcal{S}} \langle \phi(a), \phi(a') \rangle \right| \geq \tau d \right) \leq 2m \exp\left(-\frac{\tau^2 d}{2s} \right).$$

Stated another way, we have: $\tau = O(\sqrt{(s \ln m)/d})$. Following Theorem 2, in order to ensure correct decoding with high probability, we must simply argue that the codebook satisfies the subset-incoherence property with $\tau = 1/2$, meaning we should choose the encoding dimension to be $d = O(s \ln m)$.

This method of analysis is similar to that of (Plate, 2003; Gallant & Okaywe, 2013; Frady et al., 2018), who reach the same conclusion vis-à-vis linear scaling using the central limit theorem. However, our formalism is more general and is non-asymptotic.

3.1.4 Comparing Set Representations

We can estimate the size of a set by computing the norm of its encoding, where the precision of the estimate can be bounded in terms of the incoherence of ϕ. In the following discussion, we make the simplifying assumption that the codewords are all of a constant length L. Again appealing to Theorem 4, we can see that this assumption is not onerous since the codeword lengths concentrate around their expected value.

Theorem 8 *Let $\mathcal{S}$ be a set of size s. Then:*

$$s(1 - s\mu) \leq \frac{1}{L^2}\|\phi(\mathcal{S})\|_2^2 \leq s(1 + s\mu)$$

Proof The proof is by direct manipulation:

$$\begin{aligned}\frac{1}{L^2}\|\phi(\mathcal{S})\|_2^2 = \frac{1}{L^2}\langle\phi(\mathcal{S}), \phi(\mathcal{S})\rangle &= \frac{1}{L^2}\sum_{a\in\mathcal{S}}\langle\phi(a), \phi(a)\rangle + \frac{1}{L^2}\sum_{a,a'\neq a\in\mathcal{S}}\langle\phi(a), \phi(a')\rangle \\ &\leq \frac{1}{L^2}(sL^2 + s^2\mu L^2).\end{aligned}$$

The other direction is analogous. ■

Given a pair of sets $\mathcal{S}, \mathcal{S}'$ over the same alphabet, we can estimate the size of their intersection and union directly from their encoded representation.

Theorem 9 *Let $\mathcal{S}$ and $\mathcal{S}'$ be sets of size s and s' drawn from $\mathcal{A}$ and denote their encodings by $\phi(\mathcal{S})$ and $\phi(\mathcal{S}')$ respectively.*

$$|\mathcal{S} \cap \mathcal{S}'| - ss'\mu \leq \frac{1}{L^2}\langle\phi(\mathcal{S}), \phi(\mathcal{S}')\rangle \leq |\mathcal{S} \cap \mathcal{S}'| + ss'\mu$$

The proof is similar to Theorem 8 and is deferred to the appendix. Noting as well that $|\mathcal{S} \cup \mathcal{S}'| = |\mathcal{S}| + |\mathcal{S}'| - |\mathcal{S} \cap \mathcal{S}'|$, we see that we can estimate the size of the union using the previous theorem. In practice, it may be unnecessary to compute these quantities with a high degree of precision. For instance, it may only be necessary to identify sets with a large intersection-over-union. Provided the definition of "large" is somewhat loose, we can accept a higher incoherence among the codewords in exchange for reducing the encoding dimension.

3.1.5 Sparse and Low-Precision Encodings

In the previous discussion, we assumed the bundling operator was the element-wise sum. This is a natural choice when the codewords are dense or non-binary. However, the resulting encodings are of unconstrained precision which may be undesirable from a computational perspective. For the purposes of representing sets of size $\leq s$, we may truncate $\phi(\mathcal{S})$ to lie in the range $[-c, c]$, with negligible loss in accuracy provided $c = O(\sqrt{s})$. In practice, it is common to quantize the encodings more aggressively to binary precision by thresholding (Kanerva, 1994; Rahimi et al., 2017; Burrello et al., 2018; Imani et al., 2019a). In other words, we encode as $\phi(\mathcal{S}) = g_t(\mathcal{S})$, where g_t is a thresholding function that is applied coordinate-wise: $g_t(x) = 1$ if $x \geq t$ and 0 otherwise.

As a notable special case of the thresholding rule described above, we here consider encoding with *sparse* codewords. In this case, we assume that a coordinate in a codeword is non-zero with some small probability. In other words, $\phi(a)_i \sim \text{Bernoulli}(p)$, where $p \ll 1/2$. We may then bundle items by taking an element-wise sum of their codewords with threshold $t = 1$, which is equivalent to taking the element-wise maximum over the codewords. That is, $\phi(\mathcal{S}) = \max_{a \in \mathcal{S}} \phi(a)$, where the max operator is applied coordinate-wise. Noting that the max is upper bounded by the sum in this setting, the notion of incoherence is a relevant quantity and the analysis of Theorem 2 continues to apply.

This encoding procedure is essentially a standard implementation of the popular "Bloom filter" data structure for representing sets (Bloom, 1970). The conventional Bloom filter differs slightly in that the typical decoding rule is to threshold $\langle \phi(a), \phi(\mathcal{S}) \rangle$ at $\|\phi(a)\|_1$. There is a large literature on Bloom filters with applications ranging from networking and database systems to neural coding, and several schemes for generating good codewords have been proposed (Broder & Mitzenmacher, 2004; Pagh et al., 2005; Dasgupta et al., 2018). Using the random coding scheme described here, the optimal value of p can be seen to be $(\ln 2)/s$ and, to ensure the probability of a false positive is at most δ, the encoding dimension should be chosen on the order of $s \ln(1/\delta)$ (Broder & Mitzenmacher, 2004). A practical benefit of Bloom filters is that they have an efficient implementation using hash functions which does not require materializing a codebook as in methods based on random sampling. This may be beneficial when the alphabet size is large enough that storing codewords is not possible. The connections between HD computing and Bloom filters are examined in greater detail in (Kleyko et al., 2019).

We remark that this method of encoding is related to an interesting procedure known as "context dependent thinning" (CDT) which can be used to control the density of binary representations (Rachkovskij, 2001; Kleyko et al., 2018). CDT takes the logical "and" of $\phi(\mathcal{S})$ and some permutation $\sigma(\phi(\mathcal{S}))$ to obtain the thinned representation $\phi(\mathcal{S})' = \phi(\mathcal{S}) \wedge \sigma(\phi(\mathcal{S}))$. This process can be repeated until the desired density of $\phi(\mathcal{S})$ is achieved. A capacity analysis of CDT representations can be found in (Kleyko et al., 2018).

3.2 Robustness to Noise

In this section we explore the noise robustness properties of the encoding methods discussed above using the formalism of incoherence. We consider some unspecified noise process which corrupts the encoding of a set $\mathcal{S} \subset \mathcal{A}$ of size at most s according to $\tilde{\phi}(\mathcal{S}) = \phi(\mathcal{S}) + \Delta_{\mathcal{S}}$. We say $\Delta_{\mathcal{S}}$ is ρ-bounded if:

$$\max_{a \in \mathcal{A}} |\langle \phi(a), \Delta_{\mathcal{S}} \rangle| \leq \rho.$$

We are interested in understanding the conditions under which we can still decode reliably.

Theorem 10 *Suppose $\mathcal{S}$ has size $\leq s$ and $\Delta_{\mathcal{S}}$ is ρ-bounded. We can correctly decode $\mathcal{S}$ using the thresholding rule from Theorem 2 if:*

$$\frac{\rho}{L^2} + s\mu < \frac{1}{2}$$

where $L = \min_{a \in \mathcal{A}} \|\phi(a)\|_2$.

The proof is a straightforward extension of Theorem 2 and is available in the appendix. The practical implication is that there is a tradeoff between the incoherence of the codebook and robustness to noise: a higher incoherence allows for a smaller encoding dimension but at the cost of a tighter constraint on ρ. We can analyze several practically relevant noise models by placing additional restrictions on $\Delta_{\mathcal{S}}$ and by considering worst or typical case bounds on ρ. We here consider different forms of noise under constraints on $\mathcal{H}$. Our goal is to understand how the magnitude of noise that can be tolerated scales with the encoding dimension, size s of the encoded set, and size m of the alphabet. In each setting we consider a "passive" model in which the noise is sampled randomly from some distribution, and an "adversarial" model in which the noise is arbitrary and may be designed to maliciously corrupt the encodings. We again appeal to Theorem 4 to justify a simplifying assumption that the codewords are of equal length.

Lemma 11 ***Sub-Gaussian Codewords.*** *Fix a centered and σ-sub-Gaussian codebook ϕ whose codewords are of length L. Consider the passive additive white Gaussian noise model $\Delta_{\mathcal{S}} \sim \mathcal{N}(0, \sigma_\Delta^2 \mathbf{I}_d)$; that is, each coordinate is corrupted by Gaussian noise with mean zero and variance σ_Δ^2. Then, we can correctly decode with probability $1-\delta$ over random draws of $\Delta_{\mathcal{S}}$ provided:*

$$\sigma_\Delta < \frac{L}{\sqrt{2\ln(2m/\delta)}}\left(\frac{1}{2} - s\mu\right)$$

Now consider an adversarial model in which $\Delta_{\mathcal{S}}$ is arbitrary save for a constraint on the norm: $\|\Delta_{\mathcal{S}}\|_2 \leq \omega L$. Then, we can correctly decode provided:

$$\omega < \frac{1}{2} - s\mu$$

Proof Let us first consider the passive case in which $\Delta_{\mathcal{S}} \sim \mathcal{N}(0, \sigma_\Delta^2 \mathbf{I}_d)$. Fix some $a \in \mathcal{A}$. Then $\langle \phi(a), \Delta_{\mathcal{S}} \rangle \sim \mathcal{N}(0, \sigma_\Delta^2 L^2)$. By a standard tail bound on the Gaussian distribution (Wainwright, 2019) and the union bound, we have:

$$\mathbb{P}(\exists\, a \text{ s.t. } |\langle \phi(a), \Delta_{\mathcal{S}} \rangle| \geq \rho) \leq 2m \exp\left(-\frac{\rho^2}{2\sigma_\Delta^2 L^2}\right).$$

Therefore, with probability $1-\delta$, we have that $\Delta_{\mathcal{S}}$ is ρ-bounded for

$$\rho \leq \sigma_\Delta L \sqrt{2\ln(2m/\delta)}.$$

By Theorem 10 we can decode correctly if:

$$\frac{\sigma_\Delta L\sqrt{2\ln(2m/\delta)}}{L^2} + s\mu < \frac{1}{2} \Rightarrow \sigma_\Delta < \frac{L}{\sqrt{2\ln(2m/\delta)}}\left(\frac{1}{2} - s\mu\right).$$

Now consider the adversarial case in which $\|\Delta_{\mathcal{S}}\|_2 \leq \omega L$. By the Cauchy-Schwarz inequality, $|\langle \phi(a), \Delta_{\mathcal{S}} \rangle| \leq \omega L^2$. Therefore, by Theorem 10, we can decode correctly if

$$\frac{\omega L^2}{L^2} + s\mu < \frac{1}{2} \Rightarrow \omega < \frac{1}{2} - s\mu.$$

■

We again emphasize that, per Theorem 5, $\mu = O(\sqrt{(\ln m)/d})$. Since $L = O(\sqrt{d})$, we can see that we can tolerate $\sigma_\Delta \approx \sqrt{d/(\ln m)} - s$ in the passive case. We next turn to a notable special case of the above in which the codewords are dense and binary. In this case, we may assume that $\mathcal{H}$ is constrained to be integers in the range $[-s, s]$.

Lemma 12 ***Dense Binary Codewords****. Fix a codebook ϕ such that $\phi(a) \sim \{\pm 1\}^d$ for each $a \in \mathcal{A}$. Consider a passive noise model in which $\Delta_\mathcal{S} \sim \text{unif}(\{-c, ..., c\}^d)$; that is, each coordinate is shifted by an integer amount chosen uniformly at random between $-c$ and c. Then, we can correctly decode with probability $1 - \delta$ provided:*

$$c < \sqrt{\frac{d}{2\ln(2m/\delta)}}\left(\frac{1}{2} - s\mu\right)$$

Now consider an adversarial model in which we assume $\|\Delta_\mathcal{S}\|_1 \leq \omega s d$. Then we can decode correctly if:

$$\omega < \frac{1}{2s} - \mu.$$

A proof is available in the Appendix. We next consider the case of Section 3.1.5 in which the codewords are sparse and binary and the bundling operator is the element-wise maximum. We here assume that $\tilde{\phi}(\mathcal{S}) = \phi(\mathcal{S}) + \Delta_\mathcal{S}$ is truncated so that each coordinate is either 0 or $+1$.

Lemma 13 ***Sparse Binary Codewords****. Fix a codebook ϕ such that $\phi(a) \in \{0, 1\}^d$, and assume some fraction $p \ll 1/2$ of coordinates are non-zero for each $a \in \mathcal{A}$. Consider a passive noise model in which:*

$$\Delta_\mathcal{S} \sim \begin{cases} -1 & w.p.\ \frac{\theta}{2} \\ 0 & w.p.\ 1-\theta \\ +1 & w.p.\ \frac{\theta}{2}. \end{cases}$$

Then we can decode correctly with probability $1 - \delta$ provided:

$$\theta < \frac{1}{2} - 2s\mu - \sqrt{\frac{1}{2dp}\ln\frac{2m}{\delta}}.$$

Now consider an adversarial model in which $\|\Delta_\mathcal{S}\|_1 \leq \omega d$. Then we can decode correctly if $\omega < p(\frac{1}{2} - s\mu)$.

Proof Consider first the passive noise model. Fix some $\phi(a)$. Then:

$$|\langle \phi(a), \Delta_\mathcal{S} \rangle| \leq \sum_{i=1}^{d} |\phi(a)^{(i)} \Delta_\mathcal{S}^{(i)}|.$$

Treating $\phi(a)$ as a fixed vector with dp non-zero entries, the sum is concentrated in the range $dp(\theta \pm \epsilon)$, and so $\rho \leq dp(\theta + \epsilon)$ with high probability. By Chernoff/Hoeffding and the union-bound, with probability $1 - \delta$:

$$\epsilon \leq \sqrt{\frac{1}{2dp}\ln\frac{2m}{\delta}}.$$

The result is obtained by noting that $L = \sqrt{pd}$ and applying Theorem 10.

For the adversarial case, the result is obtained by again observing that $|\langle \phi(a), \Delta_{\mathcal{S}} \rangle| \leq \|\phi(a)\|_\infty \|\Delta_{\mathcal{S}}\|_1 \leq \omega d$ for any $a \in \mathcal{A}$ and applying Theorem 10. ■

4. Encoding Structures

We are often interested in representing more complex data types, such as objects with multiple attributes or "features." In general, we suppose that we observe a set of features $\mathcal{F}$ whose values are assumed to lie in some set $\mathcal{A}$. Let $\psi : \mathcal{F} \to \mathcal{H}$ be an embedding of features, and $\phi : \mathcal{A} \to \mathcal{H}$ be an embedding of values. We associate a feature with its value through use of the *binding* operator $\otimes : \mathcal{H} \times \mathcal{H} \to \mathcal{H}$ that creates an embedding for a (feature,value) pair. For a feature $f \in \mathcal{F}$ taking on a value $a \in \mathcal{A}$, its embedding is constructed as $\psi(f) \otimes \phi(a)$. A data point $\mathbf{x} = \{(f_i \in \mathcal{F}, x_i \in \mathcal{A})\}_{i=1}^n$ consists of n such pairs. For simplicity, we assume each x possesses all attributes, although our analysis also applies to the case that x possesses only some subset of attributes. The entire embedding for $\mathbf{x}$ is constructed as (Plate, 2003):

$$\phi(\mathbf{x}) = \bigoplus_{i=1}^{n} \psi(f_i) \otimes \phi(x_i) \tag{2}$$

As with sets we would typically like $\phi(\mathbf{x})$ to be *decodable* in the sense that we can recover the value associated with a particular feature, and *comparable* in the sense that $\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$ is reflective of a reasonable notion of similarity between $\mathbf{x}$ and $\mathbf{x}'$.

From a formal perspective, we require the binding operator to satisfy several properties. First, binding should be associative and commutative. That is, for all $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathcal{H}$, $(\mathbf{a} \otimes \mathbf{b}) \otimes \mathbf{c} = \mathbf{a} \otimes (\mathbf{b} \otimes \mathbf{c})$, and $\mathbf{a} \otimes \mathbf{b} = \mathbf{b} \otimes \mathbf{a}$. Second, there should exist an identity element $\mathbf{I} \in \mathcal{H}$ such that $\mathbf{I} \otimes \mathbf{a} = \mathbf{a}$ for all $\mathbf{a} \in \mathcal{H}$. Third, for all $\mathbf{a} \in \mathcal{H}$, there should exist some $\mathbf{a}^{-1} \in \mathcal{H}$ such that $\mathbf{a} \otimes \mathbf{a}^{-1} = \mathbf{I}$. These properties are equivalent to stipulating that $\mathcal{H}$ be an abelian group under $\otimes$. Furthermore, binding should distribute over bundling. That is, for any $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathcal{H}$, it should be the case that $\mathbf{a} \otimes (\mathbf{b} + \mathbf{c}) = \mathbf{a} \otimes \mathbf{b} + \mathbf{a} \otimes \mathbf{c}$. We here also require that the lengths of bound pairs are bounded, that is to say: $\max_{f \in \mathcal{F}, a \in \mathcal{A}} \|\psi(f) \otimes \phi(a)\|_2 \leq M$.

A natural choice of embedding satisfying these properties is to sample $\psi(f)$ randomly from $\{\pm 1\}^d$ and choose $\otimes$ to be the element-wise product. In this case $\psi(f)$ is its own inverse, that is $\psi(f) \otimes \psi(f) = \mathbf{I}$, and binding preserves lengths of codewords. We focus on this case here as it is intuitive, but our analysis generalizes in a straightforward way to any particular implementation satisfying the properties listed above. One can see the bound pairs satisfy various incoherence properties with high probability. For instance, we may declare the binding to be μ-incoherent if:

$$\max_{a \in \mathcal{A}} \max_{a' \in \mathcal{A}, f \in \mathcal{F}} \langle \phi(a), \psi(f) \otimes \phi(a') \rangle \leq \mu L^2$$

where $L = \min_{a \in \mathcal{A}} \|\phi(a)\|_2$. We can extend Theorem 5 to see this property is satisfied with high probability:

Theorem 14 *Fix $d, n, m \in \mathbb{Z}^+$ and $\mu \in \mathbb{R}^+$. Let ϕ be centered and σ-sub-Gaussian, $\otimes$ be the element-wise product, and $\psi(f) \sim \{\pm 1\}^d$. Then:*

$$\mathbb{P}(\exists\, a, a' \in \mathcal{A}, f \in \mathcal{F} \text{ s.t. } |\langle \phi(a), \phi(a') \otimes \psi(f) \rangle| \geq \mu L^2) \leq nm^2 \exp\left(-\frac{\kappa \mu^2 L^2}{2\sigma^2}\right)$$

where $L = \min_{a \in \mathcal{A}} \|\phi(a)\|_2$ and κ is as defined in Theorem 5.

The proof is similar to Theorem 5 and is available in the Appendix. This result is appealing because it means that the incoherence scales only logarithmically with $m \times n$ which may be large in practice. As a corollary to the previous theorem, we also obtain the following useful incoherence property:

$$\mathbb{P}(\exists\, a, a', f \neq f' \text{ s.t. } |\langle \phi(a), (\phi(a') \otimes \psi(f)) \otimes \psi^{-1}(f') \rangle| \geq \mu L^2) \leq m^2 n^2 \exp\left(-\frac{\kappa \mu^2 L^2}{2\sigma^2}\right) \tag{3}$$

where $\psi^{-1}(f)$ is the inverse of $\psi(f)$ with respect to $\otimes$. This notion of incoherence is useful for decoding representations. Along similar lines:

$$\mathbb{P}(\exists\, a, a', f \neq f' \text{ s.t. } |\langle \phi(a) \otimes \psi(f), \phi(a') \otimes \psi(f') \rangle| \geq \mu L^2) \leq m^2 n^2 \exp\left(-\frac{\kappa \mu^2 L^2}{2\sigma^2}\right) \tag{4}$$

We note that the previous statement refers to symbols associated with different attributes and thus does not require any particular incoherence assumption on the $\phi(a)$.

4.1 Decoding Structures

This representation can be decoded to recover the value associated with a particular feature. To recover the value of the i-th feature, we use the following rule:

$$\hat{x}_i = \operatorname*{argmax}_{a \in \mathcal{A}} \langle \phi(a), \phi(\mathbf{x}) \otimes \psi^{-1}(f_i) \rangle$$

where $\psi^{-1}(f)$ denotes the group inverse of $\psi(f)$. Since the binding operator is assumed to distribute over bundling, the dot-product above expands to:

$$\langle \phi(a), \phi(x_i) \rangle + \sum_{j \neq i} \langle \phi(a), (\phi(x_j) \otimes \psi(f_j)) \otimes \psi^{-1}(f_i) \rangle$$

$$\begin{cases} \geq L^2(1 - n\mu) & \text{if } x_i = a \\ \leq nL^2\mu & \text{otherwise} \end{cases}$$

where the incoherence can be bounded as as in Equation 3. Thus $\mu < 1/(2n)$ is a sufficient condition for decodability.

4.2 Comparing Structures

As with sets, we may wish to compare two structures without decoding them. As one would expect given Theorem 9, this is can be achieved by computing the dot-product between their encodings:

Theorem 15 *Let* $\mathbf{x}$ *and* $\mathbf{x}'$ *be two structures drawn from a common alphabet* $\mathcal{F} \times \mathcal{A}$*. Denote their encodings using Equation 2 by* $\phi(\mathbf{x})$ *and* $\phi(\mathbf{x}')$*. Then, if binding is* μ*-incoherent:*

$$|\mathbf{x} \cap \mathbf{x}'| - n^2\mu \leq \frac{1}{L^2}\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle \leq |\mathbf{x} \cap \mathbf{x}'| + n^2\mu$$

where $\mathbf{x} \cap \mathbf{x}'$ *is defined to be the set* $\{i : x_i = x'_i\}_{i=1}^n$*, that is, the features on which* $\mathbf{x}$ *and* $\mathbf{x}'$ *agree.*

Proof Expanding:

$$\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle = \langle \sum_{i=1}^{n} \phi(x_i) \otimes \psi(f_i), \sum_{j=1}^{n} \phi(x'_j) \otimes \psi(f_j) \rangle$$

$$= \sum_{i=1}^{n} \langle \phi(x_i) \otimes \psi(f_i), \phi(x'_i) \otimes \psi(f_i) \rangle + \sum_{i \neq j} \langle \phi(x_i) \otimes \psi(f_i), \phi(x'_j) \otimes \psi(f_j) \rangle$$

A term in the first sum is L^2 if $x_i = x'_i$ and bounded in $\pm L^2\mu$ otherwise. So the expression above is bounded as:

$$\leq L^2 |\mathbf{x} \cap \mathbf{x}'| + L^2 n^2 \mu$$

and the other direction of the inequality is analogous. ■

As a practical example, in bioinformatics it is common to search for regions of high similarity between a "reference" and "query" genome. Work in (Imani et al., 2018) and (Kim et al., 2020) explored the use HD computing to accelerate this process by encoding short segments of DNA and estimating similarity on the HD representations.

4.3 Encoding Sequences

Sequences are an important form of structured data. In this case, the feature set is simply the list of positions $\{1, 2, 3, ...\}$ in the sequence. In practical applications, we are often interested in streams of data which arrive continuously over time. Typically, real-world processes do not exhibit infinite memory and we only need to store the $n \geq 1$ most recent observations at any time. In the streaming setting, we would like to avoid needing to fully re-encode all n data points each time we receive a new sample, as would be the case using the method described above. This motivates the use of shift based encoding schemes (Kanerva, 2009; Rahimi et al., 2017; Kim et al., 2018). Let $\rho^{(i)}(\mathbf{z})$ denote a cyclic left-shift of the elements of $\mathbf{z}$ by i coordinates, and $\rho^{(-i)}(\mathbf{z})$ denote a cyclic right-shift by i coordinates. In other words: $\rho^{(1)}((z_1, z_2, \ldots, z_{d-1}, z_d)) = (z_2, z_3, \ldots, z_d, z_1)$. In shift-based encoding a sequence $\mathbf{x} = (x_1, ..., x_n)$ is represented as:

$$\phi(\mathbf{x}) = \bigoplus_{i=1}^{n} \rho^{(n-i)}(\phi(x_i)),$$

where we take $\oplus$ to be the element wise sum. Now suppose we receive symbol $n+1$ and wish to append it to $\phi(\mathbf{x})$ while removing $\phi(x_1)$. Then we may apply the rule:

$$\rho^{(1)}(\phi(\mathbf{x}) - \rho^{(n-1)}(\phi(x_1))) \oplus \phi(x_{n+1}) = \bigoplus_{i=1}^{n} \rho^{(n-i)} \phi(x_{i+1})$$

where we can additionally note that ρ is a special type of permutation and that permutations distribute over sums. However, in order to decode correctly, each $\phi(a)$ must satisfy an incoherence condition with the $\rho^{(j)}(\phi(a'))$. We can again use the randomly generated nature of the codewords to argue this is the case; however, we must here impose the additional restriction that the $\phi(a)$ be bounded, and accordingly restrict attention to the case $\phi(a) \sim \{\pm 1\}^d$.

Theorem 16 *Fix $d, m, n < d \in \mathbb{Z}^+$ and $\mu \in \mathbb{R}^+$ and let $\phi(a) \sim \{\pm 1\}^d$. Then:*

$$\mathbb{P}(\exists\, a, a' \in \mathcal{A}, i \neq 0 \text{ s.t. } |\langle \phi(a), \rho^{(i)}(\phi(a'))\rangle| \geq \mu d) \leq nm^2 \exp\left(-\frac{\mu^2 d}{4}\right)$$

Proof Fix some a, a' and i. In the case that $a \neq a'$, $\phi(a)$ and $\rho^{(i)}(\phi(a))$ are mutually independent. However, when $a = a'$, $\phi(a)$ and $\rho^{(i)}(\phi(a))$ only satisfy pairwise independence and the techniques of Theorem 5 cannot be applied. To resolve this difficulty, let $f(\phi(a)) = \langle \phi(a), \rho^{(i)}(\phi(a))\rangle$, and denote by $\phi(a)^{\backslash k}$ the vector formed by replacing the k-th coordinate in $\phi(a)$ with an arbitrary value $\in \{+1, -1\}$. Then $|f(\phi(a)) - f(\phi(a)^{\backslash k})| \leq 4$ and so by the bounded-differences inequality (McDiarmid et al., 1989):

$$\mathbb{P}(|\langle \phi(a), \rho^{(i)}(\phi(a'))\rangle| \geq \mu d) \leq 2 \exp\left(-\frac{\mu^2 d}{4}\right).$$

The result follows by the union bound. ■

Several other related methods for encoding sequential information have been proposed in the literature (Plate, 2003; Gallant & Okaywe, 2013). For an extensive discussion of these approaches as well as an interesting discussion involving sequences of infinite length, the reader is referred to (Frady et al., 2018).

4.4 Discussion and Comparison with Prior Work

We conclude our treatment of encoding and decoding discrete data with some brief discussion of our approach and its relation to antecedents in the literature. A key question addressed here and by several pieces of prior work is to bound the magnitude of crosstalk noise in terms of the encoding dimension (d), the number of items to encode (s) and the alphabet size (m). Early analysis in (Plate, 2003; Gallant & Okaywe, 2013; Kleyko et al., 2018) recovers the same asymptotic relationship as we do, but only under specific assumptions about the method used to generate the codewords and particular instantiations of the bundling and binding operators.

Work in (Frady et al., 2018) provides a significantly more general treatment which, like ours, aims to abstract away from the particular choice of distribution from which codewords are sampled and from the particular implementation of bundling and binding operator. Their approach assumes the codewords are generated by sampling each component i.i.d. from some distribution and uses the central limit theorem (CLT) to justify modeling the crosstalk noise by a Gaussian distribution. Error bounds in the non-asymptotic setting are then obtained by applying a Chernoff style bound to the resulting Gaussian distribution. This approach again recovers the same asymptotic relationship between d, s and m as us, but does not generally yield formal bounds in the non-asymptotic setting. Our approach

based on sub-Gaussianity formalizes this analysis in the non-asymptotic setting. Like us, (Frady et al., 2018) also considers the effect of noise on the HD representations, but their treatment is limited to additive white noise, whereas we address both arbitrary additive passive noise and adversarial noise.

In summary, our formalism using the notion of incoherence allows us to decouple the analysis of decoding and noise-robustness from any particular method for generating codewords and readily yields rigorous bounds in the non-asymptotic setting. Our approach is applicable to a large swath of HD computing and enables us to offer more general conditions under which thresholding based decoding schemes will succeed and of the effect of noise than is available in prior work.

5. Encoding Euclidean Data

One option for encoding Euclidean vectors is to treat them as a special case of the "structured data" considered in the preceding section. As before, we think of our data as a collection of (feature,value) pairs $\mathbf{x} = \{(f_i, x_i)\}_{i=1}^n$ with the important caveat that $x_i \in \mathbb{R}^n$. This case is more complex because the feature values may now be continuous, and because the data possesses geometric structure which is typically relevant for downstream tasks and must be preserved by encoding. We here analyze two of the most widely used methods for encoding Euclidean data and discuss general properties of structure preserving embeddings in the context of HD computing.

5.1 Position-ID Encoding

A widely-used method in practice is to quantize the raw signal to a suitably low precision and then apply the structure encoding method discussed in the previous section (Rachkovskiy et al., 2005a, 2005b; Kleyko et al., 2018; Rahimi et al., 2018).

In this approach, we first quantize the support of each feature $f \in \mathcal{F}$ into some set of m bins with centroids $a_1 < \cdots < a_m$ and assign each bin a codeword $\phi(a) \in \mathcal{H}$. However, instead of requiring the codewords to be incoherent, we now require the correlation between codewords to reflect the distance between corresponding quantizer bins. In other words $\langle \phi(a), \phi(a') \rangle$ should be monotonically decreasing in $|a - a'|$.

A simple method can be used to generate monotonic codebooks when the codewords are randomly sampled from $\{\pm 1\}^d$ (Rachkovskiy et al., 2005a; Widdows & Cohen, 2015). Fixing some feature f, the codeword for the minimal quantizer bin, $\phi(a_1)$, is generated by sampling randomly from $\{\pm 1\}^d$. To generate the codeword for the second bin, we simply flip some set of $\lceil b \rceil$ bits in $\phi(a_1)$, where:

$$b = \frac{a_2 - a_1}{a_m - a_1} \cdot \frac{d}{2}$$

The codeword for the third bin is generated analogously from the second, where we assume the bits to be flipped are sampled such that a bit is flipped at most once. Thus the codewords for the minimal and maximal bins are orthogonal and the correlation between codewords for intermediate bins is monotonically decreasing in the distance between their corresponding bin centroids.

In practice, it seems to be typical to use a single codebook for all features and for the quantizer to be a set of evenly spaced bins over the support of the data. While simple, this approach is likely to have sub-optimal rate when the features are on different scales or are far from the uniform distribution. Encoding then proceeds as follows:

$$\phi(\mathbf{x}) = \sum_{i=1}^{n} \phi(x_i) \otimes \psi(f_i)$$

where, as before $\psi \in \{\pm 1\}^d$ is a vector which encodes the index i of a feature value x_i as in the previous section on encoding sequences; hence the name "position-ID" encoding. There are several variations on this theme which are compared empirically in (Kleyko et al., 2018).

This general encoding method was analyzed by (Rachkovskiy et al., 2005b), in the specific case of sparse and binary codewords, who show it preserves the L1 distance between points in expectation but do not provide distortion bounds. We here provide such bounds using our formalism of matrix incoherence. We assume that the underlying quantization of the points is sufficiently fine that it is a low-order term that can be ignored.

Theorem 17 *Let* $\mathbf{x}$ *and* $\mathbf{x}'$ *be points in* $[0,1]^n$ *with encodings* $\phi(\mathbf{x})$ *and* $\phi(\mathbf{x}')$ *generated using the rule described above. Assume that* ϕ *satisfies* $\langle \phi(a), \phi(a') \rangle = d(1 - |a - a'|)$ *for all* $a, a' \in \mathcal{A}$*, and let* $\psi \sim \{\pm 1\}^d$*. Then, for all* $\mathbf{x}, \mathbf{x}'$*:*

$$2d(\|\mathbf{x} - \mathbf{x}'\|_1 - 2n^2\mu) \leq ||\phi(\mathbf{x}) - \phi(\mathbf{x}')||_2^2 \leq 2d(\|\mathbf{x} - \mathbf{x}'\|_1 + 2n^2\mu)$$

The proof is similar to Theorem 15 and is available in the Appendix.
The practical implication of the previous theorem is that the position-ID encoding method preserves the L1 distance between points up to an additive distortion which can be bounded by the incoherence of the codebook. Per Equation 4, $\mu = O(\sqrt{\ln(mn)/d})$. Therefore, to ensure that $\frac{1}{d}\|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|_2^2 \approx \|\mathbf{x} - \mathbf{x}'\|_1 \pm \epsilon$, the previous result implies we should choose $d = O(\frac{n^4}{\epsilon^2} \ln(nm))$. This can be relaxed to a quadratic dependence on n in exchange for a weaker pointwise bound, but in either case means the encoding method may be problematic when the dimension of the underlying data is high.

Noting that $||\phi(\mathbf{x})||_2^2 \in nd \pm n^2 d\mu$, we can see that the encodings of each point are roughly of equal norm and lie in a ball of radius at most $n\sqrt{d\mu}$, where the exact position depends on the instantiation of the codebook. Thus, we can loosely interpret the encoding procedure as mapping the data into a thin shell around the surface of a high dimensional sphere.

5.2 Random Projection Encoding

Another popular family of encoding methods embeds the data into $\mathcal{H}$ under some random linear map followed by a quantization (Rachkovskij, 2015; Imani et al., 2019c). More formally, for some $\mathbf{x} \in \mathbb{R}^n$, these embeddings take the form:

$$\phi(\mathbf{s}) = g(\mathbf{\Phi}\mathbf{x})$$

where $\mathbf{\Phi} \in \mathbb{R}^{d \times n}$ is a matrix whose rows are sampled uniformly at random from the surface of the n-dimensional unit sphere, and g is a quantizer — typically the sign function

— restricting the embedding to $\mathcal{H}$. The embedding matrix $\mathbf{\Phi}$ may also be quantized to lower precision. This encoding method has also been studied in the context of kernel approximation where it is used to approximate the angular kernel (Choromanski et al., 2017), and to construct low-distortion binary embeddings (Jacques et al., 2013; Plan & Vershynin, 2014). While the following result is well known, we here show this encoding method preserves angular distance up to an additive distortion as this fact is important for subsequent analysis.

Theorem 18 *Let $\mathcal{S}^{n-1} \subset \mathbb{R}^n$ denote the n-dimensional unit sphere. Let $\mathbf{\Phi} \in \mathbb{R}^{d \times n}$ be a matrix whose rows are sampled uniformly at random from $\mathcal{S}^{n-1}$. Let $\mathcal{X}$ be a set of points supported on $\mathcal{S}^{n-1}$. Denote the embedding of a point by $\phi(\mathbf{x}) = \text{sign}(\mathbf{\Phi}\mathbf{x})$. Then, for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, with high probability:*

$$d\theta - O(\sqrt{d}) \leq d_{ham}(\phi(\mathbf{x}), \phi(\mathbf{x}')) \leq d\theta + O(\sqrt{d})$$

where $d_{ham}(a, b)$ is the Hamming distance between a and b, defined to be the number of coordinates on which a and b differ, and $\theta = \frac{1}{\pi}\cos^{-1}(\langle \mathbf{x}, \mathbf{x}' \rangle) \in [0, 1]$ is proportional to the angle between $\mathbf{x}$ and $\mathbf{x}'$.

Proof Let $\mathbf{\Phi}^{(i)}$ denote the ith row of the matrix $\mathbf{\Phi}$. Then, the ith coordinate in the embedding of $\mathbf{x}$ can be written as $\text{sign}(\langle \mathbf{\Phi}^{(i)}, \mathbf{x} \rangle)$. The probability that the embeddings differ on their ith coordinate, that is $(\langle \mathbf{\Phi}^{(i)}, \mathbf{x} \rangle)(\langle \mathbf{\Phi}^{(i)}, \mathbf{x}' \rangle) < 0$, is exactly $\angle(\mathbf{x}, \mathbf{x}')/\pi$: the angle (in radians) between $\mathbf{x}$ and $\mathbf{x}'$ divided by π.

Therefore, the number of coordinates on which $\phi(\mathbf{x})$ and $\phi(\mathbf{x}')$ disagree is, concentrated in the range, $d(\theta \pm \epsilon)$. By Chernoff/Hoeffding, we have that with probability $1 - \delta$:

$$d\epsilon \leq \sqrt{2d \ln \frac{2}{\delta}}.$$

■

Noting that $\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle = d - 2d_{ham}(\phi(\mathbf{x}), \phi(\mathbf{x}'))$, we obtain the following simple corollary:

Corollary 19 *Let ϕ and θ be as defined in Theorem 18. Then, with high probability:*

$$d(1 - 2\theta) - O(\sqrt{d}) \leq \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle \leq d(1 - 2\theta) + O(\sqrt{d})$$

To obtain a more explicit relationship with the dot product, we can use the first-order approximation $\cos^{-1}(x) \approx (\pi/2) - x$, to obtain $\theta \approx \frac{1}{2} - \frac{1}{\pi}\langle \mathbf{x}, \mathbf{x}' \rangle$, from which we obtain:

$$d(1 - 2\theta) \approx \frac{2d}{\pi}\langle \mathbf{x}, \mathbf{x}' \rangle.$$

We emphasize that, in comparison to the position-ID method, the distortion in this case does not depend on the dimension of the underlying data which means this method may be preferable when the data dimension is large.

5.2.1 Connection with Kernel Approximation

A natural question is whether the encoding procedure described above, which preserves dot-products, can be generalized to capture more diverse notions of similarity? We can answer in the affirmative by noting that the random projection encoding method is closely related to the notion of random Fourier features which have been widely used for kernel approximation (Rahimi & Recht, 2008). The basic idea is to construct an embedding $\phi : \mathbb{R}^n \to \mathbb{R}^d$, such that $\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle \approx k(\mathbf{x}, \mathbf{x}')$, where k is a shift-invariant kernel. The construction exploits the fact that the Fourier transform of a shift-invariant kernel k is a probability measure: a well known result from harmonic analysis known as Bochner's Theorem (Rudin, 1962). The embedding itself is given by $\phi(\mathbf{x}) = \frac{1}{\sqrt{d}} \cos(\mathbf{\Phi}\mathbf{x} + \mathbf{b})$, where the rows of $\mathbf{\Phi}$ are sampled from the distribution induced by k and the coordinates of $\mathbf{b}$ are sampled uniformly at random from $[0, 2\pi]$.

Subsequent work in (Raginsky & Lazebnik, 2009) gave a simple scheme for quantizing the embeddings produced from random Fourier features to binary precision. Their construction yields an embedding $\psi : \mathbb{R}^n \to \{0, 1\}^d$ such that:

$$f_1(k(\mathbf{x}, \mathbf{x}')) - \Delta \leq \frac{1}{d} d_{ham}(\psi(\mathbf{x}), \psi(\mathbf{x}')) \leq f_2(k(\mathbf{x}, \mathbf{x}')) + \Delta$$

where $f_1, f_2 : \mathbb{R} \to \mathbb{R}$ are independent of the choice of kernel, and Δ is a distortion term. The embedding itself is constructed by applying a quantizer $Q_t(x) = \text{sign}(x + t)$ coordinate wise over the embeddings constructed from random Fourier features. In other words $\psi(\mathbf{x})_i = \frac{1}{2}(1 + Q_{t_i}(\phi(\mathbf{x})_i))$, where $t_i \sim \text{Unif}[-1, 1]$, and $\phi(\mathbf{x})$ is a random Fourier feature.

This connection is highly appealing for HD computing. The quantized random Fourier feature scheme presents a simple recipe for constructing encoding methods meeting the desiderata of HD computing while preserving a rich variety of structure in data. For instance, shift-invariant kernels preserving the L1 and L2 distance—among many others—can be approximated using the method discussed above. Furthermore, this observation provides a natural point of contact between HD computing and the vast literature on kernel methods which has produced a wealth of algorithmic and theoretical insights.

5.3 Consequences of Distance Preservation

The encoding methods discussed above are both appealing because they preserve reasonable notions of distance between points in the original data. Distance preservation is a sufficient condition to establish other desirable properties of encodings, namely preservation of neighborhood/cluster structure, robustness to various forms of noise, and in some cases, preservation of linear separability. We address the first two items here and defer the latter for our discussion of learning on HD representations. We formalize our notion of distance preservation as follows:

Definition 20 ***Distance-Preserving Embedding****: Let $\delta_{\mathcal{X}}$ be a distance function on $\mathcal{X} \subset \mathbb{R}^n$ and δ_H be a distance function on $\mathcal{H}$. We say ϕ preserves $\delta_{\mathcal{X}}$ under δ_H if, there exist functions $\alpha, \beta : \mathbb{Z}^+ \to \mathbb{R}$ such that $\beta(d)/\alpha(d) \to 0$ as $d \to \infty$, and:*

$$\alpha(d)\delta_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') - \beta(d) \leq \delta_{\mathcal{H}}(\phi(\mathbf{x}), \phi(\mathbf{x}')) \leq \alpha(d)\delta_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') + \beta(d) \tag{5}$$

for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$.

We typically wish the distance function $\delta_{\mathcal{H}}$ on $\mathcal{H}$ to be simple to compute. In practice, it is often taken to be the Euclidean, Hamming, or angular distance. The position-ID method preserves the L1 distance with δ_H the squared Euclidean distance, $\alpha(d) = 2d$, and $\beta(d) \leq n^2 \mu d$; recall that in the constructions above, μ scales inversely with d and thus $\beta(d)/\alpha(d) \to 0$. The signed random-projection method preserves the angular distance with $\alpha(d) = O(d)$, $\beta(d) = O(\sqrt{d})$, and δ_H the Hamming, angular, or Euclidean distance.

5.3.1 Preservation of Cluster Structure

In general, there is no universally applicable definition of cluster structure. Indeed, numerous algorithms have been proposed in the literature to target various reasonable notions of what constitutes a "cluster" in the data. Preservation of a distance function accords naturally with K-means like algorithms which, given a set of data $\mathcal{X} \subset \mathbb{R}^n$ compute a set of centroids $\mathcal{C} = \{\mathbf{c}_i\}_{i=1}^k$, and define associated clusters as the Voronoi cells associated with each centroid. We here adopt this notion and state that cluster structure $\mathcal{C}$ is preserved if, for any $\mathbf{x} \in \mathcal{X}$:

$$\operatorname*{argmin}_{\mathbf{c} \in \mathcal{C}} \delta_{\mathcal{X}}(\mathbf{x}, \mathbf{c}) = \operatorname*{argmin}_{\mathbf{c} \in \mathcal{C}} \delta_{\mathcal{H}}(\phi(\mathbf{x}), \phi(\mathbf{c}))$$

In other words, that the set of points bound to a particular cluster centroid does not change under the encoding. We can restate the above as requiring that, for some point $\mathbf{x}$ bound to a cluster centroid $\mathbf{c}$, it is the case that:

$$\delta_{\mathcal{H}}(\phi(\mathbf{x}), \phi(\mathbf{c})) < \delta_{\mathcal{H}}(\phi(\mathbf{x}), \phi(\mathbf{c}'))$$

for any $\mathbf{c}' \in \mathcal{C} \setminus \{\mathbf{c}\}$. From Definition 20 we have:

$$\delta_{\mathcal{H}}(\phi(\mathbf{x}), \phi(\mathbf{c}')) - \delta_{\mathcal{H}}(\phi(\mathbf{x}), \phi(\mathbf{c})) \geq \alpha(d)(\delta_{\mathcal{X}}(\mathbf{x}, \mathbf{c}') - \delta_{\mathcal{X}}(\mathbf{x}, \mathbf{c})) - 2\beta(d)$$

for any $\mathbf{x} \in \mathcal{X}$ and $\mathbf{c}, \mathbf{c}' \in \mathcal{C}$. Rearranging the expressions above we can see the desired property will be satisfied if:

$$\frac{\beta(d)}{\alpha(d)} < \min_{\mathbf{x} \in \mathcal{X}} \min_{\mathbf{c}' \neq \mathbf{c}(\mathbf{x})} \frac{1}{2}(\delta_{\mathcal{X}}(\mathbf{x}, \mathbf{c}') - \delta_{\mathcal{X}}(\mathbf{x}, \mathbf{c}(\mathbf{x}))),$$

where $\mathbf{c}(\mathbf{x}) = \operatorname{argmin}_{\mathbf{c} \in \mathcal{C}} \delta_{\mathcal{X}}(\mathbf{x}, \mathbf{c})$ denotes the center in $\mathcal{C}$ closest to $\mathbf{x}$. A sufficient condition for the existence of some d satisfying this property is that $\alpha(d)$ is monotone increasing and that $\alpha(d)$ is faster growing than $\beta(d)$. This condition is satisfied for both the random projection and position-ID encoding methods.

5.3.2 Noise Robustness

It is also of interest to consider robustness to noise in the context of encoding Euclidean data. Suppose we have a set of points, $\mathcal{X}$, in $\mathbb{R}^n$, and a distance function of interest $\delta_{\mathcal{X}}(\cdot, \cdot)$ which is preserved *à la* Definition 20. Given an arbitrary point $\mathbf{x} \in \mathcal{X}$ we consider a noise model which corrupts $\phi(\mathbf{x})$ to $\phi(\mathbf{x}) + \Delta$, where Δ is some unspecified noise process. Along the lines of Section 3.2, we say Δ is ρ-bounded if:

$$\max_{\mathbf{x} \in \mathcal{X}} |\langle \phi(\mathbf{x}), \Delta \rangle| \leq \rho$$

Suppose we wish to ensure the encodings can distinguish between all points at a distance $\leq \epsilon_1$ from $\mathbf{x}$ and all points at a distance $\geq \epsilon_2$. That is:

$$\|\phi(\mathbf{x}) + \Delta - \phi(\mathbf{x}')\| < \|\phi(\mathbf{x}) + \Delta - \phi(\mathbf{x}'')\|$$

for all $\mathbf{x}' \in \mathcal{X}$ such that $\delta_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') \leq \epsilon_1$ and all $\mathbf{x}'' \in \mathcal{X}$ such that $\delta_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') \geq \epsilon_2$. We say that such an encoding is (ϵ_1, ϵ_2)-robust.

Theorem 21 *Let $\delta_{\mathcal{X}}$ be a distance function on $\mathcal{X} \subset \mathbb{R}^n$ and suppose ϕ is an embedding preserving $\delta_{\mathcal{X}}$ under the squared Euclidean distance on $\mathcal{H}$ as described in Definition 20. Suppose Δ is ρ-bounded noise. Then ϕ is (ϵ_1, ϵ_2) robust if:*

$$\rho < \frac{\alpha(d)}{4}(\epsilon_2 - \epsilon_1) - \frac{\beta(d)}{2}.$$

Proof Fix a point $\mathbf{x}$ whose encoding is corrupted as $\phi(\mathbf{x}) + \Delta$. Then for any $\mathbf{x}', \mathbf{x}'' \in \mathcal{X}$ with $\delta_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') \leq \epsilon_1$ and $\delta_{\mathcal{X}}(\mathbf{x}, \mathbf{x}'') \geq \epsilon_2$, we have:

$$\begin{aligned}
&\|\phi(\mathbf{x}) + \Delta - \phi(\mathbf{x}'')\|_2^2 - \|\phi(\mathbf{x}) + \Delta - \phi(\mathbf{x}')\|_2^2 \\
&= \|\phi(\mathbf{x}) - \phi(\mathbf{x}'')\|_2^2 - \|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|_2^2 - 2\langle\phi(\mathbf{x}''), \Delta\rangle + 2\langle\phi(\mathbf{x}'), \Delta\rangle \\
&\geq \alpha(d)\delta_{\mathcal{X}}(\mathbf{x}, \mathbf{x}'') - \beta(d) - \alpha(d)\delta_{\mathcal{X}}(\mathbf{x}, \mathbf{x}') - \beta(d) - 4\rho \\
&\geq \alpha(d)(\epsilon_2 - \epsilon_1) - 2\beta(d) - 4\rho \ > \ 0,
\end{aligned}$$

as desired. ■

As before, we may consider passive and adversarial examples.

Additive White Gaussian Noise. First consider the case that $\mathcal{H} = \mathbb{R}^d$ and $\Delta \sim \mathcal{N}(0, \sigma_\Delta^2 \mathbf{I}_d)$; that is, each coordinate of Δ has a Gaussian distribution with mean zero and variance σ_Δ^2. Then, as before, we can note that $\langle\phi(\mathbf{x}), \Delta\rangle \sim \mathcal{N}(0, \sigma_\Delta^2\|\phi(\mathbf{x})\|_2^2)$. Then, it is very likely (four standard deviations in the tail of the normal distribution) that $\rho < 4L\sigma_\Delta$, where $L = \max_{\mathbf{x}\in\mathcal{X}} \|\phi(\mathbf{x})\|_2$. So then, we have the desired robustness property if:

$$\sigma_\Delta < \frac{\alpha(d)}{16L}(\epsilon_2 - \epsilon_1) - \frac{\beta(d)}{8L}$$

Assuming that $\alpha(d)$ is faster growing in d than L and $\beta(d)$, there will exist some encoding dimension for which we can tolerate any given level of noise. In the case of the random projection encoding scheme described above $\alpha(d) = O(d), \beta(d) = O(\sqrt{d})$ and $L = \sqrt{d}$ exactly. And so we can tolerate noise on the order of:

$$\sigma_\Delta \approx \sqrt{d}\,(\epsilon_2 - \epsilon_1) - O(1)$$

For the position-ID encoding method, $\alpha(d) = O(d)$, $L = O(\sqrt{nd})$ and $\beta(d) = O(n^2 d\mu)$, and so we can tolerate noise:

$$\sigma_\Delta \approx \sqrt{\frac{d}{n}}((\epsilon_2 - \epsilon_1) - O(n^2\mu))$$

Adversarial Noise. We now consider the case that $\mathcal{H} = \{\pm 1\}$, as in the random-projection encoding method, and Δ is noise in which some fraction $\omega \cdot d$ of coordinates in $\phi(\mathbf{x})$ are maliciously corrupted by an adversary. Since $\|\Delta\|_1 \leq \omega d$, we have, for any $\mathbf{x} \in \mathcal{X}$:

$$|\langle \phi(\mathbf{x}), \Delta \rangle| \leq \|\phi(\mathbf{x})\|_\infty \|\Delta\|_1 \leq \omega d$$

So then we can tolerate ω on the order of:

$$\omega < \frac{\alpha(d)}{4d}(\epsilon_2 - \epsilon_1) - \frac{\beta(d)}{2d}$$

In the case of the random-projection encoding method this boils down to:

$$\omega \approx (\epsilon_2 - \epsilon_1) - \frac{1}{\sqrt{d}},$$

meaning the total number of coordinates that can be corrupted is $O(d(\epsilon_2 - \epsilon_1))$.

Robustness to Input Noise. A natural question is whether the HD representations also confer any robustness to noise in the input space $\mathcal{X}$ rather than the HD space $\mathcal{H}$. In general, preservation of distance does not imply any particular robustness to input noise and the answer to this question depends on the particulars of the encoding method in question. Since a general treatment is difficult to give, we will not pursue this matter in depth at present.

6. Learning on HD Data Representations

We now turn to the question of using HD representations in learning algorithms. Our goal is to clarify in what precise sense the HD encoding process can make learning easier. We study two ways in which this can happen: the encoding process can increase the separation between classes and/or can induce sparsity. Both of these characteristics can be exploited by neurally plausible algorithms to simplify learning. Throughout this discussion, we assume access to a set of N labelled examples $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where $\mathbf{x}_i$ lies in $[0,1]^n$ and $y_i \in \mathcal{C}$ is a categorical variable indicating the class label. In general, we are interested in the case that training examples arrive in a streaming, or online, fashion, although our conclusions apply to fixed and finite data as well.

6.1 Learning by Bundling

The simplest approach to learning with HD representations is to bundle together the training examples corresponding to each class into a set of exemplars—often referred to as "prototypes"—which are then used for classification (Kleyko et al., 2018; Rahimi et al., 2018; Burrello et al., 2018). More formally, as described in Section 2, we construct the prototype $\mathbf{c}_k$ for the k-th class as:

$$\mathbf{c}_k = \bigoplus_{i \text{ s.t. } y_i = k} \phi(\mathbf{x}_i)$$

and then assign a class label for some "query" point $\mathbf{x}_q$ as:

$$\hat{y} = \operatorname*{argmax}_{k \in \mathcal{C}} \frac{\langle \mathbf{c}_k, \phi(\mathbf{x}) \rangle}{||\mathbf{c}_k||} \tag{6}$$

This approach bears a strong resemblance to naive Bayes and Fisher's linear discriminant, which are both classic simple statistical procedures for classification (Bishop, 2007). Like these methods, the bundling approach is appealing due to its simplicity. However, it also shares their weaknesses in that it may fail to separate data that is in fact linearly separable.

6.2 Learning Arbitrary Linear Separators

Linear separability is one of the most basic types of structure that can aid learning. The theory of linear models is well developed and several simple, neurally plausible, algorithms for learning linear separators are known, for instance, the Perceptron and Winnow (Rosenblatt, 1958; Littlestone, 1988). Thus, if our data is linearly separable in low-dimensional space we would like it to remain so after encoding, so that these methods can be applied. We now show formally that preservation of distance is sufficient, under some conditions, to preserve linear separability.

Theorem 22 *Let $\mathcal{X}$ and $\mathcal{X}'$ be two disjoint, closed, and convex sets of points in $\mathbb{R}^n$. Let $\mathbf{p} \in \mathcal{X}$ and $\mathbf{q} \in \mathcal{X}'$ be the closest pair of points between the two sets. Suppose ϕ preserves L2 distance on $\mathcal{X}$ under the L2 distance on $\mathcal{H}$ in the sense of Definition 20. Then, the function $f(\mathbf{x}) = \langle \phi(\mathbf{x}), \phi(\mathbf{p}) - \phi(\mathbf{q}) \rangle - \frac{1}{2}(||\phi(\mathbf{p})||_2^2 - ||\phi(\mathbf{q})||_2^2)$ is positive for all $\mathbf{x} \in \mathcal{X}$ and negative for all $\mathbf{x}' \in \mathcal{X}'$ provided:*

$$\frac{\beta(d)}{\alpha(d)} < \frac{1}{2}\|\mathbf{p} - \mathbf{q}\|_2^2.$$

Proof We first observe:

$$\langle \phi(\mathbf{x}), \phi(\mathbf{p}) - \phi(\mathbf{q}) \rangle - \frac{1}{2}\left(\|\phi(\mathbf{p})\|_2^2 - \|\phi(\mathbf{q})\|_2^2\right) = \frac{1}{2}\|\phi(\mathbf{x}) - \phi(\mathbf{q})\|_2^2 - \frac{1}{2}\|\phi(\mathbf{x}) - \phi(\mathbf{p})\|_2^2.$$

We may then use Definition 20 to obtain:

$$\begin{aligned} f(\mathbf{x}) &= \frac{1}{2}\|\phi(\mathbf{x}) - \phi(\mathbf{q})\|_2^2 - \frac{1}{2}\|\phi(\mathbf{x}) - \phi(\mathbf{p})\|_2^2 \\ &\geq \frac{\alpha(d)}{2}\|\mathbf{x} - \mathbf{q}\|_2^2 - \frac{\alpha(d)}{2}\|\mathbf{x} - \mathbf{p}\|_2^2 - \beta(d) \\ &= \alpha(d)\left(\langle \mathbf{x}, \mathbf{p} - \mathbf{q} \rangle - \frac{1}{2}\left(\|\mathbf{p}\|_2^2 - \|\mathbf{q}\|_2^2\right)\right) - \beta(d). \end{aligned}$$

By a standard proof of the hyperplane separation theorem (e.g., Section 2.5.1 of (Boyd & Vandenberghe, 2014)),

$$\langle \mathbf{x}, \mathbf{p} - \mathbf{q} \rangle - \frac{1}{2}(\|\mathbf{p}\|_2^2 - \|\mathbf{q}\|_2^2) \geq \frac{1}{2}\|\mathbf{p} - \mathbf{q}\|_2^2$$

for any $\mathbf{x} \in \mathcal{X}$, and thus $f(\mathbf{x}) > 0$ if

$$\frac{\beta(d)}{\alpha(d)} < \frac{1}{2}\|\mathbf{p} - \mathbf{q}\|_2^2.$$

the proof for $\mathbf{x} \in \mathcal{X}'$ is analogous. ■

A natural question is whether a linear separator on the HD representation can capture a *nonlinear* decision boundary on the original data? The connection with kernel methods discussed in Section 5.2.1 presents one avenue for rigorously addressing this question. As noted there, the encoding function can sometimes be interpreted as approximating the feature map of a kernel, which in turn can be used to linearize learning problems in some settings (Shawe-Taylor et al., 2004). However, a thorough examination of this question is beyond the scope of the present work.

6.2.1 Learning Sparse Classifiers on Random Projection Encodings

The random projection encoding method can be seen to lead to representations that are *sparse* in the sense that a subset of just $k \ll d$ coordinates suffice for determining the class label. This setting accords naturally with the Winnow algorithm (Littlestone, 1988) which is known to make on the order of $k \log d$ mistakes when the target function class is a linear function of $k \leq d$ variables. This can offer substantially faster convergence than the Perceptron when the margin is small. Curiously, while the Perceptron algorithm is commonly used in the HD community, we are unaware of any work using Winnow for learning.

Theorem 23 *Let $\mathcal{X}$ and $\mathcal{X}'$ be two sets of points supported on the n-dimensional unit sphere and separated by a unit-norm hyperplane $\mathbf{w}$ with margin $\gamma = \min_{\mathbf{x} \in \mathcal{X}} |\langle \mathbf{x}, \mathbf{w} \rangle|$. Let $\mathbf{\Phi} \in \mathbb{R}^{d \times n}$ be a matrix whose rows are sampled from the uniform distribution over the n-dimensional unit-sphere. Define the encoding of a point $\mathbf{x}$ by $\phi(\mathbf{x}) = \mathbf{\Phi}\mathbf{x}$. With high probability, $\mathcal{X}$ and $\mathcal{X}'$ are linearly separable using just k coordinates in the encoded space, provided:*

$$d = \Omega\left(k \exp\left(\frac{n}{2k\gamma^2}\right)\right).$$

To prove the theorem we first use the following simple Lemma:

Lemma 24 *Suppose there exists a row $\mathbf{\Phi}^{(i)}$ of the projection matrix such that $\langle \mathbf{\Phi}^{(i)}, \mathbf{w} \rangle > 1 - \gamma^2/2$. Then $\langle \mathbf{\Phi}^{(i)}, \mathbf{x} \rangle$ is positive for any $\mathbf{x} \in \mathcal{X}$ and negative for any $\mathbf{x} \in \mathcal{X}'$.*

Proof The constraint on the dot product of $\mathbf{\Phi}^{(i)}$ and $\mathbf{w}$ implies $\|\mathbf{\Phi}^{(i)} - \mathbf{w}\|^2 = \|\mathbf{\Phi}^{(i)}\|^2 + \|\mathbf{w}\|^2 - 2\langle \mathbf{\Phi}^{(i)}, \mathbf{w} \rangle < \gamma^2$. Thus for any $\mathbf{x} \in \mathcal{X}$,

$$\langle \mathbf{\Phi}^{(i)}, \mathbf{x} \rangle = \langle \mathbf{w}, \mathbf{x} \rangle + \langle \mathbf{\Phi}^{(i)} - \mathbf{w}, \mathbf{x} \rangle \geq \gamma + \langle \mathbf{\Phi}^{(i)} - \mathbf{w}, \mathbf{x} \rangle \geq \gamma - \|\mathbf{\Phi}^{(i)} - \mathbf{w}\| > 0.$$

A similar argument shows that $\langle \mathbf{\Phi}^{(i)}, \mathbf{x} \rangle$ is negative on $\mathcal{X}'$. ■

Unfortunately, the probability of randomly sampling such a direction is tiny, on the order of γ^n. However, we might instead hope to sample k vectors that are weakly correlated with $\mathbf{w}$ and exploit their cumulative effect on $\mathbf{x}$. We say a vector $\mathbf{u} \in \mathbb{R}^n$ is ρ-correlated with $\mathbf{w}$ if $\langle \mathbf{u}, \mathbf{w} \rangle \geq \rho$. We are now in a position to prove the theorem.

Proof For $\mathbf{w} \in \mathcal{S}^{n-1}$ and $\rho \in (0, 1)$, let $\mathcal{C} = \{\mathbf{u} \in S^{n-1} : \langle \mathbf{u}, \mathbf{w} \rangle \geq \rho\}$ denote the spherical cap of vectors ρ-correlated with $\mathbf{w}$. Suppose we pick vectors $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(k)}$ uniformly at random from $\mathcal{C}$. Then, with probability at least $1/2$:

$$\frac{\langle \sum_j \mathbf{u}^{(j)}, \mathbf{w} \rangle}{\| \sum_j \mathbf{u}^{(j)} \|_2} \geq 1 - \frac{1}{2k\rho^2} \tag{7}$$

To see this, note that without loss of generality we may assume $\mathbf{w} = \mathbf{e}_1$, the first standard basis vector of $\mathbb{R}^n$, and write any $\mathbf{u} \in \mathbb{R}^n$ as $\mathbf{u} = (u_1, \mathbf{u}_R)$: the first coordinate and the remaining $n-1$ coordinates. Now, let $N = \langle \sum_j \mathbf{u}^{(j)}, \mathbf{w} \rangle = \sum_j \mathbf{u}_1^{(j)} \geq k\rho$. Then:

$$\begin{aligned}\left\| \sum_j \mathbf{u}^{(j)} \right\|_2^2 &= \left(\sum_j \mathbf{u}_1^{(j)} \right)^2 + \left\| \sum_j \mathbf{u}_R^{(j)} \right\|_2^2 \\ &= N^2 + \sum_j \|\mathbf{u}_R^{(j)}\|_2^2 + \sum_{i \neq j} \langle \mathbf{u}_R^{(i)}, \mathbf{u}_R^{(j)} \rangle \\ &\leq N^2 + k + \sum_{i \neq j} \langle \mathbf{u}_R^{(i)}, \mathbf{u}_R^{(j)} \rangle.\end{aligned}$$

The last term has a symmetric distribution around zero over random samplings of the $\mathbf{u}^{(j)}$. Thus, with probability $\geq 1/2$, it is ≤ 0, whereupon

$$\frac{\langle \sum_j \mathbf{u}^{(j)}, \mathbf{w} \rangle}{\| \sum_j \mathbf{u}^{(j)} \|_2} \geq \frac{N}{\sqrt{N^2 + k}} \geq 1 - \frac{k}{2N^2} \geq 1 - \frac{1}{2k\rho^2}.$$

To ensure the quantity above is at least $1 - \gamma^2/2$, we must have:

$$\rho^2 \geq \frac{1}{k\gamma^2}.$$

It now remains to compute the probability that a vector $\mathbf{\Phi}^{(i)}$ sampled uniformly from $\mathcal{S}^{n-1}$ lies in $\mathcal{C}$, or equivalently, that $\mathbf{\Phi}_1^{(i)} \geq \rho$. Noting that we may simulate a random direction on $\mathcal{S}^{n-1}$ by sampling $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_n)$ and normalizing, we obtain the reasonable approximation: $\mathbf{\Phi}_1^{(i)} \sim \mathcal{N}(0, 1/n)$. Therefore, the probability that $\mathbf{\Phi}_1^{(i)} \geq \rho$ is on the order of $e^{-n\rho^2/2}$. So we need:

$$d = \Omega\left(k \exp\left(\frac{n}{2k\gamma^2} \right) \right)$$

■

In summary, the random projection method in tandem with the Winnow algorithm seems to be well suited to the HD setting, where sparsity can be exploited to simplify learning.

7. Conclusion

To conclude, we lay out several research directions related to HD computing we believe it would be of particular interest to further explore. There are several interesting open problems related to encoding. Our analysis established preservation of only the most basic forms of structure in data. Can encoding procedures satisfying the desiderata of HD computing be designed that capture other forms of structure? The quantized random Fourier feature construction discussed in Section 5 presents one such option, but is only applicable to structure that can be captured using a shift-invariant kernel on a Euclidean space. For instance, can we devise encoding methods that exploit low-dimensional manifold structure in the data or which are adaptive and can be learned from a particular data set?

Several recent works have claimed, based on empirical evidence, that HD computing evinces one-shot learning (Burrello et al., 2018; Imani et al., 2017; Rahimi et al., 2018) in which a single labeled example is needed to learn a generalizable classifier (Thrun, 1996; Lake et al., 2011). However, this work has focused on settings in which specialized hand-crafted features could be extracted, and it is not clear to us that existing encoding procedures would lead to one-shot classifiers absent such outside information. We would be interested to explore whether the HD representation makes one-shot learning easier in any broader sense. We expect this will necessitate the use of more sophisticated encoding procedures that can learn salient properties of a given domain. For this latter point we see dictionary learning (Olshausen & Field, 1996) as a promising avenue for developing adaptive encoding procedures. Dictionary learning is a well studied problem and can be solved using online and neurally plausible methods (Arora et al., 2015; Mairal et al., 2010) and would thus seem to be a promising avenue to address the limitations of existing encoding procedures without sacrificing the simplicity and neural plausibility of existing HD based methods.

Acknowledgements

This work was supported in part by CRISP, one of six centers in JUMP, an SRC program sponsored by DARPA, in part by an SRC-Global Research Collaboration grant, GRC TASK 3021.001, GRC TASK 2942.001, DARPA-PA-19-03-03 Agreement HR00112090036, and also NSF grants 1527034, 1730158, 1826967, 2100237, 2112167, 2052809, 2003279, 1830399, 1911095, 2028040, and 1911095.

Appendix A. Proofs of Selected Theorems

A.1 Proof of Theorem 4

Proof The result is an immediate consequence of the Hanson-Wright inequality (Hanson & Wright, 1971; Rudelson et al., 2013) which holds that, for $\mathbf{x}$ a centered, d-dimensional, σ-sub-Gaussian random vector, and $\mathbf{A} \in \mathbb{R}^{d\times d}$ an arbitrary square matrix, the quadratic form $\mathbf{x}^T\mathbf{A}\mathbf{x}$ obeys the following concentration bound:

$$\mathbb{P}(|\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbb{E}[\mathbf{x}^T\mathbf{A}\mathbf{x}]| \geq t) \leq 2\exp\left(-c\min\left(\frac{t^2}{\sigma^4\|\mathbf{A}\|_F^2}, \frac{t}{\sigma^2\|\mathbf{A}\|}\right)\right)$$

where c is a positive absolute constant, $\|\mathbf{A}\|_F^2 = \sum_{i,j}|\mathbf{A}_{ij}|^2$ is the Frobenius norm and $\|\mathbf{A}\| = \max_{\|\mathbf{x}\|\leq 1}\|\mathbf{A}\mathbf{x}\|$ is the operator norm. The result follows by taking $\mathbf{A}$ to be the $d\times d$ identity matrix, in which case $\mathbf{x}^T\mathbf{I}_d\mathbf{x} = \|\mathbf{x}\|_2^2$, and union bounding over all m symbols in the alphabet. ■

A.2 Proof of Theorem 7

Proof Fix some $a \notin \mathcal{S}$. As described in Theorem 5, the quantity $\langle\phi(a), \phi(a')\rangle$ is sub-Gaussian with parameter at most $L_{\max}^2\sigma^2$, where $L_{\max} = \max_a\|\phi(a)\|$. Then, again using

the fact that sub-Gaussianity is preserved under sums, by Hoeffding's inequality we have:

$$\mathbb{P}\left(\left|\sum_{a'\in\mathcal{S}}\langle\phi(a),\phi(a')\rangle\right|\geq\tau L^2\right)\leq 2\exp\left(-\frac{\tau^2L^4}{2sL_{\max}^2\sigma^2}\right)\leq 2\exp\left(-\frac{\kappa\tau^2L^2}{2s\sigma^2}\right)$$

where $\kappa = L^2/L_{\max}^2$. The result follows by union bounding over all m possible a. ■

A.3 Proof of Theorem 9

Proof Expanding the dot product between the two representations:

$$\begin{aligned}\frac{1}{L^2}\langle\phi(\mathcal{S}),\phi(\mathcal{S}')\rangle &= \frac{1}{L^2}\sum_{a\in\mathcal{S}\cap\mathcal{S}'}\langle\phi(a),\phi(a)\rangle + \frac{1}{L^2}\sum_{a\in\mathcal{S}}\sum_{a'\in\mathcal{S}'\setminus\{a\}}\langle\phi(a),\phi(a')\rangle\\ &\leq |\mathcal{S}\cap\mathcal{S}'| + ss'\mu.\end{aligned}$$

The other direction is analogous. ■

A.4 Proof of Theorem 10

Proof Consider some symbol $a\in\mathcal{A}$. In the event $a\in\mathcal{S}$:

$$\langle\phi(a),\phi(\mathcal{S})+\Delta_{\mathcal{S}}\rangle = \langle\phi(a),\phi(\mathcal{S})\rangle + \langle\phi(a),\Delta_{\mathcal{S}}\rangle \geq L^2 - sL^2\mu - \rho$$

and when $a\notin\mathcal{S}$:

$$\langle\phi(a),\phi(\mathcal{S})+\Delta_{\mathcal{S}}\rangle \leq sL^2\mu+\rho$$

Therefore we can decode correctly if:

$$\frac{\rho}{L^2}+s\mu<\frac{1}{2}$$

■

Proof of Lemma 12

Proof Consider first the case of passive noise. Fix some $a\in\mathcal{A}$. Noting that $\langle\phi(a),\Delta_{\mathcal{S}}\rangle$ is the sum of d terms bounded in $[-c,c]$, another application of Hoeffding's inequality and the union bound will show:

$$\mathbb{P}(\exists\, a \text{ s.t. } |\langle\phi(a),\Delta_{\mathcal{S}}\rangle|\geq\rho)\leq 2m\exp\left(-\frac{\rho^2}{2c^2d}\right).$$

Therefore, with probability $1-\delta$, we have that $\Delta_{\mathcal{S}}$ is ρ-bounded for $\rho\leq c\sqrt{2d\ln(2m/\delta)}$. Noting that $L=\sqrt{d}$ exactly, the result follows by applying Theorem 10.

Now let us consider the adversarial case in which $\|\Delta_{\mathcal{S}}\|_1 \leq \omega sd$. We first observe that $|\langle \phi(a), \Delta_{\mathcal{S}} \rangle| \leq \|\phi(a)\|_\infty \|\Delta_{\mathcal{S}}\|_1 \leq \omega sd$. Then, applying Theorem 10 we obtain:

$$\frac{\omega sd}{d} + s\mu < \frac{1}{2} \Rightarrow \omega < \frac{1}{2s} - \mu$$

as claimed. ■

Proof of Theorem 14

Proof Note first that $\|\phi(a) \otimes \psi(f)\|_2 = \|\phi(a)\|_2$. Then, fixing a, a' and f, by Hoeffding's inequality:

$$\mathbb{P}(|\langle \phi(a), \phi(a') \otimes \psi(f) \rangle| \geq \mu L^2) \leq 2 \exp\left(-\frac{L^4 \mu^2}{2\sigma^2 ||\phi(a')||_2^2}\right) \leq 2 \exp\left(-\frac{k\mu^2 L^2}{2\sigma^2}\right)$$

where we have again defined $\kappa = (\min_a \|\phi(a)\|_2^2)/(\max_{a'} \|\phi(a')\|_2^2)$. The result follows by the union bound over all $< nm^2/2$ combinations of a, a', f. ■

Proof of Theorem 17

Proof Expanding:

$$||\phi(\mathbf{x}) - \phi(\mathbf{x}')||_2^2 = ||\phi(\mathbf{x})||_2^2 + ||\phi(\mathbf{x}')||_2^2 - 2\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$$

Note first that $\|\phi(\mathbf{x})\|_2^2 = nd + \Delta$, where Δ is a mean-zero noise term due to cross-talk between the codewords. Neglecting minor errors from the ceiling function, the dot-product expands to:

$$\begin{aligned}
\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle &= \sum_{i=1}^{n} \langle \phi(x_i) \otimes \psi(f_i), \phi(x_i') \otimes \psi(f_i) \rangle + \sum_{i \neq j} \langle \phi(x_i) \otimes \psi(f_i), \phi(x_j') \otimes \psi(f_j) \rangle \\
&= \sum_{i=1}^{n} \langle \phi(x_i), \phi(x_i') \rangle + \Delta' = \sum_{i=1}^{n} d(1 - |a(x_i) - a(x_i')|) + \Delta' \\
&= d(n - \|\mathbf{x} - \mathbf{x}'\|_1) + \Delta'
\end{aligned}$$

where $a(x_i)$ is taken to be the centroid corresponding to x_i and Δ' is another noise term due to crosstalk. Putting both together and noting that $\Delta, \Delta' \leq n^2 d\mu$ we have:

$$2d(\|\mathbf{x} - \mathbf{x}'\|_1 - 2n^2\mu) \leq \|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|_2^2 \leq 2d(\|\mathbf{x} - \mathbf{x}'\|_1 + 2n^2\mu)$$

where the incoherence can be bounded as in Equation 4. ■

References

Arora, S., Ge, R., Ma, T., & Moitra, A. (2015). Simple, efficient, and neural algorithms for sparse coding. *Proceedings of The 28th Conference on Learning Theory, COLT 2015*, *40*, 113–149.

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, *39*(3), 930–945.

Bell, A., & Sejnowski, T. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*(6), 1129–1159.

Bishop, C. M. (2007). *Pattern recognition and machine learning, 5th Edition*. Information science and statistics. Springer.

Bloom, B. H. (1970). Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, *13*(7), 422–426.

Boyd, S. P., & Vandenberghe, L. (2014). *Convex Optimization*. Cambridge University Press.

Broder, A., & Mitzenmacher, M. (2004). Network applications of bloom filters: A survey. *Internet Mathematics*, *1*(4), 485–509.

Burrello, A., Schindler, K., Benini, L., & Rahimi, A. (2018). One-shot learning for ieeg seizure detection using end-to-end binary operations: Local binary patterns with hyperdimensional computing. In *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 1–4. IEEE.

Caron, S. J., Ruta, V., Abbott, L., & Axel, R. (2013). Random convergence of olfactory inputs in the drosophila mushroom body. *Nature*, *497*(7447), 113–117.

Chacron, M. J., Longtin, A., & Maler, L. (2011). Efficient computation via sparse coding in electrosensory neural networks. *Current Opinion in Neurobiology*, *21*(5), 752–760.

Choromanski, K. M., Rowland, M., & Weller, A. (2017). The unreasonable effectiveness of structured random orthogonal embeddings. In *Advances in Neural Information Processing Systems*, pp. 219–228.

Cybenko, G. (1989). Approximations by superpositions of sigmoidal functions. *Mathematics of Control, Signals, and Systems*, *2*(4), 303–314.

Dasgupta, S., Sheehan, T. C., Stevens, C. F., & Navlakha, S. (2018). A neural data structure for novelty detection. *Proceedings of the National Academy of Sciences*, *115*(51), 13093–13098.

Donoho, D. L., Elad, M., & Temlyakov, V. N. (2005). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, *52*(1), 6–18.

Frady, E. P., Kleyko, D., & Sommer, F. T. (2018). A theory of sequence indexing and working memory in recurrent neural networks. *Neural Computation*, *30*(6), 1449–1513.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202.

Gallant, S. I., & Okaywe, T. W. (2013). Representing objects, relations, and sequences. *Neural Computation*, *25*(8), 2038–2078.

Gupta, S., Imani, M., & Rosing, T. (2018). Felix: Fast and energy-efficient logic in memory. In *2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pp. 1–7. IEEE.

Hanson, D. L., & Wright, F. T. (1971). A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, *42*(3), 1079–1083.

Imani, M., Kong, D., Rahimi, A., & Rosing, T. (2017). Voicehd: Hyperdimensional computing for efficient speech recognition. In *2017 IEEE International Conference on Rebooting Computing (ICRC)*, pp. 1–8. IEEE.

Imani, M., Messerly, J., Wu, F., Pi, W., & Rosing, T. (2019a). A binary learning framework for hyperdimensional computing. In *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 126–131. IEEE.

Imani, M., Morris, J., Bosch, S., Shu, H., De Micheli, G., & Rosing, T. (2019b). Adapthd: Adaptive efficient training for brain-inspired hyperdimensional computing. In *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 1–4. IEEE.

Imani, M., Morris, J., Messerly, J., Shu, H., Deng, Y., & Rosing, T. (2019c). Bric: Locality-based encoding for energy-efficient brain-inspired hyperdimensional computing. In *Proceedings of the 56th Annual Design Automation Conference 2019*, p. 52. ACM.

Imani, M., Nassar, T., Rahimi, A., & Rosing, T. (2018). Hdna: Energy-efficient dna sequencing using hyperdimensional computing. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pp. 271–274. IEEE.

Imani, M., Rahimi, A., Kong, D., Rosing, T., & Rabaey, J. M. (2017). Exploring hyperdimensional associative memory. In *2017 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, pp. 445–456. IEEE.

Imani, M., Salamat, S., Gupta, S., Huang, J., & Rosing, T. (2019). Fach: Fpga-based acceleration of hyperdimensional computing by reducing computational complexity. In *Proceedings of the 24th Asia and South Pacific Design Automation Conference*, pp. 493–498. ACM.

Jacques, L., Laska, J. N., Boufounos, P. T., & Baraniuk, R. G. (2013). Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE Transactions on Information Theory*, *59*(4), 2082–2102.

Kanerva, P. (1994). The spatter code for encoding concepts at many levels. In *International Conference on Artificial Neural Networks*, pp. 226–229. Springer.

Kanerva, P. (1995). A family of binary spatter codes. In *International Conference on Artificial Neural Networks*, Vol. 1, pp. 517–522.

Kanerva, P. (2009). Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive Computation*, *1*(2), 139–159.

Kim, Y., Imani, M., Moshiri, N., & Rosing, T. (2020). GenieHD: Efficient dna pattern matching accelerator using hyperdimensional computing. In *2020 Design, Automation & Test in Europe Conference & Exhibition*. IEEE.

Kim, Y., Imani, M., & Rosing, T. S. (2018). Efficient human activity recognition using hyperdimensional computing. In Janowicz, K., Kuhn, W., Cena, F., Haller, A., & Vamvoudakis, K. G. (Eds.), *Proceedings of the 8th International Conference on the Internet of Things, IOT 2018, Santa Barbara, CA, USA, October 15-18, 2018*, pp. 38:1–38:6. ACM.

Kleyko, D., Rahimi, A., Gayler, R. W., & Osipov, E. (2019). Autoscaling bloom filter: controlling trade-off between true and false positives. *Neural Computing and Applications*, *32*, 1–10.

Kleyko, D., Rahimi, A., Rachkovskij, D. A., Osipov, E., & Rabaey, J. M. (2018). Classification and recall with binary hyperdimensional computing: Tradeoffs in choice of density and mapping characteristics. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(12), 5880–5898.

Lake, B., Salakhutdinov, R., Gross, J., & Tenenbaum, J. (2011). One shot learning of simple visual concepts. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 33.

Levy, S. D., & Gayler, R. (2008). Vector symbolic architectures: A new building material for artificial general intelligence. In *Conference on Artificial General Intelligence*, pp. 414–418. IOS Press.

Littlestone, N. (1988). Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, *2*(4), 285–318.

Mairal, J., Bach, F., Ponce, J., & Sapiro, G. (2010). Online learning for matrix factorization and sparse coding.. *Journal of Machine Learning Research*, *11*(1).

Masse, N. Y., Turner, G. C., & Jefferis, G. S. (2009). Olfactory information processing in drosophila. *Current Biology*, *19*(16), R700–R713.

McDiarmid, C., et al. (1989). On the method of bounded differences. *Surveys in Combinatorics*, *141*(1), 148–188.

Mitrokhin, A., Sutor, P., Fermüller, C., & Aloimonos, Y. (2019). Learning sensorimotor control with neuromorphic sensors: Toward hyperdimensional active perception. *Science Robotics*, *4*(30).

Neubert, P., Schubert, S., & Protzel, P. (2019). An introduction to hyperdimensional computing for robotics. *KI-Künstliche Intelligenz*, *33*(4), 319–330.

Olshausen, B., & Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*, 607–609.

Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, *14*(4), 481–487.

Pagh, A., Pagh, R., & Rao, S. S. (2005). An optimal bloom filter replacement. In *Proceedings of the sixteenth annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 823–829.

Parcollet, T., Ravanelli, M., Morchid, M., Linarès, G., Trabelsi, C., Mori, R. D., & Bengio, Y. (2019). Quaternion recurrent neural networks. In *International Conference on Learning Representations (ICLR)*.

Plan, Y., & Vershynin, R. (2014). Dimension reduction by random hyperplane tessellations. *Discrete & Computational Geometry*, *51*(2), 438–461.

Plate, T. (2003). *Holographic Reduced Representation: Distributed Representation for Cognitive Structures*. CSLI Lecture Notes (CSLI- CHUP) Series. CSLI Publications.

Plate, T. A. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks*, *6*(3), 623–641.

Rachkovskij, D. (2015). Formation of similarity-reflecting binary vectors with random binary projections. *Cybernetics and Systems Analysis*, *51*(2), 313–323.

Rachkovskij, D. A. (2001). Representation and processing of structures with binary sparse distributed codes. *IEEE Transactions on Knowledge and Data Engineering*, *13*(2), 261–276.

Rachkovskiy, D. A., Slipchenko, S. V., Kussul, E. M., & Baidyk, T. N. (2005a). Sparse binary distributed encoding of scalars. *Journal of Automation and Information Sciences*, *37*(6).

Rachkovskiy, D. A., Slipchenko, S. V., Misuno, I. S., Kussul, E. M., & Baidyk, T. N. (2005b). Sparse binary distributed encoding of numeric vectors. *Journal of Automation and Information Sciences*, *37*(11).

Raginsky, M., & Lazebnik, S. (2009). Locality-sensitive binary codes from shift-invariant kernels. In *Advances in Neural Information Processing Systems*, pp. 1509–1517.

Rahimi, A., Benatti, S., Kanerva, P., Benini, L., & Rabaey, J. M. (2016). Hyperdimensional biosignal processing: A case study for emg-based hand gesture recognition. In *2016 IEEE International Conference on Rebooting Computing (ICRC)*, pp. 1–8. IEEE.

Rahimi, A., Datta, S., Kleyko, D., Frady, E. P., Olshausen, B., Kanerva, P., & Rabaey, J. M. (2017). High-dimensional computing as a nanoscalable paradigm. *IEEE Transactions on Circuits and Systems I: Regular Papers*, *64*(9), 2508–2521.

Rahimi, A., Kanerva, P., Benini, L., & Rabaey, J. M. (2018). Efficient biosignal processing using hyperdimensional computing: Network templates for combined learning and classification of ExG signals. *Proceedings of the IEEE*, *107*(1), 123–143.

Rahimi, A., Tchouprina, A., Kanerva, P., Millán, J. d. R., & Rabaey, J. M. (2017). Hyperdimensional computing for blind and one-shot classification of eeg error-related potentials. *Mobile Networks and Applications*, *25*, 1–12.

Rahimi, A., & Recht, B. (2008). Random features for large-scale kernel machines. In *Advances in Neural Information Processing systems*, pp. 1177–1184.

Rosenblatt, F. (1958). The Perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408.

Rudelson, M., Vershynin, R., et al. (2013). Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, *18*.

Rudin, W. (1962). *Fourier analysis on groups*. John Wiley and Sons, Ltd.

Rumelhart, D., McClelland, J., & the PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*. MIT Press.

Sahlgren, M. (2005). An introduction to random indexing. In *Methods and applications of semantic indexing workshop at the 7th international conference on terminology and knowledge engineering*.

Salamat, S., Imani, M., Khaleghi, B., & Rosing, T. (2019). F5-hd: Fast flexible fpga-based framework for refreshing hyperdimensional computing. In *Proceedings of the 2019 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, pp. 53–62.

Schmuck, M., Benini, L., & Rahimi, A. (2019). Hardware optimizations of dense binary hyperdimensional computing: Rematerialization of hypervectors, binarized bundling, and combinational associative memory. *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, *15*(4), 1–25.

Shawe-Taylor, J., Cristianini, N., et al. (2004). *Kernel methods for pattern analysis*. Cambridge university press.

Stettler, D. D., & Axel, R. (2009). Representations of odor in the piriform cortex. *Neuron*, *63*(6), 854–864.

Thrun, S. (1996). Is learning the n-th thing any easier than learning the first?. In *Advances in Neural Information Processing Systems*, pp. 640–646.

Turner, G. C., Bazhenov, M., & Laurent, G. (2008). Olfactory representations by drosophila mushroom body neurons. *Journal of Neurophysiology*, *99*(2), 734–746.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, Vol. 48. Cambridge University Press.

Weiss, E., Cheung, B., & Olshausen, B. (2016). A neural architecture for representing and reasoning about spatial relationships. In *International Conference on Learning Representations (ICLR)*.

Widdows, D., & Cohen, T. (2015). Reasoning with vectors: A continuous model for fast robust inference. *Logic Journal of the IGPL*, *23*(2), 141–173.

Wilson, R. I. (2013). Early olfactory processing in drosophila: mechanisms and principles. *Annual Review of Neuroscience*, *36*, 217–241.

Zhang, A., Tay, Y., Zhang, S., Chan, A., Luu, A. T., Hui, S. C., & Fu, J. (2016). Beyond fully-connected layers with quaternions: Parameterization of hypercomplex multiplications with $1/n$ parameters. In *International Conference on Learning Representations (ICLR)*.