Hyperdimensional Computing and Spectral Learning

Namiko Matsumoto, Anthony Thomas, Tara Javidi, and Tajana Simunic Rosing
University of California, San Diego
{nmatsumo,ahthomas,tjavidi,tajana}@ucsd.edu

I. Introduction

Hyperdimensional (HD) computing is an emerging paradigm for machine learning based on biologically plausible models of memory. HD computing represents data as random points with i.i.d. components in a high-dimensional, low-precision, space. These points are then used as input to learning algorithms. The i.i.d. and low-precision nature of HD representations accord naturally with highly parallel hardware like FPGAs and PIM architectures and HD computing has recently gained significant popularity in the hardware community [1]–[3].

Existing work on HD computing has focused primarily on learning linear functions over the encodings. However, much data of practical interest is nonlinear, meaning that linear methods may fail to capture important structure in the data. Our goal is to assess the utility of using nonlinear learning algorithms on HD representations of data. A complication arises because nonlinear models are generally more complex computationally than simple linear methods which are traditionally used with HD computing. In this work, we show how nonlinear learning on HD representations can be formulated as a sparse convex optimization problem using the Boolean Fourier transform. We additionally present results from a preliminary empirical study motivated by this formulation designed to assess the possible practical benefits of this approach.

II. Background and Related Work

In [2], the authors use an HD encoded vector as input to a simple multilayer-perceptron with a single hidden layer. Neural network based methods may be unappealing as they introduce a multitude of new tunable hyperparameters and require non-convex optimization methods. By contrast, our approach introduces only a single hyperparameter and can be solved using efficient convex methods. Work in [4] presented simple heuristics for learning sparse classifiers on HD representations but did not consider the theoretically optimal L1 based methods considered here. Most closely related to our work is [5] who examines connections between HD computing and compressed sensing. Like ours, their approach can be formulated as an L1-penalized regression problem on a random design matrix. However, our formalism is based on the Boolean Fourier transform which has a rich theory outside of compressed sensing and explicitly addresses learning arbitrary Boolean functions over the HD representations. To the best of our knowledge, we are the first to explore the connections between HD computing and Boolean spectral learning.

A. Boolean Fourier Transform and Spectral Learning

The following reviews the bare essentials of the Boolean Fourier transform (hereafter referred to simply as the Fourier transform). Let $g(\mathbf{z})$ be a function $g : \{\pm 1\}^d \to \mathbb{R}$ and let S be a subset of $[d] = 1, ..., d$. The Fourier transform of g can be defined to be its unique representation as [6]: $g(\mathbf{z}) = \sum_{S \subseteq [d]} \hat{g}(S)\chi_S(\mathbf{z})$, where $\chi_S(\mathbf{z}) = \prod_{j \in S} z_j$ is the parity function for S. We can rewrite the above more compactly in matrix form as: $g = \mathbf{\Psi}\hat{\mathbf{g}}$, where $\mathbf{\Psi}$ is the $2^d \times 2^d$ matrix formed by stacking up all parity functions. The Fourier coefficients–$\hat{g}(S)$– can be computed as:

$$\hat{g}(S) = \mathop{\mathbb{E}}_{\mathbf{z} \sim \mathrm{Unif}(\{\pm 1\}^d)}[g(\mathbf{z})\chi_S(\mathbf{z})] = \frac{1}{2^d}\sum_{i=1}^{2^d} g(\mathbf{z}_i)\chi_S(\mathbf{z}_i) \quad (1)$$

The Fourier transform is important in learning theory because its invertibility implies learning the spectrum $\hat{g}$ and the function g are equivalent. Per Equation 1, given 2^d unique samples, we may simply compute $\hat{g}$ exactly, but this is not feasible in general. For many functions of interest the spectrum is sparse in the sense that only some $k \ll 2^d$ coefficients are nonzero so that that one may accurately estimate the spectrum from far fewer samples (see: [6] Thm. 3.29).

III. Formal Model

We assume our data is generated by some unknown and possibly nonlinear function $f : \mathcal{X} \to \{\pm 1\}$. Our goal is to learn f. We assume we have access to a set of samples $\mathcal{S} = \{(\mathbf{x}_i, f(x_i))\}_{i=1}^N$ where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^n$ and $y_i \in \{\pm 1\}$. We first map each sample x to a binary HD representation under an encoding function $\phi : \mathbb{R}^n \to \{\pm 1\}^d$. Then, we posit the existence of some function $g : \{\pm 1\}^d \to \mathbb{R}$, such that $f(\mathbf{x}) = \mathrm{sign}(g(\phi(\mathbf{x})))$ and seek to learn g by estimating its Fourier spectrum. Our approach is motivated by [7] who formulate learning as a similar convex problem, but assume the binary representation is already given, unlike our setting where the input data is Euclidean.

More formally, we assume $f(\mathbf{x}) = \mathrm{sign}(\tilde{\mathbf{\Psi}}\hat{\mathbf{g}})$ where $\mathbf{g} \in \mathbb{R}^{2^d}$, $\tilde{\mathbf{\Psi}}$ is an $N \times 2^b$ matrix such that $\tilde{\mathbf{\Psi}}_{ij} = \chi_{S_j}(\mathbf{x}_i)$. Thus, this problem amounts to the recovery of a sparse vector of coefficients using an over-complete measurement matrix ($\mathbf{\Psi}$) which can be formulated as the following convex problem:

$$\hat{\mathbf{g}}^\star = \mathop{\mathrm{argmax}}_{\hat{\mathbf{g}} \in \mathbb{R}^{2^d}} \mathbf{y}^\top \mathbf{\Psi}\hat{\mathbf{g}} - \lambda\|\hat{\mathbf{g}}\|_1 \quad (2)$$

Solving this problem exactly is still difficult since $\tilde{\mathbf{\Psi}}$ contains 2^d columns (vast in the context of HD). To obtain an inkling

of the practical benefits from this formulation of HD learning, we here exploit a well known connection between the Boolean Fourier transform and decision tree learning for which there are numerous well known algorithms. Since our interest is simply in obtaining a low-complexity representation of g– we do not care about the spectrum itself–this enables us to approximate the difficult problem above by simply learning a decision tree. The Fourier transform has well known relationship with decision trees, where the spectral and tree complexity are related by the following proposition from [6].

Prop *Let* $g : \{\pm 1\}^d \to \mathbb{R}$ *be represented by an* ℓ*-leaf, depth-*k *decision tree, and define the spectral sparsity of* g *by* $\mathsf{sparsity}(g) \triangleq |\{S \in [d] \mid \hat{g}(S) \neq 0\}|$. *Then, the spectral sparsity of* g *is bound by*

$$\mathsf{sparsity}(g) \leq \ell 2^k \leq 4^k \tag{3}$$

To obtain the binary representation of our data, we use the method of random half-spaces, which partitions $\mathcal{X}$ using a grid formed by d random hyperplanes and encodes each point using a binary string uniquely identifying each grid cell [8]. More formally: $\phi(\mathbf{x}) = \text{sign}(\mathbf{\Phi x} + \mathbf{b})$, where $\mathbf{\Phi} \in \mathbb{R}^{d \times n}$ is a matrix whose rows are sampled uniformly at random from the n-dimensional unit sphere and $\mathbf{b}$ is an intercept chosen uniformly at random from the support of $\mathcal{X}$.

IV. Empirical Results

We evaluate sparse recovery by models fit on low-dimensional data and HD encoded representations. We compare the accuracy and model complexity of decision trees with an L_1-regularized linear SVM as a baseline, similar to an approach considered in [5]. The L_1-regularized SVM is a computationally simpler model, but cannot learn as rich a space of functions. SVMs and decision trees are fit several times to select regularization constants and tree structures—number of leaves and depth—that yield sparse models whose accuracy remains within 5% of the highest observed accuracy. The sparsity captured by models is measured in terms of support size, defined as the number of nonzero coefficients of an SVM, or as $\ell 2^k$ for a depth-k, ℓ-leaf decision tree, per equation (3). To study sparse and approximated spectral learning with HD, we consider five datasets in UCI's Machine Learning Repository: *UCI HAR*, *ISOLET*, *Buzz/Tom's Hardware* (*Buzz/TH*), *Buzz/Twitter* (*Buzz/TW*), and *Adult*.

As in [5], we find HD representations enjoy significant sparsity. On our data, HD SVMs preserve sparsity at the same level as low-dimensional models and have a constant support size independent of encoding dimension. Figure 1 shows decision trees are capable of capturing *more* sparsity in HD representations than in the low-dimensional data (relative support size is < 1) despite the significantly larger HD dimension. Additionally, the support size of HD decision trees is constant w.r.t. the encoding dimension. Accounting for both accuracy and sparsity, sparse recovery with the HD decision trees requires a minimum dimension of $\approx$ 2-4,000 on "simple" (e.g. linear) datasets but of only $\approx$ 1-2,000 otherwise. In absolute terms, the support size of decision trees is larger than

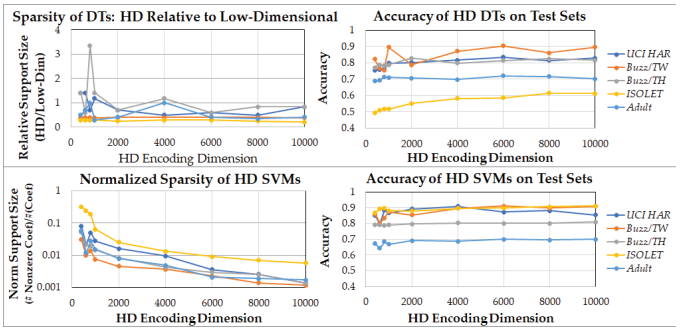

Fig. 1. Sparsity (left) and accuracy (right) of HD decision trees (DTs) (top) and HD SVMs (bottom), where the sparsity of decision trees is shown as HD support size relative to low-dimensional support size; and where the sparsity of SVMs is shown as normalized HD support size (HD support size scaled by a factor of 1/(total number of coefficients)).

SVMs; however, we emphasize that this model is able to learn a broader class of functions than the SVM.

Although HD decision trees exploit more sparsity than their low-dimensional counterparts, they may suffer from a greater propensity to overfit, particularly when data is "simple" and linearly learnable. This is evident in Figure 1 with *ISOLET*, where an SVM attains high test-time accuracy on the HD encoded data while the decision tree yields a significantly lower test-time accuracy. However, for data poorly represented by low-dimensional, linear models, such as *Adult*, our decision tree approach offers modest improvements in accuracy, e.g. by $\delta \in [1\%, 4\%]$, over the baseline HD linear SVM by approximating nonlinear "spectral learning."

Approximating spectral learning with decision trees is promising in suggesting that HD representations are sparse in the Fourier domain and that HD may be able to exploit sparse spectral learning techniques in more efficient nonlinear learning. While $L1$-regularized linear SVMs recover equal sparsity in low-dimensional and HD data, spectral learning techniques may enhance sparsity in the HD Fourier domain. This sparsity is recoverable with encoding dimensions between 1-4000, allowing spare HD learning without the extremely large dimensions (e.g. $10,000$) common in HD literature.

We note that the literature has proposed several efficient algorithms for computing the expensive product $\mathbf{\Psi}\hat{\mathbf{g}}$ when $\hat{\mathbf{g}}$ is sparse [9], [10]. In follow up work, we plan to compare our heuristic approach based on decision tree learning to the formal problem posed above and explore solving the sparse recovery problem exactly in HD representations which may alleviate the issues surrounding overfitting with decision trees. We are additionally interested in studying different embedding methods, especially in their ability to preserve or induce sparsity in HD representations and in the relationship between the geometry of the original data and the level of sparsity achievable with different embedding methods. We would be interested to study embedding methods in the context of approximate and exact solutions of the sparse recovery problem.

REFERENCES

[1] A. Burrello, K. Schindler, L. Benini, and A. Rahimi, "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)*. IEEE, 2018, pp. 1–4.

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

[3] A. Rahimi, P. Kanerva, L. Benini, and J. M. Rabaey, "Efficient biosignal processing using hyperdimensional computing: Network templates for combined learning and classification of exg signals," *Proceedings of the IEEE*, vol. 107, no. 1, pp. 123–143, 2018.

[4] M. Imani, S. Salamat, B. Khaleghi, M. Samragh, F. Koushanfar, and T. Rosing, "Sparsehd: Algorithm-hardware co-optimization for efficient high-dimensional computing," in *2019 IEEE 27th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*. IEEE, 2019, pp. 190–198.

[5] E. P. Frady, D. Kleyko, and F. T. Sommer, "Variable binding for sparse distributed representations: Theory and applications," *arXiv preprint arXiv:2009.06734*, 2020.

[6] R. O'Donnell, *Analysis of boolean functions*. Cambridge University Press, 2014.

[7] S. Negahban and D. Shah, "Learning sparse boolean polynomials," in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2012, pp. 2032–2036.

[8] L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk, "Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors," *IEEE Transactions on Information Theory*, vol. 59, no. 4, pp. 2082–2102, 2013.

[9] X. Li, J. K. Bradley, S. Pawar, and K. Ramchandran, "The spright algorithm for robust sparse hadamard transforms," in *2014 IEEE International Symposium on Information Theory*. IEEE, 2014, pp. 1857–1861.

[10] R. Scheibler, S. Haghighatshoar, and M. Vetterli, "A fast hadamard transform for signals with sublinear sparsity in the transform domain," *IEEE Transactions on Information Theory*, vol. 61, no. 4, pp. 2115–2132, 2015.