An Information-Theoretic Approach to Universal Feature Selection in High-Dimensional Inference

Shao-Lun Huang
DSIT Research Center
Tsinghua-Berkeley Shenzhen Institute
Shenzhen, China 518055
Email: shaolun.huang@sz.tsinghua.edu.cn

Anuran Makur, Lizhong Zheng, and Gregory W. Wornell
Dept. EECS and RLE
Massachusetts Institute of Technology
Cambridge, MA 02139
Email: {a_makur, lizhong, gww}@mit.edu

***Abstract*—We develop an information theoretic framework for addressing feature selection in applications where the inference task is not specified in advance and the data is from a large alphabet. We introduce a natural notion of universality for such problems, and show that locally optimal solutions are straightforward to obtain, admit natural interpretations via information geometry, have computationally efficient implementations, and represent a practically useful learning methodology. Our development also reveals the key role of Hirschfeld-Gebelein-Rényi maximal correlation and the alternating conditional expectations (ACE) algorithm in such problems.**

I. Introduction

In many applications of machine learning, there is a need to extract low-dimensional features from the available high-dimensional data, from which inference is performed. When the inference task is known in advance—and the system model is fully specified—classical statistics establishes that the appropriate features are (minimal) sufficient statistics. However, in a rapidly growing number of application scenarios, the features must be selected before the inference task is chosen. Moreover, the amount of training data available relative to its dimension is small, so that estimates of the underlying distributions are typically quite poor. In such scenarios, there is a need for methods of constructing good "universal" features of the data from the available training data. This is the problem of interest in this work.

As a motivating example, consider the following problem involving a (large) collection of consumers $\mathcal{X}$ and (large) collection of movies $\mathcal{Y}$. The distribution $P_{X,Y}(x,y)$ over $\mathcal{X}\times\mathcal{Y}$ (more specifically, $P_{Y|X}(y|x)$) captures the probability that a consumer x selects movie y. Given training data in the form of a few samples from this large joint distribution, we seek to understand what we can infer about the dominant factors governing consumer movie preferences. For example, among many other possibilities, one feature f of a consumer x's description we can extract is his/her age $u = f(x)$, and one feature g of a movie y's description we can extract is its genre $v = g(y)$. A key question we ask is: if we don't know in advance which of a number of possible features of a consumer (e.g., age, income, etc) that we might want make inferences about, what low-dimensional features of the selected movie should we select to preserve as much of the needed information as possible? In turn, we can further ask: among all possible features of consumers and movies, which are those whose relationship is expressed most strongly by the available training data, so that we can confidently predict which movies will be a good match to which consumers. As we develop, such questions are naturally addressed by formulating them as ones of universal feature selection.

A mathematical model for the problem is as follows. We begin with a Markov chain $U \leftrightarrow X \leftrightarrow Y$, associated with which is the joint distribution $P_{U,X,Y} = P_{Y|X}\,P_{U,X}$ with respect to alphabets $\mathcal{U}$, $\mathcal{X}$, and $\mathcal{Y}$. In the simplest model, we assume $P_{X,Y}$ is known, so that $P_{Y|X}$ and P_X are also known. In practice, our solution will effectively learn the aspects of $P_{X,Y}$ we require from a comparatively small set of independent, identically distributed (i.i.d.) training samples $(\tilde{x}_1,\tilde{y}_1),\ldots,(\tilde{x}_k,\tilde{y}_k)$. By contrast, during the feature design process we consider $P_{U,X}$ to be unknown; we know only that it must satisfy the marginal constraint

$$\sum_u P_{U,X}(u,x) = P_X(x). \tag{1}$$

In particular, samples of the variable U are not observed in this process. With this model, our goal is, given further i.i.d. samples $y_1,\ldots,y_n$ from $P_{Y|U=u}$ for some fixed but unknown u, determine a (real-valued) g such that $v = \sum_i g(y_i)$ is universally good for estimating u.

To quantify the notion of universality, we define a family $\mathcal{F}$ of (bivariate) distributions $P_{U,X}$—each satisfying the constraint (1)—whose elements are defined over the given (finite) alphabets $\mathcal{X}\times\mathcal{U}$. This family represents the uncertainty in the possible variables U about which we may wish to make inferences. We further define a probability measure $\mu_{\mathcal{F}}$ over this family $\mathcal{F}$ that expresses what we know about the relative likelihoods of each possible $P_{U,X}$. In turn, with respect to the family $\mathcal{F}$ and its measure $\mu_{\mathcal{F}}$, we define the universal feature of the data we seek to be

$$g^* = \underset{g:\,\mathcal{Y}\mapsto\mathbb{R}}{\arg\max}\,\mathbb{E}_{\mu_{\mathcal{F}}}\left[I(U;g(Y)\mid P_{U,X}\in\mathcal{F})\right]. \tag{2}$$

As an aside, note that the measure $\mu_{\mathcal{F}}$ can be chosen to also compensate for the fact that if there are some distributions $P_{U,X}$ in $\mathcal{F}$ for which $P_{U|X}$ is large, then we can avoid giving them undue influence in the expectation in (2) by choosing proportionally smaller weights for them, if appropriate.

978-1-5090-4096-4/17/$31.00 ©2017 IEEE

More generally, the choice of $\mathcal{F}$ and $\mu_{\mathcal{F}}$ strongly impact the feature g^* obtained from the optimization (2). In this work, we focus on particular choices that are meaningful for applications, convenient for analysis, admit a natural information-geometric interpretation, lead directly to an efficient algorithm for its solution, and which have key connections to Hirschfeld-Gebelein-Rényi (HGR) maximal correlation [3], [4], [8].

In particular, for $P_{U,X} = P_{X|U}\, P_U$ in $\mathcal{F}$ we let $\mathcal{U} = \{0, 1\}$ with P_U uniform and $I(U;X) \leq \delta$ for small $\delta > 0$.[1] Moreover, we let the measure $\mu_{\mathcal{F}}$ be uniform over the family. In this case, the (asymptotic) solution to (2) takes the form $g^*(y) = \log(P^*_{Y|U=1}(y)/P^*_{Y|U=0}(y))$, where $P^*_{Y|U=u}(y) = \sum_x P_{Y|X}(y|x)\, P^*_{X|U=u}(x)$ for $u \in \mathcal{U}$, with

$$P^*_{X|U} = \lim_{\delta \to 0} \quad \underset{\substack{P_{X|U}: \\ D(P_{X|U=1}\|P_{X|U=0}) \leq \delta}}{\arg\max} \quad D(P_{Y|U=1}\|P_{Y|U=0}), \quad (3)$$

as we will now develop, interpret, and apply.

II. The Geometry of Maximal Correlation

Our analysis will require optimizing the tradeoff between multiple (KL) divergences. We will exploit that when the distributions involved in these divergences are close to each other, local approximations can significantly simplify such optimizations. In fact, local approximations allow us to linearize the space of probability distributions, and treat it as a linear vector space. In this section, we introduce some notation for—and properties of—this vector space to facilitate our analysis.

A. The Local Geometry

Let $\mathcal{P}_{\mathcal{X}}$ denote the space of distributions on $\mathcal{X}$, where $|\mathcal{X}|$ is finite, and $\mathsf{relint}(\mathcal{P}_{\mathcal{X}})$ be the relative interior of $\mathcal{P}_{\mathcal{X}}$, which is the collection of distributions with strictly positive entries.

Definition 1 (ϵ-Neighborhood)**.** The ϵ-neighborhood of a *reference* distribution $P_{0,X} \in \mathsf{relint}(\mathcal{P}_{\mathcal{X}})$ is defined as:

$$\mathcal{N}^{(\epsilon)}(P_{0,X}) \triangleq \left\{ P_X \in \mathcal{P}_{\mathcal{X}} : \sum_{x \in \mathcal{X}} \frac{(P_X(x) - P_{0,X}(x))^2}{P_{0,X}(x)} < \epsilon^2 \right\}$$

which is the set of distributions in a χ^2-divergence ball of radius ϵ^2 around $P_{0,X}$.

Furthermore, we introduce the following notation. For each $P_X \in \mathcal{N}^{(\epsilon)}(P_{0,X})$, we write, for all $x \in \mathcal{X}$,

$$P_X(x) = P_{0,X}(x) \cdot (1 + \epsilon \cdot L(x)) \quad (4)$$

$$= P_{0,X}(x) + \epsilon \cdot \sqrt{P_{0,X}(x)} \cdot \phi(x)\,. \quad (5)$$

This defines a pair of functions

$$L(x) = \frac{1}{\epsilon} \frac{P_X(x) - P_{0,X}(x)}{P_{0,X}(x)}, \qquad \phi(x) = \sqrt{P_{0,X}(x)} L(x).$$

[1] For the family of interest, these choices for $\mathcal{U}$ and P_U are without loss of generality. Also, our family is equivalently described by the (KL) divergence constraint $D(P_{X|U=1}\|P_{X|U=0}) \leq \delta$, and corresponds to restricting attention to possible variables U that are all effectively equally detectable.

This notation will soon become convenient in our development. At this point, we only emphasize that there is a three-way one-to-one correspondence $P_X \leftrightarrow \phi \leftrightarrow L$. For example, we can rewrite the definition of the ϵ-neighborhood as

$$\mathcal{N}^{(\epsilon)}(P_{0,X}) = \left\{ P_X \leftrightarrow \phi \colon \|\phi\|^2 < 1 \right\}$$

where $\|\phi\|$ is simply the Euclidean norm of ϕ, viewed as an $|\mathcal{X}|$-dimensional vector.

We assume that for an inference problem, all the distributions of interest, including all the empirical distributions that one can possibly observe, lie in an ϵ-neighborhood of a certain $P_{0,X}$. Note that we do not assume that ϵ is small in this definition. In our mathematical analysis however, we will let $\epsilon \to 0$ and find the optimal choices of feature functions in this limiting case. It turns out that the solutions do not depend on the value of ϵ. Thus, in practice, we will apply these solutions as an approximation to the optimal choices even though ϵ is not necessarily very small.

When ϵ is small so that we look at a small neighborhood of distributions, it is well-known that the space of distributions behaves "nicely" and exhibits simple properties. Firstly, $L(\cdot)$ can be viewed as the log-likelihood ratio function

$$\log \frac{P_X(x)}{P_{0,X}(x)} = \log(1 + \epsilon L(x)) = \epsilon \cdot L(x) + o(\epsilon) \qquad \forall x \in \mathcal{X}.$$

More generally, for $P_1 \leftrightarrow L_1$ and $P_2 \leftrightarrow L_2$, we have

$$\log \frac{P_1(x)}{P_2(x)} = \epsilon \cdot (L_1(x) - L_2(x)) + o(\epsilon), \qquad \forall x \in \mathcal{X}. \quad (6)$$

Moreover, via Taylor's theorem, for any two distributions $P_1 \leftrightarrow \phi_1$ and $P_2 \leftrightarrow \phi_2$

$$D(P_1\|P_2) = \frac{\epsilon^2}{2} \cdot \underbrace{\sum_{x \in \mathcal{X}} (\phi_1(x) - \phi_2(x))^2}_{\|\phi_1 - \phi_2\|^2} + o\left(\epsilon^2\right).$$

Note that when we ignore the $o(\epsilon^2)$ term, the divergence becomes symmetric in P_1 and P_2. Furthermore, this approximation does not depend on the specific choice of reference point $P_{0,X}$ as long as it is in the neighborhood; changing the choice of $P_{0,X}$ only causes a difference in the $o(\epsilon^2)$ term.

While P_X, ϕ, and L can all be treated as vectors of dimension $|\mathcal{X}|$ for finite $|\mathcal{X}|$, to avoid confusion, we will reserve vector notation for ϕ, referring to it as an *information vector*. The squared norm of an information vector, as we have seen, corresponds to Fisher information and divergence. We will soon argue that in the context of inference problems, this squared norm corresponds to the "total amount" of information we observe. Our development then focuses on the operational meaning of the directions of these information vectors.

In this development, we locally approximate the central optimization problem in (3), i.e.,

$$\max_{P_X, Q_X \,:\, D(P_X\|Q_X) \leq \epsilon} \quad D(P_Y\|Q_Y), \quad (7)$$

with $P_X \triangleq P_{X|U=0}$ and $Q_X \triangleq P_{X|U=1}$, and $P_Y = P_{Y|U=1}$ and $Q_Y = P_{Y|U=0}$ denoting the marginal distributions on $\mathcal{Y}$

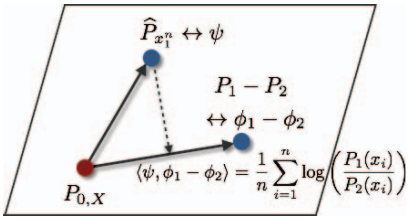

Fig. 1: Evaluating a function on a sequence of data samples is equivalent to a vector projection in the distribution space.

induced by P_X and Q_X, respectively, via $P_{Y|X}$. We find the U that maximizes the divergence between P_Y and Q_Y, then set $g(y) = (Q_Y(y)/P_Y(y)) - 1$ as the feature function. Later, we approximate this choice by $g(y) \approx \log(Q_Y(y)/P_Y(y))$, which has the interpretation of as the log-likelihood ratio suitable for detecting the most detectable target, and interpret it as finding the appropriate decomposition of the information vectors. Via (4) and (5), we then translate the optimal choices into the corresponding feature functions L.

It is useful to express the solution to a familiar binary inference problem in this vector space language to develop the relevant geometric perspective. In particular, consider a binary hypothesis testing problem in which we observe m samples, $x_1, \ldots, x_m$, drawn i.i.d. from either distribution P_1 or distribution P_2. Suppose further that both P_1 and P_2 live in a neighborhood $\mathcal{N}^{(\epsilon)}(P_{0,X})$, and recall the correspondence $P_i \leftrightarrow \phi_i \leftrightarrow L_i$ for $i = 1, 2$. Then a sufficient statistic S for this problem can be expressed in terms of the empirical distribution of the data $\hat{P}_{x_1^m} \triangleq \hat{P} \leftrightarrow \psi$ according to

$$\begin{aligned} S &= \frac{1}{m}\sum_{i=1}^{m} \log \frac{P_1(x_i)}{P_2(x_i)} = \sum_{x\in\mathcal{X}} \epsilon \cdot \hat{P}(x)(L_1(x) - L_2(x)) + o(\epsilon) \\ &= \epsilon^2 \cdot \sum_{x\in\mathcal{X}} \psi(x) \cdot (\phi_1(x) - \phi_2(x)) + o\left(\epsilon^2\right), \end{aligned}$$

where in the last equality, we use the fact that both $L_1(X)$ and $L_2(X)$ have zero mean with respect to $P_{0,X}$. Thus, evaluating a function is equivalent to computing an inner product $\langle \psi, \phi_1 - \phi_2 \rangle$, i.e., a particular projection of ψ, the vector corresponding to the entire observed information (see Fig. 1).

The inner product $\langle \psi, \phi \rangle$ between information vectors has an important interpretation for our purposes. Using the equivalence between distributions, information vectors, and functions, $P \leftrightarrow \phi \leftrightarrow L$, this can be understood as the inner product between the corresponding variations of distributions or functions. Related quantities such as the norm and projections can be similarly defined, and are repeatedly used in the sequel, and facilitate interpreting our analysis.

B. Optimal Feature Selection

We now apply this geometry to solve for g^* as defined in Section I via (3), which can be interpreted as solving for the direction in which to project our information vector. In the process, we develop how the opportunistic choice associated with (7) solves the universal problem (2) when specialized to the family and measure of interest. To do so, we exploit the connection between $\mathcal{P}_\mathcal{X}$ and $\mathcal{P}_\mathcal{Y}$, the two spaces of marginal distributions on $\mathcal{X}$ and $\mathcal{Y}$.

Starting from $P_{X,Y}$, for convenience let us first use the two marginal distributions as reference: $P_{0,X} = P_X$ and $P_{0,Y} = P_Y$ for $\mathcal{P}_\mathcal{X}$ and $\mathcal{P}_\mathcal{Y}$, respectively. We also assume that all the distributions we encounter in the following discussion, both over $\mathcal{X}$ and over $\mathcal{Y}$, belong to the corresponding ϵ-neighborhood around these two reference distributions.

A deviation in the distribution for X from P_X to $P_{X|U}$ induces a change in the distribution for Y from P_Y to $P_{Y|U}$. In the language of local geometry, we write the correspondences $P_{X|U}(\cdot|u) \leftrightarrow \phi_X$ and $P_{Y|U}(\cdot|u) \leftrightarrow \phi_Y$ using the chosen reference points, and observe a simple linear relationship between these information vectors due to the Markov structure $U \leftrightarrow X \leftrightarrow Y$. In particular, since

$$\begin{aligned} P_Y(y) &= \sum_{x\in\mathcal{X}} P_{Y|X}(y|x) \cdot P_X(x) \\ P_{Y|U}(y|u) &= \sum_{x\in\mathcal{X}} P_{Y|X}(y|x) \cdot P_{X|U}(x|u) \end{aligned}$$

for all $y \in \mathcal{Y}$, we have

$$\phi_Y(y) = \frac{1}{\sqrt{P_Y(y)}} \sum_{x\in\mathcal{X}} P_{Y|X}(y|x)\sqrt{P_X(x)} \cdot \phi_X(x). \quad (8)$$

Writing both ϕ_X and ϕ_Y as column vectors, we can equivalently express (8) in the matrix form $\phi_Y = B \cdot \phi_X$, where

$$B \triangleq \left[\sqrt{P_Y}\right]^{-1} \cdot P_{Y|X} \cdot \left[\sqrt{P_X}\right], \quad (9)$$

with $\left[\sqrt{P_X}\right]$ and $\left[\sqrt{P_Y}\right]$ denoting diagonal matrices with entries $\{\sqrt{P_X(x)} : x \in \mathcal{X}\}$ and $\{\sqrt{P_X(y)} : y \in \mathcal{Y}\}$ respectively, and $P_{Y|X}$ denoting the $|\mathcal{Y}| \times |\mathcal{X}|$ column-stochastic transition probability matrix. We refer to B as the *divergence transition matrix (DTM)* [6].

The feature selection problem with the knowledge of the target is now simple. We find ϕ_X corresponding to the target, compute its image ϕ_Y through the DTM, and choose the feature to be along ϕ_Y. This corresponds to using the log-likelihood function $\log(P_{Y|U}(\cdot|u)/P_Y(\cdot))$ as the feature. Note that we can always normalize the feature function g as $g \leftrightarrow \phi_Y/\|\phi_Y\|$, so that $\mathbb{E}\left[g(Y)^2\right] = 1$. The resulting performance of this optimal detector in terms of the error exponent for binary detection is the divergence $D(P_{Y|U}(\cdot|u)\|P_Y)$, which corresponds to the squared norm $\|\phi_Y\|^2$.

If we instead choose $g \leftrightarrow \psi \not\propto \phi_Y$, then we have a mismatched detector. Without loss of generality, we assume ψ is normalized: $\|\psi\|^2 = 1$. This mismatched detector yields a worse detection performance in that the error exponent is reduced from $\|\phi_Y\|^2$ to $|\langle \phi_Y, \psi \rangle|^2$. We can now define the performance loss factor, for a given observation model with DTM given in (9) and a choice of feature function $g \leftrightarrow \psi$, as $\nu_B(\phi_X, \psi) \triangleq |\langle \phi_Y, \psi \rangle|^2/\|\phi_X\|^2$. This has the clear operational meaning that for the binary hypothesis testing problem between $P_{X|U}(\cdot|u)$ and P_X, when we use g as the feature function, the resulting error exponent is $\nu_B(\phi_X, \psi) \cdot$

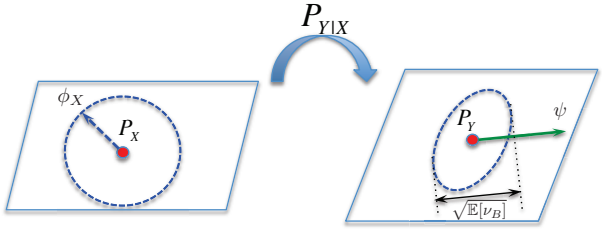

Fig. 2: The linear map from $\mathcal{P}_\mathcal{X}$ to $\mathcal{P}_\mathcal{Y}$. The unknown target ϕ_X is uniformly distributed on the surface of the unit divergence ball in $\mathcal{P}_\mathcal{X}$, and induces ψ on the divergence ellipse in $\mathcal{P}_\mathcal{Y}$. The optimal feature choice given by ψ maximizes the mean-squared inner product $\mathbb{E}[\nu_B] = \mathbb{E}\left[\langle B\phi_X, \psi\rangle^2\right]$.

$D(P_{X|U}(\cdot|u)\|P_X) + o(\epsilon^2)$. The term $D(P_{X|U}(\cdot|u)\|P_X) \leftrightarrow \|\phi_X\|^2$ is the exponent we would have if we had observed the hidden data samples X_1^m. Without loss of generality, we restrict $\|\phi_X\|^2 = 1$ in the following development. The design problem is now simply to maximize the loss factor ν_B to be as close to 1 as possible. The only remaining difficulty is that we do not know ϕ_X when we choose ψ.

It is worth noting that while a maximin formulation of universality $\max_{\psi:\,\|\psi\|^2=1} \min_{\phi_X:\,\|\phi_X\|^2=1} \nu_B(\phi_X, \psi)$ might be tempting, for all non-degenerate cases, this formulation leads to the degenerate result $\nu_B = 0$, meaning the worst-case behavior is inherently poor. Indeed, when we choose a specific ψ, nature can always adversarially choose ϕ_X such that $\phi_Y = B \cdot \phi_X$ is orthogonal to ψ. Intuitively, the feature selection process strictly losses information. If the lost part of the information happens to include what we want to detect, inference becomes impossible and we cannot hope to detect all targets after reducing the data.

Less conservatively, we optimize average-case instead of worst-case performance, seeking features that are good "on average." Specifically, we consider all ϕ_X satisfying $\|\phi_X\|^2 = 1$, which corresponds to all possible target features U such that $D(P_{X|U}(\cdot|u)\|P_X) = \frac{1}{2}\epsilon^2 + o(\epsilon^2)$. Given a probability measure over this set of targets, we then treat ϕ_X as a random vector with this law, and seek to maximize $\mathbb{E}[\nu_B(\phi_X, \psi)]$.

While in general it can be difficult or artificial to define a measure over all the different ways that the target U can be encoded in the data X, in the case of a local geometry, the dependence on this measure is rather weak. As such, we simply take ϕ_X to be uniformly distributed over the surface of unit divergence ball (see the dotted circle on $\mathcal{P}_\mathcal{X}$ in Fig. 2) defining the collection $\{\phi_X : \|\phi_X\|^2 = 1\}$. With this choice, the solution is simple. Indeed, the unit divergence ball on $\mathcal{P}_\mathcal{X}$ is mapped by the linear map B to an ellipsoid in $\mathcal{P}_\mathcal{Y}$, as also shown in Fig. 2. It is then clear that choosing ψ to be along the principal axis of this ellipsoid maximizes the average power captured. An additional convenience of optimizing this average loss factor is that the solution of choosing ψ along the principal axis is identical to the solution to the opportunistic formulation

$$\max_{\phi_X:\,\|\phi_X\|^2=1} \quad \max_{\psi:\,\|\psi\|^2=1} \quad \nu_B(\phi_X, \psi)\,. \tag{10}$$

It is convenient to re-interpret (10) in the language of divergence. Specifically, for any choice of ϕ_X, we consider for each $\epsilon > 0$ a distribution $Q_X^{(\epsilon)}(x) = P_X(x) + \epsilon\sqrt{P_X(x)}\phi_X(x)$, i.e. $Q_X^{(\epsilon)} \leftrightarrow \phi_X$, and let $Q_Y^{(\epsilon)}$ be the induced marginal distribution on $\mathcal{Y}$. Then (10) can be expressed in the form

$$\max_{\phi_X:\,\|\phi_X\|^2=1} \quad \lim_{\epsilon\to 0} \quad \frac{D(Q_Y^{(\epsilon)}\|P_Y)}{D(Q_X^{(\epsilon)}\|P_X)}, \tag{11}$$

which is also a limiting version of (7) and thus corresponds to (3). Note that in (11), $D(Q_X^{(\epsilon)}\|P_X) = \frac{1}{2}\epsilon^2 + o(\epsilon^2)$.

We can interpret (11) as finding the target U that is the most "distinguishable" from the observations of Y samples. Compared to worst-case optimization, this formulation avoids attempting to capture every possible target, and in so doing, via the linear mapping defined by the DTM, avoids locking on to some isolated individual cases, instead offering solutions that are good on average. We summarize these statements in the following proposition, which emphasizes the central role of the singular value decomposition (SVD) in computing the optimal feature; we omit the proof [6] due to space constraints.

Proposition 1 (Universal Feature Characterization)**.** *Let the optimal solutions of* (10) *be* $\phi_X^* \leftrightarrow L_1$ *and* $\psi^* \leftrightarrow L_2$*. Then,* ϕ_X^* *and* ψ^* *are the right and left singular vectors of* B *respectively, corresponding to its second largest singular value, and* L_1 *and* L_2 *are the maximal correlation functions of* X *and* Y*, respectively. Moreover, we have*

$$\frac{1}{\sqrt{P_Y(y)}}\,\psi^*(y) \propto \lim_{\epsilon\to 0} \frac{1}{\epsilon} \log \frac{Q_Y^{(\epsilon)}(y)}{P_Y(y)}, \qquad y \in \mathcal{Y},$$

where $Q_Y^{(\epsilon)}$ *is the optimal choice for* (11)*. Finally,* ψ^* *maximizes* $\mathbb{E}[\nu_B(\phi_X, \psi)]$*, where the expectation is with respect to a uniform distribution over the unit-sphere* $\{\phi_X : \|\phi_X\|^2 = 1\}$*.*

III. Efficient Computation of Universal Features

As we have developed, the optimum feature is obtained via singular vectors of B. Since the alphabets $\mathcal{X}$ and $\mathcal{Y}$ are large, directly computing the SVD of B is impractical. However, such computation can be circumvented by exploiting the special geometric structure inherent in our problem.

To see this, first note that the computation of g^* is equivalent to finding the HGR maximal correlation. Specifically, g^* is the solution g to the maximization

$$\max_{f,g} \mathbb{E}[f(X)\,g(Y)]\,. \tag{12}$$

This connection is important, because a well-known solution to (12) is given by the so-called Alternating Conditional Expectations (ACE) algorithm [1]. A description of this algorithm in our SVD notation is as follows. For a $K \times K$ real matrix A (taken to be square without loss of generality) with ordered singular values $\sigma_0 \geq \sigma_1 \geq \ldots \geq \sigma_{K-1}$ and corresponding normalized right singular vectors $u_0, u_1, \ldots, u_{K-1} \in \mathbb{R}^K$, we can find u_0 using the *power method* from numerical linear algebra [2]. We start with an arbitrary vector $\phi \in \mathbb{R}^K$, and repeatedly multiply A^TA to it. Since $A^TA = \sum_{i=0}^{K-1} \sigma_i^2 u_i u_i^T$

by the spectral theorem, and $\phi = \sum_{i=0}^{K-1} \alpha_i u_i$ for some $\alpha_i \in \mathbb{R}$ as $\{u_0, \ldots, u_{K-1}\}$ is an orthonormal basis, we can write: $\left(A^T A\right)^m \cdot \phi = \sum_{i=0}^{K-1} \sigma_i^{2m} \alpha_i u_i$. Assuming $\alpha_0 \neq 0$, as m becomes large, the component corresponding to σ_0 dominate the sum, and the resulting vector is aligned with u_0. In practice, we scale the intermediate vectors to have unit norm once every few iterations for numerical stability. The power method converges geometrically (exponentially) with ratio σ_1^2/σ_0^2. In the case $\sigma_0 = \sigma_1$ the power method outputs some linear combination of u_0 and u_1. Moreover, after computing u_0, we can compute u_1 by selecting an initial vector ϕ that is orthogonal to u_0.

Applying this method to the DTM B, we let the initial vector be $\phi \in \mathbb{R}^{|\mathcal{X}|}$ with the corresponding score function $\forall x \in \mathcal{X}$, $f(x) = \phi(x)/\sqrt{P_{0,X}(x)}$, and let $\psi = B \cdot \phi \in \mathbb{R}^{|\mathcal{Y}|}$ be the output vector with corresponding score function $\forall y \in \mathcal{Y}$, $g(y) = \psi(y)/\sqrt{P_{0,Y}(y)}$. Then, by using (9), we have, for every $y \in \mathcal{Y}$,

$$\begin{aligned} g(y) &= \frac{\psi(y)}{\sqrt{P_{0,Y}(y)}} = \frac{1}{\sqrt{P_{0,Y}(y)}} \sum_{x \in \mathcal{X}} B(x,y)\, \phi(x) \\ &= \frac{1}{\sqrt{P_{0,Y}(y)}} \sum_{x \in \mathcal{X}} \frac{P_{Y|X}(y|x)\sqrt{P_{0,X}(x)}}{\sqrt{P_{0,Y}(y)}} \sqrt{P_{0,X}(x)} f(x) \\ &= \mathbb{E}\left[f(X)|Y=y\right]. \end{aligned}$$

Hence, multiplying ϕ by B is equivalent to taking the conditional expectation of f, i.e., $\mathbb{E}\left[f(X)|Y=y\right]$. Similarly, multiplying ψ by B^T is equivalent to taking the conditional expectation $\mathbb{E}\left[g(Y)|X=x\right]$ on g. The resulting ACE algorithm for obtaining the singular vectors of B corresponding to the singular value σ_1 is as follows.

Algorithm 1 ACE Algorithm

Require: knowledge of $P_{X,Y}$

1. Initialize: randomly pick $g(y)$, $y \in \mathcal{Y}$
   Center: $g(y) \leftarrow g(y) - \mathbb{E}\left[g(Y)\right]$

repeat

2a. $f(x) \leftarrow \mathbb{E}\left[g(Y)|X=x\right]$, $\forall x \in \mathcal{X}$
2b. $g(y) \leftarrow \mathbb{E}\left[f(X)|Y=y\right]$, $\forall y \in \mathcal{Y}$
2c. Regularize: $g(y) \leftarrow g(y)/\sqrt{\mathbb{E}\left[g^2(Y)\right]}$, $\forall y \in \mathcal{Y}$

until $\mathbb{E}\left[f(X)g(Y)\right]$ stops to increase.

In Algorithm 1, the initial choice of $g(Y)$ is constrained to have zero mean. This is equivalent to setting ψ to be orthogonal to v_0, which corresponds to the constant function on $\mathcal{Y}$. This centering needs to be implemented only once at the initialization step, since we have $\mathbb{E}\left[f(X)\right] = \mathbb{E}\left[g(Y)\right] = 0$ in all of the following steps. Moreover, the regularization step 2c does not have to be performed in every iteration; it is needed only once in a while to avoid arithmetic underflow.

Finally, as we discussed, in practice we do not have available to us $P_{X,Y}$, but rather i.i.d. training samples $(\tilde{x}_1, \tilde{y}_2), \ldots, (\tilde{x}_k, \tilde{y}_k)$ from this distribution. In this case, we modify Algorithm 1. While the associated convergence analysis is subtle [7], the primary change is to replace the conditional expectations in steps 2a and 2b with their *empirical* conditional averages defined with respect to this training data. We omit the details due to space limitations.

It is important to emphasize that the resulting algorithm will generally be surprisingly effective even when the number of training samples is small and thus $\hat{P}_{X,Y}$ is poor. This is because we are not seeking to learn everything about B, only its dominant singular vectors, as a full analysis of the associated sample-complexity reveals.

IV. Concluding Remarks

There are variety of significant additional insights and extensions that space precludes including in this paper. First, there is a natural interpretation of the optimizing f^* in (12), which is also produced by the associated ACE algorithm. In particular, it has an important interpretation as the corresponding feature U of X that that is most detectable with this universal g^*.

Second, this framework extends naturally to multiple features, leading to rich generalizations of the analysis, geometry, and algorithms described in this paper. In this case, there is a collection of orthononal universal features $g_1, g_2, \ldots$, ordered by importance, and an associated orthonomal collection $f_1, f_2, \ldots$. Moreover, together with the associated singular values $\sigma_1 < \sigma_2 < \ldots$, they provide the following efficient modal decomposition of $P_{X,Y}$:

$$\frac{P_{X,Y}(x,y)}{P_X(x)\, P_Y(y)} = \sum_{i=1}^{\infty} \sigma_i\, f_i(x)\, g_i(y).$$

In practice, a relative small number of such features is often sufficient to capture the dominant behavior of this distribution.

Finally, as a preliminary investigation of potential, we applied our universal feature functions as a single-layer feature mapping in the MNIST digital recognition problem, and obtained an error rate of 2.4%. This performance is comparable to the performance of a standard two-layer convolutional neural network, and thus demonstrates that our information-theoretic framework is promising for larger scale applications.

References

[1] L. Breiman and J. H. Friedman, "Estimating Optimal Transformations for Multiple Regression and Correlation," *J. Am. Stat. Assoc.*, vol. 80, no. 391, pp. 614-619, 1985.

[2] J. W. Demmel, *Applied Numerical Linear Algebra*, 1st ed., Philadelphia: Society for Industrial and Applied Mathematics (SIAM), 1997.

[3] H. Gebelein, "Das statistische problem der Korrelation als variations-und Eigenwertproblem und sein Zusammenhang mit der Ausgleichungsrechnung," *Z. für angewandte Math. und Mech.*, vol. 21, pp. 364-379, 1941.

[4] H. O. Hirschfeld, "A connection between correlation and contingency," *Proc. Cambridge Phil. Soc.*, vol. 31, pp. 520-524, 1935.

[5] S.-L. Huang and L. Zheng, "Linear Information Coupling Problems," in *Proc. Int. Symp. Inform. Theory*, July 2012.

[6] S.-L. Huang, A. Makur, L. Zheng, and G. W. Wornell, "Universal feature for universal feature selection in high-dimensional inference: An information-theoretic framework," preprint, 2017.

[7] A. Makur, F. Kozynski, S.-L. Huang, and L. Zheng, "An efficient algorithm for information decomposition and extraction," in *Proc. Allerton Conf. Commun., Contr., Computing*, (Monticello, IL), Oct. 2015.

[8] A. Rényi, "On measures of dependence," *Acta Mathematica Academiae Scientiarum Hungarica*, vol. 10, no. 3-4, pp. 441–451, 1959.