Feature Extraction for Universal Hypothesis Testing via Rank-Constrained Optimization

Dayu Huang and Sean Meyn
Dept. of ECE and CSL, UIUC, Urbana, IL 61801, U.S.A.
dhuang8, meyn'at'illinois.edu

***Abstract*—This paper concerns the construction of tests for universal hypothesis testing problems, in which the alternate hypothesis is poorly modeled and the observation space is large. The mismatched universal test is a feature-based technique for this purpose. In prior work it is shown that its finite-observation performance can be much better than the (optimal) Hoeffding test, and good performance depends crucially on the choice of features. The contributions of this paper include:**

(i) We obtain bounds on the number of *ε-distinguishable distributions* in an exponential family.

(ii) This motivates a new framework for feature extraction, cast as a rank-constrained optimization problem.

(iii) We obtain a gradient-based algorithm to solve the rank-constrained optimization problem and prove its local convergence.

Keywords: Universal test, mismatched universal test, hypothesis testing, feature extraction, exponential family

I. Introduction

A. Universal Hypothesis Testing

In universal hypothesis testing, the problem is to design a test to decide in favor of either of two hypothesis $H0$ and $H1$, under the assumption that we know the probability distribution π^0 under $H0$, but have uncertainties about the probability distribution π^1 under $H1$. One of the applications that motivates this paper is detecting abnormal behaviors [1]: In the applications envisioned, the amount of data from abnormal behavior is limited, while there is a relatively large amount of data for normal behavior.

To be more specific, we consider the hypothesis testing problem in which a sequence of observations $Z_1^n := (Z_1, \dots, Z_n)$ from a finite observation space Z is given, where n is the number of samples. The sequence Z_1^n is assumed to be i.i.d. with marginal distribution $\pi^i \in \mathcal{P}(\mathsf{Z})$ under hypothesis Hi $(i = 0, 1)$, where $\mathcal{P}(\mathsf{Z})$ is the probability simplex on Z.

Hoeffding [2] introduced a universal test, defined using the empirical distributions and the Kullback-Leibler divergence. The empirical distributions $\{\Gamma^n : n \geq 1\}$ are defined as elements of $\mathcal{P}(\mathsf{Z})$ via,

$$\Gamma^n(A) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}\{Z_k \in A\}, \qquad A \subset \mathsf{Z}.$$

The Kullback-Leibler divergence for two probability distributions $\mu^1, \mu^0 \in \mathcal{P}(\mathsf{Z})$ is defined as,

$$D(\mu^1 \| \mu^0) = \langle \mu^1, \log(\mu^1/\mu^0) \rangle.$$

where the notation $\langle \mu, f \rangle$ denotes expectation of f under the distribution μ, i.e., $\langle \mu, f \rangle = \sum_z \mu(z) f(z)$. The Hoeffding test is the binary sequence,

$$\phi_n^{\mathsf{H}} = \mathbb{I}\{D(\Gamma^n \| \pi^0) \geq \eta\},$$

where η is a nonnegative constant. The test decides in favor of $H1$ when $\phi^{\mathsf{H}} = 1$.

It was demonstrated in [3] that the performance of the Hoeffding test is characterized by both its error exponent and the variance of the test statistics. We summarize this in Theorem 1.1. The error exponent is defined for a test sequence $\boldsymbol{\phi} := \{\phi_1, \phi_2, \dots\}$ adapted to Z_1^n as

$$J_\phi^0 := \liminf_{n\to\infty} -\frac{1}{n}\log(\pi^0\{\phi_n = 1\}),$$
$$J_\phi^1 := \liminf_{n\to\infty} -\frac{1}{n}\log(\pi^1\{\phi_n = 0\}).$$

Theorem 1.1: 1) The Hoeffding test achieves the optimal error exponent J_ϕ^1 among all tests satisfying a given constant bound $\eta \geq 0$ on the exponent J_ϕ^0, i.e., $J_{\phi^{\mathsf{H}}}^0 \geq \eta$ and

$$J_{\phi^{\mathsf{H}}}^1 = \sup\{J_\phi^1 : \ \textit{subject to } J_\phi^0 \geq \eta\},$$

2) The asymptotic variance of the Hoeffding test depends on the size of the observation space. When Z_1^n has marginal π^0, we have

$$\lim_{n\to\infty} \mathsf{Var}\,[nD(\Gamma^n \| \pi^0)] = \tfrac{1}{2}(|\mathsf{Z}| - 1).$$

Theorem 1.1 is a summary of results from [2], [3]. The second result can be derived from [4], [5], [6]. It has been demonstrated in [3] that the variance implies a drawback of the Hoeffding test, hidden in the analysis of the error exponent: Although asymptotically optimal, this test is not effective when the size of the observation space is large compared to the number of observations.

B. Mismatched Universal Test

It was demonstrated in [3] that the potentially large variance in the Hoeffding test can be addressed by using a generalization of the Hoeffding test called the *mismatched universal test*, which is based on the relaxation of KL divergence introduced in [7]. The name of the mismatched divergence comes from literature on mismatched decoding [8]. The mismatched universal test enjoys several advantages:

1) It has smaller variance.

2) It can be designed to be robust to errors in the knowledge of π^0.
3) It allows us to incorporate into the test partial knowledge about π^1 (see Lemma 2.1), as well as other considerations such as the heterogeneous cost of incorrect decisions.

The mismatched universal test is based on the following variational representation of KL divergence,

$$D(\mu\|\pi) = \sup_f \big(\langle \mu, f\rangle - \log(\langle \pi, e^f\rangle)\big) \tag{1}$$

where the optimization is taken over all functions $f\colon \mathsf{Z} \to \mathbb{R}$. The supremum is achieved by the log-likelihood ratio.

The mismatched divergence is defined by restricting the supremum in (1) to a function class $\mathcal{F}$:

$$D^{\text{MM}}_{\mathcal{F}}(\mu\|\pi) := \sup_{f\in\mathcal{F}} \big(\langle \mu, f\rangle - \log(\langle \pi, e^f\rangle)\big). \tag{2}$$

The associated mismatched universal test is defined as

$$\phi_n^{\mathsf{MM}} = \mathbb{I}\{D^{\text{MM}}(\Gamma^n\|\pi^0) \ge \eta\}.$$

In this paper we restrict to the special case of a *linear* function class: $\mathcal{F} = \big\{f_r := \sum_i^d r_i\psi_i\big\}$ where $\{\psi_i\}$ is a set of *basis* functions, and r ranges over $\mathbb{R}^d$. We assume throughout the paper that $\{\psi_i\}$ is *minimal*, i.e., $\{\mathbf{1}, \psi_1, \dots, \psi_d\}$ are linearly independent. The basis functions can be interpreted as *features* for the universal test. In this case, the definition (2) reduces to the convex program,

$$D^{\text{MM}}(\mu\|\pi) = \sup_{r\in\mathbb{R}^d} \big(\langle \mu, f_r\rangle - \log(\langle \pi, e^{f_r}\rangle).$$

The asymptotic variance of the mismatched universal test is proportional to the dimension of the function class d instead of $|\mathsf{Z}| - 1$ as seen in the Hoeffding test:

$$\lim_{n\to\infty} \mathsf{Var}\,[nD^{\text{MM}}(\Gamma^n\|\pi^0)] = \tfrac{1}{2}d,$$

when Z_1^n has marginal π^0 [3]. In this way we can expect substantial variance reduction by choosing a small d. The function class also determines how well the mismatched divergence $D^{\text{MM}}(\pi^1\|\pi^0)$ approximates the KL divergence $D(\pi^1\|\pi^0)$ for possible alternate distributions $\pi^1$and thus the error exponent of the mismatched universal test [9]. In sum, the choice of the basis functions $\{\psi_i\}$ is critical for successful implementation of the mismatched universal test. The goal of this paper is to construct algorithms to construct a suitable basis.

C. Contributions of this paper

In this paper we propose a framework to design the function class $\mathcal{F}$, which allows us to make the tradeoff between the error exponent and variance. One of the motivations comes from results presented in Section II on the maximum number of ε-*distinguishable distributions* in an exponential family, which suggests that it is possible to use approximately $d = \log(p)$ basis functions to design a test that is effective against p different distributions. In Section III we cast the feature extraction problem as a rank constrained optimization problem, and propose a gradient-based algorithm with provable local convergence property to solve it.

The construction of a basis studied in this paper is a particular case of the feature extraction problems that have been studied in many other contexts. In particular, the framework in this paper is connected to the exponential family PCA setting of [10]. The most significant difference between this work and the exponential PCA is that our framework finds features that capture the *difference* between distributions, and the latter finds features that are *common* to the distributions considered.

The mismatched divergence using empirical distributions can be interpreted as an estimator of KL divergence. To improve upon the Hoeffding test, we may apply other estimators, such as those using data dependent features [11], [12], or those motivated by source-coding techniques [13] and others [14]. Our approach is different from them in that we exploit the limited possibilities of alternate distributions.

II. Distinguishable Distributions

The quality of the approximation of KL divergence using the mismatched divergence depends on the dimension of the function class. The goal of this section is to quantify this statement.

A. Mismatched Divergence and Exponential Family

We first describe a simple result suggesting how a basis might be chosen given a finite set of alternate distributions, so that the mismatched divergence is equal to the KL divergence for those distributions:

Lemma 2.1: For any p possible alternate distributions $\{\pi^1, \pi^2, \dots, \pi^p\}$, absolutely continuous with respect to π^0, there exist $d = p$ basis functions $\{\psi_1, \dots, \psi_d\}$ such that $D^{\text{MM}}(\pi^i\|\pi^0) = D(\pi^i\|\pi^0)$ for each i. These functions can be chosen to be the log-likelihood ratios $\{\psi_i = \log(\pi^i/\pi^0)\}$. □

It is overly pessimistic to say that given p distributions we require $d = p$ basis functions. In fact, Lemma 2.2 demonstrates that if all p distributions are in the same d-dimensional *exponential family*, then d basis functions suffices. We first recall the definition of an exponential family: For a function class $\mathcal{F}$ and a distribution ν, the exponential family $\mathcal{E}(\nu, \mathcal{F})$ is defined as:

$$\mathcal{E}(\nu, \mathcal{F}) = \{\mu : \mu(z) = \frac{\nu(z)e^{f(z)}}{\langle \nu, e^f\rangle}, f \in \mathcal{F}\}.$$

We will restrict to the case of linear function class, and we say that the exponential family is d-dimensional if this is the dimension of the function class $\mathcal{F}$. The following lemma is a reinterpretation of Lemma 2.1 for the exponential family:

Lemma 2.2: Consider any $p + 1$ mutually absolutely continuous distributions $\{\pi^i : 0 \le i \le p\}$. Then $D^{\text{MM}}_{\mathcal{F}}(\pi^i\|\pi^j) = D(\pi^i\|\pi^j)$ for all $i \ne j$ if and only if $\pi^i \in \mathcal{E}(\pi^0, \mathcal{F})$ for all i.

B. Distinguishable Distributions

Except in trivial cases, there are obviously infinitely many distributions in an exponential family. In order to characterize the difference between different exponential families of different dimension, we consider a subset of distributions which we call ε-distinguishable distributions.

The motivation comes from the fact that KL divergences between two distributions are infinite if neither is absolutely continuous with respect to the other, in which case we say they are *distinguishable*. When the distributions are distinguishable, we can design a test that achieves infinite error exponent. For example, consider two distributions π^0, π^1 on $\mathsf{Z} = \{z_1, z_2, z_3\}$: $\pi^0(z_1) = \pi^0(z_2) = 0.5$; $\pi^1(z_2) = \pi^1(z_3) = 0.5$. It is easy to see that the two error exponents of the test $\phi_n(Z_1^n) = \mathbb{I}\{\Gamma^n(z_3) > 0.2\}$ are both infinite. It is then natural to ask: Given p distributions that are pairwise distinguishable, how many basis functions do we need to design a test that is effective for them?

Distributions in an exponential family must have the same support. We thus consider distributions that are approximately distinguishable, which leads to the definitions listed below: Consider the set-valued function F^ϵ parametrized by $\epsilon > 0$,

$$F^\epsilon(x) := \{z : x(z) \geq \max_z(x(z)) - \epsilon\}$$

- Two distributions π^1, π^2 are *ϵ-distinguishable* if $F(\pi^1) \setminus F(\pi^2) \neq \emptyset$ and $F(\pi^2) \setminus F(\pi^1) \neq \emptyset$.
- A distribution π is called *ϵ-extremal* if $\pi(F^\epsilon(\pi)) \geq 1-\epsilon$, and a set of distributions $\mathcal{A}$ is called ϵ-extremal if every $\pi \in \mathcal{A}$ is ϵ-extremal.
- For an exponential family $\mathcal{E}$, the integer $N(\mathcal{E})$ is defined as the maximum N such that there exists an $\epsilon_0 > 0$ such that for any $0 < \varepsilon < \epsilon_0$, there exists an ε-extremal $\mathcal{A} \subseteq \mathcal{E}$ such that $|\mathcal{A}| \geq N$ and any two distributions in $\mathcal{A}$ are ε-distinguishable.

One interpretation of the final definition is that the test using a function class $\mathcal{F}$ is effective against $N(\mathcal{E})$ distributions, in the sense that the error exponents for the mismatched universal test are the same as for the Hoeffding test, where $\mathcal{E} = \mathcal{E}(\nu, \mathcal{F})$:

Lemma 2.3: Consider a function class $\mathcal{F}$ and its associated exponential family $\mathcal{E} = \mathcal{E}(\nu, \mathcal{F})$, where ν has full support, and define $N = N(\mathcal{E}(\nu, \mathcal{F}))$. Then, there exists a sequence $\{A^{(1)}, A^{(2)}, \ldots, A^{(m)} : m \geq 1\}$, such that for each k the set $A^{(k)} \subset \mathcal{E}$ consists of N distributions,

$$D_{\mathcal{F}}^{\text{MM}}(\pi \| \pi') = D(\pi, \pi') \quad \text{for any } \pi, \pi' \in A^{(k)}$$

and

$$\lim_{k \to \infty} \min_{\substack{\pi, \pi' \in A^{(k)} \\ \pi \neq \pi'}} D_{\mathcal{F}}^{\text{MM}}(\pi \| \pi') = \infty.$$

□

Let $\mathcal{P}(d)$ denote the collection of all d-dimensional exponential families. Define $\bar{N}(d) = \max_{\mathcal{E} \in \mathcal{P}(d)} N(\mathcal{E})$. In the next result we give lower and upper bounds on $\bar{N}(d)$, which imply that $\bar{N}(d)$ depends exponentially on d:

Proposition 2.4: The maximum $\bar{N}(d) = \max_{\mathcal{E}} N(\mathcal{E})$ admits the following lower and upper bounds:

$$\bar{N}(d) \quad \geq \quad \exp\left(\lfloor \frac{d}{2} \rfloor [\log(|\mathsf{Z}|) - \log \lfloor \frac{d}{2} \rfloor - 1]\right) \tag{3}$$

$$\bar{N}(d) \quad \leq \quad \exp\left((d+1)(1 + \log(|\mathsf{Z}|) - \log(d+1))\right) \tag{4}$$

It is important to point out that $\bar{N}(d)$ is exponential in d. This answers the question asked at the beginning of this section: There exist p approximately distinguishable distributions for which we can design an effective mismatched test using approximately $\log(p)$ basis functions.

III. Feature Extraction via Rank-constrained Optimization

Suppose that it is known that the alternate distributions can take on p possible values, denoted by $\pi^1, \pi^2, \ldots, \pi^p$. Our goal is to choose the function class $\mathcal{F}$ of dimension d so that the mismatched divergence approximates the KL divergence for these alternate distributions, while at the same time keeping the variance small in the associated universal test. The choice of d gives the tradeoff between the quality of the approximation and the variance in the mismatched universal test. We assume that $0 < D(\pi^i \| \pi^0) < \infty$ for all i.[1]

We propose to use the solution to the following problem as the function class:

$$\max_{\mathcal{F}} \{\frac{1}{p} \sum_{i=1}^{p} \gamma^i D_{\mathcal{F}}^{\text{MM}}(\pi^i \| \pi^0) : \dim(\mathcal{F}) \leq d\} \tag{5}$$

where $\dim \mathcal{F}$ is the dimension of the function class $\mathcal{F}$. The weights $\{\gamma_i\}$ can be chosen to reflect the importance of different alternate distributions. This can be rewritten as the following rank-constrained optimization problem:

$$\begin{aligned} \max \quad & \tfrac{1}{p} \textstyle\sum_{i=1}^{p} \gamma^i \big(\langle \pi^i, X_i \rangle - \log(\langle \pi^0, e^{X_i} \rangle\big) \\ \text{subject to} \quad & \mathsf{rank}\,(X) \leq d \end{aligned} \tag{6}$$

where the optimization variable X is a $p \times |\mathsf{Z}|$ matrix, and X_i is the ith row of X, interpreted as a function on Z. Given an optimizer X^*, we choose $\{\psi_i\}$ to be the set of right singular vectors of X^* corresponding to nonzero singular values.

A. Algorithm

The optimization problem in (6) is not a convex problem since it has a rank constraint. It is generally very difficult to design an algorithm that is guaranteed to find a global maximum. The algorithm proposed in this paper is a generalization of the Singular Value Projection (SVP) algorithm of [15] designed to solve a low-rank matrix completion problem. It is globally convergent under certain conditions valid for matrix completion problems. However, in this prior work the objective function is quadratic; we are not aware of any prior work generalizing these algorithms to the case of a general convex objective function.

Let $h(X)$ denote the objective function of (6). Let $\mathcal{S}$ denote the set of matrices satisfying $\mathsf{rank}\,(X) \leq d$. Let $\mathcal{P}_{\mathcal{S}}$ denote the projection onto $\mathcal{S}$:

$$\mathcal{P}_{\mathcal{S}}(Y) = \arg\min\{\|Y - X\| : \mathsf{rank}\,(X) \leq d\}.$$

where we use $\|\cdot\|$ to denote the Frobenius norm. The algorithm proposed here is defined as the following iterative gradient projection:

[1] In practice the possible alternate distributions will likely take on a continuum of possible values. It is our wishful thinking that we can choose a finite approximation with p distributions, and choose d much smaller than p, and the resulting mismatched universal test will be effective against all alternate distributions. Validation of this optimism will be left to future work.

1) $Y^{k+1} = X^k + \alpha^k \nabla h(X^k)$.
2) $X^{k+1} = \mathcal{P}_{\mathcal{S}}(Y^{k+1})$.

The projection step is solved by keeping only the d largest singular values of Y^{k+1}. The iteration is initialized with some arbitrary X^0 and is stopped when the $\|X^{k+1} - X^k\| \leq \epsilon$ for some small $\epsilon > 0$.

B. Convergence Result

We can establish local convergence:

Proposition 3.1: Suppose $\bar{X}$ satisfies $\mathsf{rank}\,(\bar{X}) = d$ and is a local maximum, i.e. there exists $\delta > 0$ such that for any matrix $X \in \mathcal{S}$ satisfying $\|X - \bar{X}\| \leq \delta$, we have $h(\bar{X}) > h(X)$. Choose $\alpha^k = \alpha$ for all k where $0 < \alpha < 2/(\frac{1}{p}\max_i \gamma^i)$. Then there exists a $\delta' > 0$ such that if X^0 satisfies $\|X^0 - \bar{X}\| \leq \delta'$ and $\mathsf{rank}\,(X^0) \leq d$, then $X^k \to \bar{X}$ as $k \to \infty$. Moreover, the convergence is geometric. □

Let $\mathcal{H}$ denote the hyperplane $\mathcal{H} = \{\bar{X}W_1 + W_2\bar{X} : W_1 \in \mathbb{R}^{n\times n}, W_2 \in \mathbb{R}^{p\times p}\}$. The main idea of the proof is that near $\bar{X}$ the set $\mathcal{S}$ can be approximated by this hyperplane $\mathcal{H}$, as demonstrated in Lemma 3.2.

Lemma 3.2: There exist $\delta > 0$ and $M > 0$ such that: 1) for any $X \in \mathcal{S}$ satisfying $\|X - \bar{X}\| \leq \delta$, there exists $Z \in \mathcal{H}$ such that $\|Z - X\| \leq M\|X - \bar{X}\|^2$; 2) for any $Z \in \mathcal{H}$ satisfying $\|Z - \bar{X}\| \leq \delta$, there exists $X \in \mathcal{S}$ satisfying $\|X - Z\| \leq M\|Z - \bar{X}\|^2$.

Let $Z^k = \mathcal{P}_{\mathcal{H}}(Y^k)$, i.e., the projection of Y^k onto $\mathcal{H}$. We obtain from Lemma 3.2 that Z^k is close to X^k as follows:

Lemma 3.3: Consider any $\bar{X}$ satisfying $\mathsf{rank}\,(\bar{X}) = d$. There exist $\delta > 0$ and $M > 0$ such that if $\|Z^k - \bar{X}\| \leq \delta$, then $\|Z^k - X^k\| \leq M\|Y^k - \bar{X}\|^{\frac{3}{2}}$.

Lemma 3.4: Gradients of $h(X)$ are Lipschitz with constant $L = \frac{1}{p}\max_i \gamma^i$, i.e. $\|\nabla h(X_1) - \nabla h(X_2)\| \leq L\|X_1 - X_2\|$.

Lemma 3.5: Suppose $\bar{X}$ is a local maximum in $\mathcal{S}$ and $\mathsf{rank}\,(\bar{X}) = d$. Then $\bar{X}$ is also a local maximum in $\mathcal{H}$.

Outline of Proof of Proposition 3.1: Using standard results form optimization theory, we can prove that for any small enough $\delta > 0$, if $\|X^k - \bar{X}\| \leq \delta$ and $\alpha < \frac{2}{L}$, then $\|Z^{k+1} - \bar{X}\| \leq q\|X^k - \bar{X}\|$ for some $q < 1$ where q could depend on δ, and $\|Y^{k+1} - \bar{X}\| \leq \|X^k - \bar{X}\|$. Thus, we can choose a δ small enough so that $M\delta^{\frac{1}{2}} \leq \frac{1-q}{2}$. With this choice, we have

$$\begin{aligned}\|X^{k+1} - \bar{X}\| &\leq \|Z^{k+1} - \bar{X}\| + \|Z^{k+1} - X^{k+1}\| \\ &\leq \|Z^{k+1} - \bar{X}\| + M\delta^{\frac{1}{2}}\|Y^{k+1} - \bar{X}\| \\ &\leq (q + \tfrac{1}{2}(1-q))\|X^k - \bar{X}\|.\end{aligned}$$

Proposition 3.1 then follows from induction. ■

IV. Simulations

We consider probability distributions in an exponential family of the form $\pi^i(z) = \exp\{\sum_{k=1}^{q}\theta_{i,k}\psi_i(z) + \sum_{i=k}^{q'}\theta'_{i,k}\psi'_i(z)\}$. We first randomly generate $\{\psi_i\}$ and $\{\psi'_i\}$ to fix the model. A distribution is obtained by randomly generating $\{\theta_{i,k}\}$ and $\{\theta'_{i,k}\}$ according to uniform distributions on $[-1, 1]$ and $[-0.1, 0.1]$, respectively. In application of the algorithm presented in Section III-A, the bases $\{\psi_i\}$ and $\{\psi'_i\}$ are not given. This model can be interpreted as a perturbation to q-dimensional exponential family with basis $\{\psi_i\}$.

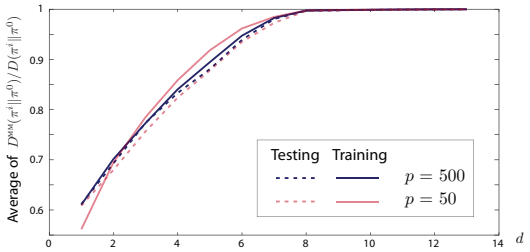

Fig. 1: Dashed curve: average of $D^{\text{MM}}(\mu^i\|\pi^0)/D(\mu^i\|\pi^0)$. Solid curve: average of $D^{\text{MM}}(\pi^i\|\pi^0)/D(\pi^i\|\pi^0)$

In the experiment we have two phases: In the feature extraction (training) phase, we randomly generate $p+1$ distributions, taken as $\pi^0, \ldots, \pi^p$. We then use our techniques in (5) with the proposed algorithm to find the function class $\mathcal{F}$. The weights γ_i are chosen as $\gamma^i = 1/D(\pi^i\|\pi^0)$ so that the objective value is no larger than 1. In the testing phase, we randomly generate t distributions, denoted by $\mu^1, \ldots, \mu^t$. We then compute the average of $D^{\text{MM}}(\mu^i\|\pi^0)/D(\mu^i\|\pi^0)$.

For the experimental results shown in Figure 1, the parameters are chosen as $q = 8$, $q' = 5$, and $t = 500$. Shown in the figure is an average of $D^{\text{MM}}(\pi^i\|\pi^0)/D(\pi^i\|\pi^0)$ (for training) as well as $D^{\text{MM}}(\mu^i\|\pi^0)/D(\mu^i\|\pi^0)$ (for testing) for two cases: $p = 50$ and $p = 500$. We observe the following:

1) The objective value increases gracefully as d increases. For $d \geq 7$, the values are close to 1.
2) The curve for training and testing are closer when p is larger, which is expected.

V. Conclusions

The main contribution of this paper is a framework to address the feature extraction problem for universal hypothesis testing, cast as a rank-constrained optimization problem. This is motivated by results on the number of easily distinguishable distributions, which demonstrates that it is possible to use a small number of features to design effective universal tests for a large number of possible distributions. We propose a gradient-based algorithm to solve the rank-constrained optimization problem, and the algorithm is proved to converge locally. Directions considered in current research include: applying the nuclear-norm heuristic [16] to solve the optimization problem (5), applying this framework to real-world data, and extension of this framework to incorporate other form of partial information.

Appendix

A. Proof of the lower bound in Proposition 2.4

We give a constructive proof of the lower bound (3) by combining ideas in Lemma A.1 and A.2.

Lemma A.1: $\bar{N}(2) \geq |\mathsf{Z}|$.

Proof: We pick the following two basis functions ψ_1, ψ_2:

$$\begin{aligned}\psi_1 &= [|\mathsf{Z}| - 1, |\mathsf{Z}| - 2, \ldots, 0], \\ \text{and } \psi_2 &= [1, 1.5, \sum_{j=0}^{2} 2^{-j}, \ldots, \sum_{j=0}^{|\mathsf{Z}|-1} 2^{-j}].\end{aligned} \tag{7}$$

For $1 \le k \le |\mathsf{Z}|$, define u^k as $u^k = \psi_1 + 2^{k-0.5}\psi_2$. Assuming without loss of generality that $\mathsf{Z} = \{1, \ldots, |\mathsf{Z}|\}$, we have $\arg\max_z u^k(z) = k$.

Now, for any $\beta > 0$, $1 \le k \le |\mathsf{Z}|$, define the distribution

$$\pi^{k,\beta}(z) = C \exp\{\beta u^k(z)\}.$$

where C is a normalizing constant. Since there are only finite choices of k, for any small enough ϵ, there exists β_0 such that for $\beta \ge \beta_0$, $\{\pi^{k,\beta}, 1 \le k \le |\mathsf{Z}|\}$ are ϵ-extremal and any two distributions in $\{\pi^{k,\beta}, 1 \le k \le |\mathsf{Z}|\}$ are ϵ-distinguishable. ■

Lemma A.2: $\bar{N}(d) \ge \binom{d}{\lfloor d/2 \rfloor}$

Proof: Take $\psi_k(z) = \mathbb{I}\{z = k\}$ for $1 \le k \le d$. ■

Outline of proof of the lower bound: The basis functions used in the construction are the Kronecker products of basis functions used for Lemma A.2 and Lemma A.1.

Let $J = \lfloor |\mathsf{Z}| / \lfloor \frac{1}{2} d \rfloor \rfloor$. Let $\bar{\psi}_1, \bar{\psi}_2$ denote the basis function defined in (7) with $|\mathsf{Z}|$ replaced by J. The basis functions used for the lower bound are given by

$$\psi_k(i + jJ) = \mathbb{I}\{j = k-1\}\bar{\psi}_1(i), \quad \text{for } 1 \le k \le \lfloor \tfrac{1}{2} d \rfloor,$$
$$\psi_{k+\lfloor d/2 \rfloor}(i + jJ) = \mathbb{I}\{j = k-1\}\bar{\psi}_2(i), \quad \text{for } 1 \le k \le \lfloor \tfrac{1}{2} d \rfloor.$$

■

B. Proof of the upper bound in Proposition 2.4

The main idea of the proof of (4) is to relate this bound to VC dimension. We first obtain an elementary upper bound.

Lemma A.3: $N(\mathcal{E}) \le \hat{N}(\mathcal{E})$, where

$$\hat{N}(\mathcal{E}) = |\{F^\epsilon(\sum_l r_l \psi_l) : r \in \mathbb{R}^d, \epsilon > 0\}|.$$

Proof: By definition if a subset A of $\mathcal{E}$ is ϵ-extremal, and any two distributions in A are ϵ-distinguishable, then for any two distributions $\pi^i, \pi^j \in A$, there exists $\epsilon_1, \epsilon_2 > 0$ such that $F^{\epsilon_1}(\log(\pi^1)) \ne F^{\epsilon_2}(\log(\pi^2))$. ■

Let $\mathbf{H}$ denote the set of all the half space in $\mathbb{R}^d$, and let $VC(\mathbf{H})$ denote the VC dimension of $\mathbf{H}$. It is known that $VC(\mathbf{H}) = d + 1$ [17, Corollary of Theorem 1].

For any finite subset B of $\mathbb{R}^d$, define $\tau(B) = |\{h \cap B : h \in \mathbf{H}\}|$. In other words, $\tau(B)$ is the number of subsets one can obtain by intersecting B with half-spaces from $\mathbf{H}$. A bound on $\tau(B)$ is given by Sauer's lemma:

Lemma A.4 (Sauer's Lemma): The following bound holds whenever $|B| \ge VC(\mathbf{H})$:

$$\tau(B) \le \left(\frac{e|B|}{VC(\mathbf{H})}\right)^{VC(\mathbf{H})}.$$

Consider any d-dimensional exponential family $\mathcal{E}$ with basis $\{\psi_l, 1 \le l \le d\}$. Define a set of function $\{y^i\} \subset \mathbb{R}^d$ via,

$$y^i_j = \psi_j(i), \qquad 1 \le i \le |Z|,\ 1 \le j \le d.$$

In other words, if we stack $\{\psi_l\}$ into a matrix so that each ψ_l is a row, then $\{y^i\}$ are the columns of this matrix. Let $B(\mathcal{E}) = \{y^i, 1 \le i \le |Z|\}$. The following lemma connects $\tau(B(\mathcal{E}))$ to $\hat{N}(\mathcal{E})$.

Lemma A.5: $\hat{N}(\mathcal{E}) \le \tau(B(\mathcal{E}))$.

Proof: For given $r \in \mathbb{R}^d$ and $\epsilon > 0$, denote $I = F^\epsilon(\sum_l r_l \psi_l)$. By the definition of F^ϵ we have $I = \{i : r^T y^i \ge \sup_z(\sum_l r_l \psi_l(z)) - \epsilon\}$. Therefore, there exists b such that $r^T y^i \ge b$ for all $i \in I$, and $r^T y^i < b$ for all $i \notin I$. That is, I is the subset of $\{y^i\}$ that lies in the half space $\{y : r^T y \ge b\}$. Thus, $\{y^i : i \in I\} \in \{h \cap B(\mathcal{E}) : h \in H\}$. Since this holds for any element in $\{F^\epsilon(\sum_l r_l \psi_l) : r \in \mathbb{R}^d, \epsilon > 0\}$, we obtain the result. ■

Proof of the upper bound: We obtain (4) on combining Lemma A.3, Lemma A.4 and Lemma A.5, together with the identity $VC(\mathbf{H}) = d + 1$. ■

Acknowledgment: This research was partially supported by AFOSR under grant AFOSR FA9550-09-1-0190 and NSF under grants NSF CCF 07-29031 and NSF CCF 08-30776. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the AFOSR or NSF.

REFERENCES

[1] D. E. Denning, "An intrusion-detection model," *IEEE Trans. Softw. Eng.*, vol. 13, no. 2, pp. 222 – 232, 1987.

[2] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Statist.*, vol. 36, pp. 369 – 401, 1965.

[3] J. Unnikrishnan, D. Huang, S. Meyn, A. Surana, and V. Veeravalli, "Universal and composite hypothesis testing via mismatched divergence," submitted for publication. [Online]. Available: http://arxiv.org/abs/0909.2234

[4] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *Ann. Math. Statist.*, vol. 9, pp. 60 – 62, 1938.

[5] B. S. Clarke and A. R. Barron, "Information-theoretic asymptotics of Bayes methods," *IEEE Trans. Inf. Theory*, vol. 36, no. 3, pp. 453 – 471, May 1990.

[6] I. Csiszár and P. C. Shields, "Information theory and statistics: A tutorial," *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 4, pp. 417 – 528, 2004.

[7] E. Abbe, M. Médard, S. Meyn, and Z. Lizhong, "Finding the best mismatched detector for channel coding and hypothesis testing," in *Information Theory and Applications Workshop, 2007*, 29 Feb. 2007, pp. 284 – 288.

[8] N. Merhav, G. Kaplan, A. Lapidoth, and S. S. Shitz, "On information rates for mismatched decoders," *IEEE Trans. Inf. Theory*, vol. 40, no. 6, pp. 1953 – 1967, Nov. 1994.

[9] D. Huang, J. Unnikrishnan, S. Meyn, V. Veeravalli, and A. Surana, "Statistical SVMs for robust detection, supervised learning, and universal classification," in *IEEE Information Theory Workshop on Networking and Information Theory*, Jun. 2009, pp. 62 – 66.

[10] M. Collins, S. Dasgupta, and R. E. Schapire, "A generalization of principal component analysis to the exponential family," in *Advances in Neural Information Processing Systems*, vol. 14. MIT Press, 2001, pp. 617–624.

[11] Q. Wang, S. R. Kulkarni, and S. Verdú, "Divergence estimation of continuous distributions based on data-dependent partitions," *IEEE Trans. Inf. Theory*, vol. 51, no. 9, pp. 3064 – 3074, Sep. 2005.

[12] W. Qing, S. R. Kulkarni, and S. Verdú, "Divergence estimation for multidimensional densities via -nearest-neighbor distances," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2392 – 2405, May 2009.

[13] J. Ziv and N. Merhav, "A measure of relative entropy between individual sequences with application to universal classification," *IEEE Trans. Inf. Theory*, vol. 39, no. 4, pp. 1270 – 1279, Jul. 1993.

[14] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," Department of Statistics, UC Berkeley, Tech. Rep. 764, Jan. 2007.

[15] R. Meka, P. Jain, and I. S. Dhillon, "Guaranteed rank minimization via singular value projection," 2009. [Online]. Available: http://www.citebase.org/abstract?id=oai:arXiv.org:0909.5457

[16] M. Fazel, H. Hindi, and S. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *Proceedings of the american control conference*, vol. 6, 2001, pp. 4734 – 4739.

[17] C. J. C. Burges, "A tutorial on Support Vector Machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121 – 167, Jun. 1998.