Statistical SVMs for robust detection, supervised learning, and universal classification.

Dayu Huang, Jayakrishnan Unnikrishnan, Sean Meyn, Venugopal Veeravalli
Department of Electrical and Computer Engineering
and the Coordinated Science Laboratory,
University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
dhuang8, junnikr2, meyn, vvv'at'illinois.edu

Amit Surana
United Technologies Research Center,
411 Silver Lane, East Hartford, CT.
SuranaA'at'utrc.utc.com

ABSTRACT

The *support vector machine* (SVM) has emerged as one of the most popular approaches to classification and supervised learning. It is a flexible approach for solving the problems posed in these areas, but the approach is not easily adapted to noisy data in which absolute discrimination is not possible. We address this issue in this paper by returning to the statistical setting. The main contribution is the introduction of a *statistical support vector machine* (SSVM) that captures all of the desirable features of the SVM, along with desirable statistical features of the classical likelihood ratio test. In particular, we establish the following:

(i) The SSVM can be designed so that it forms a continuous function of the data, yet also approximates the potentially discontinuous log likelihood ratio test.

(ii) Extension to universal detection is developed, in which only one hypothesis is labeled (a semi-supervised learning problem).

(iii) The SSVM generalizes the robust hypothesis testing problem based on a moment class.

Motivation for the approach and analysis are each based on ideas from information theory. A detailed performance analysis is provided in the special case of i.i.d. observations.

This research was partially supported by NSF under grant CCF 07-29031, by UTRC, Motorola, and by the DARPA ITMANET program. 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 NSF, UTRC, Motorola, or DARPA.

I. INTRODUCTION

Consider the binary hypothesis testing problem in which a single observation Z is drawn from some observation space Z. The two hypotheses are labeled H_0 and H_1. Here and in virtually all of the literature we restrict to tests that are determined by a decision region $\mathsf{Z}_1 \subset \mathsf{Z}$ such that H_1 is declared to be true if and only if $Z \in \mathsf{Z}_1$. Equivalently, the test is expressed as the binary value $\phi(Z)$, where ϕ is the indicator function on Z_1.

In the Bayesian setting it is assumed that Z has distribution π^1 under H_1, and π^0 otherwise. The log-likelihood ratio is defined by the logarithm of the Radon-Nykodim derivative, $L = \log(d\pi^1/d\pi^0)$. Given a threshold $c \in \mathbb{R}$, the log-likelihood ratio test (LRT) declares H_1 to be true if and only if $L(Z) \geq c$. That is, $\mathsf{Z}_1 = \{z \in \mathsf{Z} : L(z) \geq c\}$, and hence $\phi(z) = \mathbb{I}\{L(z) \geq c\}$.

An alternative approach that is becoming increasingly popular instead insists on perfect discrimination: Suppose that there exists a set Z_1 such $Z \in \mathsf{Z}_1$ under H_1, and $Z \in \mathsf{Z}_1^c$ otherwise. To construct an effective test it is assumed that a family of functions $\mathcal{F}$ is given, and a test is sought among the class of indicators $\phi(z) = \mathbb{I}\{f(z) \geq c\}$, where c is a scalar threshold. In [1] the following optimization problem is posed to construct a test which is optimal over this class:

$$\Delta^* = \max_{f \in \mathcal{F}} \inf_{z_1, z_0} (f(z_1) - f(z_0))$$

where the infimum is over all training data $\{z_1\}$ observed under H_1, and $\{z_0\}$ observed under H_0. If $\Delta^* > 0$, then a maximizer f^* will yield a test that discriminates perfectly. It is hoped that the discrimination is robust in the sense that $\Delta^* \gg 0$. In this case we can conclude that for some $c \in \mathbb{R}$,

$$f^*(Z) \gg c \text{ under } H_1, \quad \textit{and} \quad f^*(Z) \ll c \text{ under } H_0. \quad (1)$$

This process is known as the support vector machine, or SVM. Tests of this form may be found in the earlier independent work of Vapnik and Chervonenkis [2] and Cover [3].

If perfect discrimination is not possible, then the optimization problem to choose f^* is modified (e.g., via a penalty-function — see [4], [1], [5], [6]).

The motivation for the criterion (1) can be found in the LRT. Suppose that the modeling assumptions of the SVM hold, so that perfect discrimination is possible. Take any pair of probability distributions π^1 and π^0 that are mutually singular, and crudely model the data by assuming that $Z \sim \pi^i$ under H_i. Then, the criterion (1) is satisfied optimally using the log likelihood ratio:

$$L(Z) = +\infty \text{ under } H_1, \quad \textit{and} \quad L(Z) = -\infty \text{ under } H_0. \quad (2)$$

In this way the SVM framework can be embedded in a statistical framework.

However, this embedding obscures the true value of the SVM framework. In typical settings where SVMs are applied, the data is of very high dimension, and the construction of a

prior based on training data may be infeasible. Consider for example the problem of distinguishing handwritten numbers — By restricting to two numbers, say 7 and 8, this is a binary hypothesis testing problem. It is not unreasonable to argue that perfect discrimination is possible: Contained in any instance of the number 8 is a 'cross' near the center, which is very unlikely to appear in the written 7. It is not obvious how to construct a useful prior distribution on each digit. On the other hand, the SVM framework is appealing conceptually and a numerically convenient alternative that yields an effective test for this particular application and many others.

Moreover, the flexibility in choice of the function class $\mathcal{F}$ has tremendous value. Through choice of $\mathcal{F}$, prior information can be encoded. For example, if the number 8 is rotated slightly then it remains an 8. If the function f^* is continuous, then it will be robust to this and similar deformations.

One goal of this paper is to create a more cohesive bridge between SVM and statistical classification. Within the restricted setting of binary classification, we obtain a test that defines a separating hyperplane in the space of probability measures, as in the classical Bayesian setting. However, the function f that defines the hyperplane can be taken to be continuous even when the log-likelihood ratio is discontinuous, or extended valued. In the general statistical support vector machine (SSVM) framework we have a family $\mathcal{F}$ of functions from Z to $\mathbb{R}$, and we search for a function $f^* \in \mathcal{F}$ that separates 'optimally' π^0 and π^1. The formulation of optimality will depend upon the context, and is based on standard formulations from statistical hypothesis testing. Throughout much of the paper we restrict to a linear function class: We assume that real-valued functions $\{\psi_1, \dots, \psi_d\}$ are given, define $f_r = \sum r_i \psi_i$ for $r \in \mathbb{R}^d$, and $\mathcal{F} = \{f_r : r \in \mathbb{R}^d\}$. We use ψ to denote the column vector $(\psi_1, \dots, \psi_d)^{\intercal}$, so that $f_r = r^{\intercal}\psi$.

The SSVM detector is developed in a sequential hypothesis testing framework, in which the observations are obtained as an infinite sequence $\boldsymbol{Z} = (Z_1, Z_2, \dots)$ evolving in Z. A sequence of tests is defined by a sequence of functions $\phi_n \colon \mathsf{Z}^n \to \{0, 1\}$, assumed to be of the form, for some function $f \colon \mathsf{Z} \to \mathbb{R}$, and some scalar c,

$$\phi_n(Z_1, \dots, Z_n) = \mathbb{I}\Big\{\frac{1}{n}\sum_{t=1}^{n} f(Z_t) \ge c\Big\}. \tag{3}$$

The problem is to choose the function f to obtain the most effective test. The SSVM shares many desirable features of the usual SVM. In particular, the test can be constructed so that it forms a continuous function of the data, which ensures that it is robust to small perturbations of the data. In addition, robustness to noise is built into the construction of an optimal test.

In the sequential hypothesis testing problem with i.i.d. observations we obtain finer results:

(i) Optimization of the detector (3) over a linear function class was considered previously in [7]. The optimal test is obtained by solving a finite-dimensional convex program.

(ii) Performance of the detector is addressed using a bound on divergence (i.e. relative entropy) called the *mismatched divergence*. The mismatched divergence shares many properties with ordinary divergence.

(iii) A universal detector is introduced, based on the mismatched divergence, that can be designed to form a continuous function of the observations. It achieves the same asymptotic performance as the detector in (i) when $\mathcal{F}$ is a linear class, without knowlege of π^1.
Expressions for the asymptotic mean and variance are obtained that mirror those known for the standard Bayesian setting [8], [9].

(iv) When $\mathcal{F}$ is taken to be the logarithmic class, $\mathcal{F} = \{\log(\sum r_i \psi_i) : r \in \mathbb{R}^d\}$ with $\psi_1 \equiv 1$, then the universal detector in (iii) coincides with one of the min-max algorithms introduced in [10], [11]. The bound in (ii) coincides with the "worst-case divergence subject to moment constraints" obtained in this prior work.

While the original SVM framework is not based on statistics, there is a considerable body of work on the statistical analysis of the SVM and its refinements. The survey [12] provides a full statistical analysis of SVM techniques, and surveys several refinements. See also the more recent monograph [6]. Much of this prior work is restricted to an i.i.d. setting as in this paper.

The mismatched divergence developed in this paper was first introduced in [7]. Other generalizations of divergence that include mismatched divergence are introduced in [13]. Csiszár's f-divergence is also similar in spirit to the mismatched divergence developed here [14].

The SSVM introduced in this paper is as flexible as the SVM approach, or any of its refinements. Its value lies in its natural performance analysis via information theoretic methods, which leads to the exact asymptotic bias and variance results obtained in Sec. III.

II. Sequential hypothesis testing

A. Notation and essentials

We let Z denote a subset of Euclidean space, and let $\mathcal{P}(\mathsf{Z})$ denote the space of (Borel) probability distributions on Z. For a measurable function $f \colon \mathsf{Z} \to \mathbb{R}$ and $\pi \in \mathcal{P}(\mathsf{Z})$ we denote the mean $\int f(z)\pi(dz)$ by $\pi(f)$, or by $\langle \pi, f \rangle$ when we wish to emphasize the convex-analytic setting. For two probability distributions μ^0, $\mu^1 \in \mathcal{P}(\mathsf{Z})$ the divergence is expressed,

$$D(\mu^1 \| \mu^0) = \langle \mu^1, \log(d\mu^1 / d\mu^0) \rangle \tag{4}$$

The *divergence set* is defined by $\mathcal{Q}_\alpha(\pi) = \{\mu \in \mathcal{P}(\mathsf{Z}) : D(\mu \| \pi) < \alpha\}$, for $\pi \in \mathcal{P}(\mathsf{Z})$ and $\alpha > 0$.

Suppose that $\boldsymbol{Z}$ is a sequence taking values in Z. Throughout the paper it is assumed that $\boldsymbol{Z}$ is i.i.d., and we consider an asymptotic setting for performance evaluation. This provides a convenient setting for analysis. However, the decision algorithms and computational algorithms obtained in this paper are effective under much more general conditions.

The empirical distributions $\{\Gamma^n : n \geq 1\}$ are elements of $\mathcal{P}(\mathsf{Z})$ defined by

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

where we adopt the convention that $\mu(A)$ represents the measure of set A under μ. In the binary hypothesis testing problem it is assumed that Γ^n approximates π^0 or π^1, depending on which of the two hypotheses is true. In the classical setting in which these observations are i.i.d., the LLR test is optimal for any n, with respect to Bayesian or Neyman-Pearson criteria [15]. This can be expressed in terms of the empirical distributions as follows, with $\{c_n\}$ a sequence of thresholds:

$$\phi^{\text{LRT}}(Z_1^n) = \mathbb{I}\{\langle \Gamma^n, L\rangle > c_n\} \tag{5}$$

When Z is finite there is also the universal test in which π^1 need not be specified. At time n, based on the n observations, this is expressed using a threshold δ_n,

$$\phi^{\text{UNI}}(Z_1^n) = \mathbb{I}\{\Gamma^n \notin \mathcal{Q}_{\delta_n}(\pi^0)\} \tag{6}$$

See [16], [17], [11].

The asymptotic performance criteria are the two error rates, defined for a test sequence $\phi := \{\phi_1, \phi_2, \dots\}$ via,

$$J_\phi^0 := \liminf_{n\to\infty} -\frac{1}{n}\log(\pi^0\{\phi_n(Z_1,\dots,Z_n) = 1\}), \tag{7}$$

$$J_\phi^1 := \liminf_{n\to\infty} -\frac{1}{n}\log(\pi^1\{\phi_n(Z_1,\dots,Z_n) = 0\})\,. \tag{8}$$

The asymptotic Neyman-Pearson (N-P) criterion of Hoeffding [18] is described as follows: For a given constant bound $\eta \geq 0$ on the false-alarm exponent, an optimal test is the solution to,

$$\beta^*(\eta) = \sup\{J_\phi^1 : \ \textit{subject to } J_\phi^0 \geq \eta\}\,, \tag{9}$$

where the supremum is over all test sequences $\boldsymbol{\phi}$ (see also Csiszár et. al. [19], [20].)

The universal test (6) with $\delta_n \equiv \eta$ is optimal for the finite alphabet model. It was shown in [18] that the simpler LRT (5) is also optimal for some threshold c independent of n, chosen so that the hyperplane $\mathcal{H} = \{\mu : \langle \mu, L\rangle = c\}$ separates the two divergence sets $\mathcal{Q}_\eta(\pi^0)$ and $\mathcal{Q}_{\beta^*(\eta)}(\pi^1)$, as shown in Fig. 1.

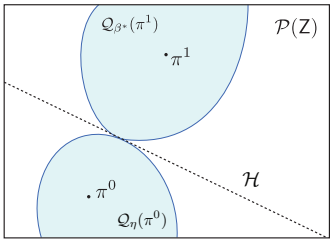

Fig. 1. Geometric solution to the Neyman Pearson problem.

The SSVM is defined as a relaxation of the Neyman Pearson formulation of hypothesis testing. Given a function class, denoted $\mathcal{F}$, the optimization problem (9) is solved over all tests in the class Φ, where Φ is the collection of tests of the form $\phi(Z_1^n) = \mathbb{I}\{\langle \Gamma^n, f\rangle > c\}$, with $c \in \mathbb{R}$, and with $f \in \mathcal{F}$. The value of this optimization problem is denoted,

$$\beta^{\text{MM}*}(\eta) := \sup\{J_\phi^1 : \ \textit{subject to } J_\phi^0 \geq \eta, \phi \in \Phi\}\,. \tag{10}$$

Several characterizations of $\beta^{\text{MM}*}$ are contained in Prop. 2.2.

To analyze the resulting test, and to obtain a smoothed universal test, we introduce next a relaxation of divergence.

B. Mismatched divergence

Relative entropy can be expressed as the convex dual of the log moment generating function as follows,

$$D(\mu\|\pi) = \sup\big(\mu(f) - \Lambda_\pi(f)\big)$$

where $\Lambda_\pi(f) = \log\big(\pi(e^f)\big)$, and the supremum is over all measurable functions $f\colon \mathsf{Z} \to \mathbb{R}$ satisfying $\Lambda_\pi(f) < \infty$. Based on a given function class $\mathcal{F}$, made up of a collection of functions f satisfying this requirement, we restrict the supremum to this class to obtain the *mismatched divergence*,

$$D^{\text{MM}}(\mu\|\pi) = \sup_{f\in\mathcal{F}}\big(\mu(f) - \Lambda_\pi(f)\big) \tag{11}$$

It is called the mismatched divergence since the framework is parallel to the mismatched decoder of [21]. We denote $\mathcal{Q}_\alpha^{\text{MM}}(\pi) = \{\mu \in \mathcal{P}(\mathsf{Z}) : D^{\text{MM}}(\mu\|\pi) < \alpha\}$.

In one application considered in further detail below, we obtain a universal test as the relaxation of (6):

$$\phi^{\text{UNI MM}}(Z_1^n) = \mathbb{I}\{\Gamma^n \notin \mathcal{Q}_{\delta_n}^{\text{MM}}(\pi^0)\} \tag{12}$$

Prop. 3.1 establishes that this test achieves $J_\phi^1 \geq \beta^{\text{MM}*}(\eta)$, and satisfies the constraint $J_\phi^0 \geq \eta$, when $\delta_n \equiv \eta$. However, the bias result in Prop. 3.2 suggests that a time-varying sequence will have better performance for a finite time horizon.

In the remainder of this section we consider two general function classes. We first show that, for a particular choice of $\mathcal{F}$, this relaxation of divergence captures the solution to a robust hypothesis testing problem.

1) Robust hypothesis testing: We say that $\mathcal{F}$ is a *logarithmic class* if it is expressed $\mathcal{F} = \{f_r = \log(\sum r_i\psi_i) : r \in \mathbb{R}^d$ and $f_r\colon \mathsf{Z} \to \mathbb{R}\}$. The following result follows from Theorem 2 of [10].

Proposition 2.1: Suppose that Z is compact, the functions $\{\psi_i\}$ are continuous, that $\psi_1 \equiv 1$, and suppose that $\mathcal{F}$ is a logarithmic class based on these functions. Then the mismatched divergence can be expressed,

$$D^{\text{MM}}(\mu\|\pi^0) = \inf_{\pi\in\mathbb{P}} D(\mu\|\pi)$$

where $\mathbb{P}$ denotes the "moment class" $\mathbb{P} = \{\pi : \pi(\psi) = \pi^0(\psi)\}$. Consequently, the decision region for the universal test (12), with $\delta_n = \eta$ for each n, corresponds to the robust decision region introduced in [10]:

$$Q_\eta^{\text{MM}}(\pi^0) = Q_\eta(\mathbb{P}) := \{\mu : \inf_{\pi\in\mathbb{P}} D(\mu\|\pi) < \eta\}.$$

2) Linear class: Here and throughout the remainder of the paper we restrict to the special case in which $\mathcal{F}$ is a finite dimensional linear class. In this case the "mismatched error exponent" $\beta^{\text{MM}*}(\eta)$ can be expressed as the solution to a finite dimensional convex program.

We assume that $\{\psi_i : 1 \leq i \leq d\}$ are measurable functions, and denote

$$\mathcal{F} = \{f_r = \sum r_i\psi_i : r \in \mathbb{R}^d\} \tag{13}$$

The following non-degeneracy conditions are imposed:

(A1) For any non-zero $v \in \mathbb{R}^d$,

$$\Sigma_0 v = 0 \Longrightarrow \Sigma_1 v \neq 0 \tag{14}$$

where the covariance matrices are defined by

$$\Sigma_i = \pi^i(\psi\psi^{\intercal}) - \pi^i(\psi)\pi^i(\psi^{\intercal}), \qquad i = 0, 1.$$

(A2) Function class $\mathcal{F}$ satisfies

$$D^{\text{MM}}(\pi^1 \| \pi^0) < \infty \text{ and } D^{\text{MM}}(\pi^0 \| \pi^1) < \infty$$

Note that (A1) implies that $\mathcal{F}$ does not include any non-zero constant functions. Condition (A2) implies that perfect discrimination is impossible. That is, if $f \in \mathcal{F}$ satisfies, for some $c \in \mathbb{R}$,

$$f \geq c \;\; a.e.\; [\pi^1] \quad \textit{and} \quad f \leq c \;\; a.e.\; [\pi^0]$$

Then $c = 0$ and $f \equiv 0$.

Proposition 2.2: Suppose that (A1) and (A2) hold. Then $\beta^{\text{MM}*}(\eta) < \infty$ for each $\eta \in (0, D^{\text{MM}}(\pi^1 \| \pi^0))$, and the following representations hold:

(i) $\beta^{\text{MM}*}(\eta) = \inf\{D^{\text{MM}}(\mu \| \pi^1) : D^{\text{MM}}(\mu \| \pi^0) \leq \eta\}$.

(ii) For some $\varrho > 0$:

$$\beta^{\text{MM}*}(\eta) = \max_{f \in \mathcal{F}} \Big(-\Lambda_{\pi^1}(-\varrho f) - \varrho(\Lambda_{\pi^0}(f) - \eta) \Big) \tag{15}$$

(iii) An optimizer $f^* \in \mathcal{F}$ to (15) defines an optimal test $\phi^* \in \Phi$, of the form $\phi^*(Z_1^n) = \mathbb{I}\{\langle \Gamma^n, f^* \rangle > c^*\}$ for some threshold $c^* \in \mathbb{R}$. It is optimal in the sense that it achieves the value $\beta^{\text{MM}*}(\eta)$, defined in (10).

III. Universal hypothesis testing

In this section, we consider the universal test based on the linear function class.

A. Mismatched universal detector

The universal test statistic is denoted

$$U(\Gamma^n) = D^{\text{MM}}(\Gamma^n \| \pi^0). \tag{16}$$

A universal test that is optimal, in the sense that it achieves the error rate $\beta^{\text{MM}*}(\eta)$, can be expressed in terms of this test statistic. It is expressed $\phi^{\text{UNI MM}}(Z_1^n) = \mathbb{I}\{\Gamma^n \notin \mathcal{Q}_\eta^{\text{MM}}(\pi^0)\}$, or equivalently $\phi^{\text{UNI MM}}(Z_1^n) = \mathbb{I}\{U(\Gamma^n) \geq \eta\}$. To establish optimality we take a closer look at Prop. 2.2.

The proof of Prop. 2.2 is based on the geometry in $\mathcal{P}(\mathsf{Z})$ illustrated in Fig. 2. The hyperplane $\mathcal{H} = \{\mu : \langle \mu, f^* \rangle = c\}$ separates the sets $\mathcal{Q}_\eta^{\text{MM}}(\pi^0)$, $\mathcal{Q}_{\beta^{\text{MM}*}(\eta)}^{\text{MM}}(\pi^1)$. Based on this interpretation we can can conclude that the optimal error exponent has the characterization,

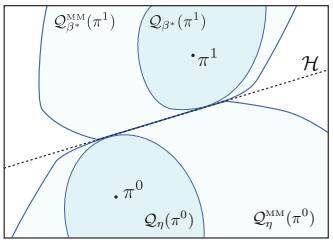

Fig. 2. Solution to the mismatched Neyman Pearson problem.

$$\beta^{\text{MM}*}(\eta) = \inf\{\beta : \mathcal{Q}_\eta^{\text{MM}}(\pi^0) \cap \mathcal{Q}_\beta^{\text{MM}}(\pi^1) \neq \emptyset\}$$

From this we can establish optimality of the universal detector.

Proposition 3.1: For $\eta \in (0, D^{\text{MM}}(\pi^1 \| \pi^0))$, with $\beta^{\text{MM}*}(\eta) < \infty$, the universal detector $\phi^{\text{UNI MM}}$ is optimal in the sense that it achieves the error rate $\beta^{\text{MM}*}(\eta)$.

We consider next the asymptotic bias and variance of the universal test statistic with i.i.d. observations.

B. Asymptotic bias and variance

By construction, the asymptotic mean of the universal test statistic is zero under H_0:

$$\lim_{n\to\infty} U(\Gamma^n) = \lim_{n\to\infty} \mathsf{E}[U(\Gamma^n)] = 0.$$

where the first limit holds with probability one. The variability of $U(\Gamma^n)$ is most naturally addressed through its asymptotic variance and bias when $\boldsymbol{Z}$ has marginal distribution π^0. We have computed these values in the special case in which the alphabet is finite.

Proposition 3.2: Suppose $\boldsymbol{Z}$ is drawn i.i.d. from a finite set Z with marginal π^0 and assume $\Sigma_0 > 0$. Then the universal statistic has bias of order n^{-1} and variance of order n^{-2}, and the normalized asymptotic values have simple, explicit forms:

$$\lim_{n\to\infty} n\mathsf{E}[U(\Gamma_n)] = \tfrac{1}{2}d \tag{17}$$

$$\lim_{n\to\infty} n^2\mathsf{E}[\big(U(\Gamma_n) - \mathsf{E}[U(\Gamma_n)]\big)^2] = \tfrac{1}{2}d \tag{18}$$

The expression for the asymptotic bias (17) can be deduced from the corresponding result in [8, p. 8, Section III.C] in the context of universal Bayesian hypothesis testing. Although the connection is not straightforward, a special case of (17) is easily derived from [8]: Suppose that $d = |\mathsf{Z}| - 1$, so that $\{1, \psi_1, \ldots, \psi_d\}$ is a maximal linearly independent set of functions. In this case $D^{\text{MM}} = D$, which gives

$$\lim_{n\to\infty} n\mathsf{E}[D(\Gamma_n \| \pi^0)] = \tfrac{1}{2}(|\mathsf{Z}| - 1).$$

Moreover, the variance expression given in eq. (18) implies the following expression for the asymptotic variance of the universal test:

$$\lim_{n\to\infty} n^2\mathsf{E}[\big(D(\Gamma_n \| \pi^0) - \mathsf{E}[D(\Gamma_n \| \pi^0)]\big)^2] = \tfrac{1}{2}(|\mathsf{Z}| - 1).$$

Consequently, the universal test based on ordinary divergence will be unreliable when the time horizon n is less than the square root of the alphabet size $|\mathsf{Z}|$. The proposed universal SSVM provides a tractable solution to this problem where the user can control bias and variance by choosing the dimensionality of the function class.

The proof of Prop. 3.2 is based on the following application of Taylor's theorem and the strong law of large numbers.

Lemma 3.3: Let $\{X_i : i = 1, 2, \ldots\}$ be a zero-mean i.i.d. sequence taking values in a compact convex set $\mathsf{X} \subset \mathbb{R}^N$, containing the origin θ as a relative interior point.

Suppose that the function $g : \mathbb{R}^N \mapsto \mathbb{R}$ satisfies $g(\theta) = 0$, $\sup_{x \in \mathsf{X}} |g(x)| \leq \overline{g} \in \mathbb{R}$, and $\nabla g(\theta)^{\intercal} X_i \equiv 0$ for all i. Further suppose that there is a compact set K containing θ as a relative interior point such that the gradient $\nabla g(x)$ and the Hessian $\nabla^2 g(x)$ are continuous over K. The set K also satisfies $-\frac{1}{n}\log \mathsf{P}\{S^n \notin K\} > 0$, where $S^n = \frac{1}{n}\sum_{i=1}^n X_i$, $n \geq 1$. Denote $M = \nabla^2 g(\theta)$, and $\Pi = \mathsf{E}[X_1(X_1)^{\intercal}]$. Then,

(i) $\lim_{n\to\infty} n\mathsf{E}[g(S^n)] = \frac{1}{2}\text{trace}(M\Pi)$
(ii) $\lim_{n\to\infty} n^2\mathsf{E}[(g(S^n) - \mathsf{E}[g(S^n)])^2] = \frac{1}{2}\text{trace}(M\Pi M\Pi)$

Prop. 3.2 is established by applying Lemma 3.3 to the function $g(x) \equiv U(x + \pi^0)$ with $\mathsf{X} = \mathcal{P}(\mathsf{Z}) - \pi^0$, $X_i = (\mathbb{I}_{z_1}(Z_i), \mathbb{I}_{z_2}(Z_i), \dots \mathbb{I}_{z_N}(Z_i))^{\intercal} - \pi^0$, $S^n = \Gamma^n - \pi^0$, $K = \{x \in \mathcal{P}(\mathsf{Z}) - \pi^0 : |x_i| \le \delta, \text{ for all } i \text{ such that } \pi_i^0 > 0\}$ where $\mathsf{Z} = \{z_1, z_2, \dots, z_N\}$ and $0 < \delta < \min\{\pi_i^0 : \pi_i^0 > 0\}$.

The bulk of the work in the remainder of the proof of Prop. 3.2 is the computation of derivatives of g.

C. Generalized Pinsker's Inequality

We conclude with a bound on the mismatched divergence that generalizes Pinsker's inequality.

The total variation norm distance between $\mu, \pi \in \mathcal{P}(\mathsf{Z})$ is defined by the supremum,

$$\|\mu - \pi\|_{\text{TV}} := \sup_A |\mu(A) - \pi(A)|$$

Pinsker's inequality [22] provides a lower bound on the divergence in terms of this norm: For any two probability measures

$$D(\mu\|\pi) \ge 2(\|\mu - \pi\|_{\text{TV}})^2 \tag{19}$$

The mismatched divergence has an equally simple lower bound. For any function $f\colon \mathsf{Z} \to \mathbb{R}$, the *span norm* is defined by $\|f\|_{\infty,\text{SP}} = (\sup f(z)) - (\inf f(z))$.

Proposition 3.4 (Generalized Pinsker's Inequality): For any two probability measures,

$$D^{\text{MM}}(\mu\|\pi) \ge 2\sup\left(\frac{\mu(f_r) - \pi(f_r)}{\|f_r\|_{\infty,\text{SP}}}\right)^2 \tag{20}$$

where the supremum is over all non-zero $r \in \mathbb{R}^d$.

Prop. 3.4 generalizes the classical inequality of Pinsker. For any $A \in \mathcal{B}(\mathsf{Z})$, take $d = 1$, $r = 1$, and let $\psi_1(z) = \mathbb{I}_A(z)$. In this case $\|f_r\|_{\infty,\text{SP}} = 1$, and Prop. 3.4 gives,

$$D(\mu\|\pi) \ge D^{\text{MM}}(\mu\|\pi) \ge 2|\mu(A) - \pi(A)|^2.$$

This implies Pinsker's inequality since $A \in \mathcal{B}(\mathsf{Z})$ was arbitrary.

IV. Conclusions

The main conclusion of this paper is that effective tests can be constructed that capture desirable features of both SVM and Bayesian approaches. Although performance analysis is restricted to an i.i.d. setting, the resulting tests are applicable in any setting in which the standard SVM approach can be applied.

There are many directions for future research:

(i) Based on Prop. 2.2 we can capture the 'machine' aspect of the SVM: A version of the stochastic gradient or stochastic Newton-Raphson method is introduced in [7] to compute the optimal test that achieves $\beta^{\text{MM}*}(\eta)$. Analysis in a non i.i.d. setting is lacking.

(ii) Extensions of these results can be formulated for the more general setting in which the parameterization is linear, but infinite dimensional. Suppose that $K\colon \mathsf{Z} \times \mathsf{Z} \to \mathbb{R}$ is bounded and continuous, and for any $\gamma \in \mathcal{P}(\mathsf{Z})$ denote $f_\gamma(z) = \int K(z_0, z)\gamma(dz_0)$. Computation of f^* remains feasible using a variant of the steepest ascent algorithm introduced in [11].

(iii) Applications to channel coding, and to change detection are topics of current research.

References

[1] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[2] V. Vapnik and A. Chervonenkis, "A note on a class of perceptrons," *Automat. Remote Control*, vol. 25, pp. 103–109, 1964.

[3] T. M. Cover, "Geometric and statistical properties of systems of linear inequalities with applications to pattern recognition," *IEEE Transactions on Elec. Comp.*, vol. EC-14, 1965.

[4] K. Bennett and O. Mangasarian, "Robust linear programming discrimination of two linearly inseparable sets," *Optimization Methods and Software*, vol. 1, pp. 23–34, 1992.

[5] H. Xu, S. Mannor, and C. Caramanis, "Robustness and regularization of support vector machines," McGill University," Unpublished report, 2008.

[6] I. Steinwart and A. Christmann, *Support Vector Machines*. Springer Publishing Company, 2008.

[7] E. Abbe, M. Medard, S. Meyn, and L. Zheng, "Finding the best mismatched detector for channel coding and hypothesis testing," *Information Theory and Applications Workshop, 2007*, pp. 284–288, 29 2007-Feb. 2 2007.

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

[9] H. Chernoff, "On the distribution of the likelihood ratio," *Ann. Math. Statistics*, vol. 25, pp. 573–578, 1954.

[10] J. Huang, C. Pandit, S. Meyn, and V. V. Veeravalli, "Extremal distributions in information theory and hypothesis testing (invited.)," in *In Proc. IEEE Information Theory Workshop, San Antonio, TX*, October 24-29 2004, pp. 76–81.

[11] C. Pandit and S. P. Meyn, "Worst-case large-deviations with application to queueing and information theory," *Stoch. Proc. Applns.*, vol. 116, no. 5, pp. 724–756, May 2006.

[12] S. Boucheron, O. Bousquet, and G. Lugosi, "Theory of classification: a survey of some recent advances," *ESAIM: P&S*, vol. 9, pp. 323–375, nov 2005. [Online]. Available: http://dx.doi.org/doi/10.1051/ps:2005018

[13] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *CoRR*, vol. abs/0809.0853, 2008.

[14] I. Csiszár, "Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems," *Ann. Statist.*, vol. 19, pp. 2032–2066, 1991.

[15] H. V. Poor, *An introduction to signal detection and estimation*, 2nd ed., ser. Springer Texts in Electrical Engineering. New York: Springer-Verlag, 1994, a Dowden & Culver Book.

[16] O. Zeitouni and M. Gutman, "On universal hypotheses testing via large deviations," *IEEE Trans. Inform. Theory*, vol. 37, no. 2, pp. 285–290, 1991.

[17] ——, "Correction to: "On universal hypotheses testing via large deviations"," *IEEE Trans. Inform. Theory*, vol. 37, no. 3, part 1, p. 698, 1991.

[18] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Statist.*, vol. 36, pp. 369–408, 1965.

[19] I. Csiszár, G. Katona, and G. Tusnády, "Information sources with different cost scales and the principle of conservation of entropy," in *Proc. Colloquium on Information Theory (Debrecen, 1967), Vol. I.* Budapest: János Bolyai Math. Soc., 1968, pp. 101–128.

[20] I. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedures," *Statistics & Decisions. International Journal for Statistical. Supplemental Issue # 1.*, pp. 205–237, 1984.

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

[22] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes*. Moscow, U.S.S.R.: Izv. Akad. Nauk, 1960.