[11] Z. Zhang and R. W. Yeung, "A non-Shannontype conditional inequality of information quantities," *IEEE Trans. Inf. Theory*, vol. 43, pp. 1982–1986, 1997.

[12] ——, "On characterization of entropy function via information inequalities," *IEEE Trans. Inf. Theory*, vol. 44, pp. 1440–1452, 1998.

On Minimizing Distortion and Relative Entropy

Michael P. Friedlander and Maya R. Gupta, *Member, IEEE*

***Abstract*—A common approach for estimating a probability mass function w when given a prior q and moment constraints given by $Aw \leq b$ is to minimize the relative entropy between w and q subject to the set of linear constraints. In such cases, the solution w is known to have exponential form. We consider the case in which the linear constraints are noisy, uncertain, infeasible, or otherwise "soft." A solution can then be obtained by minimizing both the relative entropy and violation of the constraints $Aw \leq b$. A penalty parameter σ weights the relative importance of these two objectives. We show that this penalty formulation also yields a solution w with exponential form. If the distortion is based on an ℓ_p norm, then the exponential form of w is shown to have exponential decay parameters that are bounded as a function of σ. We also state conditions under which the solution w to the penalty formulation will result in zero distortion, so that the moment constraints hold exactly. These properties are useful in choosing penalty parameters, evaluating the impact of chosen penalty parameters, and proving properties about methods that use such penalty formulations.**

***Index Terms*—Convex optimization, cross-entropy, exact penalty, function, inverse problem, relative entropy, Kullback–Leibler distance, maximum entropy, moment constraint.**

I. Introduction

Consider the problem of estimating a probability mass function $w \in [0,1]^k$ given a strictly positive prior $q \in [0,1]^k$. A useful restriction is that w must satisfy a set of moment constraints: if a random variable X is drawn according to w, the probability mass function w must satisfy

$$E_w[f_i(X)] \leq b_i \tag{1}$$

where $b_1, b_2, \ldots, b_m$ are the required moments, E_w is the expectation operator, and the functions f_i are (possibly nonlinear) transformations of the random variable X. (Typically, (1) is expressed as an equation; however, it is not much more difficult to treat the more general inequality case, as we do here.) The expectation operator is linear in w, so that we may compactly express the set of m moment constraints as a set of linear equations $Aw \leq b$, where the columns of $A \in \mathbb{R}^{m \times k}$ represent the transformations of a random variable X drawn according to w, and $b = (b_1, b_2, \ldots, b_m) \in \mathbb{R}^m$.

Manuscript received January 6, 2004; revised December 22, 2004. The work of M. P. Friedlander was supported by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, U.S. Department of Energy, under Contract W-31-109-Eng-38, and by the National Science and Engineering Research Council of Canada. The work of M. R. Gupta was supported in part by the National Science Foundation under Grant CCR-0073050.

M. P. Friedlander is with the Department of Computer Science, University of British Columbia, Vancouver, BC V6T 1Z4, Canada (e-mail: mpf@cs.ubc.ca).

M. R. Gupta is with the Department of Electrical Engineering, University of Washington, Seattle, WA 98195 USA (e-mail: gupta@ee.washington.edu).

Communicated by A. B. Nobel, Associate Editor for Pattern Recognition, Statistical Learning, and Inference.

Digital Object Identifier 10.1109/TIT.2005.860448

A standard approach for estimating w is to minimize the relative entropy function

$$\mathcal{I}(w;q) = \sum_{j=1}^{k} w_j \log \frac{w_j}{q_j} \tag{2}$$

over the constrained probability simplex

$$\mathbf{1}^T w = 1, \quad w \geq 0 \tag{3}$$

$$Aw \leq b \tag{4}$$

(see, for example, [1]–[3]). The symbol $\mathbf{1}$ denotes a vector of ones; its length is determined by context. Often the prior is the uniform distribution (i.e., $q = \frac{1}{k}\mathbf{1}$), and in that case the minimization of (2) is equivalent to maximizing entropy.

In practice, the data may be noisy or uncertain, or the constraints (3)–(4) may be *infeasible* and admit no solution. In such cases, a more appropriate estimate of the probability mass function may then be the solution w of

$$\begin{aligned} &\underset{w}{\text{minimize}}\ \mathcal{I}(w;q) + \sigma D(Aw - b) \\ &\text{subject to } \mathbf{1}^T w = 1, \qquad w \geq 0 \end{aligned} \tag{5}$$

where $D : \mathbb{R}^m \to \mathbb{R}$ is a convex function that measures the distortion in satisfying (4), and σ is a positive scalar used to balance the trade-off between minimizing relative entropy and distortion from the required moments. The parameter σ can be set in response to expected noise in measurements of A or b, and may depend on the units of measurement of the distortion function compared with the bits of relative entropy.

In this correspondence, we characterize the analytic form of the minimizing probability mass function of (5). Its solution has an exponential form for any convex distortion D (see Section III). When the distortion function is based on the ℓ_p norm, a minimizer of (5) has a rate of decay that is bounded by a function of σ, and moreover, the moment constraints will hold *exactly* for all σ over a finite threshold value (see Section IV). The special case of the ℓ_1 penalty is computationally important in practice and serves to illustrate the more general case. We discuss it in more detail in Section IV-D.

The estimate's bounded exponential-decay property ensures that no single observation (the columns of A) receives an arbitrarily large or small relative weight, regardless of the number of observations or their specific values. The bound on the exponential decay described in Theorem 4.2 translates into a bound on the ratio between the largest and the smallest components of the weight vector. We use the bounded exponential-decay property to prove asymptotic behavior of an asymmetric nonparametric neighborhood learning method that trades-off solving the linear interpolation equations with maximizing entropy. This application of (5), and the properties of its solution, are discussed in detail in Section V.

In Section VI, we discuss two extensions to our approach. First, the given analysis of (5) can be extended easily to the space of continuous probability distributions, and we note the assumptions required. Second, we consider a variation of the minimum relative entropy function with reversed arguments—the parameters that normally describe the prior become the variables—and give a source-coding interpretation.

Both the relative entropy problem (2)–(4) and the penalty function problem (5) belong to the class of *convex* optimization problems, which are characterized by a convex objective function and a convex polyhedral feasible region. Their convex structure makes them especially suitable for numerical solution by a variety of interior-point solvers for nonlinear optimization, such as KNITRO [4], LOQO [5], and MOSEK

[6] (though the problem may first have to be reformulated, as discussed above). We focus on the analytic solution of these problems.

A. Other Work on Minimizing Relative Entropy

The exponential form of the solutions to minimum relative entropy problems with equality constraints is a classic result. Proofs can be found in Kullback [2, Theorem 2.1] and in Cover and Thomas [7, Ch. 11]. Similar problems arise in rate-distortion theory, and solutions in that framework have also been shown to have an exponential form. (For an overview of rate-distortion results, see [7] and [8].)

Statistical inference based on minimum relative (or maximum) entropy with constraints is now standard (see [9], [2], [10], [1]). For example, relative entropy arises in statistics as the expected logarithm of the likelihood ratio, and its minimization has applications in hypothesis testing [2]. The results we present are applicable to large-deviations theory; for example, Sanov's theorem relies on solving for distributions that minimize a relative entropy [11]. Maximum entropy plays a role in statistical physics (see [12], [13]), and our results may illuminate certain problems in that field. Minimizing relative entropy is a common way to solve a wide variety of ill-posed problems (see [14], [3], [15]–[18]). There has been recent interest in approaches that do not necessarily treat (4) as a *hard* constraint (see, for example, [19], [20], and [21]).

Campbell [19] derives an analytic solution for a special case of (5). He considers a single equality constraint $a^T w = b$, where $a \in \mathbb{R}^k$, $b \in \mathbb{R}$, and the distortion function is based on the ℓ_1 norm. In that case, $D(a^T w - b) = |\sum_{j=1}^k a_j w_j - b|$. Campbell's analysis assumes that the uncertain moment b lies within the convex hull of the event set, so that b must satisfy $\min_i a_i \leq b \leq \max_i a_i$. In effect, the constraints (3)–(4) must be feasible.

Our analysis extends Campbell's work in several useful ways. First, multiple inequality constraints are allowed, so that there may be m moment constraints. Second, we give results for any convex distortion function D and derive more extensive results for specific distortion functions. Third, the moments b (now in $\mathbb{R}^m$) need not lie in the convex hull of the data vectors specified by the columns of A—a solution continues to exist even if the constraints (3)–(4) are infeasible.

B. Definitions

Much of our theoretical development is focused on a distortion function based on the ℓ_p norm, defined by

$$\|x\|_p = \begin{cases} \left(\sum_j |x_j|^p\right)^{1/p} & \text{if } 1 \leq p < \infty \\ \max_j |x_j| & \text{if } p = \infty. \end{cases}$$

We denote the jth component of a vector x by x_j. Let x^+ denote the positive part of a vector x, so that $(x)_j^+ = \max(0, x_j)$. Let a_{ij} denote the ijth element (ith row, jth column) of A.

The ℓ_p norm is convex and continuous, but not differentiable everywhere. In practice, the ℓ_1 and ℓ_∞ norms are most useful because there are well-known techniques for reformulating optimization problems with these functions into smooth optimization problems with equivalent solutions (see, for example, [22, Theorem 4.8] and [23, Sec. 4.B.3]). We use such a technique in Section IV-D.

Let ζ and y be the Lagrange multipliers associated with the first and second linear constraints of (3) and (4), respectively, and let z be the Lagrange multiplier associated with the bound constraint on w. A minimizer w^*, together with its associated Lagrange multipliers ζ^*, y^*, and z^*, must satisfy the Karush–Kuhn–Tucker (KKT) conditions.

Definition 1.1 (First-Order KKT Optimality Conditions): The 4-tuple (w^*, ζ^*, y^*, z^*) is a first-order KKT point of the optimization problem defined by (2)–(4) if the following hold:

$$1^T w^* = \mathbf{1} \tag{6a}$$

$$\nabla_w \mathcal{I}(w^*; q) + \zeta^* \mathbf{1} + A^T y^* = z^* \tag{6b}$$

$$\min(b - Aw^*, y^*) = 0 \tag{6c}$$

$$\min(w^*, z^*) = 0. \tag{6d}$$

Conditions (6c) and (6d) are shorthand, respectively, for feasibility and complementarity conditions that can be expressed explicitly as

$$\{b - Aw^* \geq 0,\ y^* \geq 0\} \text{ and } \{(Aw^*)_i = b_i \text{ or } y_i^* = 0\}$$

and

$$\{w^* \geq 0,\ z^* \geq 0\} \text{ and } \{w_i^* = 0 \text{ or } z_i^* = 0\}.$$

II. The Exponential Form

Before we consider the penalty-function formulation of the minimum relative entropy problem, we examine the formulation in which the constraints (3)–(4) are imposed explicitly, as so-called *hard* constraints

$$\begin{aligned} &\underset{w}{\text{minimize}}\ \mathcal{I}(w; q) \\ &\text{subject to } \mathbf{1}^T w = 1 \\ &\qquad\qquad Aw \leq b. \end{aligned} \tag{7}$$

We have anticipated a strictly positive solution w^* and disregarded the nonnegativity constraint $w \geq 0$. When computing a numerical solution in practice, however, the constraint would usually be kept explicit.

We establish in this section that the solution of (7) is exponential; its parameters are given by the columns of A and by y, the Lagrange multipliers associated with the second constraint. (The Lagrange multipliers ζ of the first constraint can be eliminated.) As discussed in Section I, this result is not new, and it can be derived from a variety of perspectives. Kullback [2, Theorem 2.1] derives the required result for continuous probability distributions and equality constraints. Vital to his approach are the assumptions that there exists a feasible solution to the constraints and that both the solution and the prior must be *generalized probability densities*. We make the analogous assumptions that the prior q is strictly positive, and that there exists a strictly positive w that satisfies the constraints of (7). The latter assumption is known in the optimization literature as either *Slater's constraint qualification* or *Slater's condition*. With this assumption, the first-order optimality conditions of (7) are in fact both necessary and sufficient (see, for example, [24], [25]).

Assumption 2.1 (Slater's Condition): There exists a feasible point w in the relative interior of the domain of $\mathcal{I}(w; q)$. In other words, there exists a w such that

$$w > 0, \quad \mathbf{1}^T w = 1, \quad Aw \leq b.$$

The KKT conditions for (7) are a special case of (6). The solution w^* of (7) must be positive, so that (6d) implies that $z^* = 0$. An optimal point of (7), together with its associated Lagrange multipliers, must therefore satisfy

$$\mathbf{1}^T w^* = 1 \tag{8a}$$

$$\nabla_w \mathcal{I}(w^*; q) + \zeta^* \mathbf{1} + A^T y^* = 0 \tag{8b}$$

$$\min(b - Aw^*, y^*) = 0. \tag{8c}$$

Theorem 2.2 (Exponential Form): Suppose that Slater's condition holds. Then there exist Lagrange multipliers y^* and ζ^* (corresponding to the constraints $Aw \leq b$ and $\mathbf{1}^T w = 1$, respectively) that satisfy (8), and (7) is solved by the vector with components

$$w_j^* = \frac{u_j}{\sum_{i=1}^k u_i}, \qquad j = 1, \ldots, k \tag{9a}$$

where

$$u_j = q_j \exp\left(- \sum_{\ell \in \mathcal{A}} a_{\ell j} y_\ell^* \right) \tag{9b}$$

and $\mathcal{A} = \{i | (Aw^*)_i = b_i\}$.

Proof: By Slater's condition, the feasible set of (7) is nonempty. Because $q > 0$, the definition of relative entropy implies that the level set $\{w | \mathcal{I}(w; q) \leq \mathcal{I}(w_0; q)\}$ is closed and bounded. The strict convexity of $\mathcal{I}$ therefore implies that there exists a unique solution w^* to (7). Slater's condition is sufficient to guarantee that there exist Lagrange multipliers ζ^* and y^* such that (w^*, ζ^*, y^*) satisfies the KKT conditions (8) (see [25, Sec. 5.2.3]).

Note that $\nabla_w \mathcal{I}(w; q)_j = 1 + \log(w_j / q_j)$. Solve the jth equation of (8b) for w_j^* to obtain

$$w_j^* = q_j \exp\left(-\zeta^* - (A^T y^*)_j - 1 \right). \tag{10}$$

Sum (10) over all j, and use (8a) to obtain

$$\sum_{j=1}^k q_j \exp\left(-\zeta^* - (A^T y^*)_j - 1 \right) = 1.$$

Hence, ζ^* must satisfy

$$\zeta^* = \log\left(\sum_{j=1}^k q_j \exp\left\{ -(A^T y^*)_j - 1 \right\} \right). \tag{11}$$

Replacing ζ^* in (8b) with (11), and subsequently solving for w_j^*, we arrive at

$$w_j^* = \frac{q_j \exp\left\{ -(A^T y^*)_j \right\}}{\sum_{i=1}^k q_i \exp\left\{ -(A^T y^*)_i \right\}}. \tag{12}$$

However, note that (8c) implies that $y_i^* = 0$ if $i \notin \mathcal{A}$, so that (12) can be rewritten as (9a)–(9b), as required. □

III. Penalty Formulation

It may be that the constraints (3)–(4) are infeasible, or that the constraint $Aw \leq b$ need not be solved exactly. For example, the data may be known to be noisy, such that $b = b_{\text{true}} + n$, where n is some known or unknown noise; or the mean constraint may be uncertain; or fidelity to the prior q may be highly important relative to the constraint. These cases may be captured by introducing the set of constraints $Aw \leq b$ into the objective via a penalty function as done in (5).

In Lemma 3.1, we show that the solution to (5) will have an exponential form for a wide class of convex distortion functions D. The minimizer of (5) will depend uniquely on the penalty parameter σ. Therefore, we can denote its parameterized solution by $w^*(\sigma)$. A consequence of Lemma 3.1 is that $w^*(\sigma)$ is *also* the unique solution of (7), but with the mean constraint given by $Aw = b(\sigma)$, where $b(\sigma) \stackrel{\text{def}}{=} Aw^*(\sigma)$.

Lemma 3.1 (Exponential Form: Penalty Formulation): Suppose that D is convex and that $D(Aw - b)$ achieves its minimum for all w such that $Aw \leq b$. The following properties then hold:

1) the problem (5) has a unique solution $w^* > 0$;
2) w^* is also a unique solution to (7) with $b \equiv \bar{b}$, where $\bar{b}$ is any vector such that $\bar{b} \geq Aw^*$;
3) the unique solution w^* has an exponential form defined by (9), where y^* is the Lagrange multiplier of (7) corresponding to the constraints $Aw \leq \bar{b}$.

Proof:

(Part 1) Over the compact set defined by the constraints $\mathbf{1}^T w = 1$ and $w \geq 0$, D is convex and $\mathcal{I}$ is strictly convex. Therefore, a unique minimizer w^* of (5) exists, and $w^* > 0$.

(Part 2) Let $\bar{b} \stackrel{\text{def}}{=} Aw^*$, and let $\bar{w}$ be a minimizer of (7) with the constraint $Aw \leq \bar{b}$. (Such a minimizer must exist because the constraint is feasible, i.e., w^* is feasible.) Because $\bar{w}$ solves (7), it satisfies the constraint $A\bar{w} \leq \bar{b}$, and thus $D(A\bar{w} - \bar{b}) = D(Aw^* - \bar{b})$, the minimum value of D.

Further, it must be that $\mathcal{I}(w^*; q) = \mathcal{I}(\bar{w}; q)$. Otherwise, if $\mathcal{I}(w^*; q) < \mathcal{I}(\bar{w}; q)$, $\bar{w}$ could not solve (7), because w^* would be feasible and have lower relative entropy, which is a contradiction. Similarly, if $\mathcal{I}(w^*; q) > \mathcal{I}(\bar{w}; q)$, w^* could not be the minimizer of (5). Because w^* satisfies the constraint $Aw^* \leq \bar{b}$, and $\mathcal{I}(w^*; q) = \mathcal{I}(\bar{w}; q)$, w^* must be a minimizer of (7). Moreover, the minimizer of (7) is unique, so that $w^* = \bar{w}$, as required.

(Part 3) Parts 1 and 2 imply that (7) satisfies Slater's condition. Therefore, Theorem 2.2 applies, and because w^* is the unique solution of (7), it must satisfy (9), as required. □

IV. Exact Penalty Formulation

Consider the distortion function

$$D(Aw - b) = \| (Aw - b)^+ \|_p. \tag{13}$$

With this distortion, the penalty-function formulation (5) is *exact* if (3)–(4) is feasible: for a finitely large penalty parameter σ, the solution of (5) has zero distortion, so that the constraints (3)–(4) are satisfied exactly. Exact penalty functions play an important role in the modeling of continuous optimization problems; they have rich theoretical properties, and when the norm is polyhedral (i.e., $p = 1$ or $p = \infty$), they are computationally practical because they can be reformulated as polyhedral constraints. The use of penalty functions constitutes a particular approach: by augmentation of the objective function to include a penalty on the constraint violation, a constrained (and possibly difficult) problem can be transformed into an unconstrained (and easier, we hope) problem. Exact penalty functions were first analyzed by Pietrzykowski [26], and later by Bertsekas [27], Fletcher [28], and Han and Mangasarian [22], among others.

The objective of (5) is convex for any $1 \leq p \leq \infty$, but it is not everywhere differentiable. When the norm is polyhedral, (5) can be reformulated as an equivalent and smooth problem, and in that case, the corresponding first-order KKT conditions (see Definition 1.1) can be applied to find optimal solutions. Moreover, a variety of algorithms for smooth, constrained optimization could then be used to numerically solve the smooth reformulations. We discuss one such reformulation for $p = 1$ in Section IV-D. In general, there exists a rich theory of optimization for nonsmooth functions, and in Theorem 4.2 we derive a result analogous to Theorem 2.2 for (5) when the distortion function is given by the more general ℓ_p norm.

A. Nonsmooth Optimality Concepts

We summarize in this section some of the optimality concepts from nonsmooth optimization that we need for our analysis. Our treatment follows the approach of [28]. The vector g is a *subgradient* of the convex function $f(x) : \mathbb{R}^n \to \mathbb{R}$ at x if

$$f(x+p) \geq f(x) + g^T p$$

for all $p \in \mathbb{R}^n$. The subgradient is normal to a supporting hyperplane of f at x. The set of all subgradients

$$\partial f(x) \stackrel{\text{def}}{=} \{g | f(x+p) \geq f(x) + g^T p \text{ for all } p \in \mathbb{R}^n\}$$

is the *subdifferential* of f at x. When f is differentiable at x, there is only a single supporting hyperplane at that point, and the subgradient is unique and corresponds to $\nabla f(x)$.

Definition 4.1 (Nonsmooth First-Order Optimality): A triple (w^*, ζ^*, y^*) is a first-order optimal point of (5) if it satisfies the following conditions:

$$\mathbf{1}^T w^* = 1 \tag{14a}$$

$$\nabla_w \mathcal{I}(w^*; q) + \zeta^* \mathbf{1} + A^T y^* = 0 \tag{14b}$$

where $y^* \in \partial(\sigma \|(Aw^* - b)^+\|_p)$. [Note that we have anticipated a strictly positive solution w^*, so that there is no need for a condition analogous to (6d).]

Comparing (8) with (14), we see how $y^* \in \partial(\sigma \|(Aw^* - b)^+\|_p)$ may be interpreted as a kind of Lagrange multiplier for the constraint $Aw \leq b$ implied by the penalty function (13). Note that for any given σ, $Aw^* \leq b$ may or may not hold with strict inequality.

The *dual norm* is particularly important for the study of exact penalty functions, and will be useful in our analysis. For any norm $\|y\|_p$ in $\mathbb{R}^n$, the corresponding dual norm is defined as

$$\|y\|_d = \sup_{\|x\|_p \leq 1} y^T x.$$

For any p and d such that $1/p + 1/d = 1$, the ℓ_p and ℓ_d norms are duals of each other.

B. Solution With Exact Penalization: Inequality Constraints

Theorem 4.2 shows that the penalty function formulation (5) always has a solution with the form specified in (9). The parameter y^* is the Lagrange multiplier associated with the (relaxed) moment constraints $Aw \leq b$, and it is computed as a by-product of the solution of (5); its norm is bounded as a function of σ.

Theorem 4.2 (Exponential Form: Exact Penalty): The solution to (5), where D is defined by (13), is a vector with components

$$w_j^* = \frac{q_j \exp\left\{-(A^T y^*)_j\right\}}{\sum_{i=1}^k q_i \exp\left\{-(A^T y^*)_i\right\}}, \qquad j = 1, \ldots, k$$

where the parameter $y^* \in \partial(\sigma \|(Aw^* - b)^+\|_p)$.

Proof: The feasible set of (5) is nonempty, and the level sets $\{w | \mathcal{I}(w; q) \leq \mathcal{I}(w_0; q)\}$ are closed and bounded. Hence, the strict convexity of $\mathcal{I}$ implies that there exists a unique solution w^* to (5). Moreover, the constraint $\mathbf{1}^T w = 1$ is linear, so that, by [28, Theorem 14.6.1], there exist a vector $y^* \in \partial(\sigma \|(Aw^* - b)^+\|_p)$ and a scalar ζ^* such that (w^*, ζ^*, y^*) satisfies the first-order optimality conditions (14).

The form of the solution w^* can be derived in the same manner as (12). □

Note that the condition $y^* \in \partial(\sigma \|(Aw^* - b)^+\|_p)$ imposes an implicit bound on the magnitude of y^*. With the definition of the dual norm, the subdifferential of $\sigma \|(Aw - b)^+\|_p$ can be equivalently stated (see [28, Ch. 14]) as

$$\partial(\sigma \|(Aw - b)^+\|_p) = \{\, y | y^T (Aw - b) = \sigma \|(Aw - b)^+\|_p, \; 0 \leq y, \|y\|_d \leq \sigma \,\}. \tag{15}$$

The very last condition in (15) guarantees a bound on the norm of y^*. This is a critical part of our analysis: because a bound on the norm of y^* now exists (given by σ), we can derive an *a priori* bound on the exponential decay of the solution w^* (expression (12) shows the relationship between y^* and w^*). We use this property in Section V.

The exact penalty function formulation is exact in the following sense: for all penalty-parameter values greater than a certain threshold value, KKT points of (7) are also stationary points of its exact penalty-function formulation.

Theorem 4.3 (Exact Penalization): Let w^* be a solution of (7), with corresponding Lagrange multipliers ζ^* and y^* (see Theorem 2.2). Then for every $\sigma > \|y^*\|_d$, w^* is also a minimizer of (5) where D is defined in (13).

Proof: This result follows immediately from [28, Theorem 14.3.1]. □

C. Solution With Exact Penalization: Equality Constraints

Section IV-B discusses properties of exact penalization of inequality constraints on the moments $Aw \leq b$. For completeness, and because we will refer to them in Section V, we specialize Theorems 4.2 and 4.3 to the penalization of *equality* moment constraints $Aw = b$.

For equality constraints, a penalty needs to be applied to components of Aw that are positive *or* negative. In that case, the distortion function (13) is instead defined as

$$D(Aw - b) = \|Aw - b\|_p. \tag{16}$$

(Both the positive and negative parts of $Aw - b$ are considered.)

The following two corollaries parallel Theorems 4.2 and 4.3, and give properties of the solution of (5) when the distortion function is defined by (16). The proofs can be derived as a special case of the proofs for Theorems 4.2 and 4.3. The vital difference is that the solution of (5) (now with equality constraints) satisfies (14), but now $y^* \in \partial(\sigma \|Aw^* - b\|_p)$. This subdifferential can be equivalently stated (see [28, Ch. 14]) as

$$\partial(\sigma \|Aw - b\|_p) = \{\, y | y^T (Aw - b) = \sigma \|Aw - b\|_p, \|y\|_d \leq \sigma \,\}.$$

A bound on $\|y^*\|_d$ continues to hold, but nonnegativity of y^* is no longer required.

Corollary 4.4 (Exponential Form: Exact Penalty): The solution of (5), where D is defined by (16), is a vector with components

$$w_j^* = \frac{q_j \exp\left\{-(A^T y^*)_j\right\}}{\sum_{i=1}^k q_i \exp\left\{-(A^T y^*)_i\right\}}, \quad j = 1, \ldots, k$$

where the parameter $y^* \in \partial(\sigma \|Aw^* - b\|_p)$.

Corollary 4.5 (Exact Penalization): Let w^* be a solution of (7) with constraint $Aw = b$, and with corresponding Lagrange multipliers ζ^* and y^* (see Theorem 2.2). Then for every $\sigma > \|y^*\|_d$, w^* is also a minimizer of (5), where D is defined by (16).

D. The ℓ_1 Penalty Function

When the penalty function (13) is based on the ℓ_1 norm, its objective is discontinuous over sets of hyperplanes. However, a common technique is to reformulate it as an equivalent and smooth problem (see, for example, [22, Theorem 4.8] and [23, Sec. 4.2.3]). The ℓ_1 penalty function can be implemented so that it is computationally efficient, and it also allows us to further characterize the solution described in Theorem 4.2.

We introduce a pair of *elastic variables* $r, s \geq 0$, and rewrite (5) as

$$\begin{aligned} \underset{w,r,s}{\text{minimize}} \;& \mathcal{I}(w;q) + \sigma \mathbf{1}^T r \\ \text{subject to} \;& \mathbf{1}^T w = 1 \\ & Aw + r - s = b \\ & r, s \geq 0 \end{aligned} \tag{17}$$

where we use $D(Aw-b) = \|Aw-b\|_1$. The solution of (17) is a 5-tuple $(w^*, r^*, s^*, \zeta^*, y^*)$ that satisfies the first-order KKT conditions

$$\mathbf{1}^T w^* = 1 \tag{18a}$$

$$Aw^* + r^* - s^* = b \tag{18b}$$

$$\nabla_w \mathcal{I}(w^*;q) + \zeta^* \mathbf{1} + A^T y^* = 0 \tag{18c}$$

$$\min(r^*, \sigma\mathbf{1} - y^*) = 0 \tag{18d}$$

$$\min(s^*, y^*) = 0. \tag{18e}$$

The last two conditions (18d)–(18e) imply that their arguments are nonnegative, so that $0 \leq y^* \leq \sigma \mathbf{1}$. This pair of inequalities can be restated as

$$0 \leq y^*, \quad \|y^*\|_\infty \leq \sigma. \tag{19}$$

Note that the ℓ_1 and ℓ_∞ norms are duals of each other, so that (19) is equivalent to $0 \leq y^*, \|y^*\|_d \leq \sigma$ [see (15)].

Lemma 4.6 (Complementarity of r and s): Suppose that the 5-tuple $(w^*, r^*, s^*, \zeta^*, y^*)$ is a solution of (18) with $\sigma > 0$. Then r^* and s^* are componentwise complementary; that is, $r_i^* s_i^* = 0$, for $i = 1, \ldots, k$.

Proof: Set $z^r = \sigma\mathbf{1} - y^*$ and $z^s = y^*$. Then

$$z^r + y^* = \sigma \mathbf{1} > 0 \tag{20}$$

because $\sigma > 0$ by hypothesis. Note that (18d)–(18e) imply that

$$r_i^* z_i^r = s_i^* y_i^* = 0. \tag{21}$$

Now suppose that $r_i^* > 0$. Multiplying the ith component of (20) by r_i^* yields $r_i^* y_i^* > 0$. Hence $y_i^* > 0$, and from (21), we have $s_i^* = 0$. By analogous argument, $s_i^* > 0$ implies $r_i^* = 0$. Then $r_i^* s_i^* = 0$, as required. □

Theorem 4.7 (Exponential Form: ℓ_1 Penalty): Let σ be a positive constant. Suppose that $(w^*, r^*, s^*, \zeta^*, y^*)$ is a solution of (17). Let $\mathcal{A}, \mathcal{A}_+, \mathcal{A}_-$ be index sets such that

$$\begin{aligned} (Aw^*)_i &= b, \qquad i \in \mathcal{A} \\ (Aw^*)_i &> b, \qquad i \in \mathcal{A}_+ \\ (Aw^*)_i &< b, \qquad i \in \mathcal{A}_-. \end{aligned}$$

Then

$$w_j^* = \frac{u_j}{\sum_{j=1}^k u_j}$$

where

$$u_j = q_j \exp\left(-\sum_{i \in \mathcal{A}} a_{ij} y_i^* - \sigma \sum_{i \in \mathcal{A}_-} a_{ij} \right).$$

Proof: Because (17) and (5) are equivalent, w^* must have the form specified by (12).

Now we consider the values that each y_i^* may have. If $i \in \mathcal{A}_+$, then $(Aw^*)_i > b$, and (18b) together with (18d) implies that $0 \leq r_i^* < s_i^*$. By Lemma 4.6, we must have $r_i^* = 0$. Then from (18e) we deduce that $y_i^* = 0$. By analogous argument, $y_i^* = \sigma$ for $i \in \mathcal{A}_-$. For $i \in \mathcal{A}$, $(Aw^*)_i = b$, and by (18b), $r_i^* = s_i^*$. Lemma 4.6 then implies $r_i^* = s_i^* = 0$, and so we have from (18d)–(18e) that $\sigma \geq y_i^* \geq 0$. In summary, we may now write $\sum_{i=1}^k a_{ij} y_i^*$ as

$$a_j^T y^* = \sum_{i \in \mathcal{A}} a_{ij} y_i^* + \sigma \sum_{i \in \mathcal{A}_-} a_{ij} \tag{22}$$

for each $j = 1, \ldots, k$. Substituting (22) into (9), we see that w_j^* has the required form. □

V. Application to Statistical Supervised Learning

An estimation method can be practically and philosophically justified as a method of induction if it can be proved that, in some sense, the generated estimates converge to the "truth" (see Pierce [29] and Kneale [30]). For statistical estimation methods this criteria can be formalized as *consistency*. The results developed in Sections III and IV play a vital role in the proof of statistical consistency of the linear interpolation with maximum entropy (LIME) nonparametric statistical learning algorithm [31], [32].

LIME is a nonparametric neighborhood method that determines how to weight near-neighbors of a test point by solving a mean-constrained problem with a maximum entropy penalty on the weights. LIME has been shown to perform significantly better than other standard neighborhood methods in a situation of high bias [33], and to be a useful method for estimating pipeline integrity [34], estimating look-up tables for color management [35], and estimating custom color enhancements based on sample color transformations [36].

In particular, Corollaries 4.4 and 4.5 are needed to prove that the LIME method generates a sequence of estimates of a random variable that converge to its true expected value. In light of the exponential form given by Lemma 3.1, the LIME method can be interpreted as a data-adaptive exponential kernel. The results and proof presented here may also be useful for proving statistical consistency of other asymmetric kernel estimators.

In the application of supervised statistical learning methods it is assumed that each test point X can be associated with a neighborhood of k (out of a total of n) sample pairs (X_i, Y_i), $i = 1, \ldots, k$. Each X_i is a random feature vector (with m components), and each Y_i is a random associated scalar. The aim is to estimate the true value of Y that is associated with the test point X; the estimate is based on the n sample pairs (X_i, Y_i). The neighborhood size k might be fixed, might depend on the total number of neighbors n, or might otherwise be adaptive.

To prove the consistency of LIME, we make the following common statistical learning assumptions. The n sample pairs (X_i, Y_i) and the test pair (X, Y) are all drawn independently and are identically distributed with a joint distribution $P_{X,Y}$. Let k be a function of n, and gather the $k(n)$ nearest-neighbors $X_1, X_2, \ldots, X_{k(n)}$ of X into the columns of the $m \times k(n)$ matrix $\boldsymbol{X}_{k(n)}$. Similarly, we form the $k(n)$-vector $\boldsymbol{Y}_{k(n)}$ from the associated scalars $Y_1, Y_2, \ldots, Y_{k(n)}$. Let the ith component of $w_{k(n)}$ be the weight on the ith nearest neighbor.

Let $H(w)$ be the Shannon entropy function, so that $H(w_{k(n)}) = \mathcal{I}(w_{k(n)}; \frac{1}{k(n)}\mathbf{1})$. The LIME algorithm solves an optimization problem

similar to (5) that combines the relaxed moment constraint with the maximum-entropy criteria, and finds the vector of weights $w^*_{k(n)}$ that solve

$$\begin{aligned} &\underset{w\in\mathbb{R}^{k(n)}}{\text{minimize}} \; -H(w)+\sigma D\left(\boldsymbol{X}_{k(n)}w-X\right) \\ &\text{subject to } \mathbf{1}^T w=1, \quad w\geq 0. \end{aligned} \tag{23}$$

The function $D=\|\boldsymbol{X}_{k(n)}w-X\|_p$ measures the distortion in satisfying the moment constraint. The positive scalar parameter σ balances the tradeoff between maximizing entropy and satisfying the moment constraint. It is fixed in the LIME objective and might have been determined by cross-validation, by an estimate of signal-to-noise ratio, or it might simply be set arbitrarily large to focus the objective on minimizing the distortion D.

Lemma 3.1 implies that the LIME weight vector $w^*_{k(n)}$ has an exponential form, and we may therefore interpret the LIME weights as a data-adaptive kernel with exponential shape that tends to center around the test point X. Learning kernels are usually symmetric around the test point. Recent reviews of statistical learning methods, Hastie *et al.* [37] and Kulkarni *et al.* [38], discuss kernel estimators in more depth. The simplest neighborhood weighting is the Fix and Hodges k-nearest-neighbor algorithm [39], which assigns uniform weights $u_{k(n)}\stackrel{\text{def}}{=}\frac{1}{k(n)}\mathbf{1}$ to the $k(n)$ data points in a neighborhood around the test point X. The resulting estimate of Y is the equally weighted average of the sample outputs, given by $\hat{Y}=\boldsymbol{Y}^T_{k(n)}u_{k(n)}$. As is well known, such uniform weights maximize entropy when there are no additional constraints on the distribution.

The LIME approach balances satisfaction of the moment constraints and equalization of the weights, with a tradeoff specified by σ. At one extreme, with σ small, the LIME objective (23) focuses on maximizing entropy, and the LIME method is similar to the k-nearest-neighbor method. At the other extreme, with σ large, emphasis is placed on finding weights on the training samples that place the center of mass of the k-nearest neighbors close to the point X, so that the LIME weights (approximately) satisfy the constraint $\boldsymbol{X}_{k(n)}w_{k(n)}=X$. Note that, together with the requirements that $\mathbf{1}^T w_{k(n)}=1$ and $w_{k(n)}\geq 0$, the constraint $\boldsymbol{X}_{k(n)}w_{k(n)}=X$ could be infeasible. We conjecture, based on simulation results in [33], that the distortion term D in the objective of (23) helps to reduce estimation bias. On the other hand, the estimation variance is lowered by requiring the weights to satisfy the maximum entropy objective (and thus be close to uniform).

We show that the LIME estimates $\hat{Y}_{k(n)}\stackrel{\text{def}}{=}\boldsymbol{Y}^T_{k(n)}w^*_{k(n)}$ are consistent in the following standard sense: if Y satisfies $E(Y^s)<\infty$ for $s>1$, then $E(\hat{Y}_{k(n)}|X)\to E(Y|X)$ in L^s as $n\to\infty, k(n)\to\infty$, and $k(n)/n\to 0$, where the expectations are over the training set and test point. (See, for example, Devroye, *et al.* [40, Ch. 6].) The proof relies on the following result (stated as a corollary by Stone [41]), which is sufficient to show consistency for nonparametric neighborhood estimators:

Theorem 5.1 (Stone [41, Corollary 2]): Let $u_{k(n)}$ be a consistent sequence of probability weights. Suppose that $w_{k(n)}$ is a sequence of probability weights such that $w_{k(n)}\leq M u_{k(n)}$ for some constant $M>1$ and for all n. Then $w_{k(n)}$ is consistent.

The proof of LIME's consistency is nontrivial because of its adaptive, asymmetric kernel. With Stone's result, we can prove LIME's consistency using Corollary 4.4.

Theorem 5.2 (LIME Consistency): Let $w^*_{k(n)}$ be the pmf that solves the LIME minimization problem (23) for X and its $k(n)$ near-neighbors $\boldsymbol{X}_{k(n)}$. Suppose that all training and test feature vectors are random variables drawn iid from a distribution with bound $\|X\|_p\leq\alpha$, for some constant $\alpha>0$. Then the sequence of weights $w^*_{k(n)}$ is consistent as $n\to\infty, k(n)\to\infty$, and $k(n)/n\to 0$.

Proof: Let $u_{k(n)}$ be the k-nearest-neighbor uniform weights so that $u_{k(n)}=\frac{1}{k(n)}\mathbf{1}$. As proved by Stone ([41, Corollary 3]), the k-nearest-neighbor uniform weights (for near-neighbors ranked by ℓ_p distance) are a consistent sequence of probability weights under the standard assumptions $n\to\infty, k(n)\to\infty$, and $k(n)/n\to 0$.

By construction, $\mathbf{1}^T w^*_{k(n)}=1$ and $w^*_{k(n)}\in[0,1]^{k(n)}$ for all n, and this $w^*_{k(n)}$ is a sequence of probability weights. It remains to establish that there exists a finite $M>1$ such that $w^*_{k(n)}\leq M u_{k(n)}$ for all $n\geq 1$, so that under Theorem 5.1, $w^*_{k(n)}$ is consistent.

Because q is uniform ($q_i=1/k(n)$), it follows from Corollary 4.4 that the jth LIME weight is given by

$$(w^*_{k(n)})_j=\frac{\frac{1}{k(n)}\exp\left\{-\left(\boldsymbol{X}^T_{k(n)}y^*\right)_j\right\}}{\sum_{i=1}^{k(n)}\frac{1}{k(n)}\exp\left\{-\left(\boldsymbol{X}^T_{k(n)}y^*\right)_i\right\}} \tag{24}$$

for $j=1,\ldots,k(n)$, where y^* is the vector of Lagrange multipliers of (23).

To show that $w^*_{k(n)}\leq M u_{k(n)}$, we show that $\|w^*_{k(n)}\|_\infty\leq M/k(n)$, or equivalently, that $k(n)\|w^*_{k(n)}\|_\infty\leq M$ for some finite M. From Corollary 4.4, $\|y^*\|_d\leq\sigma$. Hence, the boundedness of $\|x\|_p$ and the Hölder inequality imply that

$$|(\boldsymbol{X}^T_{k(n)}y^*)_j|=|X_j^T y^*|\leq\|X_j\|_p\|y^*\|_d\leq\alpha\|y^*\|_d$$

for some positive constant α and for all $j=1,\ldots,n$. Therefore,

$$\begin{aligned} k(n)\|w^*_{k(n)}\|_\infty &\leq k(n)\max_j\frac{\frac{1}{k(n)}\exp\left\{-\left(\boldsymbol{X}^T_{k(n)}y^*\right)_j\right\}}{\frac{1}{k(n)}\sum_{i=1}^{k(n)}\exp\left\{-\left(\boldsymbol{X}^T_{k(n)}y^*\right)_i\right\}} \\ &\leq\frac{\max_j\exp\left\{-\left(\boldsymbol{X}^T_{k(n)}y^*\right)_j\right\}}{\min_j\exp\left\{-\left(\boldsymbol{X}^T_{k(n)}y^*\right)_j\right\}} \\ &\leq\exp\left(2\max_j\left(\boldsymbol{X}^T_{k(n)}y^*\right)_j\right) \\ &\leq\exp(2\alpha\|y^*\|_d) \\ &\leq\exp(2\alpha\sigma)\equiv M. \end{aligned}$$

Because α and σ are both positive, $M>1$, as required by Theorem 5.1. □

Note that Theorem 5.2 guarantees the consistency of the LIME weights for any positive value of the penalty parameter σ, regardless of the feasibility of the moment constraint $\boldsymbol{X}_{k(n)}w_{k(n)}=X$. If the constraint is feasible (i.e., it admits probability weights that satisfy that equation), then Corollary 4.5 asserts that for σ large enough, the constraint will be satisfied exactly.

VI. Extensions

In this section, we discuss two extensions, first the continuous version of problem (4), and then the case in which the relative-entropy arguments are reversed.

A. Continuous Density Functions

We have presented results for discrete minimizers of (5), but the mathematical development and results are analogous when the problem is defined for continuous distributions q and w. We restate the problem

and required assumptions for the continuous case, and explain why the development is parallel.

Suppose Q and W are probability measures in Euclidean space $\mathbb{R}^n$ that are absolutely continuous with respect to Lebesgue measure, such that the respective Radon-Nikodym derivatives (density functions) q and w exist. The relative entropy between the densities w and q is now given by $\mathcal{I}(w;q) = \int_{\mathbb{R}^n} w(x) \log w(x)/q(x) dx$. Then the discrete problem (5) has the following continuous analog:

$$\begin{aligned} &\underset{w}{\text{minimize}} \ \mathcal{I}(w;q) + \sigma D(e) \\ &\text{subject to} \int_{\mathbb{R}^n} w(x)\, dx = 1 \end{aligned}$$

where e is an m-vector of errors of the m moment constraints such that $e_i = \int_{\mathbb{R}^n} w(x) f_i(x)\, dx - b_i$ for given continuous functions $f_i(x)$ and scalars $b_i, i = 1, \ldots, m$.

The continuous problem above differs from (5) in that the relative entropy is now defined over a continuous domain. Notably, the distortion D remains a function of a set of discrete errors corresponding to the error for each constraint; it could be defined by either (13) or (16), respectively, depending on whether the moment constraints are inequalities or equalities. Then, given the modified assumptions for the continuous case, there are no significant differences in terms of the optimization problem.

Analogous forms of the results given in this paper all hold based on the same mathematical development. For example, application of Lemma 3.1 applied to the continuous case yields

$$w^*(x) = \frac{q(x) \exp\left(-\sum_{i=1}^m f_i(x) y_i^*\right)}{\int_{\mathbb{R}^n} q(x) \exp\left(-\sum_{i=1}^m f_i(x) y_i^*\right)\, dx}.$$

The continuous analogues to Theorems 4.2 and 4.3 imply that $\|y^*\|$ is bounded.

B. Reversed Arguments in Relative Entropy

A minimum relative entropy problem is usually formulated as "Given q, find w that minimizes $\mathcal{I}(w;q)$ subject to some constraints on w." However, one might be interested in reversing the arguments of the relative entropy and saying, "Given w, find q that minimizes $\mathcal{I}(w;q)$ subject to some constraints on q." Čencov discusses the relationship of this variant to the maximum-likelihood problem and establishes conditions for the existence of a minimizer [42, pp. 115, 323–334]. We note that the discrete version of the variation has the following source-coding interpretation: "Given w, find the probability mass function q that results in an efficient code that, on average, requires the fewest additional bits $\mathcal{I}(w;q)$ to code identically and independently distributed random variables drawn from source w, subject to some constraints on q" (see [7, Theorem 5.4.3].

Consider the minimum relative entropy problem that reverses the objective function arguments of (7)

$$\begin{aligned} &\underset{q}{\text{minimize}} \ \mathcal{I}(w;q) \\ &\text{subject to} \ \mathbf{1}^T q = 1 \\ &\qquad\qquad Aq \le b. \end{aligned} \tag{25}$$

A similar approach to that in Theorem 2.2 is used to derive an expression for each component of the minimizing probability mass function of (25)

$$q_j^* = \frac{w_j}{\zeta^* + (A^T y^*)_j}, \qquad j = 1, \ldots, k. \tag{26}$$

Analogous to (10), ζ^* and y^* are the Lagrange multipliers associated with the first constraint and the set of constraints of $Aq \le b$ of (25). Notably, (26) shows that the exponential form of the minimizer is lost when the arguments are reversed.

The reverse-argument problem specified in (25) can also be reformulated with soft constraints as a penalty problem, and the approach given earlier in this paper can be used again to show that the result will be unique and have the form specified by (26).

In general, the logic of Lemma 3.1 is applicable to the broad set of hard-constraint problems that have been relaxed by using a penalty formulation in which a strictly convex function is minimized with a convex penalty.

ACKNOWLEDGMENT

The authors would like to thank L. Lorne Campbell, Andrew B. Nobel, Santosh Srivastava, and an anonymous referee for their careful readings of this paper and its revisions. The readers' valuable comments and suggestions helped to clarify the description and led to several extensions and many changes. Santosh Srivastava's exceptionally close reading at the last moment led to many detailed corrections. They are also grateful to Robert M. Gray and Richard A. Olshen for their knowledgeable input.

REFERENCES

[1] J. Shore and R. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inf. Theory*, vol. IT-33, no. 1, pp. 26–37, Jan. 1980.

[2] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959.

[3] I. Csiszár, "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems," *Ann. Statist.*, vol. 19, no. 4, pp. 2032–2066, 1991.

[4] R. Byrd, M. E. Hribar, and J. Nocedal, "An interior point method for large scale nonlinear programming," *SIAM J. Optim.*, vol. 9, no. 4, pp. 877–900, 1999.

[5] D. Shanno and R. Vanderbei, "An interior-point algorithm for nonconvex nonlinear programming," *Comput. Optim. Applicat.*, vol. 13, pp. 231–252, 1999.

[6] E. D. Andersen. (2003) Mosek. [Online]. Available: http://www.mosek.com/documentation.html

[7] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[8] R. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[9] E. T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, vol. 106, pp. 620–630, 1957.

[10] J. P. Burg, "Maximum entropy spectral analysis," in *Proc. 37th Annu. Int. Meet. Society of Exploratory Geophysics*, Oklahoma City, OK, 1967.

[11] I. Csiszár, "Sanov property, generalized I-projection and a conditional limit theorem," *Ann. Prob.*, vol. 12, no. 3, pp. 768–793, 1984.

[12] P. A. Mello and N. Kumar, *Quantum Transport in Mesoscopic Systems: Complexity and Statistical Fluctuations: A Maximum-Entropy Viewpoint (Mesoscopic Physics and Nanotechnology)*. Cambridge, U.K.: Oxford Univ. Press, 2004.

[13] W. T. G. Jr and P. W. Milonni, *Physics and Probability, Essays in Honor of E. T. Jaynes*. Cambridge, U.K.: Cambridge Univ. Press, 1993.

[14] J. Navaza, "The use of nonlocal constraints in maximum-entropy electron density reconstruction," *Acta Crystallographica*, pp. 212–223, 1986.

[15] T. Jaakkola, M. Meila, and T. Jebara, "Maximum entropy discrimination," *Adv. Neural Inf. Process. Syst.*, vol. 12, 1999.

[16] J.-F. Bercher, G. LeBesnerais, and G. Demoment, "The maximum entropy on the mean method, noise, and sensitivity," *Maximum Entropy and Bayesian Methods*, pp. 223–232, 1996.

[17] H. Gzyl and Y. Velasquez, "Maxentropic interpolation by cubic splines with possibly noisy data," in *Proc. Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 20th Int. Workshop*, Gif-sur-Yvette, France, Jul. 2001, pp. 216–228.

[18] J. O. Katz and D. McCormick, *The Encyclopedia of Trading Strategies*. New York: McGrawTrade, 2000.

[19] L. L. Campbell, "Minimum cross-entropy estimation with inaccurate side information," *IEEE Trans. Inf.. Theory*, vol. 45, no. 7, pp. 2650–2652, Nov. 1999.

[20] G. L. Besenerais, J.-F. Bercher, and G. Demoment, "A new look at entropy for solving linear inverse problems," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1565–1577, Jul. 1999.
[21] I. Csiszár, F. Gamboa, and E. Gassiat, "MEM pixel correlated solution for generalized moment and interpolation problems," *IEEE Trans. Inf. Theory*, vol. 45, no. 7, pp. 2253–2270, Nov. 1999.
[22] S.-P. Han and O. L. Mangasarian, "Exact penalty functions in nonlinear programming," *Math. Prog.*, vol. 17, pp. 251–269, 1979.
[23] P. E. Gill, W. Murray, and M. H. Wright, *Practical Optimization*. San Diego, CA: Academic, 1981.
[24] O. L. Mangasarian, *Nonlinear Programming*. ser. Classics in Applied Mathematics, G. Golub, Ed. Philadelphia, PA: SIAM, 1994, vol. 10. Originally published: New York, McGraw-Hill, 1969.
[25] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York: Cambridge Univ. Press, 2004.
[26] T. Pietrzykowski, "An exact potential method for constrained maxima," *SIAM J. Numer. Anal.*, vol. 6, pp. 262–304, 1969.
[27] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*. New York: Academic, 1982.
[28] R. Fletcher, *Practical Methods of Optimization*, 2nd ed. New York: Wiley, 1987.
[29] C. S. Peirce, *The Philosophy of Peirce: Selected Writings*. London, U.K.: Jarrolds, 1956.
[30] W. Kneale, *Probability and Induction*. Oxford, U.K.: Clarendon, 1949.
[31] M. Gupta, "An Information Theory Approach to Supervised Learning," Ph.D. dissertation, Stanford Univ., Stanford, CA, 2003.
[32] M. R. Gupta, R. M. Gray, and R. A. Olshen, "Nonparametric supervised learning with linear interpolation and maximum entropy," *IEEE Trans. Pattern Anal. Machine Intell.*. Available [Online] at ee.washington.edu/research/guptalab/publications.html, to be published.
[33] M. R. Gupta and R. M. Gray, "Reducing bias in supervised learning," in *Proc. IEEE Workshop on Statistical Signal Processing*, St. Louis, MO, Sep. 2003, pp. 482–485.
[34] D. O'Brien, M. Gupta, and R. M. Gray, "Analysis and classification of internal pipeline images," in *Proc. Int. Conf. Image Proc.*, Barcelona, Spain, Sep. 2003.
[35] M. R. Gupta, "Inverting color transforms," in *Proc. SPIE Conf. Computational Imaging*, vol. 5299, San Jose, CA, Jan. 2004, pp. 83–93.
[36] M. R. Gupta, S. Upton, and J. Bowen, "Simulating the effect of illumination using color transformations," in *Proc. SPIE Conf. Computational Imaging*, vol. 5674, San Jose, CA, Jan. 2005, pp. 248–258.
[37] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York: Springer-Verlag, 2001.
[38] S. Kulkarni, G. Lugosi, and S. S. Venkatesh, "Learning pattern classification—A survey," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2178–2206, Oct. 1998.
[39] E. Fix and J. L. Hodges, "Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties," uSAF School of Aviation Medicine, TX, Tech. Rep. 4, 1951.
[40] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.
[41] C. Stone, "Consistent nonparametric regression," *Ann. Statist.*, vol. 5, no. 4, pp. 595–645, 1977.
[42] N. N. Čencov, *Statistical Decision Rules and Optimal Inference*. Providence, RI: Amer. Math. Soc., 1982. Translated from Russian by the Israel Program for Scientific Translations. Translation edited by L. J. Leifman.

Error Exponents for Finite-Hypothesis Channel Identification

Patrick Mitran, *Student Member, IEEE,* and Aleksandar Kavčić, *Member, IEEE*

***Abstract*—We consider the problem of designing optimal probing signals for finite-hypothesis testing. Equivalently, we cast the problem as the design of optimal channel input sequences for identifying a discrete channel under observation from a finite set of known channels. The optimality criterion that we employ is the exponent of the Bayesian probability of error. In our study, we consider a feedforward scenario where there is no feedback from the channel output to the signal selector at the channel input and a feedback scenario where the past channel outputs are revealed to the signal selector.**

In the feedforward scenario, only the type of the input sequence matters and our main result is an expression for the error exponent in terms of the limiting distribution of the input sequence. In the feedback case, we show that when discriminating between two channels, the optimal scheme in the first scenario is simultaneously the optimal time-invariant Markov feedback policy of *any* order.

***Index Terms*—Bayesian hypothesis testing, classification, detection theory, Chernoff's theorem, error exponent, feedback, method of types, sequential detection, signal selection, waveform selection.**

I. Introduction

In traditional hypothesis testing, we are given a set of hypotheses $\mathcal{H}$. For each hypothesis $h \in \mathcal{H}$, we know the probability law for an observable variable Y, i.e., we know $P_Y^h[y] \triangleq P_{Y \mid H}[y \mid h]$. We make n observations $y_1^n = [y_1, y_2, \ldots, y_n]$ and based on these observations, we need to infer the hypothesis $h \in \mathcal{H}$. This is a well-known problem with well-known solutions in the context of Bayesian and Neyman–Pearson decision making [14]. Furthermore, the type-II error exponent (for Neyman–Pearson) or average error exponent (for Bayesian) detection is well known and may be derived by the method of types [5], [4] or large deviation theory [8].

Let us now suppose that the observable variable Y is obtained as the response to an input variable X, which we control. In particular, each hypothesis h, drawn from a finite set of hypotheses $\mathcal{H}$, may be viewed as a memoryless channel $P_{Y \mid X}^h[y \mid x] \triangleq P_{Y \mid X,H}[y \mid x, h]$. We refer to this type of problem as a *finite-hypothesis channel identification* or *channel detection* problem. The objective is to choose a set of input signals $x_1^n = [x_1, x_2, \ldots, x_n]$ according to some policy. We will consider two broad classes of policies.

Open-Loop Policies: We transmit these n signals x_1^n and only after all signals are transmitted, we observe the n responses $y_1^n = [y_1, \ldots, y_n]$. We make a decision on $h \in \mathcal{H}$ after we observe all outputs y_1^n.

Feedback Policies: This case may be described as follows. At time $t = 1$, an input x_1 is chosen according to some policy and sent over the channel. Based on the observation of the response output y_1, a new input x_2 is chosen. The signal x_2 is transmitted and a response y_2 is observed. Based on knowledge of x_1, x_2, y_1, and y_2, an input x_3 is

Manuscript received June 4, 2005; revised September 13, 2005. The materiral in this correspondence was presented in part at the IEEE International Symposium on Information Theory, Chicago, IL, June/July 2004, and the International Symposium on Information Theory and Its Applications, Parma, Italy, October 2004.

The authors are with the Division of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138 USA (e-mail: mitran@deas.harvard.edu; kavcic@hrl.harvard.edu).

Communicated by X. Wang, Associate Editor for Detection and Estimation.
Digital Object Identifier 10.1109/TIT.2005.860468