Finding the best mismatched detector for channel coding and hypothesis testing

Emmanuel Abbe
EECS and LIDS, MIT
eabbe@mit.edu

Muriel Médard
EECS and LIDS, MIT
medard@mit.edu

Sean Meyn
ECE and CSL, UIUC
meyn@uiuc.edu

Lizhong Zheng
EECS and LIDS, MIT
lizhong@mit.edu

***Abstract*— The mismatched-channel formulation is generalized to obtain simplified algorithms for computation of capacity bounds and improved signal constellation designs. The following issues are addressed:**

(i) For a given finite dimensional family of linear detectors, how can we compute the best in this class to maximize the reliably received rate? That is, what is the *best* mismatched detector in a given class?

(ii) For computation of the best detector, a new algorithm is proposed based on a stochastic approximation implementation of the Newton-Raphson method.

(iii) The geometric setting provides a unified treatment of channel coding and robust/adaptive hypothesis testing.

I. Introduction

Consider a discrete memoryless channel (DMC) with input sequence $\boldsymbol{X}$ on X, output sequence $\boldsymbol{Y}$ on Y, with $\mathsf{Y} = \mathbb{C}$, and X a subset of $\mathbb{R}$ or $\mathbb{C}$. It is assumed that a channel density exists,

$$\mathsf{P}\{Y(t) \in dy \mid X(t) \in x\} = P_{Y|X}(y \mid x)dy,$$

$x \in \mathsf{X}$, $y \in \mathsf{Y}$. We consider in this paper the reliably received transmission rate over a non-coherent DMC using a mismatched detector, following the work of Csiszar, Kaplan, Lapidoth, Merhav, Narayan and Shamai [1], [2], [3]. The basic problem statement can be described as follows: We are given a codebook consisting of e^{nR} codewords of length n, where R is the input rate in *nats*. The ith codeword is denoted $\{X_t^i : 1 \leq t \leq n\}$. We let Γ_n^i denote the associated empirical distribution for the joint input-output process,

$$\Gamma_n^i = n^{-1} \sum_{t=1}^{n} \delta_{X_t^i, Y_t}. \tag{1}$$

Suppose that a function of two variables, $F \colon \mathsf{X} \times \mathsf{Y} \to \mathbb{R}$ is given. A linear detector based on this function is described as follows: The codeword i^* is chosen as the maximizer,

$$i^* = \arg\max_i \langle \Gamma_n^i, F \rangle. \tag{2}$$

For a random codebook this is the Maximum Likelihood decoder when F is the log likelihood ratio (LLR). In the mismatched detector the function F is arbitrary.

There is ample motivation for considering general F: First, this can be interpreted as the typical situation in which side information at the receiver is imperfect. Also, in some cases the LLR is difficult to evaluate so that simpler functions can be used to balance the tradeoff between complexity and performance.

In Merhav et. al. [1] and Csiszar [2] the following bound on the reliably received rate is obtained, known as the *generalized mutual information*: $I_{\text{GMI}}(P_X; F) :=$

$$\min\{D(\Gamma \| P_X \otimes P_Y) : \langle \Gamma, F \rangle = \gamma,\ \Gamma_2 = P_Y\}, \tag{3}$$

where Γ_i denotes the ith marginal of the distribution Γ on $\mathsf{X} \times \mathsf{Y}$, and $\gamma = \langle P_{XY}, F \rangle$. Under certain conditions the generalized mutual information coincides with the reliably received rate using (2).

In this paper we provide a geometric derivation of this result based on Sanov's theorem. This approach is related to Csiszar's *Information Geometry* [3], [4], [5], which is itself a close cousin of the geometry associated with Large Deviations Theory [6], [7], [8], [9], [10], [11].

Based on this geometry we answer the following question: Given a basis $\{\psi_j : 1 \leq j \leq d\}$ of functions on the product space $\mathsf{X} \times \mathsf{Y}$, and the resulting family of detectors based on the collection of functions $\{F_\alpha = \sum_1^d \alpha_i \psi_i\}$, how can we obtain the best reliably received rate over this class? To compute α^* we construct a new recursive algorithm that is a stochastic approximation implementation of the Newton Raphson method. The algorithm can be run at the receiver *blindly* - without knowledge of $\boldsymbol{X}$ or the channel statistics.

The geometry surveyed here is specialized to the low SNR setting in the companion paper [12].

II. Geometry in information theory

We begin with the following general setting: $\boldsymbol{Z}$ denotes an i.i.d. sequence taking values in a compact set Z, and the empirical distributions $\{\Gamma_n : n \geq 1\}$ are defined as the sequence of discrete probability measure on $\mathcal{B}(\mathsf{Z})$,

$$\Gamma_n(A) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{I}\{Z_k \in A\}, \qquad A \in \mathcal{B}(\mathsf{Z}). \tag{4}$$

For a given probability measure π and constant $\beta > 0$ we denote the divergence sets $\mathcal{Q}_\beta(\pi) = \{\Gamma : D(\Gamma \| \pi) < \beta\}$ and $\mathcal{Q}_\beta^+(\pi) = \{\Gamma : D(\Gamma \| \pi) \leq \beta\}$.

The logarithmic moment generating function (log-MGF) for a given probability measure π is defined for any bounded measurable function G via

$$\Lambda(G) = \log\Big(\int e^{G(z)} \pi(dz)\Big). \tag{5}$$

We write Λ_π when we wish to emphasize the particular probability used in this definition.

The geometry emphasized in this paper is based on Proposition 2.1 that expresses divergence as the convex dual of the log-MGF. For a proof see [6].

Proposition 2.1: The relative entropy $D(\pi^1 \| \pi^0)$ is the solution to the convex program (6), where the supremum is over all bounded continuous functions on Z:

$$\sup \ -\Lambda_{\pi_0}(F) \quad \text{subject to} \ \ \pi^1(F) \geq 0\,. \tag{6}$$

It is also the solution to the following unconstrained convex program,

$$D(\pi^1 \| \pi^0) = \sup\{\pi^1(F) - \Lambda_{\pi_0}(F)\} \tag{7}$$

That is, the relative entropy is the convex dual of Λ_{π_0}.

We conclude this section with a brief sketch of the proof to set the stage for the characterizations that follow.

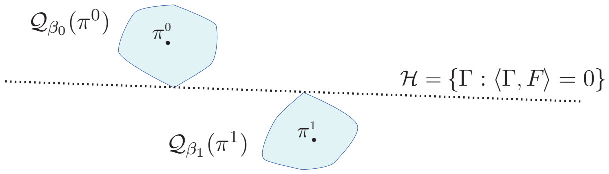
$\mathcal{Q}_{\beta_0}(\pi^0)$
π^0
$\mathcal{H} = \{\Gamma : \langle \Gamma, F \rangle = 0\}$
π^1
$\mathcal{Q}_{\beta_1}(\pi^1)$

Fig. 1. The entropy neighborhoods $\mathcal{Q}_{\beta_1}(\pi^1)$ and $\mathcal{Q}_{\beta_0}(\pi^0)$ lie on opposite sides of the hyperplane $\mathcal{H}$.

For a given function F we denote $\mathcal{H} = \{\Gamma : \langle \Gamma, F \rangle = 0\}$. We interpret $\mathcal{H}$ as a hyperplane, and consider the two 'half spaces', $\mathcal{H}_- := \{\Gamma \in \mathcal{M} : \langle \Gamma, F \rangle \leq 0\}$ and $\mathcal{H}_+ := \{\Gamma \in \mathcal{M} : \langle \Gamma, F \rangle \geq 0\}$. Suppose that $\mathcal{H}$ separates π^0 and π^1 as shown in Figure 1. Assuming also that F is feasible for (6), we must have $\pi^1(F) \geq 0$ and $\pi^0(F) \leq 0$. That is, $\pi^1 \in \mathcal{H}_+$ and $\pi^0 \in \mathcal{H}_-$. Evidently $D(\pi^1 \| \pi^0) \geq \beta_0$, where $\mathcal{Q}_{\beta_0}(\pi^0)$ is the largest divergence set contained in $\mathcal{H}_-$.

If Γ^{0*} lies on the intersection of the hyperplane and the divergence set shown, $\Gamma^{0*} \in \mathcal{H} \cap \mathcal{Q}_{\beta_0}(\pi^0)$, then by definition $D(\Gamma^{0*} \| \pi^0) = \beta_0$. Moreover, Sanov's Theorem (in the form of Chernoff's bound) gives $D(\Gamma^{0*} \| \pi^0) \geq -\Lambda(\theta_0 F)$ for any $\theta_0 \geq 0$. Taking $\theta_0 = 1$ we obtain $D(\Gamma^{0*} \| \pi^0) \geq -\Lambda(F)$. This shows that the value of (6) is a lower bound on the divergence.

To see that this lower bound can be approximately arbitrarily closely, consider a continuous approximation to the log likelihood ratio $L = \log(d\pi^1/d\pi^0) - D(\pi^1 \| \pi^0)$.

A. Hypothesis testing

In a similar fashion we can obtain a solution to the Neyman-Pearson hypothesis testing problem in terms of a supremum over linear tests.

Consider the binary hypothesis testing problem based on a finite number of observations from a sequence $\boldsymbol{Z} = \{Z_t : t = 1, \ldots\}$, taking values in the set X. It is assumed that, conditioned on either of the hypotheses H_0 or H_1, these observations are independent and identically distributed (i.i.d.). The marginal probability distribution on X is denoted π^j under hypothesis H_j for $j = 0, 1$. The goal is to classify a given set of observations into one of the two hypotheses.

For a given $n \geq 1$, suppose that a decision test ϕ_n is constructed based on the finite set of measurements $\{Z_1, \ldots, Z_n\}$. This may be expressed as the characteristic function of a subset $A_1^n \subset \mathsf{X}^n$. The test declares that hypothesis H_1 is true if $\phi_n = 1$, or equivalently, $(Z_1, Z_2, \ldots, Z_n) \in A_1^n$.

The performance of a *sequence* of tests $\boldsymbol{\phi} := \{\phi_n : n \geq 1\}$ is reflected in the error exponents for the type-I error probability and type-II error probability, defined respectively by,

$$\begin{aligned} J_{\phi} &:= -\liminf_{n \to \infty} \frac{1}{n} \log(\mathsf{P}_{\pi^0}\{\phi_n(Z_1, \ldots, Z_n) = 1\}), \\ I_{\phi} &:= -\liminf_{n \to \infty} \frac{1}{n} \log(\mathsf{P}_{\pi^1}\{\phi_n(Z_1, \ldots, Z_n) = 0\}), \end{aligned} \tag{8}$$

The asymptotic N-P criterion of Hoeffding [13] is described as follows: For a given constant bound $R \geq 0$ on the false-alarm exponent, an optimal test is the solution to,

$$\beta^* = \sup\{I_{\phi} : \ \textit{subject to} \ J_{\phi} \geq R\}\,, \tag{9}$$

where the supremum is over all test sequences $\boldsymbol{\phi}$ (see also Csiszár et. al. [14], [15].)

It is known that the optimal value of the exponent I_{ϕ} is described as the solution to a convex program involving relative entropy [7], and that the solution leads to optimal tests defined in terms of the empirical distributions. The form of an optimal test is as follows:

$$\phi_N = \mathbb{I}\{\Gamma_N \in \mathcal{A}\}. \tag{10}$$

Under general conditions on the set $\mathcal{A}$, we can conclude from Sanov's Theorem that the error exponents are given by, $I_{\phi} = \inf\{D(\Gamma \| \pi^1) : \Gamma \in \mathcal{A}^c\}$ and $J_{\phi} = \inf\{D(\Gamma \| \pi^0) : \Gamma \in \mathcal{A}\}$.

One choice of $\mathcal{A}$ is the half-space $\mathcal{A} = \{\Gamma : \Gamma(F) \geq 0\}$ for a given function F. On optimizing over all F that give a feasible test we obtain a new characterization of the solution to the Neyman-Pearson problem. The proof of Proposition 2.2 is similar to the proof of Proposition 2.1, based on Figure 1.

Proposition 2.2: For a given $R > 0$ there exists $\varrho > 0$ such that the solution β^* to the Neyman-Pearson hypothesis testing problem is the solution to the convex program,

$$\sup \quad -\Lambda_{\pi^1}(-\varrho F) - (\Lambda_{\pi_0}(F) + R)\varrho, \tag{11}$$

where the supremum is over continuous and bounded functions. The value of (11) is achieved with the function $F^* = -\kappa + t\log(d\pi^1/d\pi^0)$ with t and κ constant, $(t = (1+\varrho)^{-1})$. □

B. Capacity

Using a random code book with distribution P_X combined with maximum likelihood decoding, the reliably received rate is the mutual information $I(X;Y) = D(P_{XY} \| P_X \otimes P_Y)$. Hence, the mutual information is the solution to the convex program (6) with $\pi^1 = P_{XY}$ and $\pi^0 = P_X \otimes P_Y$. The function F appearing in (6) is a function of two variables since π^0 and π^1 are joint distributions on $\mathsf{X} \times \mathsf{Y}$.

III. A GEOMETRIC VIEW OF THE MISMATCHED CHANNEL

In this section we consider the reliably received rate based on a given linear test.

Suppose that F is given with $\gamma := \langle P_{XY}, F\rangle > 0$ and $\langle P_X \otimes P_Y, F\rangle < \gamma$. Sanov's Theorem implies a lower bound on the pair-wise probability of error using the detector (2), and applying a union bound we obtain the following bound on the reliably received rate:

$$I_{\text{LDP}}(P_X; F) = \min\{D(\Gamma \| P_X \otimes P_Y) : \langle \Gamma, F\rangle = \gamma\}. \qquad (12)$$

We have $I_{\text{LDP}}(P_X; F) \leq I_{\text{GMI}}(P_X; F)$ since the typicality constraint $\Gamma_2 = P_Y$ is absent in (12).

For the given function F and $y \in \mathsf{Y}$ we define $F_y \colon \mathsf{X} \to \mathbb{R}$ by $F_y(x) = F(x,y)$, and for any $\theta \in \mathbb{R}$ the log MGF is denoted,

$$\Lambda_{P_X}(\theta F_y) := \log \int e^{\theta F(x,y)} P_X(dx)\,. \qquad (13)$$

Proposition 3.1: For a given function $F \colon \mathsf{X} \times \mathsf{Y} \to \mathbb{R}$, define the new function,

$$F^\circ(x,y) = \theta^* F(x,y) - \Lambda_{P_X}(\theta^* F_y), \qquad (14)$$

where θ^* is chosen so that the following identity holds,

$$\int e^{\theta^* F(x,y) - \Lambda_{P_X}(\theta^* F_y)} F(x,y) P_X(dx) P_Y(dy) = \gamma.$$

Then, the generalized mutual information $I_{\text{GMI}}(P_X; F)$ can be expressed in the following three equivalent forms:

(a) $\sup\limits_{G(y)} I_{\text{LDP}}(P_X; F + G)$,
(b) $I_{\text{LDP}}(P_X; F^\circ)$,
(c) $\theta^*\gamma - \int \Lambda_{P_X}(\theta^* F_y) P_Y(dy)$.

Proof: The proof of (a) will be contained in the full paper.

To prove (b) and (c) we construct a Lagrangian relaxation of the convex program (3) that defines $I_{\text{GMI}}(P_X; F)$. The Lagrangian is denoted $\mathcal{L}(\Gamma) =$

$$D(\Gamma \| P_X \otimes P_Y) - \theta\langle \Gamma, F - \gamma\rangle \\ - \theta_0(\langle \Gamma, 1\rangle - 1) - \langle \Gamma_2 - P_Y, G\rangle$$

where θ is the Lagrange multiplier corresponding to the mean constraint, θ_0 corresponds to to the constraint that Γ has mass one, and the function G corresponds to the typicality constraint, $\Gamma_2 = P_Y$. The dual functional is defined as the infimum, $\Psi(\theta_0, \theta, G) = \inf\limits_{\Gamma \geq 0} \mathcal{L}(\Gamma)$. A bit of calculus gives the optimizer $\Gamma^*(dx, dy) = e^{\theta F(x,y) - \Pi + G(y)} P_X(dx) P_Y(dy)$, where $\Pi \in \mathbb{R}$ is a constant. The next task is to understand the Lagrange multipliers.

To find G we simply integrate over x to obtain,

$$P_Y(dy) = \int_{x \in \mathsf{X}} \Gamma^*(dx, dy) \\ = \Big(\int_{x \in \mathsf{X}} e^{\theta F(x,y)} P_X(dx)\Big) e^{-\Pi + G(y)} P_Y(dy)$$

Hence we have for a.e. y $[P_Y]$,

$$e^{-\Pi + G(y)} = \Big(\int_{x \in \mathsf{X}} e^{\theta F(x,y)} P_X(dx)\Big)^{-1} = e^{-\Lambda_{P_X}(\theta F_y)}.$$

We thereby obtain a formula for the optimizer,

$$\Gamma^*(dx, dy) = e^{\theta F(x,y) - \Lambda_{P_X}(\theta F_y)} P_X(dx) P_Y(dy)\,,$$

as well as the value, $D(\Gamma^* \| P_X \otimes P_Y) = \langle \Gamma^*, \theta F - \Lambda_{P_X}(\theta F_y)\rangle$. If Γ^* is feasible so it optimizes the original convex program, then the mean of F under Γ^* is equal to γ, and we obtain the desired expression (c) for the generalized mutual information $I_{\text{GMI}}(P_X; F)$.

The identity (b) then follows from the definition of the relaxation which gives $D(\Gamma^* \| P_X \otimes P_Y) =$

$$\min\{D(\Gamma \| P_X \otimes P_Y) - \theta^*\langle \Gamma, F^\circ - \gamma^\circ\rangle \\ - \theta_0^*(\langle \Gamma, 1\rangle - 1)\}$$

where F° is given in (14) and γ° is the mean of F° under P_{XY}. By restricting to only those Γ that are probability measures with the common mean $\Gamma(F^\circ) = \gamma^\circ$ we obtain the convex program,

$$\begin{array}{rl} \min & D(\Gamma \| P_X \otimes P_Y) \\ \text{subject to} & \langle \Gamma, F^\circ\rangle = \gamma^\circ \end{array}$$

This is the rate function $I_{\text{LDP}}(P_X; F^\circ)$ defined in (12). □

IV. OPTIMIZATION OF A LINEAR FAMILY OF DETECTORS

We now consider methods for computing an effective test. We restrict to functions in a linear class of the form $\{F_\alpha = \sum \alpha_i \psi_i : \alpha \in \mathbb{R}^d\}$ where $\{\psi_i : 1 \leq i \leq d\}$ are given. We obtain recursive algorithms based on the Newton-Raphson method.

Our main goal is to set the stage for the construction of sample-path algorithms (popularly called 'machine learning'). The resulting algorithms are similar to a technique introduced in [8] to solve a robust Neyman-Pearson hypothesis testing problem, and recent independent work of Basu et. al. [17] concerning risk-sensitive optimal control.

A. Hypothesis testing

We first consider approximation techniques for the Neyman-Pearson hypothesis testing problem.

For a given false-alarm exponent $R > 0$ we seek the best test in this class: It must be feasible, so that $\mathcal{Q}_R^+(\pi^0) \subset \mathcal{H}_-$, and subject to this constraint we seek the largest exponent $\beta(F_\alpha) = \min\{D(\Gamma \| \pi^1) : \Gamma \in \mathcal{H}\}$. Define $\mathcal{H}_-(\alpha) := \{\Gamma \in \mathcal{M} : \langle \Gamma, \alpha^{\scriptscriptstyle T}\psi\rangle \leq 0\}$ for $\alpha \in \mathbb{R}^d$, and let $\mathcal{G}_- \subset \mathcal{M}$ the intersection of all possible acceptance regions for H_0,

$$\mathcal{G}_- = \bigcap \big\{\mathcal{H}_-(\alpha) : \mathcal{Q}_R^+(\pi^0) \subset \mathcal{H}_-(\alpha),\ \alpha \in \mathbb{R}^d\big\}$$

Then, $\beta^* = \min\{D(\Gamma \| \pi^1) : \Gamma \in \mathcal{G}_-\}$. Alternatively, as in Proposition 2.2 we can show that β^* is the solution to the d-dimensional convex program,

$$\begin{array}{rl} \min & \Lambda_{\pi^1}(-\varrho F_\alpha) + \varrho \Lambda_{\pi_0}(F_\alpha) \\ \text{subject to} & \pi^0(F_\alpha) \leq \pi^1(F_\alpha) \end{array} \qquad (15)$$

The first order condition for optimality of α, ignoring the inequality constraint $\pi^0(F_\alpha) \leq \pi^1(F_\alpha)$, is $\nabla_\alpha\{\Lambda_{\pi^1}(-\varrho\alpha^{\scriptscriptstyle T}\psi) + \varrho\Lambda_{\pi_0}(\alpha^{\scriptscriptstyle T}\psi)\} = 0$. This can be expressed

$$-\varrho\check{\pi}^1(\psi) + \varrho\check{\pi}^0(\psi) = 0 \in \mathbb{R}^d, \qquad (16)$$

where for any function G,

$$\check{\pi}^1(G) = \frac{\pi^1(e^{-\varrho F_\alpha} G)}{\pi^1(e^{-\varrho F_\alpha})}, \quad \check{\pi}^0(G) = \frac{\pi^0(e^{F_\alpha} G)}{\pi^0(e^{F_\alpha})}\,. \tag{17}$$

On multiplying through by $\alpha^{*\intercal}$ and dividing by ϱ, with α^* a solution to (16), we then have $-\check{\pi}^1(\alpha^{*\intercal}\psi) + \check{\pi}^0(\alpha^{*\intercal}\psi) = 0$.

Recall that if F^0_α optimizes (11) then $F^0_\alpha - \gamma$ has the same value for any $\gamma \in \mathbb{R}$. We define F^* as the normalized function, $F^* = \alpha^{*\intercal}\psi - \check{\pi}^0(\alpha^{*\intercal}\psi)$ so that $\check{\pi}^1(F^*) = \check{\pi}^0(F^*) = 0$. We then necessarily have, by appealing to convexity of the log moment generating functions,

$$\pi^0(F^*) = \tfrac{d}{d\theta}\Lambda_{\pi_0}(\theta F^*)\Big|_{\theta=0} \le \tfrac{d}{d\theta}\Lambda_{\pi_0}(\theta F^*)\Big|_{\theta=1} = 0,$$

$$\pi^1(F^*) = \tfrac{d}{d\theta}\Lambda_{\pi^1}(\theta F^*)\Big|_{\theta=0} \ge \tfrac{d}{d\theta}\Lambda_{\pi^1}(\theta F^*)\Big|_{\theta=-\varrho} = 0.$$

Consequently we have feasibility, $\pi(F^*) \le 0 \le \pi^1(F^*)$.

For fixed ϱ we obtain corresponding exponents,

$$\begin{aligned} R_\varrho &= -\Lambda_{\pi_0}(F^*) = -\Lambda_{\pi_0}(\alpha^{*\intercal}\psi) + \check{\pi}^0(\alpha^{*\intercal}\psi), \\ \beta^*_\varrho &= -\Lambda_{\pi^1}(-\varrho F^*) = -\Lambda_{\pi^1}(-\varrho\alpha^{*\intercal}\psi) - \varrho\check{\pi}^0(\alpha^{*\intercal}\psi) \end{aligned}$$

The Hessian of the objective function in (15) can be expressed,

$$\nabla^2_\alpha\{\Lambda_{\pi^1}(-\varrho\alpha^\intercal\psi) + \varrho\Lambda_{\pi_0}(\alpha^\intercal\psi)\} = \varrho^2\check{\Sigma}^1 + \varrho\check{\Sigma}^0 \tag{18}$$

where the matrices $\{\check{\Sigma}^i\}$ are the covariance of ψ under the respective distributions $\{\check{\pi}^i\}$:

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

Hence, the Newton-Raphson method to estimate α^* is given by the recursion,

$$\alpha_{t+1} = \alpha_t - [\varrho\check{\Sigma}^1_{\alpha_t} + \check{\Sigma}^0_{\alpha_t}]^{-1}[-\check{\pi}^1_{\alpha_t}(\psi) + \check{\pi}^0_{\alpha_t}(\psi)], \tag{19}$$

$t \ge 0$, with $\alpha_0 \in \mathbb{R}^d$ given as initial condition.

B. Capacity

The rate function (12) can be expressed in terms of the log-MGF via,

$$I_{\text{LDP}}(P_X; F_\alpha) = \theta^*\gamma_\alpha - \Lambda_{\pi^0}(\theta^* F_\alpha) \tag{20}$$

Since $\theta^* F_\alpha = \theta^*\alpha^\intercal\psi$ we arrive at the convex program,

$$I_{\text{LDP}}(P_X; \psi) = \max\{\alpha^\intercal\bar{\psi} - \Lambda_{\pi^0}(\alpha^\intercal\psi) : \alpha \in \mathbb{R}^d\},$$

where $\bar{\psi} = \langle P_{XY}, \psi\rangle$.

Although it is in principle possible to find a test maximizing the generalized mutual information I_{GMI}, the resulting optimization problem is complex and requires significant channel information. A simpler algorithm is obtained if we minimize the bound I_{LDP}; there is no loss of optimality if the function the function class $\{F_\alpha = \sum \alpha_i\psi_i : \alpha \in \mathbb{R}^d\}$ is normalized so that F°_α defined in (14) is zero for each α.

V. Simulation-based computation

Optimization via simulation is compelling:

(i) It is convenient in complex models since we avoid numerical integration.

(ii) It offers the opportunity to optimize based on real data.

(iii) One obstacle in solving (15) is that this nonlinear program must be solved for several values of $\varrho > 0$ until an appropriate value of R is discovered. This is largely resolved using simulation techniques since parallelization is straightforward.

A. Hypothesis testing

We introduce here a stochastic approximation algorithm intended to approximate the Newton-Raphson recursion (19).

Let $\{\gamma_t, \epsilon_t\}$ denote a pair of positive scalar sequences. In the following algorithm $\nabla^2_t \ge 0$ is an estimate of the Hessian (18) and ϵ_t is chosen so that for some constant $\epsilon > 0$,

$$\epsilon_t \le \epsilon, \qquad \text{and} \quad [\epsilon_t I + \nabla^2_t] \ge \epsilon I, \qquad t \ge 0.$$

The sequence $\{\gamma_t\}$ is the gain for the stochastic approximation, such as $\gamma_t = \gamma_0/(t+1)$, $t \ge 0$, with $\gamma_0 > 0$ given.

```
For $t$ = 1 to runlength
▷ Draw independent samples from $\{\pi^i\}$ and evaluate $\psi(X^i)$: $X^0_t \sim \pi^0$, $X^1_t \sim \pi^1$, $\psi^0_t = \psi(X^0_t)$, $\psi^1_t = \psi(X^1_t)$.
▷ Estimate the mean of $e^{\alpha^\intercal\psi}$ under $\pi^0$ and the mean of $e^{-\varrho\alpha^\intercal\psi}$ under $\pi^1$:
    $\check{\eta}^0_{t+1} = \check{\eta}^0_t + \gamma_t\big(e^{\alpha^\intercal_t\psi^0_t} - \check{\eta}^0_t\big)$
    $\check{\eta}^1_{t+1} = \check{\eta}^1_t + \gamma_t\big(e^{-\varrho\alpha^\intercal_t\psi^1_t} - \check{\eta}^1_t\big)$    (21)
▷ Estimate the un-normalized twisted mean of $\psi$ under $\check{\pi}^0$ and $\check{\pi}^1$:
    $\Psi^0_{t+1} = \Psi^0_t + \gamma_t\big(e^{\alpha^\intercal_t\psi^0_t}\psi^0_t - \Psi^0_t\big)$
    $\Psi^1_{t+1} = \Psi^1_t + \gamma_t\big(e^{-\varrho\alpha^\intercal_t\psi^1_t}\psi^1_t - \Psi^1_t\big)$    (22)
▷ Estimate the un-normalized twisted covariance of $\psi$ under $\check{\pi}^0$ and $\check{\pi}^1$:
    $\check{\Sigma}^0_{t+1} = \check{\Sigma}^0_t + \gamma_t\big(e^{\alpha^\intercal_t\psi^0_t}\psi^0_t(\psi^0_t)^\intercal - \check{\Sigma}^0_t\big)$
    $\check{\Sigma}^1_{t+1} = \check{\Sigma}^1_t + \gamma_t\big(e^{-\varrho\alpha^\intercal_t\psi^1_t}\psi^1_t(\psi^1_t)^\intercal - \check{\Sigma}^1_t\big)$    (23)
▷ Estimate the Hessian, $\nabla^2_t = \varrho\,\check{\Sigma}^1_t/\check{\eta}^1_t + \check{\Sigma}^0_t/\check{\eta}^0_t$.
▷ Newton Raphson recursion, $\alpha_{t+1} - \alpha_t =$
    $-\gamma_t\Big[\epsilon_t I + \nabla^2_t\Big]^{-1}\Big[-\frac{e^{-\varrho\alpha^\intercal_t\psi^1_t}}{\check{\eta}^1_t}\psi^1_t + \frac{e^{\alpha^\intercal_t\psi^0_t}}{\check{\eta}^0_t}\psi^0_t\Big]$
```

We illustrate this technique with an example: π^1 is uniform on $[0,1]$, and π^0 is the distribution of the mth root of a realization of π^1 for a fixed constant $m > 0$. Shown in Figure 2 are plots of E_r vs. R for $m = 5$, with various choices of $\psi\colon \mathbb{R}^d \to \mathbb{R}^d$. These results were obtained using the stochastic Newton-Raphson method.

Figure 3 is intended to illustrate the transient behavior of this stochastic algorithm. The trajectories of the estimates of E_r and R appear highly deterministic. This is the typical behavior seen in all of the experiments we have run for this model.

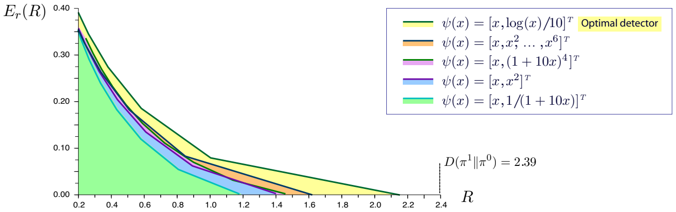

Fig. 2. Solutions to the Neyman-Pearson hypothesis testing problem for various linear tests. The marginal distribution of X^1 was taken uniform on $[0,1]$, and $X^0 = \sqrt[5]{X^1}$.

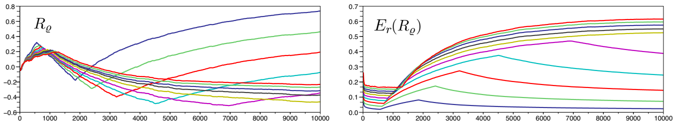

Fig. 3. Evolution of the transient phase of the stochastic Newton-Raphson method. These plots show estimates of R_ϱ and $E(R_\varrho)$ as a function of time using $\psi = [(1+10x)^{-1}, (1+10x)^{-2}, \ldots, (1+10x)^{-6}]$, with ϱ equally spaced among ten points from 0.1 to 1.

B. Capacity

We close with some recent results on computation of the best mismatched detector. Our goal is to maximize the objective function $J(\alpha) := I_{\text{LDP}}(P_X; F_\alpha)$ defined in (20),

$$\begin{aligned}\nabla J &= \langle P_{XY}, \psi\rangle - \langle \Gamma^\alpha, \psi\rangle \\ \nabla^2 J &= \langle \Gamma^\alpha, \psi\psi^T\rangle - \langle \Gamma^\alpha, \psi\rangle\langle \Gamma^\alpha, \psi\rangle^T\end{aligned}$$

where for any set $A \in \mathcal{B}(\mathsf{X}\times\mathsf{Y})$,

$$\Gamma^\alpha(A) := \frac{\langle P_X \otimes P_Y, e^{F_\alpha}\mathbb{I}_A\rangle}{\langle P_X \otimes P_Y, e^{F_\alpha}\rangle}$$

Hence a stochastic Newton-Raphson algorithm can be constructed in analogy with the Neyman-Pearson problem. Results for the AWGN channel are shown in Figure 4 for a special case in which the LLR lies in the quadratic function class obtained with $\psi = (x^2, y^2, xy)^T$.

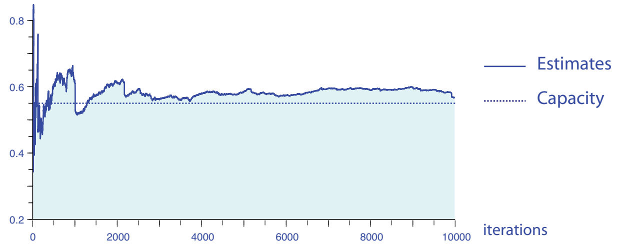

Fig. 4. Capacity estimates for the AWGN channel with SNR=2.

VI. Conclusions

In this paper we have relied on the geometry surrounding Sanov's Theorem to obtain a simple derivation of the mismatched coding bound. The insight obtained combined with stochastic approximation provides a powerful new computational tool. Based on these results we are currently considering the following:

(i) Signal constellation and code design based on a finite dimensional detector class $\{F_\alpha\}$.

(ii) Maximization of the error exponent for a given transmission rate, and a given detector.

(iii) Applications to MIMO channels and networks to simultaneously resolve coding and resource allocation in complex models.

Acknowledgment

Research supported in part by Motorola RPS #32 and ITMANET DARPA RK 2006-07284.

References

[1] 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.

[2] I. Csiszár and P. Narayan, "Channel capacity for a given decoding metric," *IEEE Trans. Inform. Theory*, vol. 41, no. 1, pp. 35–43, 1995.

[3] I. Csiszár and J. Körner, "Graph decomposition: a new key to coding theorems," *IEEE Trans. Inform. Theory*, vol. 27, no. 1, pp. 5–12, 1981.

[4] I. Csiszár, "The method of types," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, pp. 2505–2523, 1998, information theory: 1948–1998.

[5] S. Borade and L. Zheng, "I-projection and the geometry of error exponents," in *Proceedings of the Forty-Fourth Annual Allerton Conference on Communication, Control, and Computing, Sept 27-29, 2006*, UIUC, Illinois, USA, 2006.

[6] A. Dembo and O. Zeitouni, *Large Deviations Techniques And Applications*, 2nd ed. New York: Springer-Verlag, 1998.

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

[8] 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.

[9] 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.

[10] J. Huang, S. Meyn, and M. Medard, "Error exponents for channel coding and signal constellation design," *IEEE J. Selected Areas in Comm.*, vol. 24, no. 8, pp. 1647–, 2006, special issue on NONLINEAR OPTIMIZATION OF COMMUNICATION SYSTEMS. Guest Editors M. Chiang, S. H. Low, Z.-Q. Luo, N. B. Shroff, and W. Yu.

[11] I. Kontoyiannis, L. Lastras-Moñtano, and S. Meyn, "Relative entropy and exponential deviation bounds for general Markov chains," in *Proceedings of the 2005 IEEE International Symposium on Information Theory*, Sept. 2005, pp. 1563–1567.

[12] E. Abbe, M. Médard, S. Meyn, and L. Zheng, "Geometry of mismatched decoders," 2007, submitted to IEEE International Symposium on Information Theory.

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

[14] 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.

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

[16] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: John Wiley & Sons, Inc., 1968.

[17] A. Basu, T. Bhattacharyya, and V. S. Borkar, "A learning algorithm for risk-sensitive cost," Tata Institute for Fundamental Research," Unpublished report available at `http://math.iisc.ernet.in/˜eprints/`, 2006.