Geometry of Mismatched Decoders

Emmanuel Abbe
Massachusetts Institute of Technology
Laboratory for Information and Decision Systems
Cambridge, MA 02139
eabbe@mit.edu

Lizhong Zheng
Massachusetts Institute of Technology
Laboratory for Information and Decision Systems
Cambridge, MA 02139
lizhong@mit.edu

Sean Meyn
University of Illinois at Urbana-Champaign
Coordinated Science Laboratory
Urbana, IL 61801
meyn@uiuc.edu

Muriel Medard
Massachusetts Institute of Technology
Laboratory for Information and Decision Systems
Cambridge, MA 02139
medard@mit.edu

***Abstract*— "THIS PAPER IS ELIGIBLE FOR THE STUDENT PAPER AWARD".**
Reliable transmission over a discrete-time memoryless channel with a decoding metric that is not necessarily matched to the channel or not optimal (mismatched decoding) is considered. We address the question of comparing the performance of different mismatched metrics. A geometrical interpretation of this problem is presented, in particular, the very noisy case is shown to reduce to a tractable geometrical projection problem, for which an analytic solution is found.

I. Introduction

For a Discrete Memoryless Channel, in order to achieve the largest rate given by the Shannon capacity, it is necessary to use an optimal decoding rule. In [2],[8], the use of arbitrary decoding rules, possibly suboptimal, are investigated. A mismatched decoding rule represents a realistic model for time-varying channels, for which the transitions probabilities are unknown to the receiver. Suboptimal decoding metrics also appear when computational issues dictate a given decoding metric. No matter what the motivation is, we refer to suboptimal metric as mismatched metric.

In [2],[4],[8], coding theorems for mismatched DMC's are presented. "The random coding capacity" of mismatched decoders is known (and is shown to be a strict lower bound to the general mismatched capacity, as examples of non-randomly generated codebooks achieving higher rates have been found).

In this paper, we aim to give a geometrical picture of mismatched decoding. We start by considering the general case, expressing it as a projection problem with respect to the I-projection (cf. [1],[3]) and then solve analytically the very noisy case, showing how the I-projection problem maps to a L_2-projection problem with an appropriate inner product (i.e an appropriate measure), leaving us with a clear geometrical picture.

II. Problem Statement

Let $\mathcal{X}$, $\mathcal{Y}$ be finite sets, $P_X \in M_1(\mathcal{X})$ (a probability distribution on $\mathcal{X}$) and $P_{Y|X} \in M_1(\mathcal{Y}|\mathcal{X})$. We consider a DMC with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, and transition probabilities $P_{Y|X}$. We generate a code book with M codewords of length n, $\mathcal{C}(n) = \{x_1, \dots, x_M\}$, drawn i.i.d. according to P_X^n. We denote by P_Y, the induced marginal distribution on $\mathcal{Y}$, i.e. $P_Y(y) = \sum_{x\in\mathcal{X}} P_{Y|X}(y|x)P_X(x)$. Therefore, if a codeword, say x_1, is transmitted and if y is the received message, the joint distribution of (x_1, y) is given by $P_{Y|X} \odot P_X$, which we will also denote by μ_J, and the joint distribution of (x_i, y) for $i \neq 1$ is given by $P_X \times P_Y$, which we will also denote by μ_P. For any vector v in any alphabet, we denote by P_v the empirical distribution function (or type) of v.

Upon receiving y, the decoder looks for the elements x_i that maximizes a given function $F(x_i, y)$,

$$\hat{x} = \arg \max_{1\leq i\leq M} F(x_i, y).$$

If $F(x_i, y) = P\{y|x_i\}$ or equivalently $F(x_i, y) = \frac{1}{n} \log \frac{P\{y|x_i\}}{P\{y\}}$, this is the usual ML decoding. Note that $\frac{1}{n} \log \frac{P\{y|x_i\}}{P\{y\}} = \mathbb{E}_{P_{x_i,y}} \log \frac{\mu_J}{\mu_P}$. If the receiver uses the ML decoder with $Q_{Y|X}$ instead of $P_{Y|X}$, the objective function becomes $\mathbb{E}_{P_{x_i,y}} \log \frac{Q_{Y|X}}{Q_Y}$. We will consider decoding rule depending only on the type $P_{x_i,y}$ of (x_i, y), so that the decoding rule can be expressed as $F(P_{(x_i,y)})$. We will be interested in linear decoders, like the ML decoder, but instead of having to compute the expectation of $\log \frac{\mu_J}{\mu_P}$, we will consider an arbitrary function F on $\mathcal{X} \times \mathcal{Y}$. The first goal is to set a geometrical understanding of this problem, this will be tackled in the next section. We then consider a set of functions $\{f_i\}_{i=1}^k$ on $\mathcal{X} \times \mathcal{Y}$ and look for the linear combination of these functions $F = \sum_{i=1}^k \alpha_i f_i$, that will achieve the largest rate. We will solve this problem in the very noisy case and give a

complete geometrical picture, which also helps understanding the general problem.

III. Information Geometric Formulation

A. Review of Information Geometry

In $\mathbb{R}^N$, $N \geq 1$, an hyper-plane is described by all points x satisfying a set of $1 \leq i \leq N$ linear equations of the form $f_i^T \cdot x = \alpha_i$, with $f_i \in \mathbb{R}^N$ and $\alpha_i \in \mathbb{R}$. We refer to the f_i's as being the normal directions, since for any point x in the hyper-plane, the new hyper-plane given by all points y satisfying $y = x + \sum_i \lambda_i f_i$ for some λ_i's in $\mathbb{R}$, is orthogonal (with respect to the euclidean inner product). Moreover, the projection of a point onto an hyper-plane belongs to the intersection with the normal hyper-plane.

We now consider $M_1(\mathcal{X} \times \mathcal{Y})$ instead of $\mathbb{R}^N$. We denote by $\mathcal{Z}$ an arbitrary alphabets (which can be $\mathcal{X}$ or $\mathcal{X} \times \mathcal{Y}$).

Definition 1: Let $k \geq 1$, $i \in \{1, \ldots, k\}$, $f_i : \mathcal{Z} \to \mathbb{R}$ (*normal directions*) and $\alpha_i \in \mathbb{R}$ (*positions*). A linear family $\mathcal{L}_{f_i,\alpha_i}$ in $M_1(\mathcal{Z})$ is defined by

$$\mathcal{L}_{f_i,\alpha_i} = \{P \in M_1(\mathcal{Z}) | \forall 1 \leq i \leq k, \mathbb{E}_P f_i = \alpha_i\}.$$

Definition 2: Let $k \geq 1$, $i \in \{1, \ldots, k\}$, $f_i : \mathcal{Z} \to \mathbb{R}$ (*directions*) and $P_0 \in M_1(\mathcal{Z})$. An exponential family $\mathcal{E}_{P_0,f_i}$ in $M_1(\mathcal{Z})$ is defined by

$$\mathcal{E}_{P_0,f_i} = \{P \in M_1(\mathcal{Z}) | \exists \lambda \in \mathbb{R}^k \text{ s.t. } P = P_0 e^{\sum_{i=1}^k \lambda_i f_i}/c(\lambda)\},$$

where $c(\lambda) = \sum_{z \in \mathcal{Z}} P_0(z) e^{\sum_{i=1}^k \lambda_i f_i(z)}$.

The linear families will be pictured in a similar way as the hyper-planes in the euclidean geometry, the f_i's can also be interpreted as normal directions, not with respect to another linear family, but with respect to an exponential family. Let $\mathcal{L}_{f,\alpha}$ be a linear family passing through a point P_0 and $\mathcal{E}_{P_0,f}$ (which we also denote by $P(\lambda)$) its "normal" exponential family passing through P_0. We then have similar properties as in the euclidean setting, involving the divergence instead of the euclidean norm. Namely, for any $Q \in \mathcal{E}_{P_0,f}$, one has from the I-projection property (cf. [3])

$$\arg\min_{P \in \mathcal{L}_f} D(P||Q) = P_0.$$

For $P \in \mathcal{L}_{f,\alpha}$, if one defines the one-dimensional linear family $P(t) = tP + (1-t)P_0$ (which is clearly contained in $\mathcal{L}_{f,\alpha}$), and if $\lambda \in \mathbb{R}$, we then have

$$\mathbb{E}\partial_t \log P(t) \partial_\lambda \log P(\lambda)|_{t=\lambda=0} = 0,$$

i.e. the Fisher inner product makes these two curves orthogonal at P_0 (cf. [1]). This result can be equivalently stated in terms of the divergence, the function

$$\mathcal{L}_f \ni P_1 \mapsto D(P||Q) - D(P||P_1) - D(P_1||Q)$$

defines a notion of angles (with positive or negative signs characterizing obtuse or acute angles), it achieves 0 at $P_1 = P_0$, corresponding to an analogue of the pythagorean theorem. Thus, the divergence exhibits similarities with the squared euclidean distance.

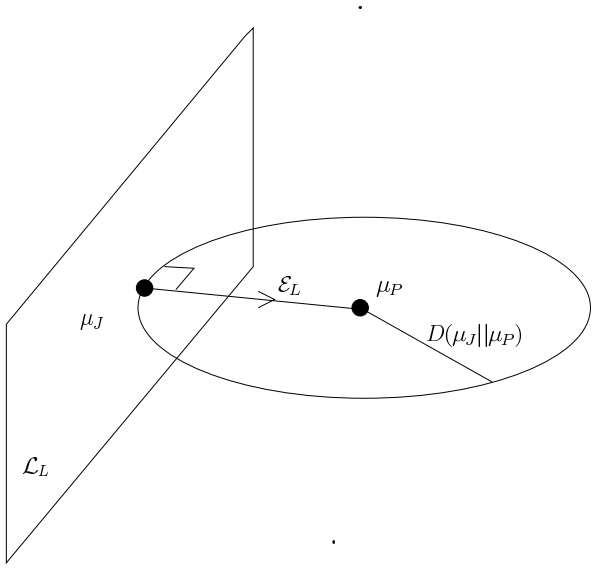

Fig. 1. Orthogonality property of the ML-decoder

B. Mismatch Geometry

We now consider the decoding rule finding the index i that maximizes $F(P_i)$. If x_1 is sent and y is received, denoting $P_i = P_{x_i,y}$, an error occurs if the following event happens

$$\mathrm{E} \equiv \{\exists i \neq 1 \text{ s.t. } F(P_i) > F(P_1)\}.$$

Note that if F is such that for any $\gamma \in \mathbb{R}$ (for which $\{F \geq \gamma\}$ is not empty), $\mu_J \in \{F \geq \gamma\}$, then for any γ, we have

$$\mathrm{E} \subset \{F(P_1) < \gamma\} \cup \cup_{i \neq 1} \{F(P_i) \geq \gamma\},$$

and using union bound

$$\mathbb{P}\{\mathrm{E}\} \leq \mathbb{P}\{F(P_1) < \gamma\} + \min(M\mathbb{P}\{F(P_2) \geq \gamma\}, 1), \quad (1)$$

defining $R = \frac{\log M}{n}$ and denoting by E_r the error exponent, we get from Sanov's theorem,

$$E_r \geq \min\left[\inf_{F(P_1)<\gamma} D(P_1||\mu_J), |\inf_{F(P_2)\geq\gamma} D(P_2||\mu_P) - R|^+\right]. \quad (2)$$

We now examine the information geometry of the log-likelihood decoding rule. As we mentioned before, maximizing (over i) $P\{y|x_i\}$, is the same as maximizing $\mathbb{E}_{P_{x_i,y}} \log \frac{\mu_J}{\mu_P}$. So in this case $\{P : F(P) = \gamma\}$ is the linear family $\{P : \mathbb{E}_P L = \gamma\}$, where

$$L = \log \frac{\mu_J}{\mu_P},$$

which is orthogonal to $\mathcal{E}_{\mu_J,L} = \mathcal{E}_{\mu_P,L} = \mu_J^s \mu_P^{(1-s)}/c(s)$. Choosing now $\gamma = \mathbb{E}_{\mu_J^+} L = D(\mu_J^+||\mu_P)$, where μ_J^+ refers to a arbitrarily close measure to μ_J in oder to keep both exponent in (2) positive, we get a capacity for the ML decoding rule given by $D(\mu_J||\mu_P)$, as shown in figure 1. Note that it is a particularity of the ML-decoder that the projection of μ_P onto the linear family of orthogonal direction given by L, is precisely μ_J (see figure 1). As a general fact, if

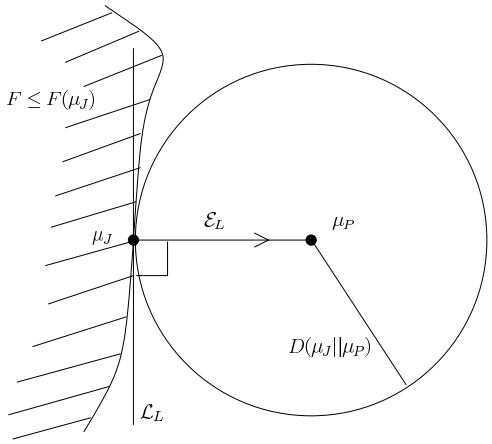

Fig. 2. Decoder achieving full capacity

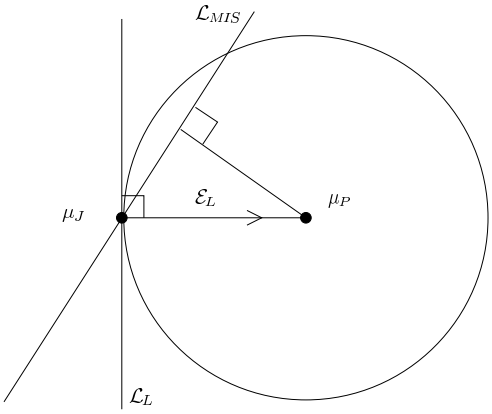

Fig. 3. Mismatched linear decoder

$$\{F \geq F(\mu_J)\} \supset B_D(\mu_P, D(\mu_J||\mu_P)),$$

(where $B_D(\mu_P, D(\mu_J||\mu_P) = \{P : D(P||\mu_P) \leq D(\mu_J||\mu_P)\}$) then, using the decoding rule $\arg\min_i F(\hat{P}_i)$, the capacity is $D(\mu_J||\mu_P)$ (cf. figure 2). This tells us in particular that a decoding rule that declares i if P_i is the only type within a small neighborhood around μ_J (like a norm-ball for some norm, e.g. L^∞), will achieve capacity if the neighborhood can be shrunk as much as desired around μ_J, or equivalently when the radius of the ball gets as small as desired. We now investigate the case in which F is given by a linear family, which is not the log-likelihood, from now on, F denotes the orthogonal direction of the linear family, i.e. the decoding rule looks for i maximizing $\mathbb{E}_{P_i}F$. Geometrically, this means that the orthogonal direction is not given by L, as pictured in 3. In that case, the previous general fact do not apply. Previously, if a type appears typical, it has in particular typical marginals, this is because in these cases the dominant types tends to μ_J. But now, this is no longer true. However, the event $\bar{\bar{P}}_i \in B_D(P_Y, \nu)$ is a probability 1 event (by Sanov's theorem, by taking ν as small as desired, we can make the probability of this event arbitrarily close to 1 in the exponential scale), therefore we can include this event under the probabilities in (1) and the error probability still tends to 0 under this condition, getting the following exponent and bound on the capacity

$$C > \inf_{\substack{P:\mathbb{E}_P F \geq \mathbb{E}_{\mu_J} F \\ \bar{\bar{P}}=P_Y}} D(P||\mu_P).$$

In [8], this bound is shown to be achievable and also tight when restricted to code books drawn from a random ensemble, we will denote it by C_{LM}. We now draw the geometrical picture at C_{LM}. First, observe that $\bar{\bar{P}} = P_Y$ is a linear family given by

$$\forall 1 \leq k \leq |\mathcal{Y}| - 1, \quad \mathbb{E}_P \delta_k = P_Y(k), \tag{3}$$

where $\delta_k(i,j) = \delta_{\mathcal{X},k}(i,j)$ equals 1 if $j = k$ and 0 otherwise. Therefore C_{LM} is the distance (KL-divergence) between μ_P and its projection onto the linear family resulting of the intersection between $\{P : \mathbb{E}_P F = \mathbb{E}_{\mu_J}\}$ and (3), which lies on the orthogonal exponential family. Thus, as the orthogonal exponential family passing through μ_P is

$$\mu_P \exp(tF + \sum_i \lambda_i \delta_i) c(t, \lambda)$$

the projection is the result of following equations

$$\mathbb{E}_{\mu_P \exp(tF + \sum_i \lambda_i \delta_i) c(t,\lambda)} F = \mathbb{E}_{\mu_J} F \tag{4}$$

$$\forall j = 1, \ldots, |\mathcal{Y}| - 1,$$

$$\sum_k \mu_P(k,j) \exp(tF(k,j) + \lambda_j) c(t,\lambda) = P_Y(j). \tag{5}$$

At that level of generality, the problem becomes a numerical problem when trying to solve it. It is possible to express it as a projection of L with respect to the Fisher inner product, but does not change the problem in a more tractable form.

IV. Very Noisy Geometry

We recall that

$$L = \log \frac{\mu_J}{\mu_P},$$

and we define $L_j = L(\cdot, j)$ for $j \in \mathcal{Y}$. We have

$$\mu_J = \mu_P \frac{\mu_J}{\mu_P} = \mu_P e^{\log \frac{\mu_J}{\mu_P}} = \mu_P(1 + L) + o(L),$$

and

$$P_{X|Y=j} = (1 + L_j) P_X + o(L_j),$$

where $o(f)$ for a function f means $o(\sup_i f(i))$.
By very noisy, we mean that μ_J and μ_P are close and $P_{X|Y=j}$ and P_X are close for each j, formally, one has to think as a family of channels indexed by a parameter ϵ, such as the exponential family connecting μ_J and μ_P, and we are interested in a first order taylor expansion (in ϵ) at μ_P (it is important that the approximation is uniform in ϵ, i.e. we want the L_j to tend uniformly to zero). In what follows, we skip the parameter ϵ and directly treat the L_j's as our small parameter, claiming that all approximation we will make leave us with an $o(L)$ approximation.
Thus, we approximate μ_J by $\mu_P(1 + L)$ requiring

$$\mathbb{E}_{\mu_P} L = 0,$$

and $P_{X|Y=j}$ by $(1+L_j)P_X$, with

$$\mathbb{E}_{P_X} L_j = 0, \quad \forall j$$

We in turn approximate $\mu_P \exp(tF + \sum_i \lambda_i \delta_i) c(t,\lambda)$ by $\mu_P(1 + tF + \sum_i \lambda_i \delta_i)$, with $\mathbb{E}_{\mu_P}(tF + \sum_i \lambda_i \delta_i) = 0$. With this, (4) becomes

$$\mathbb{E}_{\mu_P}(1 + tF + \sum_i \lambda_i \delta_i)F = \mathbb{E}_{\mu_P}(1+L)F$$

and (5) becomes a

$$\mathbb{E}_{\mu_P}\delta_k(1 + tF + \sum_i \lambda_i \delta_i) = P_Y(k).$$

Note that for $P \in M_1(\mathcal{Z})$ and $V \in M_0(\mathcal{Z})$ (a signed measure on $\mathcal{Z}$ integrating to 0),

$$\frac{\partial^2}{\partial \theta^2} D(P + \theta V || P) = \mathbb{E}_P \frac{V^2}{P^2},$$

and we approximate the divergence in our problem by

$$D(\mu_P + \mu_P(tF + \sum_k \lambda_k \delta_k)) \approx \frac{1}{2}\mathbb{E}_{\mu_P}(tF + \sum_k \lambda_k \delta_k)^2.$$

In addition, we can always shift our function F to $\tilde{F}$, by subtracting $\mathbb{E}_{P_X} F_j$ to each component F_j, so that $\mathbb{E}_{p_X} \tilde{F}_j = 0$. Geometrically this corresponds to projecting F in an orthogonal direction to the δ_k's.
We now define

$$\langle f, g \rangle = \mathbb{E}_{\mu_P} fg,$$

and our projection problem described in (4), (5), reduces to

$$t||\tilde{F}||^2 \quad = \quad \langle \tilde{F}, L \rangle. \tag{6}$$

which gives

$$t \quad = \quad \frac{\langle \tilde{F}, L \rangle}{||\tilde{F}||^2}. \tag{7}$$

Since the divergence reduces to

$$\frac{1}{2} t^2 ||\tilde{F}||^2,$$

we are left with the following expression for the capacity.
Proposition:
In the very noisy case, the mismatch capacity is given by

$$\frac{1}{2}\frac{\langle \tilde{F}, L \rangle^2}{||\tilde{F}||^2}.$$

Above expression is the norm squared of the projection of L onto the linear family orthogonal to $\tilde{F}$, with respect to the μ_P-inner product. In figure 4, we summarize the several steps we have performed:

- The very noisy assumptions implies that

  $$\langle L, \delta_k \rangle = 0$$

  for any k. Thus L belongs to the inclined plane of the figure.
- Projecting F to $\tilde{F}$, restrict ourself to the intersection between the two planes. Then, projecting μ_P, onto the intersection gives P^{**} which is clearly closer (at a smaller divergence) to μ_P than P^*, obtained without the second marginal condition, i.e. achieving $\inf_{P:\, \mathbb{E}_P F \geq \mathbb{E}_{\mu_J} F} D(P||\mu_P)$.
- The very noisy setting allow us to deal with all previous steps in $L_2(\mu_J)$, projecting L onto $\tilde{F}$.
- The fact that P^{**} and μ_J are distinct shows the gap to the non-mismatched capacity

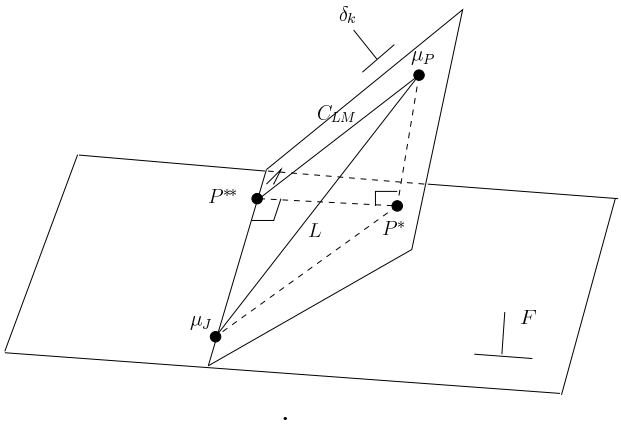

Fig. 4. Mismatched capacity in the very noisy case

We know from Cauchy-Schwartz

$$\frac{1}{2}\frac{\langle \tilde{F}, L \rangle^2}{||\tilde{F}||^2} \leq \frac{1}{2}||L||^2,$$

with equality only if $\tilde{F} = L$. If $F = \sum_{i=1}^k \alpha_i f_i$ we project L onto $\sum_{i=1}^k \alpha_i \tilde{f}_i$, i.e.

$$\alpha_i^* = \langle \tilde{f}_i', L \rangle,$$

where the $\tilde{f}_i'$ are a Gram-schmidt expansion of the $\tilde{f}_i$'s and the capacity is given by $\frac{1}{2}\sum_i \alpha_i^{*2}$. Therefore, one can sequentially improve the performance of the decoder by adding μ_P-orthogonal decoding functions.

V. Extensions

1. We have seen that the mismatched problem can be formulated as a projection problem. In the general setting, with respect to the Fisher inner product and in the very noisy case with respect to the $L_2(\mu_P)$ inner-product. It seems natural to study the "non-noisy" case, and see to what geometry the problem is mapped. However, defining a very "non-noisy" channel requires some thinking by its own. It would be interesting to have the picture in both extreme cases, to understand how the randomness of the channel affects the decoding rules.

2. In a mismatched situation, the optimization over the input distribution requires that both the channel and the mismatched metric is known at the receiver. Assume now that the transmitter only knows that the receiver will use a mismatch metric belonging to a given neighborhood (thought to be a "small

set" in order to use similar techniques as in the very noisy case). For each input distribution, the transmitter can identify which mismatch direction is worse within this neighborhood. One can then find the input distribution maximizing the worse mismatched capacity, ensuring to achieve the largest achievable rate (even if the decoder uses the worse mismatch metric).

ACKNOWLEDGMENT

We would like to thank Professor Telatar, whose comments have contributed to improve the paper's presentation. The second author thanks Professor Shamai for suggesting using information geometry to study this problem.

REFERENCES

[1] S. Amari and H. Nagaoka, "Methods of Information Geometry", *American Mathematical Society*, January 2001.

[2] I. Csiszar and P. Narayan, "Channel capacity for a given decoding metric," *IEEE Trans. Inform. Theory*, 41(1):3543, 1995.

[3] I. Csiszar and G. Tusnady, "Information geometry and alternating minimization procedures," *Statistics & Decisions. International Journal for Statistical*. Supplemental Issue # 1., pages 205237, 1984.

[4] I. Csiszir and J. Korner, Graph decomposition: A new key to coding theorems, *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 5-12, Jan. 1981.

[5] I. Csiszar, "Information Theory and Statistics: a Tutorial", *Foundations and Trends in Communications and Information Theory*, editor S. Verdu, vol. 1, issue 4, 2004.

[6] A. Dembo and O. Zeitouni, "Large Deviations Techniques And Applications," *Springer- Verlag*, New York, second edition, 1998.

[7] Robert G. Gallager, "Information Theory and Reliable Communication," *John Wiley & Sons*, Inc., New York, NY, USA, 1968.

[8] G. Kaplan, A. Lapidoth, S. Shamai, and N. Merhav. "On information rates for mismatched decoders", *IEEE Trans. Inform. Theory*, 40(6):19531967, 1994.