I-Projection and the Geometry of Error Exponents

Shashi Borade and Lizhong Zheng

***Abstract*— We present a geometric approach, using I-projections [4], for analyzing error exponents in various information theory problems—e.g., hypothesis-testing, source coding, and channel coding. By illuminating on the hidden geometrical structure, it also clarifies the distribution of the log likelihood for correct and incorrect codewords. Calculating the error exponent for very noisy channels becomes essentially trivial now. We also prove tightness of Gallager's formula for error exponent of the expurgated ensemble.**

I. Introduction

A large number of information theoretic problems can be written as optimization of the Kullback-Liebler divergence. This includes most calculations of channel capacities, rate-distortion functions, and the reliability functions. With a handful of famous exceptions, most of these calculations can only be carried out numerically. Especially, the results of many multi-user information theory problems are given in the form of high dimensional optimizations, with little effort spent in finding the structure of the solutions.

As an example of such divergence minimization problem, the calculation of the error exponent is well known to have two different forms of solutions. The original solutions by Gallager *et. al.* [6], [8] were derived using Chernoff bound and similar techniques. These results take the form of optimization over the input distribution and a scalar parameter ρ. While these results are concise and relatively easy to compute, it is sometimes hard to capture the intuition behind the derivations. In comparison, Csiszar and Korner took a conceptually more tractable approach, using large deviation to study the decoding errors in a discrete memoryless channel (DMC), with the results directly in the form of divergence minimizations. The solution to these minimization problems has an important operational meaning— they characterize the typical error event. This approach is much more intuitive and thus widely used in a variety of information theory problems. However, this is a high dimensional optimization problem, over the space of channel realizations instead of a scalar parameter, which is often harder to solve.

A natural question is that since Gallager's result solves the same problem with scalar optimization instead of the high dimensional divergence minimization required by Csiszar and Korner, then is there a general structure of the solutions to the later problem? While this question is partially answered in the exercises of [5], we try to address this problem using the methods of information geometry in this paper. We hope that by revealing the geometric structure of the error exponent problem, we can obtain more insights to the more general information theoretic divergence minimization problems.

The study of information geometry started as early as 1920's by Fisher. Amari and Nagaoka offers a comprehensive text [1] on the subject. The key idea is to think the set of all probability distributions $\mathcal{P}(\mathcal{Z})$ on alphabet $\mathcal{Z}$, as a manifold in $\mathcal{R}^{|\mathcal{Z}|}$, and to view a probability model, defined as a set of parameterized distributions $\{\bar{P}_\theta, \theta \in \mathcal{R}^m\}$, as a sub-manifold of $\mathcal{P}(\mathcal{Z})$, where $\bar{P}_\theta$ denotes a distribution (or a point in $\mathcal{R}^{|\mathcal{Z}|}$) satisfying $\sum_{z\in\mathcal{Z}} \bar{P}_\theta(z) = 1$. By properly defining the geometric structure of the manifold, one can establish a correspondence between some quantities of statistical interest and some geometric concepts—Fisher information as local metric, KL divergence as a generalization of distance, etc.. Natural concepts raised from this approach include *linear families* and *exponential families*, which sometimes can be thought as "*orthogonal subspaces*" in $\mathcal{P}(\mathcal{Z})$. This can then be used to define a notion of *projection*, which is directly related to the divergence minimization problems.

In the following of this paper, we start by giving a very brief discussion of the geometric structure used. We assume no previous knowledge of differential geometry by the readers, and try to focus on the motivation of the formulation instead of the detailed calculations. It is worth pointing out that many of the results derived using the method of information geometry can in fact be obtained from direct algebraic calculations. Thus much of the value of the approach lies in the simplicity of the solution and the new insights it brings. Throughout the paper, we use $\approx$ to denote exponential approximation.

II. Exponential Family and I-Projection

In this section, we summarize the main ideas in [1] very briefly, and introduce several concepts used in error exponent calculations. The basic idea of information geometry is to think a parameterized family of distributions,

$$\{\bar{P}_\theta(\cdot), \quad \theta \in \mathcal{R}^m\}$$

as a manifold, where each point denotes a distribution over $\mathcal{Z}$. The family is often called a *probability model*, and denoted as $\mathcal{M}$. The tangent space at a point $p \in \mathcal{M}$ is denoted as T_p, which has dimension m.

At each point p, the parameter θ gives a natural coordinate system on T_p. Let $\theta = [\theta_1, \theta_2, \ldots, \theta_m]$ and consider the tangent plane $T_{\bar{P}_\theta}$. By perturbing only component θ_i,

$$\bar{P}_{[\theta_1,\ldots\theta_{i-1},\theta_i,\theta_{i+1}\ldots,\theta_n]} \to \bar{P}_{[\theta_1,\ldots,\theta_i+\delta\theta_i,\ldots,\theta_n]}$$

This research is supported in part by NSF-Career award CCF-0347395, and by AFOSR under grant FA9550-06-1-0156.

Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139. {spb,lizhong}@mit.edu

gives a direction to move away from $\bar{P}_\theta$, and we denote that direction by the following vector (in $\mathcal{R}^{\mathcal{Z}}$)

$$(\partial_i)_{\bar{P}_\theta} = \frac{\partial}{\partial \theta_i} \bar{P}_\theta,$$

The collection $\{(\partial_i)_{\bar{P}_\theta}, i = 1, \ldots, m\}$ provides a set of basis vectors of the tangent space $T_{\bar{P}_\theta}$.

In addition to these local geometric structures, as we would like to talk about the "*distance*" between distributions over the entire probability model, it is necessary to introduce a notion of *affine connection*, which relates the tangent spaces at different points. Let $p \equiv \bar{P}_\theta \in \mathcal{M}$, and for a small perturbation by vector $\Delta\theta$, let $q \equiv \bar{P}_{\theta+\Delta\theta}$ be in the neighborhood of p. A connection is a map

$$\Pi_{p,q} : T_p \to T_q,$$

which is affine in T_p and perturbation $\Delta\theta$. For a vector $u \in T_p$, its image $\Pi_{p,q}(u) \in T_q$ is called the '*parallel translation*' of u, from p to q.

Now for a distribution r that is "*far away*" from p, one can apply the parallel translations, one after other, along a curve connecting p and r, thus relate the tangent spaces at any two points.

Due to linearity, to specify a connection, it is sufficient to specify the image of a set of basis vectors of T_p. For a parameterized probability model defined above, it is often convenient to consider the connection that maps $(\partial_i)_p$ to $(\partial_i)_q$. The following two examples of probability manifolds are often used in this paper.

Example1: Linear Family

Consider the family

$$\mathcal{L} \equiv \left\{ \bar{P}_\theta(\cdot) : \bar{P}_\theta = c + \sum_{i=1}^{m} \theta_i f_i \right\} \text{ where } c, f_1, \cdots f_m \in \mathcal{R}^{|\mathcal{Z}|}$$

and w.l.o.g. we assume $\sum_z c(z) = 1$ and $\sum_z f_i(z) = 0$ for all i, so that resulting $\bar{P}_\theta$ is indeed a distribution. Now

$$(\partial_i)_{\bar{P}_\theta} = \frac{\partial}{\partial \theta_i} \bar{P}_\theta = f_i \tag{1}$$

for $i = 1, \ldots, m$ are a set of basis vectors at point $\bar{P}_\theta$. Now $\sum_z f_i(z) = 0$ implies that for every i,

$$\sum_z (\partial_i)_{\bar{P}_\theta}(z) = 0 \tag{2}$$

which ensures that any tangent vector for this manifold is also tangent to the probability simplex for alphabet $\mathcal{Z}$. A natural parallel translation for this model is simply the identity map,

$$(\partial_i)_q = (\partial_i)_p = f_i(\cdot) \Rightarrow \Pi_{p,q}(u) = u$$

for any $p, q \in \mathcal{L}$. This connection of identity map is called the linear connection. Notice that the parallel translation of a tangent vector $u \in T_p$ along any curve connecting p to q will be the same. We say in this case that the linear family $\mathcal{L}$ is *flat* w.r.t. the linear connection.■

Example2: Exponential family

Consider the exponential family defined as

$$\mathcal{E} \equiv \left\{ \bar{P}_\theta(z) : \bar{P}_\theta(z) = \exp\left[c(z) + \sum_{i=1}^{m} \theta_i f_i(z) - \psi(\theta) \right] \right\}$$

where $\psi(\theta)$ arises due to a normalization factor that we will explain later.

In order to compare with the previous example, we consider an embedding that takes a distribution $p \in \mathcal{P}(\mathcal{Z})$, and maps it to $\log p \in \mathcal{R}^{|\mathcal{Z}|}$. The exponential family is thus mapped into a sub-manifold of $\mathcal{R}^{|\mathcal{Z}|}$, which is similar to the linear family of the previous example. For this new log-image manifold, let $(\partial_i)_p^{(e)}$ denote a basis vector of the tangent space at p, to distinguish it from $(\partial_i)_p$ for the original manifold. There is a simple relation between these two sets of basis vectors at $p \equiv \bar{P}_\theta$:

$$(\partial_i)_p^{(e)} = \frac{\partial}{\partial \theta_i} \log \bar{P}_\theta = \frac{1}{\bar{P}_\theta} \frac{\partial}{\partial \theta_i} \bar{P}_\theta = \frac{1}{\bar{P}_\theta} (\partial_i)_p \tag{3}$$

Now to define a connection on $\mathcal{E}$ called as the *exponential connection*: we first use (3) to map the tangent vector $(\partial_i)_p$ to its corresponding $(\partial_i)_p^{(e)}$, then translate it to $(\partial_i)_q^{(e)}$ using a linear connection discussed ahead, then use reverse of (3) to map it back to $(\partial_i)_q$ on the original probability model. Such mapping of all basis vectors at T_p defines the exponential connection from T_p to T_q.

By creating connection this, although the exponential family $\mathcal{E}$ is different from the linear family $\mathcal{L}$, its log-image becomes similar. The symbol $(\partial_i)^{(e)}$ should be thought as a representation the tangent vector (∂_i) in the original space. Equation (2) now takes the form

$$0 = \sum_z (\partial_i)_{\bar{P}_\theta}(z) = \sum_z \bar{P}_\theta(z) (\partial_i)^{(e)}_{\bar{P}_\theta}(z) = E_{\bar{P}_\theta}\left[\frac{\partial \log \bar{P}_\theta(\mathbf{z})}{\partial \theta_i} \right] \tag{4}$$

To satisfy this constraint at both p and q, the linear map from $(\partial_i)_p^{(e)}$ to $(\partial_i)_q^{(e)}$ should be defined as

$$(\partial_i)_q^{(e)} = (\partial_i)_p^{(e)} - \mathbf{1} \cdot E_q\left[\frac{\partial \log \bar{P}_\theta(\mathbf{z})}{\partial \theta_i} \right]$$

where $\mathbf{1}$ denotes a vector of all ones. This map differs from the identity map by a constant shift. Specializing to the exponential family, this yields the following connection from tangent space at $p \equiv \bar{P}_\theta$ to tangent space at $q \equiv \bar{P}_{\hat{\theta}}$

$$(\partial_i)^{(e)}_{\bar{P}_\theta} = f_i - \mathbf{1} \cdot \frac{\partial}{\partial \theta_i} \psi_\theta \longrightarrow (\partial_i)^{(e)}_{\bar{P}_{\hat{\theta}}} = f_i - \mathbf{1} \cdot \frac{\partial}{\partial \theta_i} \psi_{\hat{\theta}}$$

The key observation is that if we apply this connection one after other to neighboring points on the log image along a curve, connecting the images of p and q, the resulting constant shifts depend only on the end points, but not the which curve we use to connect them. Thus, the induced map in the original probability space $(\partial_i)_p \to (\partial_i)_q$ also does not depend on the curve. We therefore say exponential family is flat w.r.t. exponential connection. ■

The concepts of exponential connection and exponential family are highly relevant for developing the notions of

orthogonality and distance for probability manifolds. At any given point p, a (possibly different) notion of metric, $\langle u, v\rangle_p$ for $u, v \in T_p$, can be defined (similar to an inner product space). However, when a connection $\Pi_{p,q}$ is defined, it is not always the case that

$$\langle u, v\rangle_p = \langle \Pi_{p,q}(u), \Pi_{p,q}(v)\rangle_q \tag{5}$$

is satisfied. A connection that satisfies (5) is called a *Riemannian connection*. It can be shown that for a given metric, there exists a unique Riemannian connection; and that a Euclidean coordinate system exists if and only if the Riemannian connection is flat. Thus (5) and flatness lead to the simplest geometric structure.

Unfortunately, in the study of probability models, a slightly more complicated geometric structure is involved. In such models, it is often desirable to use the Fisher metric,

$$g_{ij}(\theta) \equiv \langle(\partial_i), (\partial_j)\rangle_{\bar{P}_\theta} \equiv E_{\bar{P}_\theta}\left[\frac{\partial}{\partial\theta_i}\log\bar{P}_\theta(\mathbf{z})\frac{\partial}{\partial\theta_j}\log\bar{P}_\theta(\mathbf{z})\right]$$

Fisher information has a strong operational meaning in parametric estimation. The only caveat is that it does not give rise to a Euclidean geometry.

If one insist to develop a notion of global distance and orthogonality on the probability model, (5) has to be relaxed. We would like to find a pair of connections Π^1 and Π^2, s.t.

$$\langle u, v\rangle_p = \langle \Pi^1_{p,q}(u), \Pi^2_{p,q}(v)\rangle_q$$

We say in this case Π^1 and Π^2 are dual connections. It turns out that if we use Fisher metric, then linear connection and exponential connection are dual to each other. Consequently, the Fisher metric defined over a probability model often involves mapping one vector with linear connection, and the other with exponential connection.

One important application of this concept is as follows. Suppose $u, v \in T_p$ are locally orthogonal, i.e., the Fisher metric $\langle u, v\rangle_p = 0$. Now if we "*extend*" u and v w.r.t. linear and exponential connections, respectively, as shown in Figure 1, the resulting curves are geodesics, which can be thought as "*straight lines*", w.r.t. to corresponding connections. The two curves are in a sense orthogonal to each other, which is made precise in the following Theorem from [4].

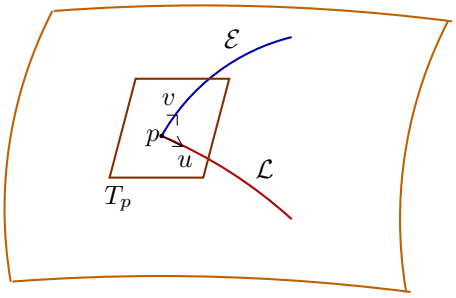


Fig. 1. "*Orthogonal Spaces*": exponential and linear families

Theorem 1: Given a function $f : \mathcal{Z} \to \mathcal{R}^K$ and a constant α, consider a linear family

$$\mathcal{L}_{f,\alpha} \equiv \{q : E_q[f(\mathbf{z})] = \alpha\}$$

and an exponential family for the same f through distribution p is defined as

$$\mathcal{E}_{f,p} \equiv \left\{q : \ q(z) = \frac{p(z)\cdot\exp\left(\sum_{i=1}^m \theta_i f_i(z)\right)}{k(\theta)}, \theta \in \mathcal{R}^m\right\}$$

where $k(\theta) = \sum_z p(z)\exp\left(\sum_{i=1}^m \theta_i f_i(z)\right)$ is the normalization factor. We then have

$$q^* \equiv \arg\min_{q\in\mathcal{L}_{f,\alpha}} D(q\|p) \in \mathcal{E}_{f,p}$$

Moreover, for any $q \in \mathcal{L}_{f,\alpha}$,

$$D(q\|p) = D(q\|q^*) + D(q^*\|p) \tag{6}$$

This result can be thought as finding the projection of p on the linear family $\mathcal{L}_{f,a}$, which is called by Csiszar as the *I-projection*. Above relation in (6) is called the Pythagorean relation for I-projection.

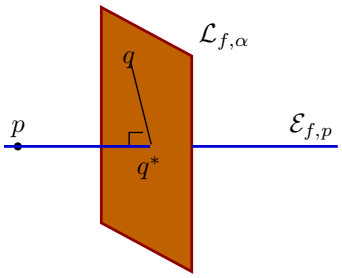


Fig. 2. I-projection theorem

I-projection can be widely used in error exponent problems. The case that f is a scalar function, i.e., $K = 1$, is particularly useful, since I-projection in this case reduces the divergence minimization problem into a search over scalar parameter θ. The following example shows that the Chernoff bound, which is often used in Gallager's derivations of the error exponents, is directly related to this geometric picture.

Example: Chernoff Bound

Let $\mathbf{z}^n = [\mathbf{z}_1, \ldots, \mathbf{z}_n]$ be drawn i.i.d. from distribution p. The event $\frac{1}{n}\sum_{i=1}^n f(\mathbf{z}_i) \geq \alpha$ can be rewritten as $\mathbf{z}^n$ taking an empirical distribution q such that $E_q[f] \geq \alpha$.

$$\Pr\left(\frac{1}{n}\sum_{i=1}^n f(\mathbf{z}_i) \geq \alpha\right) \approx \exp\left[-n \min_{q:\ E_q[f]\geq\alpha} D(q\|p)\right]$$

By the I-projection theorem, we know that $q \in \mathcal{E}_{f,p}$ i.e.,

$$q(z) = \frac{p(z)\cdot e^{\theta f(z)}}{k(\theta)}$$

where θ is chosen to satisfy $E_q[f] = \alpha$.

$$\begin{aligned}\text{Now} \quad D(q\|p) &= E_q\left[\log\frac{q(\mathbf{z})}{p(\mathbf{z})}\right] \\ &= E_q\left[\theta f(\mathbf{z})\right] - \log k(\theta) \\ &= \theta\alpha - \log\left(\sum_z p(z)\cdot e^{\theta f(z)}\right)\end{aligned}$$

which gives the same exponent as the familiar Chernoff bound. Thus showing that although an upper bound, Chernoff's bound is exponentially tight.

A. Binary hypothesis testing

In the rest of this section, we will focus on a particular kind of exponential family, the one connecting two given distributions. The first use of this kind of exponential family is for binary hypothesis testing.

Consider $\mathbf{z}^n$, drawn i.i.d from distribution p_0 under hypothesis H_0, and from p_1 under hypothesis H_1. The Neyman-Pearson test makes the decision by comparing the average log-likelihood ratio (LLR) $\frac{1}{n}\sum_{i=1}^n L(\mathbf{z}_i)$ (where $L(z) = \log\frac{p_1(z)}{p_0(z)}$) to a threshold α. The two types of error events have probability

$$\mathsf{Pr}(H_0 \to H_1) \approx \exp\left[-n \min_{q:\ E_q[L]\geq\alpha} D(q\|p_0)\right] \quad (7)$$

$$\mathsf{Pr}(H_1 \to H_0) \approx \exp\left[-n \min_{q:\ E_q[L]\leq\alpha} D(q\|p_1)\right] \quad (8)$$

I-projection implies that the optimum q for the two optimizations above, lie on $\mathcal{E}_{L,p_0}$ and $\mathcal{E}_{L,p_1}$, respectively. Since $L(\cdot)$ is the LLR function between p_1 and p_0, these two exponential families are in fact the same, which we write as

$$\mathcal{E}_{p_0,p_1} = \left\{p : p(z) = \frac{p_0(z)\exp[tL(z)]}{k(t)} = \frac{p_1^t(z)p_0^{1-t}(z)}{k(t)}\right\}$$

where t is the scalar exponential parameter $\theta \in \mathcal{R}$. Thus the solutions of (7) and (8) are indeed the same distribution p_{t^*}, where t^* is chosen such that $E_{p_{t^*}}[L] = \alpha$. The two exponents are then given by $D(p_{t^*}\|p_0)$ and $D(p_{t^*}\|p_1)$, respectively. For convenience, we usually limit the range of t to be within $[0,1]$. Clearly, $t=0$ corresponds to p_0 and $t=1$ to p_1. We can thus visualize the exponential family as a straight line connecting p_0 and p_1.

There are four quantities that are particularly important for this family:

- t is the exponential parameter
- $\eta \equiv E_{p_t}[L]$, is the average log-likelihood ratio corresponding to t
- $\psi \equiv \log k(t) = \log\sum_z p_1^t(z)p_0^{1-t}(z)$ is the log normalization factor
- $D(p_t\|p_0)$, the K-L divergence corresponding to t.

The following relations between these quantities (depicted in Figure 3) are easy to verify

$$\frac{\partial\psi}{\partial t} = \eta \quad (9)$$

$$\frac{\partial D(p_t\|p_0)}{\partial\eta} = t \quad (10)$$

$$t\cdot\eta = D(p_t\|p_0) + \psi \Rightarrow D(p_t\|p_0) = t\eta - \psi \quad (11)$$

Similarly, one gets $D(p_t\|p_1) = (t-1)\eta - \psi$ (12)

Note from (10) that the exponential parameter t signifies the sensitivity of divergence w.r.t. average log-likelihood ratio η.

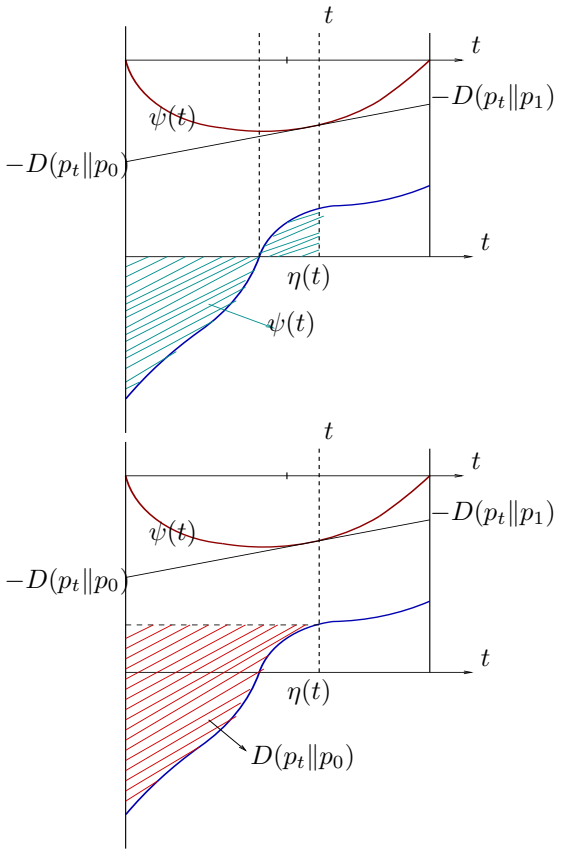


Fig. 3. The intersects of the tangent to ψ-curve, at $t=0$ and $t=1$, give exponents of the two errors in hypothesis testing. Blue region denotes $\psi(t)$—the integral of η w.r.t. t. The red region denotes $D(p_t\|p_0)$—the integral of t w.r.t η. Adding these two regions gives a rectangle of area $t\cdot\eta = \psi + D(p_t\|p_0)$.

Now recall the definition of Fisher information for this one-dimensional exponential family:

$$g_t = E_{p_t}\left[\left(\frac{\partial}{\partial t}\log p_t(\mathbf{z})\right)^2\right] = \text{variance of } L(\mathbf{z})$$

One can check that derivative of $\eta(t)$ curve equals g_t.

$$\frac{\partial^2\psi}{\partial t^2} = \frac{\partial\eta}{\partial t} = \frac{\partial}{\partial t}\left[\sum_z p_t(z)L(z)\right] = g_t$$

and similarly $\frac{\partial^2 D(p_t\|p_0)}{\partial\eta^2} = \frac{\partial}{\partial\eta}t = 1/g_t$

This gives a simple relation between $D(p_t\|p_0)$ and the Fisher information as

$$D(p_t\|p_0) = \int\int \frac{1}{g_{\tilde{t}}} d\tilde{\eta}^2 = \int_0^t s g_s ds \quad (13)$$

$$\text{Similarly, we get } D(p_t\|p_1) = \int_t^1 (1-s)g_s ds \quad (14)$$

Remarks: Similar to the results in [11], (13) gives a relation between an information theoretic quantity and an estimation theoretic quantity. However, this result involves a double integral. In fact, we believe there is no close connection between the two results. We have found that (13) also has an operational meaning in terms of a multi-layered source code.

The simplest case of (13) is when g_t remains constant along $\mathcal{E}_{p_0,p_1}$. This is a good approximation when p_0 is very close to p_1, which corresponds to a very noisy hypothesis testing problems. Recalling the definition of a metric connection in (5), one can thus think of the very noisy approximation as approximating the exponential connection as a Riemannian connection. Not surprisingly, this simplifies the problem. In this case, the double integral gives a simple quadratic relation

$$D(p_t\|p_0) = \frac{1}{2}gt^2, \qquad D(p_t\|p_1) = \frac{1}{2}g(1-t)^2 \tag{15}$$

We will revisit this relation when deriving the error exponent for the very noisy channels.

B. Error exponents for source coding

We now apply the I-projection theorem to the simple problem of error exponent for fixed length source coding. Let $\mathbf{z}^n$ be drawn i.i.d. with distribution P. Error can happen only when the empirical distribution of $\mathbf{z}^n$ (also called as its '*type*') is Q such that $H(Q) \geq R$. All other sequences can be encoded correctly with rate R, as their total number is $\approx \exp(nR)$. Hence by Sanov's theorem, source coding error exponent $E(R)$ is

$$\begin{aligned} E(R) &= \min_{Q:\ H(Q)\geq R} D(Q\|P) && (16)\\ &= \min_{Q:\ \log|\mathcal{Z}|-D(Q\|U)\geq R} D(Q\|P) && (17)\\ &= \min_{Q:\ D(Q\|U)\leq \log|\mathcal{Z}|-R\equiv\hat{R}} D(Q\|P) && (18)\end{aligned}$$

where U denotes the uniform distribution on the source alphabet $\mathcal{Z}$. Here I-projection can be used, although this is not a standard I-projection problem of projecting a distribution on a linear family (it is projecting P on a 'sphere' around U as shown in Fig. II-B).

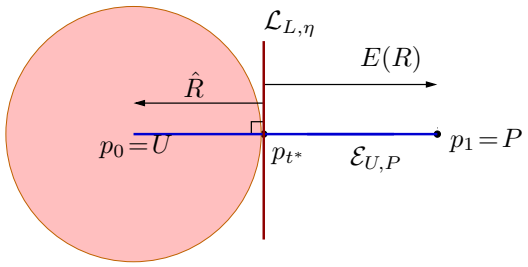


Fig. 4. Using I-projection for projecting on a 'sphere'

Consider a feasible distribution Q (satisfying $D(Q\|U) \leq \hat{R}$), which lies outside the exponential family $\mathcal{E}_{U,P}$ connecting $p_0 = U$ to $p_1 = P$. Let its average log-likelihood ratio (LLR) be $\mathrm{E}_Q[L] = \eta$, i.e. $Q \in \mathcal{L}_{L,\eta}$. Let $p_t \in \mathcal{E}_{U,P}$ such that it also lies on $\mathcal{L}_{L,\eta}$. Now pulling Q to p_t reduces its divergence from both U and P due to I-projection theorem. Hence the minimum in (18) must be attained at $p_{t^*} \in \mathcal{E}_{U,P}$ such that $D(p_{t^*}\|U) = \hat{R}$ and $D(p_{t^*}\|P) = E(R)$. Thus 'distance' i.e. divergence from one end U is a function of the rate R and that from the other end P gives the error exponent. Thus the exponential family clarifies how exactly increasing R (i.e. reducing $\hat{R}$) causes an increase in $E(R)$.

III. Random Coding exponent for DMC

Now as a more advanced application of I-projection, we consider the problem of random coding exponent for a DMC. Error exponents of discrete memoryless channels with random coding were analyzed in the classic work of Gallager [6] and later in [5], [7] and others. More recently, [9] derived these results using large deviation theory and Lagrange multipliers. We use the same random i.i.d. coding formulation[1] as in [9].

A random i.i.d. code of length n and rate R (nats/symbol) consists of $\exp(nR)$ codewords of length n. For transmitting message $k \in \{1, 2\cdots, \exp(nR)\}$, the codeword k denoted by $\mathbf{x}^n(k)$ is transmitted. Symbols of every codeword $\mathbf{x}^n(k)$ are chosen i.i.d. with distribution $P_{\mathbf{x}}$. Output of the channel takes values from the finite set $\mathcal{Y}$. The channel transition probability is denoted by $P_{\mathbf{y}|\mathbf{x}}$, that is, $P_{\mathbf{y}|\mathbf{x}}(y|x)$ specifies the probability of observing output $y \in \mathcal{Y}$ given input $x \in \mathcal{X}$.

Without loss of generality, we assume that message 1 was transmitted. Channel memoryless implies that distribution of output sequence $\mathbf{y}^n$ conditioned on the input $\mathbf{x}^n(1)$ is

$$P(\mathbf{y}^n|\ \mathbf{x}^n(1)) = \prod_{i=1}^{n} P_{\mathbf{y}|\mathbf{x}}(\mathbf{y}_i|\mathbf{x}_i(1)) \tag{19}$$

$$\Rightarrow \quad P(\mathbf{y}^n, \mathbf{x}^n(1)) = \prod_{i=1}^{n} P_{\mathbf{x}\mathbf{y}}(\mathbf{x}_i(1), \mathbf{y}_i) \tag{20}$$

where $P_{\mathbf{x}\mathbf{y}}$ denotes the joint distribution $P_{\mathbf{x}}P_{\mathbf{y}|\mathbf{x}}$. Last step followed because symbols in $\mathbf{x}^n(1)$ are generated i.i.d. with distribution $P_{\mathbf{x}}$. Hence the pair $(\mathbf{x}^n(1), \mathbf{y}^n)$ of the correct codeword and the output sequence is an i.i.d. sequence generated by distribution $P_{\mathbf{x}\mathbf{y}}$. Let the corresponding marginal distribution of $\mathbf{y}$ be denoted by $P_{\mathbf{y}}$

$$P_{\mathbf{y}}(y) = \sum_{x\in\mathcal{X}} P_{\mathbf{x}\mathbf{y}}(x, y)$$

We can also write $P_{\mathbf{x}\mathbf{y}}$ as $P_{\mathbf{y}}P_{\mathbf{x}|\mathbf{y}}$, where $P_{\mathbf{x}|\mathbf{y}} = \frac{P_{\mathbf{x}}P_{\mathbf{y}|\mathbf{x}}}{P_{\mathbf{y}}}$ denotes the reverse channel from $\mathbf{y}$ to $\mathbf{x}$.

Since the codewords are generated independently, the output sequence is independent of any incorrect codeword $\mathbf{x}^n(j)$, where $j \neq 1$. Hence the pair $(\mathbf{x}^n(j), \mathbf{y}^n)$ of the incorrect codeword and the output sequence is an i.i.d. sequence generated by distribution $P_{\mathbf{x}}P_{\mathbf{y}}$.

$$P(\mathbf{x}^n(j),\ \mathbf{y}^n) = \prod_{i=1}^{n} P_{\mathbf{x}}(\mathbf{x}_i(j))\ P_{\mathbf{y}}(\mathbf{y}_i)$$

A. Error exponent conditioned on the output type

We now analyze the error probability when the output sequence $\mathbf{y}^n$ has type $Q_{\mathbf{y}}$. This analysis will give us the error exponent $E_r(R, Q_{\mathbf{y}})$ at rate R conditioned on the output type $Q_{\mathbf{y}}$. Since the number of output types is polynomial in n, the overall error exponent can be obtained later by a minimization over $Q_{\mathbf{y}}$.

[1] Although we only consider random i.i.d. codes in this paper, error exponents for randomly chosen fixed composition codes can be also obtained on similar lines using I-projection. Also, the error exponents for List-of-L decoding can be obtained on these lines.

When the received output type is $Q_{\mathbf{y}}$, our space of possible joint $(\mathbf{x},\mathbf{y})$-types is all the distributions which ensure that marginal distribution of $\mathbf{y}$ is $Q_{\mathbf{y}}$. It is easy to check that this space of distributions is a linear family. With little abuse of notation, we denote this family by $\mathcal{L}_{Q_{\mathbf{y}}}$. Any point in this family has the form $Q_{\mathbf{y}}W_{\mathbf{x}|\mathbf{y}}$, for some reverse channel type $W_{\mathbf{x}|\mathbf{y}}$. Divergence between a point in this family and the distribution $P_{\mathbf{xy}}$ (related to correct input-output pair) is

$$\begin{aligned} & D(Q_{\mathbf{y}}W_{\mathbf{x}|\mathbf{y}}\|P_{\mathbf{xy}}) = && (21)\\ & \sum_{\mathbf{x},\mathbf{y}} Q_{\mathbf{y}}(y)W_{\mathbf{x}|\mathbf{y}}(x|y)\log\frac{Q_{\mathbf{y}}(y)W_{\mathbf{x}|\mathbf{y}}(x|y)}{P_{\mathbf{xy}}(x,y)} && (22)\\ = & \sum_{\mathbf{x},\mathbf{y}} Q_{\mathbf{y}}(y)W_{\mathbf{x}|\mathbf{y}}(x|y)\log\frac{Q_{\mathbf{y}}(y)}{P_{\mathbf{y}}(y)}\frac{W_{\mathbf{x}|\mathbf{y}}(x|y)}{P_{\mathbf{x}|\mathbf{y}}(x|y)} && (23)\\ = & D(Q_{\mathbf{y}}\|P_{\mathbf{y}}) + D(Q_{\mathbf{y}}W_{\mathbf{x}|\mathbf{y}}\|Q_{\mathbf{y}}P_{\mathbf{x}|\mathbf{y}}) && (24)\\ \geq & D(Q_{\mathbf{y}}\|P_{\mathbf{y}}) && (25)\end{aligned}$$

The last step is met with equality when $W_{\mathbf{x}|\mathbf{y}} = P_{\mathbf{x}|\mathbf{y}}$. Hence, projection of $P_{\mathbf{xy}}$ on this linear family is given by $Q_{\mathbf{y}}P_{\mathbf{x}|\mathbf{y}}$. Thus only the marginal distribution is changed from $P_{\mathbf{y}}$ to $Q_{\mathbf{y}}$ but the reverse channel type is the same as $P_{\mathbf{x}|\mathbf{y}}$.

On similar lines, divergence between a point in this family and the distribution $P_{\mathbf{x}}P_{\mathbf{y}}$ (corresponding to incorrect input and output pair) equals

$$D(Q_{\mathbf{y}}W_{\mathbf{x}|\mathbf{y}}\|P_{\mathbf{y}}P_{\mathbf{x}}) = D(Q_{\mathbf{y}}\|P_{\mathbf{y}}) + D(Q_{\mathbf{y}}W_{\mathbf{x}|\mathbf{y}}\|Q_{\mathbf{y}}P_{\mathbf{x}}) \quad (26)$$

Thus the projection of $P_{\mathbf{y}}P_{\mathbf{x}}$ on $\mathcal{L}_{Q_{\mathbf{y}}}$ is given by changing the $\mathbf{y}$-marginal to $Q_{\mathbf{y}}$ and keeping the reverse channel type the same as $P_{\mathbf{x}}$.

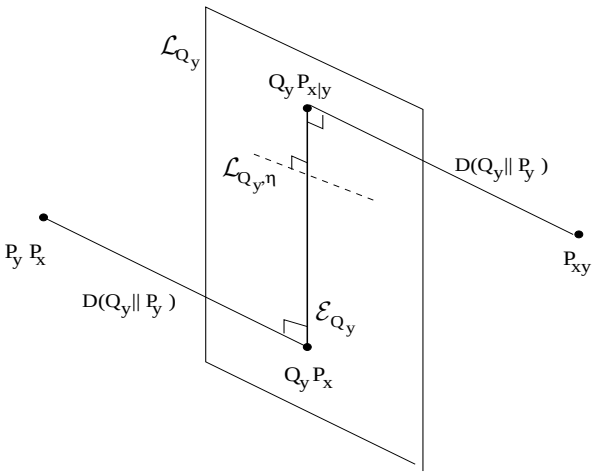


Fig. 5. Linear family $\mathcal{L}_{Q_{\mathbf{y}}}$ having marginal distribution $Q_{\mathbf{y}}$ in the space of joint distributions on $(\mathbf{x},\mathbf{y})$.

Now we show that Maximum-Likelihood decoder can also be thought as a Maximum-LLR decoder. It is because for given output sequence $\mathbf{y}^n$ of type $Q_{\mathbf{y}}$, the decoded message $\hat{m}$ by the ML decoder is

$$\begin{aligned}\hat{m} &= \arg\max_{1\leq k\leq e^{nR}} P(\mathbf{y}^n|\mathbf{x}^n(k))\\ &= \arg\max_{1\leq k\leq e^{nR}} \sum_{i=1}^{n}\log P_{\mathbf{y}|\mathbf{x}}(\mathbf{y}_i|\mathbf{x}_i(k)) \quad \text{(memorylessness)}\\ &= \arg\max_{1\leq k\leq e^{nR}} \sum_{i=1}^{n}\log\frac{P_{\mathbf{x}|\mathbf{y}}(\mathbf{x}_i(k)|\mathbf{y}_i)}{P_{\mathbf{x}}(\mathbf{x}_i(k))} \qquad \text{(Baye's rule)}\\ &= \arg\max_{1\leq k\leq e^{nR}} \sum_{i=1}^{n}\log\frac{P_{\mathbf{x}|\mathbf{y}}(\mathbf{x}_i(k)|\mathbf{y}_i)}{P_{\mathbf{x}}(\mathbf{x}_i(k))}\frac{Q_{\mathbf{y}}(\mathbf{y}_i)}{Q_{\mathbf{y}}(\mathbf{y}_i)}\end{aligned}$$

Dividing this by n gives

$$\hat{m} = \arg\max_{1\leq k\leq e^{nR}} \sum_{x,y} Q_{\mathbf{y}}(y)W^k_{\mathbf{x}|\mathbf{y}}(x|y)\log\frac{Q_{\mathbf{y}}(y)P_{\mathbf{x}|\mathbf{y}}(x|y)}{Q_{\mathbf{y}}(y)P_{\mathbf{x}}(x)}$$

where $Q_{\mathbf{y}}W^k_{\mathbf{x}|\mathbf{y}}$ denotes the joint type of the k'th codeword and output sequence $(\mathbf{x}^n(k),\mathbf{y}^n)$. Thus ML decoding is equivalent to decoding the codeword which gives largest normalized log-likelihood-ratio between the two distributions $Q_{\mathbf{y}}P_{\mathbf{x}|\mathbf{y}}$ and $Q_{\mathbf{y}}P_{\mathbf{x}}$. Recalling the notation in previous section, let the joint distribution $Q_{\mathbf{y}}P_{\mathbf{x}|\mathbf{y}}$ be denoted by p_1 and $Q_{\mathbf{y}}P_{\mathbf{x}}$ be denoted by p_0. Their log-likelihood ratio be denoted by $L(\mathbf{x},\mathbf{y})$.

$$L(\mathbf{x},\mathbf{y}) = \log\frac{p_1(\mathbf{x},\mathbf{y})}{p_0(\mathbf{x},\mathbf{y})} = \log\frac{P_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})}{p_{\mathbf{x}}(\mathbf{x})}$$

The LLR decoder chooses the codeword with the largest LLR score.

$$\hat{m} = \arg\max_{1\leq k\leq e^{nR}} \mathrm{E}_{Q_{\mathbf{y}}W^k_{\mathbf{x}|\mathbf{y}}}[L(\mathbf{x},\mathbf{y})] \equiv \arg\max_{1\leq k\leq e^{nR}} S_k$$

where $W^k_{\mathbf{x}|\mathbf{y}}$ is the reverse channel type for pair $(\mathbf{x}^n(k),\mathbf{y}^n)$ and $S_k \equiv \mathrm{E}_{Q_{\mathbf{y}}W^k_{\mathbf{x}|\mathbf{y}}}[L(\mathbf{x},\mathbf{y})]$ is the LLR score of the k'th codeword. Note that all scores are random variables depending on the channel noise and codewords. Error happens if score of S_1 of the correct codeword is less than score S_j of any incorrect codeword. To analyze the error exponent, we should know the distribution of these score variables. We need to find the nature of the dominant reverse channel type $W^k_{\mathbf{x}|\mathbf{y}}$ which causes error.

Consider a sub-family of $\mathcal{L}_{Q_{\mathbf{y}}}$ where the expectation of $L(\mathbf{x},\mathbf{y})$ equals η. This is a linear family within $\mathcal{L}_{Q_{\mathbf{y}}}$ and we denote it by $\mathcal{L}_{Q_{\mathbf{y}},\eta}$. It corresponds to the dashed line in Fig. 5. Applying I-projection theorem, we get that projection of p_0 (or p_1) on $\mathcal{L}_{Q_{\mathbf{y}},\eta}$ is given by $Q_{\mathbf{y}}P^{(t)}_{\mathbf{x}|\mathbf{y}}$ for some $t\in[0,1]$, where the reverse channel $P^{(t)}_{\mathbf{x}|\mathbf{y}}$ for any given $y\in\mathcal{Y}$ is

$$\begin{aligned} P^{(t)}_{\mathbf{x}|\mathbf{y}}(x|y) &= \frac{P^t_{\mathbf{x}|\mathbf{y}}(x|y)P^{1-t}_{\mathbf{x}}(x)}{k_y(t)} && (27)\\ \text{where} \quad k_y(t) &= \sum_{x\in\mathcal{X}} P^t_{\mathbf{x}|\mathbf{y}}(x|y)P^{1-t}_{\mathbf{x}}(x) && (28)\end{aligned}$$

Superscript of $P^{(t)}_{\mathbf{x}|\mathbf{y}}(x|y)$ is in bracket to distinguish it from $P^t_{\mathbf{x}|\mathbf{y}}(x|y)$, the t^{th} power of $P_{\mathbf{x}|\mathbf{y}}(x|y)$. Conditioned on output type $Q_{\mathbf{y}}$, an incorrect codeword is generated i.i.d. with distribution $P_{\mathbf{x}}$. Hence applying Stein's lemma and Sanov's

theorem gives the following exponent for the score S_j of a wrong codeword being η (conditioned on output type $Q_{\mathbf{y}}$). It is obtained by optimizing the reverse channel type $W^j_{\mathbf{x}|\mathbf{y}}$.

$$\lim_{n\to\infty} -\frac{\log P(S_j \geq \eta|\ Q_{\mathbf{y}})}{n} = \tag{29}$$

$$\min_{W^j_{\mathbf{x}|\mathbf{y}}} D(Q_{\mathbf{y}} W^j_{\mathbf{x}|\mathbf{y}} \| p_0) = D(Q_{\mathbf{y}} P^{(t)}_{\mathbf{x}|\mathbf{y}} \| p_0) \tag{30}$$

$$= D(Q_{\mathbf{y}} P^{(t)}_{\mathbf{x}|\mathbf{y}} \| Q_{\mathbf{y}} P_{\mathbf{x}}) \tag{31}$$

$$\text{where } t \text{ satisifes } \eta = \mathrm{E}_{Q_{\mathbf{y}} P^{(t)}_{\mathbf{x}|\mathbf{y}}}[L(\mathbf{x},\mathbf{y})] \ \equiv \eta_{Q_{\mathbf{y}}}(t) \tag{32}$$

Thus the dominating manner in which a wrong codeword's score S_j crosses η is when the corresponding reverse channel type $W^j_{\mathbf{x}|\mathbf{y}}$ lies on the exponential family $\mathcal{E}_{Q_{\mathbf{y}}}$ defined as

$$\mathcal{E}_{Q_{\mathbf{y}}} = \{Q_{\mathbf{y}} P^{(t)}_{\mathbf{x}|\mathbf{y}} \text{ for } t \in [0,1]\}$$

This family connects $Q_{\mathbf{y}} P_{\mathbf{x}|\mathbf{y}}$ and $Q_{\mathbf{y}} P_{\mathbf{x}}$ within $\mathcal{L}_{Q_{\mathbf{y}}}$ (see Fig. 5).

Similar steps can be repeated for the score S_1 of the correct codeword. Conditioned on the output type $Q_{\mathbf{y}}$, symbols of the correct codeword are chosen i.i.d. with distribution $P_{\mathbf{x}|\mathbf{y}}$. Again apply Sanov's theorem and Stein's lemma to optimize the reverse channel type $W^1_{\mathbf{x}|\mathbf{y}}$ for correct codeword.

$$\lim_{n\to\infty} -\frac{\log P(S_1 \leq \eta_{Q_{\mathbf{y}}}(t)|\ Q_{\mathbf{y}})}{n} = \tag{33}$$

$$\min_{W^1_{\mathbf{x}|\mathbf{y}}} D(Q_{\mathbf{y}} W^1_{\mathbf{x}|\mathbf{y}} \| p_1) = D(Q_{\mathbf{y}} P^{(t)}_{\mathbf{x}|\mathbf{y}} \| p_1) \tag{34}$$

$$= D(Q_{\mathbf{y}} P^{(t)}_{\mathbf{x}|\mathbf{y}} \| Q_{\mathbf{y}} P_{\mathbf{x}|\mathbf{y}}) \tag{35}$$

Thus the dominating manner in which the correct codeword's score S_1 is smaller than η is when the its reverse channel type $W^1_{\mathbf{x}|\mathbf{y}}$ is on the same exponential family $\mathcal{E}_{Q_{\mathbf{y}}}$.

By union bound, the exponent of probability of any wrong codeword's score crossing the threshold $\eta_{Q_{\mathbf{y}}}(t)$ is given by $[D(Q_{\mathbf{y}} P^{(t)}_{\mathbf{x}|\mathbf{y}} \| Q_{\mathbf{y}} P_{\mathbf{x}}) - R]^+$, where $[\mathsf{x}]^+ = \mathsf{x} \cdot \mathbf{1}_{\{\mathsf{x}>0\}}$. Since all codewords, are drawn independently of each other, the exponent $E(t, R, Q_{\mathbf{y}})$ for the joint probability of $S_1 < \eta_{Q_{\mathbf{y}}}(t)$ and $S_j \geq \eta_{Q_{\mathbf{y}}}(t)$ (for some $j \neq 1$) is sum of the exponents for these independent events.

$$E(t, R, Q_{\mathbf{y}}) \tag{36}$$

$$\equiv -\lim_{n\to\infty} \frac{\log P(S_1 \leq \eta_{Q_{\mathbf{y}}}(t),\ S_j \geq \eta_{Q_{\mathbf{y}}}(t) \mid Q_{\mathbf{y}})}{n} \tag{37}$$

$$= D(Q_{\mathbf{y}} P^{(t)}_{\mathbf{x}|\mathbf{y}} \| Q_{\mathbf{y}} P_{\mathbf{x}|\mathbf{y}}) + [D(Q_{\mathbf{y}} P^{(t)}_{\mathbf{x}|\mathbf{y}} \| Q_{\mathbf{y}} P_{\mathbf{x}}) - R]^+ \tag{38}$$

The error exponent $E_r(R, Q_{\mathbf{y}})$ conditioned on $Q_{\mathbf{y}}$ is obtained by minimizing the above expression over the LLR $\eta_{Q_{\mathbf{y}}}(t)$ or equivalently minimizing it over t. This minimization corresponds to finding the LLR which dominates the error event conditioned on $Q_{\mathbf{y}}$.

Using (11) and (12), we can prove the following:

$$D(Q_{\mathbf{y}} P^{(t)}_{\mathbf{x}|\mathbf{y}} \| Q_{\mathbf{y}} P_{\mathbf{x}}) = D(Q_{\mathbf{y}} P^{(t)}_{\mathbf{x}|\mathbf{y}} \| p_0) = t\eta_{Q_{\mathbf{y}}}(t) - \psi_{Q_{\mathbf{y}}}(t) \tag{39}$$

$$\text{where } \ \psi_{Q_{\mathbf{y}}}(t) \equiv \sum_{y\in\mathcal{Y}} Q_{\mathbf{y}}(y) \log k_y(t) \tag{40}$$

$$\text{Similarly, } \ D(Q_{\mathbf{y}} P^{(t)}_{\mathbf{x}|\mathbf{y}} \| Q_{\mathbf{y}} P_{\mathbf{x}|\mathbf{y}}) = D(Q_{\mathbf{y}} P^{(t)}_{\mathbf{x}|\mathbf{y}} \| p_1) \tag{41}$$

$$= (t-1)\eta_{Q_{\mathbf{y}}}(t) - \psi_{Q_{\mathbf{y}}}(t) \tag{42}$$

Recall that for a given y, $k_y(t)$ is the normalization constant for the reverse channel $P^{(t)}_{\mathbf{x}|\mathbf{y}}$.

Now let us minimize $E(t, R, Q_{\mathbf{y}})$ over t to obtain $E_r(R, Q_{\mathbf{y}})$. Let $\hat{t}$ be the solution to equation

$$D(Q_{\mathbf{y}} P^{(\hat{t})}_{\mathbf{x}|\mathbf{y}} \| Q_{\mathbf{y}} P_{\mathbf{x}}) = D(Q_{\mathbf{y}} P^{(\hat{t})}_{\mathbf{x}|\mathbf{y}} \| p_0) = R \tag{43}$$

For any $t < \hat{t}$, the exponent $E(t, R, Q_{\mathbf{y}}) > E(\hat{t}, R, Q_{\mathbf{y}})$. Hence the optimum solution lies in $[\hat{t}, 1]$. Since $D(Q_{\mathbf{y}} P^{(t)}_{\mathbf{x}|\mathbf{y}} \| Q_{\mathbf{y}} P_{\mathbf{x}}) \geq R$ for t in this range,

$$[D(Q_{\mathbf{y}} P^{(t)}_{\mathbf{x}|\mathbf{y}} \| Q_{\mathbf{y}} P_{\mathbf{x}}) - R]^+ = D(Q_{\mathbf{y}} P^{(t)}_{\mathbf{x}|\mathbf{y}} \| Q_{\mathbf{y}} P_{\mathbf{x}}) - R$$

$$\Rightarrow \quad E(t, R, Q_{\mathbf{y}}) = D(Q_{\mathbf{y}} P^{(t)}_{\mathbf{x}|\mathbf{y}} \| Q_{\mathbf{y}} P_{\mathbf{x}|\mathbf{y}}) + D(Q_{\mathbf{y}} P^{(t)}_{\mathbf{x}|\mathbf{y}} \| Q_{\mathbf{y}} P_{\mathbf{x}}) - R$$

$$= (2t-1)\eta_{Q_{\mathbf{y}}}(t) - 2\psi_{Q_{\mathbf{y}}}(t) - R \qquad \text{(from (39) and (42))}$$

Differentiating this w.r.t. t and equating it to 0 gives,

$$(2t-1) g_{Q_{\mathbf{y}}}(t) + 2\eta_{Q_{\mathbf{y}}}(t) - 2\psi'_{Q_{\mathbf{y}}}(t) = 0$$

$$\Rightarrow \qquad (2t-1) g_{Q_{\mathbf{y}}}(t) = 0 \qquad \text{(because } \psi'_{Q_{\mathbf{y}}}(t) = \eta_{Q_{\mathbf{y}}}(t))$$

where $\psi'_{Q_{\mathbf{y}}}$ denotes derivative of $\psi_{Q_{\mathbf{y}}}$ w.r.t. t. Since the Fisher information $g_{Q_{\mathbf{y}}}(t)$ is strictly positive, the optimum $t^* = 1/2$ provided $1/2 \in [\hat{t}, 1]$. Otherwise if $1/2 < \hat{t}$, then $E(t, R, Q_{\mathbf{y}})$ is strictly increasing in $[\hat{t}, 1]$ because its derivative $(2t-1) g_{Q_{\mathbf{y}}}(t)$ is always positive. Then the optimum t^* equals $\hat{t}$. Thus in either case, the optimum $t^* \geq 1/2$.

This phenomenon reflects the union bound constraint $\rho \leq 1$ in Gallager's analysis. In fact the ρ in that analysis and t in this analysis are related as $t = 1/(1+\rho)$.

Thus we get, the error exponent conditioned on output type $Q_{\mathbf{y}}$

$$E_r(R, Q_{\mathbf{y}}) \tag{44}$$

$$= \ D(Q_{\mathbf{y}} P^{(\hat{t})}_{\mathbf{x}|\mathbf{y}} \| Q_{\mathbf{y}} P_{\mathbf{x}|\mathbf{y}}) \quad \text{if} \ \ \hat{t} \geq 1/2 \tag{45}$$

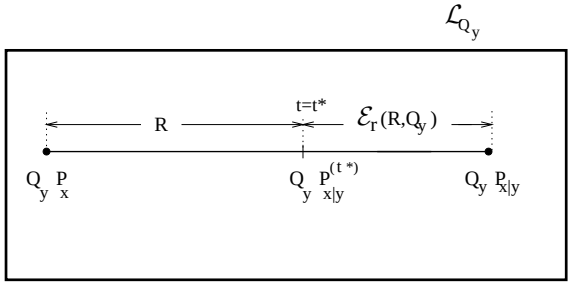
$$E_r(R, Q_{\mathbf{y}}) \quad (\text{if} \quad \hat{t} \leq 1/2), \tag{46}$$

$$= \ D(Q_{\mathbf{y}} P^{(1/2)}_{\mathbf{x}|\mathbf{y}} \| Q_{\mathbf{y}} P_{\mathbf{x}|\mathbf{y}}) + D(Q_{\mathbf{y}} P^{(1/2)}_{\mathbf{x}|\mathbf{y}} \| Q_{\mathbf{y}} P_{\mathbf{x}}) - R \tag{47}$$

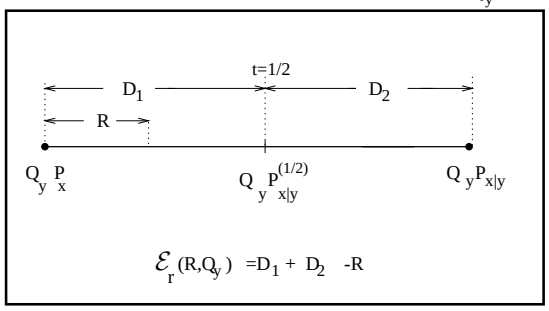
where $\hat{t}$ is the solution to $D(Q_{\mathbf{y}} P^{(\hat{t})}_{\mathbf{x}|\mathbf{y}} \| Q_{\mathbf{y}} P_{\mathbf{x}}) = R$. This solution is depicted in the Figure below.

Note from (43) that higher value of $\hat{t}$ means R is large and vice versa. Hence the case of $\hat{t} \geq 1/2$ corresponds to high (enough) rates R and vice versa. The above bound thus (re)derives the following phenomenon in [10]:

1) The dominant cause of error for high (enough) rates (i.e. $\hat{t} \geq 1/2$) is when a large number of incorrect codewords can be confused with the correct one.
2) The dominant cause of error for lower rates ($\hat{t} < 1/2$) is when a single incorrect codeword is confused with the correct one.



(a) $t^* = \hat{t} \geq 1/2$



(b) $t^* = 1/2 > \hat{t}$

Fig. 6. Geometric interpretation of error exponent conditioned on output type $Q_{\mathbf{y}}$. The rectangular frame in each figure denotes the linear family $\mathcal{L}_{Q_{\mathbf{y}}}$. The solid line in it represents the exponential family $\mathcal{E}_{Q_{\mathbf{y}}}$. The distance between two points corresponds to their divergence.

The expression in (45) also provides an upper bound to the actual random coding exponent $E_r(R, Q_{\mathbf{y}})$. This bound is related to the sphere-packing exponent. This expression is an upper bound because it is equivalent to relaxing the constraint $t^* \geq 1/2$ (or $\rho < 1$) by assuming t^* always equals $\hat{t}$.

Remark 1: Note that for each output letter y, the dominant reverse channel type $W_{\mathbf{x}|\mathbf{y}=y}$ lies on the exponential family (in the space of distributions on $\mathcal{X}$) connecting $P_{\mathbf{x}}$ and $P_{\mathbf{x}|\mathbf{y}=y}$. Analysis in this section trivially shows the following coupling phenomenon between the dominant reverse channel type for all output letters y. The exponential parameter t is the same for the reverse channel type from each letter y, thus causing a coupling between these reverse channel-types. Thus the dominant reverse channels for all output letters are equally tilted.

B. Optimizing over output type

Previously we found the error exponent conditioned on the output type $Q_{\mathbf{y}}$. Since the output sequence is generated i.i.d. according to $P_{\mathbf{y}}$, the exponent of observing type $Q_{\mathbf{y}}$ is given by $D(Q_{\mathbf{y}} \| P_{\mathbf{y}})$. Hence the overall exponent of error corresponding to output type $Q_{\mathbf{y}}$ is given by

$$E_r(R, Q_{\mathbf{y}}) + D(Q_{\mathbf{y}} \| P_{\mathbf{y}})$$

The effective error exponent is obtained by minimizing the above expression over all $Q_{\mathbf{y}}$.

$$E_r(R) = \min_{Q_{\mathbf{y}}} \; D(Q_{\mathbf{y}} \| P_{\mathbf{y}}) + E_r(R, Q_{\mathbf{y}}) \tag{48}$$

Let the optimum (or dominating) $Q_{\mathbf{y}}$ be denoted by $Q^*_{\mathbf{y}}$. For a symmetric channel (like the BSC) with uniform input distribution $P_{\mathbf{x}}$, conditional error exponent $E_r(R, Q_{\mathbf{y}})$ is independent of $Q_{\mathbf{y}}$. Hence the $Q^*_{\mathbf{y}}$ equals $P_{\mathbf{y}}$. Thus previous subsection is enough to understand the symmetric channel case.

However, for non-symmetric channels, the optimum $Q^*_{\mathbf{y}}$ need not be simply $P_{\mathbf{y}}$. Let the joint type which dominates the error event be given by $Q^*_{\mathbf{y}} P^{(t^*)}_{\mathbf{x}|\mathbf{y}}$, where $Q^*_{\mathbf{y}}$ optimizes (48) and conditioned on $Q^*_{\mathbf{y}}$, the reverse channel type $P^{(t^*)}_{\mathbf{x}|\mathbf{y}}$ achieves $E(R, Q^*_{\mathbf{y}})$ in (47,45). We saw previously that dominating error event conditioned on the output type happens when the reverse channel type lies on the exponential family $\{P^{(t)}_{\mathbf{x}|\mathbf{y}} \text{ for } t \in [0,1]\}$ of reverse channels. This family connects the trivial reverse channel $P_{\mathbf{x}}$ and actual reverse channel $P_{\mathbf{x}|\mathbf{y}}$. It turns out that the dominating output type $Q^*_{\mathbf{y}}$ also has such interpretation in terms of a certain exponential family. Refer to appendix for a simple proof based on I-projection.

Theorem 2: Consider the exponential family connecting $p_1 = P_{\mathbf{xy}}$ to $p_0 = Q^*_{\mathbf{y}} P_{\mathbf{x}}$. The joint type $Q^*_{\mathbf{y}} P^{(t^*)}_{\mathbf{x}|\mathbf{y}}$ dominating the error event lies on this exponential family (see Fig. 7).

$$\begin{aligned} Q^*_{\mathbf{y}}(y) P^{(t^*)}_{\mathbf{x}|\mathbf{y}}(x|y) &= \frac{p_1^{t^*}(x,y) p_0^{1-t^*}(x,y)}{k(t^*)} \\ \text{where } \; k(t^*) &= \sum_{x,y} p_1^{t^*}(x,y) p_0^{1-t^*}(x,y) \end{aligned}$$

Thus although $Q^*_{\mathbf{y}}$ need not be the $\mathbf{y}$-marginal throughout this exponential family, it is indeed the $\mathbf{y}$-marginal at the optimum t^* on this exponential family. Also note that reverse channel at any t on this exponential family is the same as $P^{(t)}_{\mathbf{x}|\mathbf{y}}$ seen before. Recall that $P^{(t)}_{\mathbf{x}|\mathbf{y}}$ was the reverse channel seen in previous subsection (where $\mathbf{y}$-marginal was fixed to $Q_{\mathbf{y}}$) for exponential family $\mathcal{E}_{Q_{\mathbf{y}}}$ connecting $Q_{\mathbf{y}} P_{\mathbf{x}}$ and $Q_{\mathbf{y}} P_{\mathbf{x}|\mathbf{y}}$ within $\mathcal{L}_{Q_{\mathbf{y}}}$.

Recalling $P^{(t^*)}_{\mathbf{x}|\mathbf{y}}(x|y) = \frac{P^{t^*}_{\mathbf{x}|\mathbf{y}}(x|y) P^{1-t^*}_{\mathbf{x}}(x)}{k_y(t^*)}$ from (27) and plugging this in the above Theorem gives

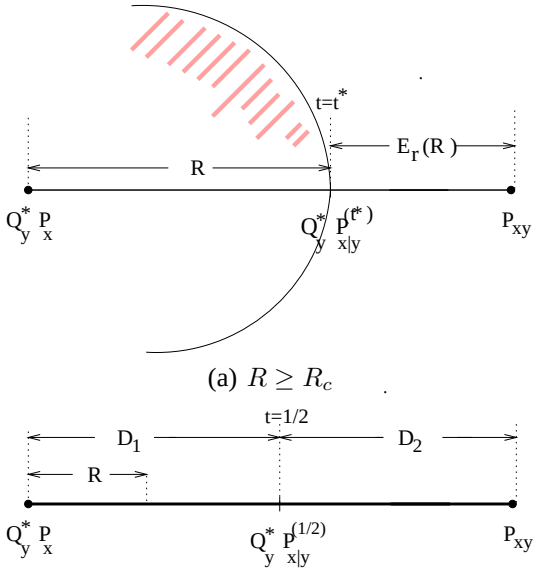
$$\begin{aligned} \left(Q^*_{\mathbf{y}}(y)\right)^{t^*} &\propto (P_{\mathbf{y}}(y))^{t^*} k_y(t^*) && (49) \\ \Rightarrow \quad Q^*_{\mathbf{y}}(y) &\propto P_{\mathbf{y}}(y) k_y^{1/t^*}(t^*) && (50) \\ &= P_{\mathbf{y}} \left(\sum_x P^{t^*}_{\mathbf{x}|\mathbf{y}}(x|y) P^{1-t^*}_{\mathbf{x}}(x) \right)^{1/t^*} && (51) \\ \text{(by Baye's rule)} \qquad &= \left(\sum_x P_{\mathbf{x}}(x) P^{t^*}_{\mathbf{y}|\mathbf{x}}(y|x) \right)^{1/t^*} && (52) \end{aligned}$$

This is the same solution as [9] for the dominant output type $Q^*_{\mathbf{y}}$ for error events. To emphasize the dependence of $Q^*_{\mathbf{y}}$ on t^*, let it be denoted by $Q^{(t^*)}_{\mathbf{y}}$. The optimum t^* equals $1/2$ for rates below the critical rate given by

$$R_c = D(Q^{(1/2)}_{\mathbf{y}} P^{(1/2)}_{\mathbf{x}|\mathbf{y}} \| Q^{(1/2)}_{\mathbf{y}} P_{\mathbf{x}}) \tag{53}$$

For higher rates, t^* is the solution to

$$D(Q^{(t^*)}_{\mathbf{y}} P^{(t^*)}_{\mathbf{x}|\mathbf{y}} \| Q^{(t^*)}_{\mathbf{y}} P_{\mathbf{x}}) = R \tag{54}$$



(a) $R \geq R_c$

(b) $R \leq R_c$

Fig. 7. Geometric interpretation of the dominant $Q_{\mathbf{y}}$ and the random coding exponent $E_r(R)$. Here plane of the paper represents the space of all joint distributions. The solid line in each figure represents the exponential family connecting its two ends. The distance between two points corresponds to their divergence. Figure (a) above reminds us of the figure for source coding error exponent.

Since t^* and $Q_{\mathbf{y}}^{(t^*)}$ are dependent on each other through (52) and (54), a closed for expression cannot be given for either of them. However, iterating between (52) and (54) should converge to the optimum t^* and $Q_{\mathbf{y}}^{(t^*)}$. This gives a Blahut-Arimoto-type algorithm for finding error exponent at $R > R_c$.

To summarize, the error exponent is given by

$$E_r(R) = D(Q_{\mathbf{y}}^{(1/2)} P_{\mathbf{x}|\mathbf{y}}^{(1/2)} \| P_{\mathbf{xy}}) + \quad (55)$$

$$D(Q_{\mathbf{y}}^{(1/2)} P_{\mathbf{x}|\mathbf{y}}^{(1/2)} \| Q_{\mathbf{y}}^{(1/2)} P_{\mathbf{x}}) - R \quad \text{for } R < R_c \quad (56)$$

$$= (Q_{\mathbf{y}}^{(t^*)} P_{\mathbf{x}|\mathbf{y}}^{(t^*)} \| P_{\mathbf{xy}}) \text{ for } R \geq R_c \quad (57)$$

C. Very noisy channels

We define a very noisy channel for which the distribution $P_{\mathbf{x}}$ is very close to $P_{\mathbf{x}|\mathbf{y}=y}$ for each given $y \in \mathcal{Y}$. Equivalently for each given $y \in \mathcal{Y}$, the Fisher information $g_y(t)$ is constant for the exponential family of $\mathcal{X}$-distributions joining $P_{\mathbf{x}|\mathbf{y}=y}$ and $P_{\mathbf{x}}$. Let this constant Fisher information for output y be denoted by g_y. For a given output type $Q_{\mathbf{y}}$ this implies,

$$D(Q_{\mathbf{y}} P_{\mathbf{x}|\mathbf{y}}^{(t)} \| Q_{\mathbf{y}} P_{\mathbf{x}}) = \sum_{y \in \mathcal{Y}} Q_{\mathbf{y}}(y) D(P_{\mathbf{x}|\mathbf{y}=y}^{(t)} \| P_{\mathbf{x}})$$

$$= \sum_{y \in \mathcal{Y}} Q_{\mathbf{y}}(y) \frac{g_y t^2}{2}$$

(denoting $p_0 = P_{\mathbf{x}}$ & $p_1 = P_{\mathbf{x}|\mathbf{y}=y}$ and using (15))

$$= t^2 \left(\sum_{y \in \mathcal{Y}} \frac{Q_{\mathbf{y}}(y) g_y}{2} \right)$$

$$\equiv t^2 C_{Q_{\mathbf{y}}}$$

where $C_{Q_{\mathbf{y}}}$ is a shorthand for $\left(\sum_{y \in \mathcal{Y}} \frac{Q_{\mathbf{y}}(y) g_y}{2} \right)$. Substituting $t = 1$ shows that $C_{Q_{\mathbf{y}}}$ equals $D(Q_{\mathbf{y}} P_{\mathbf{x}|\mathbf{y}} \| Q_{\mathbf{y}} P_{\mathbf{x}})$. Similarly using (15), we can show that

$$D(Q_{\mathbf{y}} P_{\mathbf{x}|\mathbf{y}}^{(t)} \| Q_{\mathbf{y}} P_{\mathbf{x}|\mathbf{y}}) = \sum_{y \in \mathcal{Y}} Q_{\mathbf{y}}(y) D(P_{\mathbf{x}|\mathbf{y}=y}^{(t)} \| P_{\mathbf{x}|\mathbf{y}=y}) = (1-t)^2 C_{Q_{\mathbf{y}}}$$

Recalling that the error exponent conditioned on $Q_{\mathbf{y}}$ for $R \geq R_c$ is given by

$$E_r(R, Q_{\mathbf{y}}) = D(Q_{\mathbf{y}} P_{\mathbf{x}|\mathbf{y}}^{(t)} \| Q_{\mathbf{y}} P_{\mathbf{x}|\mathbf{y}}) = (1-t)^2 C_{Q_{\mathbf{y}}}$$

where t satisfies $R = D(Q_{\mathbf{y}} P_{\mathbf{x}|\mathbf{y}}^{(t)} \| Q_{\mathbf{y}} P_{\mathbf{x}}) = t^2 C_{Q_{\mathbf{y}}}$

Eliminating t from the two equations gives,

$$E_r(R, Q_{\mathbf{y}}) = C_{Q_{\mathbf{y}}} \left(1 - \sqrt{R/C_{Q_{\mathbf{y}}}} \right)^2 = (\sqrt{C_{Q_{\mathbf{y}}}} - \sqrt{R})^2$$

The overall error exponent is obtained by optimizing over the output type

$$E_r(R) = \min_{Q_{\mathbf{y}}} D(Q_{\mathbf{y}} \| P_{\mathbf{y}}) + (\sqrt{C_{Q_{\mathbf{y}}}} - \sqrt{R})^2$$

Choosing $Q_{\mathbf{y}} = P_{\mathbf{y}}$ for capacity achieving $P_{\mathbf{y}}$ gives the approximate[2] solution in [6].

Since $E_r(R, Q_{\mathbf{y}})$ depends on $Q_{\mathbf{y}}$ only through its effect on $C_{Q_{\mathbf{y}}} = \mathrm{E}_{Q_{\mathbf{y}}}[g_y/2]$, the optimum $Q_{\mathbf{y}}^*$ will be on the exponential family of $\mathcal{Y}$-distributions going through $P_{\mathbf{y}}$ corresponding to the function $f(y) = g_y (= 2D(P_{\mathbf{x}|\mathbf{y}=y} \| P_{\mathbf{x}}))$. This again follows by the I-projection theorem. This family is denoted by $\mathcal{E}_{g_y, P_{\mathbf{y}}}$ (see Fig. 8). Thus optimum $Q_{\mathbf{y}}^*$ is of the form

$$Q_{\mathbf{y}}^*(y) = \frac{1}{k(\theta)} P_{\mathbf{y}}(y) \exp(\theta g_y)$$

for some $\theta \in \mathcal{R}$ where $k(\theta)$ is the normalization constant. However, this property may not be true for a general (not very noisy) channel.

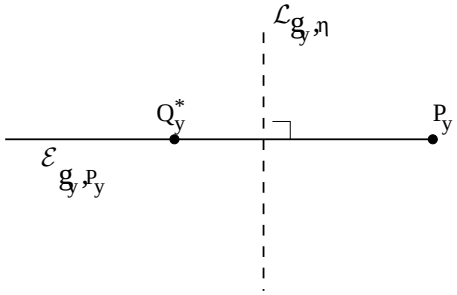


Fig. 8. In the space of distribution on $\mathcal{Y}$: Solid line shows the exponential family $\mathcal{E}_{g_y, P_{\mathbf{y}}}$ and the dashed line is an orthogonal linear family $\mathcal{L}_{g_y, \eta} = \{Q_{\mathbf{y}} | \mathrm{E}_{Q_{\mathbf{y}}}[g_{\mathbf{y}}] = \eta\}$.

[2] In fact, for Gallager's definition of very noisy channel, substituting $P_{\mathbf{y}}$ instead of the optimum $Q_{\mathbf{y}}$ gives the same answer in his very noisy limit.

IV. Error exponent of the expurgated ensemble

Since I-projection gives us a method to address high-dimensional optimizations, let us address the error exponent of Gallager's expurgated ensemble. We first create a i.i.d. random code of rate R, generated using input distribution P. With some abuse of notation, $\mathbf{x}^n$ denotes the correct codeword $\mathbf{x}^n(1)$ and $\mathbf{z}^n$ denotes an incorrect codeword $\mathbf{x}^n(i)$ and $Q_{\mathbf{xz}}$ denotes the joint type of the pair $(\mathbf{x}^n, \mathbf{z}^n)$. Let $W(\cdot|xz) \in \mathcal{P}(\mathcal{Y})$ denote the channel type where correct input $\mathbf{x} = x$ and incorrect input $\mathbf{z} = z$.

From this random code, we throw away the "*bad*" codeword pairs which are "*too close*". More specifically, we expurge codeword pairs such that $D(Q_{\mathbf{xz}}\|P \otimes P) > R$, where $P \otimes P$ denotes independent $\mathbf{x}$ and $\mathbf{z}$, with marginal distribution P for both. Since the fraction of such codewords will be exponentially small, this rule guarantees that rate of the expurgated code is not smaller than R. After expurgation all codeword pairs will satisfy $D(Q_{\mathbf{xz}}\|P \otimes P) \leq R$. The error exponent $E_{ex}(R)$ of this expurgated code equals

$$\min_{Q_{\mathbf{xz}}:\ D(Q_{\mathbf{xz}}\|P\otimes P)\leq R} E(Q_{\mathbf{xz}}) + (D(Q_{\mathbf{xz}}\|P \otimes P) - R)$$

where $E(Q_{\mathbf{xz}})$ denotes the error exponent for a codeword pair with type $Q_{\mathbf{xz}}$. The second term is due to union bound and the fact that exponent of observing the joint type $Q_{\mathbf{xz}}$ in an i.i.d. random code equals $D(Q_{\mathbf{xz}}\|P \otimes P)$. Now let us analyze $E(Q_{\mathbf{xz}})$, which is given by

$$E(Q_{\mathbf{xz}}) = \min_{W(\cdot|\mathbf{xz}):\ \text{error}} \sum_{x,z} Q_{\mathbf{xz}}(x,z) D\left(W(\cdot|xz)\|P_{\mathbf{y}|x}(\cdot)\right)$$

where $P_{\mathbf{y}|x}(\cdot)$ denotes the actual channel distribution from the correct input and error happens when log-likelihood of the correct codeword is smaller than that of the wrong codeword, that is,

$$\begin{aligned} \sum_{x,z,y} Q_{\mathbf{xz}}(x,z) W(y|xz) \log \frac{P_{\mathbf{y}|\mathbf{x}}(y|x)}{P_{\mathbf{y}|\mathbf{x}}(y|z)} &\leq 0 \\ \text{i.e.} \qquad \mathrm{E}_{Q_{\mathbf{xz}}W(\cdot|\mathbf{xz})}\left[L(y|x,z)\right] &\leq 0 \end{aligned}$$

where $L(y|x,z)$ is a shorthand for the log-likelihood ratio $\log \frac{P_{\mathbf{y}|\mathbf{x}}(y|x)}{P_{\mathbf{y}|\mathbf{x}}(y|z)}$. Thus calculating $E(Q_{\mathbf{xz}})$ involves minimizing a weighted average of $D\left(W(\cdot|xz)\|P_{\mathbf{y}|x}(\cdot)\right)$ with a constraint on the weighted average of log-likelihood ratio. Similar to the previous section, I-projection implies that optimum channel $W(\cdot|xz) \in \mathcal{P}(\mathcal{Y})$ for any pair (x,z) lies on the exponential family connecting the channel from correct input $P_{\mathbf{y}|\mathbf{x}}(\cdot|x)$ to the channel from incorrect input $P_{\mathbf{y}|\mathbf{x}}(\cdot|z)$ as below

$$W^t(\cdot|xz) \propto P^t_{\mathbf{y}|\mathbf{x}}(\cdot|x) \cdot P^{1-t}_{\mathbf{y}|\mathbf{x}}(\cdot|z) \ \ \forall\ x,z$$

Similar to Remark 1, the exponential parameter t is the same for each pair (x,z). Hence finding $E(Q_{\mathbf{xz}})$ only needs a scalar optimization:

$$E(Q_{\mathbf{xz}}) = \min_{\hat{t}:\text{error}} \sum_{x,z} Q_{\mathbf{xz}}(x,z) D\left(W^{\hat{t}}(\cdot|xz)\|P_{\mathbf{y}|x}(\cdot)\right) \quad (58)$$

where error happens for $\hat{t}$ such that $\mathrm{E}_{Q_{\mathbf{xz}}W^{\hat{t}}(\cdot|\mathbf{xz})}\left[L(y|x,z)\right] \leq 0$.

We compare this approach to Gallager's analysis of the expurgated ensemble in [6]. There the error probability for codeword pair (x^n, z^n) is bounded as follows:

$$\begin{aligned} \Pr(\text{error from } x^n \text{ to } z^n) &= \sum_{y^n:\ P_{\mathbf{y}|\mathbf{x}}(y^n|z^n)\geq P_{\mathbf{y}|\mathbf{x}}(y^n|x^n)} P_{\mathbf{y}|\mathbf{x}}(y^n|x^n) \\ &\leq \sum_{y^n} P_{\mathbf{y}|\mathbf{x}}(y^n|x^n) \sqrt{\frac{P_{\mathbf{y}|\mathbf{x}}(y^n|z^n)}{P_{\mathbf{y}|\mathbf{x}}(y^n|x^n)}} \end{aligned}$$

An exercise in [5] also starts with the same trick and gets the same error exponent as in [6]. This trick of square-root is equivalent to substituting $\hat{t} = 1/2$ as the minima in (58). It is not clear why the minimum should always be attained at $1/2$.

However, at $R = 0$, the expurgation constraint $D(Q_{\mathbf{xz}}\|P \otimes P) \leq 0$ implies $Q_{\mathbf{xz}} = P \otimes P$, which is a symmetric distribution in (x,z). For symmetric $Q_{\mathbf{xz}}$, it is easy to show that $\hat{t} = 1/2$ should attain the minimum. Although for $R > 0$, it is not obvious. Nonetheless, we have shown that even for $R > 0$, the minimum should be attained at $1/2$. This proves that Gallager's formula for expurgated ensemble is tight, which further strengthens their conjecture in [8] about tightness of this bound for any code (at small enough R).

A similar analysis can be used for more high-dimensional problems such as expurgation for List-of-L decoding, which clarifies what it means for a tuple of codewords to be "*too close*" to each other, i.e. which codeword-lists are more likely to cause errors.

Appendix: Proof of Theorem 2

The optimum output type is given by $Q^*_{\mathbf{y}}$ and the error dominating joint type is $Q^*_{\mathbf{y}}P^{(t^*)}_{\mathbf{x}|\mathbf{y}}$. Using (48), the error exponent is given by

$$\begin{aligned} E_r(R) &= D(Q^*_{\mathbf{y}}\|P_{\mathbf{y}}) + E(R, Q^*_{\mathbf{y}}) \\ &= D(Q^*_{\mathbf{y}}\|P_{\mathbf{y}}) + D(Q^*_{\mathbf{y}}P^{(t^*)}_{\mathbf{x}|\mathbf{y}}\|Q^*_{\mathbf{y}}P_{\mathbf{x}|\mathbf{y}}) \\ &+ \left[D(Q^*_{\mathbf{y}}P^{(t^*)}_{\mathbf{x}|\mathbf{y}}\|Q^*_{\mathbf{y}}P_{\mathbf{x}}) - R\right]^+ \\ &= D(Q^*_{\mathbf{y}}P^{(t^*)}_{\mathbf{x}|\mathbf{y}}\|P_{\mathbf{xy}}) + \left[D(Q^*_{\mathbf{y}}P^{(t^*)}_{\mathbf{x}|\mathbf{y}}\|Q^*_{\mathbf{y}}P_{\mathbf{x}}) - R\right]^+ \end{aligned}$$

This is a non-decreasing function in $D(Q^*_{\mathbf{y}}P^{(t^*)}_{\mathbf{x}|\mathbf{y}}\|P_{\mathbf{xy}})$ and $D(Q^*_{\mathbf{y}}P^{(t^*)}_{\mathbf{x}|\mathbf{y}}\|Q^*_{\mathbf{y}}P_{\mathbf{x}})$. Similar to previous subsection, we can show that a ML decoder is equivalent to a LLR decoder for $p_1 = P_{\mathbf{xy}}$ to $p_0 = Q^*_{\mathbf{y}}P_{\mathbf{x}}$. Now assume to the contrary that the dominating joint type $Q^*_{\mathbf{y}}P^{(t^*)}_{\mathbf{x}|\mathbf{y}}$ does not lie on the exponential family joining p_1 and p_0. Now move $Q^*_{\mathbf{y}}P^{(t^*)}_{\mathbf{x}|\mathbf{y}}$ to $Q'_{\mathbf{y}}P^{(t')}_{\mathbf{x}|\mathbf{y}}$ which is on the exponential family and has the same expected LLR as $Q^*_{\mathbf{y}}P^{(t^*)}_{\mathbf{x}|\mathbf{y}}$. Thus by I-projection theorem

$$\begin{aligned} D(Q'_{\mathbf{y}}P^{(t')}_{\mathbf{x}|\mathbf{y}}\|P_{\mathbf{xy}}) &< D(Q^*_{\mathbf{y}}P^{(t^*)}_{\mathbf{x}|\mathbf{y}}\|P_{\mathbf{xy}}) \\ \text{and} \quad D(Q'_{\mathbf{y}}P^{(t')}_{\mathbf{x}|\mathbf{y}}\|Q^*_{\mathbf{y}}P_{\mathbf{x}}) &< D(Q^*_{\mathbf{y}}P^{(t^*)}_{\mathbf{x}|\mathbf{y}}\|Q^*_{\mathbf{y}}P_{\mathbf{x}}) \end{aligned}$$

Moreover,

$$\begin{aligned} D(Q'_{\mathbf{y}}P^{(t')}_{\mathbf{x}|\mathbf{y}}\|Q^*_{\mathbf{y}}P_{\mathbf{x}}) &= D(Q'_{\mathbf{y}}P^{(t')}_{\mathbf{x}|\mathbf{y}}\|Q'_{\mathbf{y}}P_{\mathbf{x}}) + D(Q'_{\mathbf{y}}\|Q^*_{\mathbf{y}}) \\ &\geq D(Q'_{\mathbf{y}}P^{(t')}_{\mathbf{x}|\mathbf{y}}\|Q'_{\mathbf{y}}P_{\mathbf{x}}) \\ \Rightarrow \quad D(Q'_{\mathbf{y}}P^{(t')}_{\mathbf{x}|\mathbf{y}}\|Q'_{\mathbf{y}}P_{\mathbf{x}}) &< D(Q^*_{\mathbf{y}}P^{(t^*)}_{\mathbf{x}|\mathbf{y}}\|Q^*_{\mathbf{y}}P_{\mathbf{x}}) \end{aligned}$$

Thus replacing $Q^*_{\mathbf{y}}P^{(t^*)}_{\mathbf{x}|\mathbf{y}}$ by $Q'_{\mathbf{y}}P^{(t')}_{\mathbf{x}|\mathbf{y}}$ gives a smaller exponent, which contradicts the optimality of $Q^*_{\mathbf{y}}P^{(t^*)}_{\mathbf{x}|\mathbf{y}}$.

REFERENCES

[1] S. Amari and H. Nagaoka, *Methods of Information Geometry*, Oxford University Press, 2000.

[2] Y.Pawitan, *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford Science Publications, 2001

[3] M. Murray and J. Rice, *Differntial Geometry and Statistics*, Chapman & Hall/CRC, 1993

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

[5] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, 1981.

[6] R. Gallager,"A Simple Derivation of the Coding Theorem and Some Application", *IEEE Trans. on Information Theory*, Vol. 11, No. 1, pp. 3-18, Jan. 1965.

[7] R. Gallager, "A Random Coding Bound on Fixed Composition Codes," course notes for Information Theory, May 1992.

[8] C. E. Shannon, R. G. Gallager, and E. R. Berlekamp, "Lower Bounds to Error Probability for Coding on Discrete Memoryless Channels", *Information and Control*, pp. 65-103, December 1966.

[9] A. Montanari and D. Forney, "On exponential error bounds for random codes on the DMC," unpublished manuscript.

[10] D. Forney, "On exponential error bounds for random codes on the BSC," unpublished manuscript.

[11] D. Guo, S. Shamai and S. Verdu, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inform. Theory*, vol. 51, pp. 1261-1282, April 2005.