

Euclidean Information Theory

Shashi Borade Lizhong Zheng

Abstract

Many problems in information theory involve optimizing the Kullback-Leibler (KL) divergence between probability distributions. Since KL divergence is difficult to analyze, these optimizations are often intractable. We simplify these problems by assuming the distributions of interest to be close to each other. Under this assumption, the KL divergence behaves like a squared Euclidean distance. We demonstrate that with this simplification, we can shed new insights to the structure of the optimal solution to some information theoretic problems. In particular, this approach helps us to obtain single letter solutions to some broadcast problems.

1 Introduction

A variety of information theory problems involve optimizing KL divergence. Understanding the structure of the optimum solution is helpful for characterizing the achievable regions as well as the converse bounds. For example, it could be helpful in converting a multi-letter characterization of a capacity region into a single-letter characterization. Moreover, even when a single-letter characterization is available, it is often implicit in the form of an optimization over different distributions. Knowledge of these optimum distributions gives additional insights on the capacity region and design of good codes. Similar optimizations also arise in rate-distortion problems and error exponent analysis.

However, we do not have a systematic approach for finding the optimum solution in general. The main source of difficulty is that the KL divergence is not a metric in the space of probability distributions. In fact, the collection of distributions in general, form a manifold, which invalidates interpreting the KL divergence as a distance between distributions. A natural way to simplify this general scenario is to restrict our attention to a local neighborhood of distributions, in which the manifold behaves like a Euclidean space and the KL divergence behaves like the Euclidean metric. The goal of this paper is to explore the use of this approach to solve some network information theory problems.

Note that an upper bound on the KL divergence between distributions P and Q is obtained using the bound $\ln(1+t) \geq t - \frac{t^2}{2}$. We assume $P, Q > 0$, i.e., all entries of P and Q are strictly positive.

$$\begin{aligned} D(P\|Q) &= -\sum_i P_i \ln\left(1 + \frac{Q_i - P_i}{P_i}\right) \leq -\sum_i P_i \left(\frac{Q_i - P_i}{P_i} - \frac{(Q_i - P_i)^2}{2P_i^2}\right) \\ &= 0 + \sum_i \frac{(Q_i - P_i)^2}{2P_i} \triangleq \frac{1}{2}\|Q - P\|_P^2 \end{aligned}$$

where the summation is over all symbols of the distribution and $\|a\|_b^2$ denotes the squared norm of a , weighted with b as its weight vector: $\|a\|_b^2 \triangleq \sum_i \frac{a_i^2}{b_i}$ for $b > 0$. This bound on $D(P\|Q)$ is tight when $P \approx Q$. Moreover, it remains tight even if the subscript P in $\|Q -$

$P\|_P^2$ (the weight vector for the squared norm) is changed to another nearby distribution $\hat{P} \approx P$. That is, the difference between $\|Q - P\|_P^2$ and $\|Q - \hat{P}\|_{\hat{P}}^2$ is negligible compared to either term when $P \approx \hat{P}$. Thus we can view the weight vector as only dependent on the neighborhood of distributions. The divergence between any pair of distributions in this neighborhood has the same weight vector for its Euclidean approximation. In particular, this implies $D(P\|Q) \approx D(Q\|P)$ when $P \approx Q$.

With this simplification, we study the problem of source coding with a helper [1] and the degraded broadcast channel problem [2]. Our solutions shed some new light on these already solved problems. We then consider the broadcast problem with degraded message sets [3] for two or more users. Although this problem is open in general for more than two users, a single-letter characterization of the capacity region is possible with our simplification. We also obtain some new insights on the capacity region in terms of the singular value decomposition (SVD) of certain matrices, which depend on the channels involved.

Now let us write mutual information in this approximation. Let X, Y be a pair of discrete random variables with marginal distributions P_X, P_Y . Let W denote the probability transition matrix, *i.e.*, $W_{Y|x}(\cdot) = P_{Y|X}(\cdot|x)$. Now note that

$$I(X; Y) = \sum_x P_X(x) D(W_{Y|x} \| P_Y) = \mathbb{E}_{P_X} [D(W_{Y|X} \| P_Y)] \quad (1)$$

$$\approx \frac{1}{2} \mathbb{E}_{P_X} [\|W_{Y|X} - P_Y\|_{P_0}^2] \quad (2)$$

This approximation is tight when the conditional distributions and P_Y are in the neighborhood of P_0 , that is, $W_{Y|x} \approx P_0$ for each $x \in \mathcal{X}$ and hence $P_Y \approx P_0$. Since P_Y is the average of $W_{Y|x}$ under P_X , the above expression for mutual information looks like (half of) the ‘variance’ of the conditional distributions $W_{Y|x}$. Although, instead of the usual Euclidean norm, we are taking the weighted Euclidean norm according to P_0 .

Remark: The capacity achieving output distribution P_Y^* has a very intuitive geometric interpretation now. Recall that for every input x used in the capacity achieving distribution, $D(W_{Y|x} \| P_Y^*)$ equals capacity [4]. Under our simplification, this means that (half of) the weighted Euclidean norm $\|W_{Y|x} - P_Y^*\|_{P_0}^2$ equals capacity for all those inputs. Thus P_Y^* is the ‘circum-center’ of the polygon formed with different $P_{Y|x}$ and the channel capacity equals half the squared ‘circum-radius’ of this polygon¹. Figure 1 shows this result for a channel with a ternary input alphabet $\mathcal{X} = \{1, 2, 3\}$.

2 Source coding with a helper

Consider a pair of correlated memoryless sources (X, Y) . As before, let P_X, P_Y and W denote their marginal distributions and probability transition matrix. The decoder is interested in the lossless reconstruction of Y , but the available rate from Y , $R_y \leq H(Y)$. Hence the R_y link alone is not sufficient for the lossless reconstruction of Y .

The helper node observes X (see Fig. 2) and summarizes it into U under a rate constraint on this summary $I(U; X) \leq R_x$. Once U is available at the decoder, the additional rate required from Y equals $H(Y|U)$ (using Slepian-Wolf Coding). The optimal

¹The quotation marks denote that we are talking about a weighted squared norm according to P_0 instead of the standard Euclidean norm. Although, if P_0 is the uniform distribution, then the weighted squared norm is simply a multiple of the standard (unweighted) Euclidean norm and the channel capacity is related to the standard Euclidean circum-radius of this polygon.

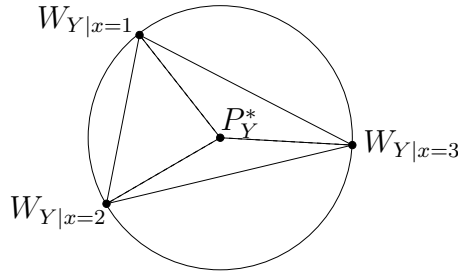


Figure 1: Geometric interpretation of channel capacity. Optimum output distribution P_Y^* is ‘equidistant’ from every conditional output distribution.

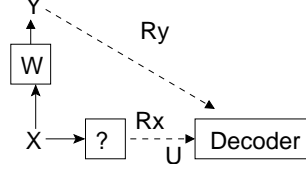


Figure 2: Under rate constraint R_x , summarize X though U .

tradeoff between R_y and the helper rate R_x is given by the following optimization.

$$R_y^* = \min_{U-X-Y: I(U;X) \leq R_x} H(Y|U) = H(Y) - \max_{U-X-Y: I(U;X) \leq R_x} I(U;Y) \quad (3)$$

For using our Euclidian framework, we assume a vanishing data-rate from X to the decoder ($I(U; X) = R_x \approx 0$) and assume the distributions $P_{X|u}$ to be close to each other for different u . Now from Eq. (2),

$$I(U; X) \approx \frac{1}{2} \mathbb{E}_{P_U} [\|P_{X|U} - P_X\|_{P_X}^2] \quad \& \quad I(U; Y) \approx \frac{1}{2} \mathbb{E}_{P_U} [\|P_{Y|U} - P_Y\|_{P_Y}^2] \quad (4)$$

More precisely, for any fixed choice of P_U and $P_{X|U}$, let $\theta_u = P_{X|u} - P_X$. Now we define a family of random variables $U^{(\epsilon)}$ indexed by ϵ such that $P_{U^{(\epsilon)}} = P_U$ and $P_{X|U^{(\epsilon)}} = P_X + \epsilon \theta_U$.

$$\begin{aligned} \text{Define } R_x^{(\epsilon)} &\triangleq I(U^{(\epsilon)}; X) \quad \& \quad R_y^{(\epsilon)} \triangleq I(U^{(\epsilon)}; Y) \\ \text{and } r_x &\triangleq \lim_{\epsilon \rightarrow 0} \frac{R_x^{(\epsilon)}}{\epsilon^2} \quad \& \quad r_y \triangleq \lim_{\epsilon \rightarrow 0} \frac{R_y^{(\epsilon)}}{\epsilon^2}. \end{aligned}$$

Then Euclidean approximation implies, $r_x = \frac{1}{2} \mathbb{E}_{P_U} [\|\theta_U\|_{P_X}^2] \quad \& \quad r_y = \frac{1}{2} \mathbb{E}_{P_U} [\|W' \theta_U\|_{P_Y}^2]$

where $W' \theta_U = W'(P_{X|U} - P_X) = P_{Y|U} - P_Y$. Thus Eq. (4) is a good approximation if distributions $P_{X|u}$ are close to each other for different u . Hence we will study the tradeoff between r_x and r_y , as an approximation for the tradeoff in Eq. (3). This corresponds to the slope of the R_y - R_x tradeoff when $R_x \approx 0$.

Now note that the weighted squared norm $\|a\|_b^2$ can be converted into a standard Euclidian norm as follows.

$$\|a\|_b^2 = \sum_i \frac{a_i^2}{b_i} = \|[b^{-1/2}] \cdot a\|^2$$

where $[b^{-1/2}]$ denotes a diagonal matrix whose i 'th diagonal entry equals $b_i^{-1/2}$. With this observation, the optimization in Eq. (3) is equivalent to,

$$\max_{P_U, \theta_U: \mathbb{E}_{P_U} [\|[P_X^{-1/2}] \cdot \theta_U\|^2] \leq 2r_x} \|[P_Y^{-1/2}] \cdot W' \cdot \theta_U\|^2 \quad (5)$$

Substituting $\phi_U = [P_X^{-1/2}] \theta_U$ converts this problem to

$$\max_{P_U, \phi_U: \mathbb{E}_{P_U} [\|\phi_U\|^2] \leq 2r_x} \mathbb{E}_{P_U} [\|B \cdot \phi_U\|^2] \quad \text{where } B \triangleq [P_Y^{-1/2}] \cdot W' \cdot [P_X^{1/2}] \quad (6)$$

We call B as the *divergence translation matrix*, since it transforms the divergence between X distributions to that between Y distributions. If P_X and P_Y are uniform distributions, then this matrix is equivalent to W , the channel matrix itself.

This optimization is a standard problem in linear algebra. Its solution depends on the SVD of B , which has the following property. It is proved using the data-processing theorem.

Lemma 1 *Let $\sigma_1, \sigma_2 \dots$ denote the singular-values of B in descending order and the corresponding singular vectors be $v_1, v_2 \dots$. Then the largest singular-value $\sigma_1 = 1$ and $v_1 = P_X^{1/2}$, which denotes element-wise square-root of vector P_X .*

If there were no constraints on ϕ_u , the optimal choice of each ϕ_u should be along v_1 , the singular vector with the largest singular-value².

However, it turns out that v_1 is an infeasible direction for $\phi_u = [P_X^{-1/2}] \theta_u$, where $\theta_u = P_{X|u} - P_X$. Note that θ_u lies along the probability simplex, it satisfies

$$\sum \theta_u(\cdot) = 0 \Leftrightarrow v_1' \cdot [P_X^{-1/2}] \theta_u = v_1' \phi_u = 0 \Rightarrow v_1 \perp \phi \in \text{span}(v_2, v_3 \dots)$$

This means that linear combinations of $\{v_2, v_3 \dots\}$ correspond to all feasible θ_u directions along the simplex. Since $\phi \in \text{span}(v_2, v_3 \dots)$, the optimal ϕ_u lies along v_2 , the *feasible* direction with the largest singular value. This implies,

$$r_y \leq \sigma_2^2 r_x \quad \text{i.e.,} \quad I(U^{(\epsilon)}; Y) \leq \sigma_2^2 I(U^{(\epsilon)}; X) \quad \text{for } \epsilon \ll 1$$

with equality achieved as $\epsilon \rightarrow 0$ when ϕ_u are chosen along v_2 .

Thus σ_2^2 equals the slope of the optimal R_y vs. R_x curve at the $R_x = 0$ intercept. Note that $\sigma_2 \leq 1$ reflects the intuition that the R_x link is not as effective as R_y for conveying Y and its effectiveness is characterized by σ_2^2 . It is somewhat surprising that this effectiveness has no direct relation with the mutual information $I(X; Y)$, but instead, it depends on the second largest singular-value σ_2 of the divergence translation matrix.

It is more interesting to consider the multi-letter problem. In the following, we study the 2-letter case and general K -letter case follows on the same lines.

Since the source pair XY is distributed i.i.d. over time, the two-letter distributions and channels are given by $P_{X_1 X_2} = P_X \otimes P_X$, $P_{Y_1 Y_2} = P_Y \otimes P_Y$ & $P_{Y_1 Y_2 | X_1 X_2} = W \otimes W$, where \otimes denotes the Kronecker product. For the 2-letter case, we are interested in the 2-letter version of the optimization in Eq. (3).

$$\max_{U-X_1 X_2-Y_1 Y_2: I(U; X_1 X_2) \leq 2R_x} I(U; Y_1 Y_2) \quad (7)$$

Again, with the local assumption on distributions, and substituting $\theta_u = P_{X_1 X_2 | u} - P_{X_1 X_2}$, we can rewrite this as

$$\max_{P_U, \phi_U: \mathbb{E}_{P_U} [\|\phi_U\|^2] \leq 2r_x} \mathbb{E}_{P_U} [\|B^{(2)} \cdot \phi_U\|^2] \quad \text{where, } \phi_u = [P_{X_1 X_2}^{-1/2}] \theta_u \quad (8)$$

$$\text{and } B^{(2)} \triangleq [P_{Y_1 Y_2}^{-1/2}] \cdot (W \otimes W)' \cdot [P_{X_1 X_2}^{1/2}] = B \otimes B \quad (9)$$

is the divergence translation matrix for this 2-letter case.

²This is essentially like multiantenna beamforming for maximum power gain—putting all the “power” along the largest eigenvector.

Lemma 2 *Let v_i and v_j denote two singular vectors of B with singular-values σ_i and σ_j . Then $v_i \otimes v_j$ is an singular vector of $B^{(2)}$ and its singular-value is $\sigma_i \sigma_j$.*

Again, the largest singular-value equals 1 corresponding to the singular vector $v_1 \otimes v_1$, which is an infeasible direction. The singular vectors $v_1 \otimes v_2$ and $v_2 \otimes v_1$ correspond to the second largest singular-value σ_2 .

Recalling $v_1 = P_X^{1/2}$ to notice that ϕ_u along $v_1 \otimes v_2$ translates to θ_u along $P_X \otimes \alpha$, where $\alpha = [P_X^{1/2}]v_2$. Similarly, $v_2 \otimes v_1$ translates to $\alpha \otimes P_X$. Thus the optimal $P_{X_1 X_2 | u^{(\epsilon)}}$ for any $u^{(\epsilon)}$ looks like

$$\begin{aligned} P_{X_1 X_2 | u^{(\epsilon)}} &= P_{X_1 X_2} + \epsilon \theta_u \\ &= P_X \otimes P_X + \epsilon_1 P_X \otimes \alpha + \epsilon_2 \alpha \otimes P_X \quad (\text{for some } \epsilon_1, \epsilon_2 \ll 1) \\ &\approx (P_X + \epsilon_1 \alpha) \otimes (P_X + \epsilon_2 \alpha) \end{aligned}$$

The last step follows by adding $\epsilon_1 \epsilon_2 (\alpha \otimes \alpha)$, which is of smaller order than the other terms. This result says that the conditional distribution $P_{X_1 X_2 | u^{(\epsilon)}}$ is independent over time; thus single letter solution is optimal. This fact is proved in [1] under general conditions. However, our proof based on linear algebra uses a different approach.

For any given number of letters K , the same results hold true as ϵ goes to zero. It is worth mentioning that we are fixing the number of letters K and let ϵ go to zero. This is different from fixing ϵ and letting K go to infinity, where the Euclidean approximations are not clearly justified. These comments also hold true for the next sections.

3 Degraded broadcast channel

After studying the helper problem, the degraded broadcast channel problem follows on similar lines. Consider the physically degraded broadcast (Fig. 3) from X to Y and Z .

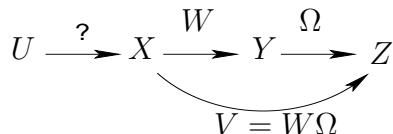


Figure 3: Physically degraded broadcast channel

It is known [2] that the achievable rates are $R_c = I(U; Z)$ for common information to both receivers and $R_p = I(X; Y | U) = I(X; Y) - I(U; Y)$ for private information to Y , where U satisfies the Markov relation $U - X - Y - Z$. The capacity region is given by

$$R_c^* = \max_{U-X-Y-Z: I(X; Y | U) \geq R_p} I(U; Z) = \max_{U-X-Y-Z: I(U; Y) \leq I(X; Y) - R_p} I(U; Z) \quad (10)$$

In the following, we fix the X distribution at P_X and only focus on the choice of U , which means that $I(X; Y) - R_p$ is a fixed constraint (say γ) on $I(U; Y)$.

For applying Euclidean approach, we need local assumptions.

- We can assume channel W to be very noisy, which implies that all Y distributions of interest are close to each other and hence all Z distributions of interest are also close to each other.

- Alternatively, even for a not very noisy channel, we can assume that $P_{X|u}$ are close for different u . This corresponds to low common information rate, $R_c \approx 0$. Like in the helper problem, this yields the slope of the R_p vs. R_c tradeoff at $R_c = 0$.

Using either of these assumptions and defining $\theta_u = P_{Y|u} - P_Y$, we rewrite Eq. (10) as,

$$\max_{P_U, \theta_U: \mathbb{E}_{P_U}[\| [P_Y^{-1/2}] \cdot \theta_U \|^2] \leq 2\gamma} \| [P_Z^{-1/2}] \cdot \Omega' \cdot \theta_U \|^2 \quad (11)$$

This is the same as Eq. (5) in the helper problem. Repeating those same steps, we can again show that the multi-letter problem reduces to the single-letter case. This is particularly useful for obtaining converse results.

Note however that in this optimization, we are optimizing over $P_{Y|u}$ instead of $P_{X|u}$ (since $\theta_u = P_{Y|u} - P_Y$). Since $P_{Y|u}$ should be a feasible output distribution for channel W , it lies in a convex set (the convex hull of $\{W_{Y|x} : x \in \mathcal{X}\}$). For sufficiently small R_c , the solutions based on SVD (related to v_2 of the divergence translation matrix) are feasible choices. Thus the slope of the R_p - R_c tradeoff near $R_c = 0$ equals σ_2^2 , the second largest singular-value squared. This holds true for the general degraded broadcast problem even if it is not very noisy.

For the case when W is very noisy, we can increase $R_c = I(U; Z)$ by extending θ_u further along the direction corresponding to v_2 until $P_{Y|u}$ reaches the boundary of its feasible set. Then to further increase R_c , the choice of θ_u will move along this boundary. The resulting R_p - R_c tradeoff should hence be piecewise linear in shape.

4 Broadcast with degraded message sets

Now consider the situation when Z is not a degraded version of Y . For using our Euclidean framework, we assume that W and V are both very noisy. That is, for every input x , the conditional distributions $W_{Y|x}$ and $V_{Z|x}$ are close to some Q_y and Q_z . In matrix notation,

$$W = \mathbf{1} \otimes Q'_y + \epsilon \cdot \Theta \quad \& \quad V = \mathbf{1} \otimes Q'_z + \epsilon \cdot \Phi \quad (12)$$

where $\mathbf{1} \otimes Q'_y$ denotes a matrix whose every row equals Q_y and Θ and Φ are fixed matrices with every row summing to 0. Receiver Z wants to decode a common message at rate

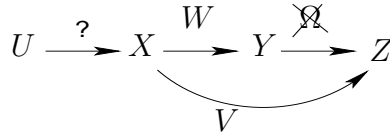


Figure 4: General broadcast channel

R_c and Y wants to decode this common message as well as a private message at rate R_p . The capacity region in this case is given by [3]

$$R_c^* = \max_{U-X-(YZ): I(X;Y|U) \geq R_p} \min\{I(U; Z), I(U; Y)\} \quad (13)$$

As before, we fix the X distribution at P_X and only focus on the choice of U , which means that $I(X; Y) - R_p$ is a fixed constraint (say γ) on $I(U; Y)$. Hence it is sufficient to solve the following optimization.

$$\max_{U-X-(YZ): I(U; Y) \leq \gamma} I(U; Z) \quad (14)$$

Under the very noisy assumptions, this is equivalent to solving

$$P_U, \phi_U: \max_{\mathbb{E}_{P_U}[\|\phi_U\|^2] \leq 2\gamma} \mathbb{E}_{P_U} [\|B_{y \rightarrow z} \cdot \phi_U\|^2] \quad (15)$$

$$\text{where, } B_{y \rightarrow z} = [P_Z^{1/2}] \cdot (V'W'^{-1}) \cdot [P_Y^{1/2}] \quad (16)$$

As before, the divergence translation matrix $B_{y \rightarrow z}$ has a singular-vector $P_Y^{1/2}$ with singular-value 1. However, since $V'W'^{-1}$ is not a probability transition matrix like Ω' before, $B_{y \rightarrow z}$ could have singular-values larger than 1.

Assuming v_i is such a singular vector of $B_{y \rightarrow z}$ with singular-value $\sigma_i \geq 1$, we should choose ϕ_u along v_i for small enough γ . The optimum $I(U; Z)$ in this case satisfies $I(U; Y) \leq \gamma \leq I(U; Z)$, so the common information R_c is bottlenecked at $R_c = I(U; Y)$. Hence for small enough R_c , the R_c - R_p tradeoff is $R_c + R_p = I(X; Y)$ as $R_p = I(X; Y|U)$.

If all singular values of $B_{y \rightarrow z}$ are upper bounded by 1 with $\sigma_1 = 1$, then the common information is bottlenecked by Z and the slope of R_c - R_p tradeoff for small R_c equals σ_2^2 , where σ_2 is second largest singular-value of $B_{y \rightarrow z}$. We should mention that even for a non-degraded broadcast channel, all singular values of $B_{y \rightarrow z}$ could upper bounded by 1. For example, this happens when W is a BSC and V is an asymmetric binary channel [6].

4.1 Multi-letter case

Lets define the divergence translation matrices from X to Y and from X to Z .

$$B_{x \rightarrow y} = [P_Y^{-1/2}]W'[P_X^{1/2}] \quad \& \quad B_{x \rightarrow z} = [P_Z^{-1/2}]V'[P_X^{1/2}]$$

and note that $B_{y \rightarrow z} = B_{x \rightarrow z}B_{x \rightarrow y}^{-1}$.

Let $\{\mu_1, \mu_2 \dots\}$ and $\{\nu_1, \nu_2 \dots\}$ denote the singular-values of $B_{x \rightarrow y}$ and $B_{x \rightarrow z}$ in descending order and let $\{g_1, g_2 \dots\}$ and $\{h_1, h_2 \dots\}$ denote the corresponding singular vectors (respectively). It can be verified that the largest singular values $\mu_1 = \nu_1 = 1$ and $g_1 = h_1 = P_X^{1/2}$. Moreover, since W and V are very noisy as in Eq. (12), all other singular values are of the order $O(\epsilon)$.

Now using Lemma 2 for the 2-letter case, the singular-values of $B_{x \rightarrow y}^{(2)}$ can be divided into three classes: $\mu_1\mu_1 = 1$, $\{\mu_1\mu_i = \mu_i = O(\epsilon) \text{ for } i \neq 1\}$, and $\{\mu_i\mu_j = O(\epsilon^2) \text{ for } i, j \neq 1\}$. The corresponding singular vectors are $g_1 \otimes g_1$, $\{g_i \otimes g_1 \ \& \ g_1 \otimes g_i\}$ and $\{g_i \otimes g_j\}$.

Choosing $\theta_u = P_{X_1X_2|u} - P_{X_1X_2}$ in the direction corresponding to $g_1 \otimes g_1$ is infeasible due to the simplex constraint. Choosing θ_u along the direction corresponding $g_1 \otimes g_i$ and $g_i \otimes g_1$ causes an $O(\epsilon)$ change in $P_{Y_1Y_2|u}$ and hence an $O(\epsilon^2)$ change in $I(U; Y)$. On the other hand, choosing θ_u along the direction corresponding $g_i \otimes g_j$ causes an $O(\epsilon^2)$ change in $P_{Y_1Y_2|u}$ and hence an $O(\epsilon^4)$ change in $I(U; Y)$, which is negligible compared to $O(\epsilon^2)$.

Now the only relevant singular vectors, $g_i \otimes g_1$ and $g_1 \otimes g_i$, correspond to θ_u of the form $P_Y \otimes \alpha_i$ and $\alpha_i \otimes P_Y$, where $\alpha_i \stackrel{\Delta}{=} [P_Y^{1/2}]v_i$ for $i \neq 1$. Thus optimal $P_{Y_1Y_2|u}$ looks like

$$\begin{aligned} P_{Y_1Y_2|u} &= P_{Y_1Y_2} + \epsilon_1(P_Y \otimes \hat{\alpha}) + \epsilon_2(\tilde{\alpha} \otimes P_Y) = P_Y \otimes P_Y + \epsilon_1(P_Y \otimes \hat{\alpha}) + \epsilon_2(\tilde{\alpha} \otimes P_Y) \\ &\approx (P_Y + \epsilon_1\hat{\alpha}) \otimes (P_Y + \epsilon_2\tilde{\alpha}) \end{aligned}$$

for some $\epsilon_1, \epsilon_2 \ll 1$. Here $\hat{\alpha}$ and $\tilde{\alpha}$ are some linear combinations of $\{\alpha_2, \alpha_3 \dots\}$. The last step followed by adding $\epsilon_1\epsilon_2(\hat{\alpha} \otimes \tilde{\alpha})$, which is of smaller order than the other terms.

Thus the conditional distribution $P_{Y_1Y_2|u}$ is independent over time; hence a single letter solution is optimal even in this non-degraded broadcast channel. This result is also

useful for proving the converse. For a general 2-user broadcast channel, the optimality of single-letter characterization was proved in [3] using information theoretic techniques. Their technique does not extend to more than 2 receivers for broadcast with degraded message sets. However, our analysis can be applied for any number receivers as long as all channels are assumed to be very noisy and a single-letter characterization of the capacity region could be obtained. In particular, this can prove the capacity conjecture in [5] when the multilevel broadcast networks are very noisy.

Three user example: Consider a 3 user broadcast, where Z_1 and Z_2 want the common message at rate R_c and Y wants the common message as well as a private message at rate R_p .

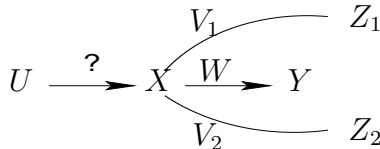


Figure 5: General 3-user broadcast channel: W, V_1, V_2 denote channels to Y, Z_1 and Z_2 .

The optimal R_c - R_p tradeoff in this case is obtained by solving,

$$\max_{U-X-(Y Z_1 Z_2): I(U;Y) \leq \gamma} \min\{I(U; Z_1), I(U; Z_2)\} \quad \text{where, } \gamma = I(X; Y) - R_p \quad (17)$$

Under the Euclidean approximation, this simplifies to

$$\max_{P_U, \phi_U: \mathbb{E}_{P_U} [\|\phi_U\|^2] \leq 2\gamma} \min \{ \mathbb{E}_{P_U} [\| B_{y \rightarrow z_1} \cdot \phi_U \|^2], \mathbb{E}_{P_U} [\| B_{y \rightarrow z_2} \cdot \phi_U \|^2] \}$$

where, $B_{y \rightarrow z_i} = [P_{Z_i}^{1/2}] \cdot (V_i' W'^{-1}) \cdot [P_Y^{1/2}]$

Here we have to choose a direction of ϕ_u to maximize the minimum of two quadratic forms. The solution need not be directly related to singular vectors of $B_{y \rightarrow z_1}$ and $B_{y \rightarrow z_2}$, nonetheless, it is just a quadratic optimization. It is particularly easy to solve when the singular-vectors of $B_{y \rightarrow z_1}$ and $B_{y \rightarrow z_2}$ are aligned. See [6] for a specific example of this kind.

References

- [1] I. Csiszar and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, 1981.
- [2] R. Gallager, Capacity and coding for degraded broadcast channels, *Probl. Pered. Inform.*, vol. 10, no. 3, pp. 3-14, 1974.
- [3] K. Marton and J. Korner, "General broadcast channels with degraded message sets," *IEEE Trans. Inform. Theory*, vol. 23, no. 1, pp. 60-64, Jan. 1977.
- [4] R. Gallager, *Information Theory and Reliable Communication*, Wiley, 1968.
- [5] S. Borade, L. Zheng and M. Trott, "Multilevel broadcast networks," IEEE Int. Symp. on Info. Theory, Nice, France, June, 2007.
- [6] S. Borade and L. Zheng, "Euclidean information theory," preprint.