Linear Information Coupling Problems

Shao-Lun Huang
Department of EECS
Massachusetts Institute of Technology
Cambridge, MA 02139-4307
Email: shaolun@mit.edu

Lizhong Zheng
Department of EECS
Massachusetts Institute of Technology
Cambridge, MA 02139-4307
Email: lizhong@mit.edu

***Abstract*—Many network information theory problems face the similar difficulty of single letterization. We argue that this is due to the lack of a geometric structure on the space of probability distribution. In this paper, we develop such a structure by assuming that the distributions of interest are close to each other. Under this assumption, the K-L divergence is reduced to the squared Euclidean metric in an Euclidean space. Moreover, we construct the notion of coordinate and inner product, which will facilitate solving communication problems. We will also present the application of this approach to the point-to-point channel and the general broadcast channel, which demonstrates how our technique simplifies information theory problems.**

I. Introduction

Since Shannon introduced the notion of capacity sixty years ago, finding the capacity of channels and networks are core problems in information theory. The analyzation of capacity answers the problem that how many bits can be transmitted through a communication network, and also provides many insights in engineer problems [1]. However, for general problems, there is no systematic way to obtain optimal single-letter solutions. By cleverly picking auxiliary random variables, we sometimes can prove the constructed single-letter solutions are optimal for some problems. But, when this fails, we cannot tell whether it is because we have not tried hard enough or the problem itself does not have an optimal single-letter solution.

The difficulty of obtaining optimal single letter solutions comes from the fact that, most of the information theoretical quantities, such as entropy, mutual information, and error exponents, are all special cases of the Kullback-Leibler (K-L) divergence. The K-L divergence is a measure of distance between two probability distributions. However, in multi-terminal communication problems, there are multiple input and output distributions, and we usually need to deal with problems in high dimensional probability spaces. In these cases, describing problems only with the distance measure is cumbersome, and solving these information theory problems turns out to be extremely hard even with numerical aids. Therefore, we need more geometric structures to describe the problems, such as inner products. This is however difficult, as the K-L divergence between two distributions, $D(P\|Q)$, is not symmetric between P and Q, and the K-L divergence is in general not a valid metric. Thus, the space of probability distributions is not a linear vector space but a manifold [2], when the K-L divergence behaves as the distance measure.

In this paper, we present an approach [3] to simplify the problems with the assumption that the distributions of interest are close to each other. With this assumption, the manifold formed by distributions can be approximated by a tangent plane, which is an Euclidean space. Moreover, the K-L divergence will behave as the squared Euclidean metric between distributions in this Euclidean space. Therefore, we obtain the notion of coordinate, inner product, and orthogonality in a linear metric space, to describe information theory problems. Moreover, we will demonstrate in the rest sections that, the linear structure constructed from out local approximation will transfer information theory problems to linear algebra problems, which can be solved easily. In particular, we show that a systematic approach can be used to solve the single-letterization problems. We apply this to the general broadcast channel, and obtain new insights of the optimality of the existing solutions [4]-[7]

The rest of this paper is organized as follows. We introduce the notion of local approximation in section II, and show that the K-L divergence can be approximated as the squared Euclidean metric. In section III and IV, we present the application of our local approximation to the point-to-point channel and the general broadcast channel, respectively. We will illustrate how the information theory problems become simple linear algebra problems when applying our technique.

II. The Local Approximation

The key step of our approach is to use a local approximation of the Kullback-Leibler (K-L) divergence. Let P and Q be two distributions over the same alphabet $\mathcal{X}$. We assume that $Q(x) = P(x) + \epsilon J(x)$, for some small value ϵ, then the K-L divergence can be written, with second order Taylor expansion, as

$$\begin{aligned} D(P\|Q) &= -\sum_x P(x)\log\frac{Q(x)}{P(x)} \\ &= -\sum_x P(x)\log\left(1+\epsilon\cdot\frac{J(x)}{P(x)}\right) \\ &= \frac{1}{2}\epsilon^2\cdot\sum_x\frac{1}{P(x)}J^2(x)+o(\epsilon^2). \end{aligned}$$

We denote $\sum_x J^2(x)/P(x)$ as $\|J\|_P^2$, which is the weighted norm square of the perturbation vector J. It is easy to verify here that replacing the weight in this norm by Q only results in

a $o(\epsilon^2)$ difference. That is, up to the first order approximation, the weights in the norm simply indicate the neighborhood of distributions where the divergence is computed. As a consequence, $D(P\|Q)$ and $D(Q\|P)$ are considered as equal up to the first order approximation.

For convenience, we define the weighted perturbation vector as

$$L(x) \triangleq \frac{1}{\sqrt{P(x)}} J(x), \quad \forall x,$$

or in vector form $L \triangleq \left[\sqrt{P}^{-1}\right] J$, where $\left[\sqrt{P}^{-1}\right]$ represents the diagonal matrix with entries $\left\{\sqrt{P(x)}^{-1},\ x \in \mathcal{X}\right\}$. This allows us to write $\|J\|_P^2 = \|L\|^2$, where the last norm is simply the Euclidean norm.

With this definition of the norm on the perturbations of distributions, we can generalize to define the corresponding notion of inner products. Let $Q_i(x) = P(x) + \epsilon \cdot J_i(x), \forall x, i = 1, 2$, we can define

$$\langle J_1, J_2 \rangle_P \triangleq \sum_x \frac{1}{P(x)} J_1(x) J_2(x) = \langle L_1, L_2 \rangle,$$

where $L_i = \left[\sqrt{P}^{-1}\right] J_i$, for $i = 1, 2$. From this, notions of orthogonal perturbations and projections can be similarly defined. The point here is that we can view a neighborhood of distributions as a linear metric space, and define notions of orthonormal basis and coordinates on it.

III. The Point to Point Channel

We start by using this local geometric structure to study the point-to-point channels to demonstrate the new insights we can obtain from this approach, even on a well-understood problem. It is well-known that the capacity problem is

$$\max_{P_X} I(X; Y), \tag{1}$$

but this is in fact a single letter solution of the coding problem

$$\max_{U \to X^n \to Y^n} \frac{1}{n} I(U; Y^n), \tag{2}$$

for some discrete random variable U, such that $U \to X^n \to Y^n$ forms a Markov chain.

Now, to apply the local approximations, instead of solving (2). we study a slightly different problem

$$\max_{U \to X^n \to Y^n : \frac{1}{n} I(U; X^n) \le \frac{1}{2}\epsilon^2} \frac{1}{n} I(U; Y^n). \tag{3}$$

We call the problem (3) as the *linear information coupling problem*. The only difference between (2) and (3) lies in the constraint $\frac{1}{n} I(U; X^n) \le \frac{1}{2}\epsilon^2$ on (3). That is, instead of trying to find how many bits in total that we can send through the given channel, we ask the question of how efficiently we can send a thin layer of information through this channel. One advantage of (3) is that it allows easy single letterization as we will demonstrate in the following. In fact, the step of single-letterization, namely, form (2) to (1), is the difficult step of most network problems. For these problems, the approach we used for the point-to-point problem can not be applied. What we will show in the rest of this section is that there is an alternative approach to do the well-known steps [1] to go from (2) to (1), and this new approach based on the geometric structures can be applied to more general problems. For simplicity, in this paper, we assume that the marginal distribution P_{X^n} is given, and is an i.i.d. distribution over the n letters[1], so that we can focus on finding U and the conditional distribution $P_{X^n|U}$ optimizing (3).

First, we solve the single-letter version, namely $n = 1$, of this problem. Observing that we can write the constraint as

$$I(U; X) = \sum_u P_U(u) \cdot D(P_{X|U}(\cdot|u) \| P_X) \le \frac{1}{2}\epsilon^2.$$

This implies that for each value of u, the conditional distribution $P_{X|U=u}$ is a local perturbation from P_X, that is, $P_{X|U=u} = P_X + \epsilon \cdot J_u$.

Next, using the notation that $L_u = \left[\sqrt{P}^{-1}\right] J_u$, for each value of u, we observe that

$$\begin{aligned} P_{Y|U=u} &= W P_{X|U=u} \\ &= W P_X + \epsilon \cdot W J_u \\ &= P_Y + \epsilon \cdot W[\sqrt{P_X}] L_u, \end{aligned}$$

where the channel applied to an input distribution is simply viewed as the channel matrix W, of dimension $|\mathcal{Y}| \times |\mathcal{X}|$, multiplying the input distribution as a vector. At this point, we have reduced both the spaces of input and output distributions as linear spaces, and the channel acts as a linear transform between these two spaces. The information coupling problem can be rewritten as, ignoring the $o(\epsilon^2)$ terms:

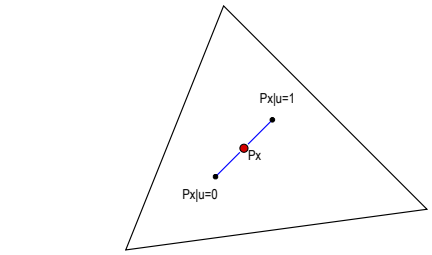
$$\begin{aligned} \max .\ & \sum_u P_U(u) \cdot \|W J_u\|_{P_Y}^2, \\ \text{subject to: } & \sum_u P_U(u) \cdot \|J_u\|_{P_X}^2 = 1, \end{aligned}$$

or equivalently in terms of Euclidean norms,

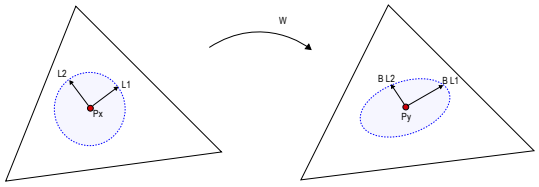
$$\begin{aligned} \max .\ & \sum_u P_U(u) \cdot \| \left[\sqrt{P_Y}^{-1}\right] W \left[\sqrt{P_X}\right] \cdot L_u\|^2, \\ \text{subject to: } & \sum_u P_U(u) \cdot \|L_u\|^2 = 1. \end{aligned}$$

This problem of linear algebra is simple. We need to find the joint distribution $U \to X \to Y$ by specifying the P_U and the perturbations J_u for each value of u, such that the marginal constraint on P_X is met, and also these perturbations are the most visible at the Y end, in the sense that multiplied by the channel matrix, WJ_u's have large norms. This can be readily solved by setting the weighted perturbation vectors L_u's to be along the input (right) singular vectors of the matrix $B \triangleq \left[\sqrt{P_Y}^{-1}\right] W \left[\sqrt{P_X}\right]$ with large singular values. Moreover, the

[1]This assumption can be proved to be "without loss of the optimality" for some cases [3]. In general, it requires a separate optimization, which is not the main issue addressed in this paper. To that end, we also assume that the given marginal P_{X^n} has strictly positive entries.



(a)



(b)

Fig. 1. (a) Choice of P_U and $P_{X|U}$ to maintain the marginal P_X. (b) Divergence Transition Map as a linear map between two spaces, with right and left singular vectors as orthonormal bases.

choice of P_U has no effect in the optimization, and might be taken as binary uniform for simplicity. This is illustrated in Figure 1(a).

We call the matrix B the *divergence transition matrix* (DTM). It maps divergence in the space of input distributions to that of the output distributions. The singular value decomposition (SVD) structure of this linear map has a critical role of our analysis. It can be shown that the largest singular value of B is 1, corresponding to an input singular vector $\left[\sqrt{P_X}, x \in \mathcal{X}\right]^T$, which is orthogonal to the simplex of probability distributions. This is not a valid choice for perturbation vectors. However, all vectors orthogonal to this vector, or equivalently, all linear combinations of other singular vectors are valid choices of the perturbation vectors L_u. Thus, the optimum of the above problem is achieved by setting L_u to be along the singular vector with the second largest singular value.

This can be visualized as in Figure 1(b), the orthonormal bases for the input and output spaces, respectively, according to the right and left singular vectors of B. The key point here is that while $I(U;X)$ measures how many bits of information is modulated in X, depending on how they are modulated, in terms of which direction the corresponding perturbation vector is, these bits have different levels of visibility at the Y end. This is a quantitative way to show why viewing a channel as a bit-pipe carrying uniform bits is a bad idea.

Moreover, recalling that the data processing inequality tells that, from the Markov chain $U \to X \to Y$, the mutual informations have the relation $I(U;X) \geq I(U;Y)$. Let us assume that the second largest singular value of B is $\sigma \leq 1$, then the above derivations imply that $\sigma^2 \cdot I(U;X) \geq I(U;Y)$. Thus, we actually come up with a stronger result than the data processing inequality, and the equality can be achieved by setting the perturbation vector to be along the right singular vector of B, with the second largest singular value.

The most important feature of the linear information coupling problem is that the single-letterization (3) is simple. To illustrate the idea, we consider a 2-letter version of the point-to-point channel:

$$\max_{U \to X^2 \to Y^2 : \frac{1}{2} I(U;X^2) \leq \frac{1}{2}\epsilon^2} \frac{1}{2} I(U;Y^2). \tag{4}$$

Let P_X, P_Y, W, and B be the input and output distributions, channel matrix, and the DTM, respectively for the single letter version of the problem. Then, the 2-letter problem has $P_X^{(2)} = P_X \otimes P_X$, $P_Y^{(2)} = P_Y \otimes P_Y$, and $W^{(2)} = W \otimes W$, where $\otimes$ denotes the Kronecker product. As a result, the new DTM is $B^{(2)} = B \otimes B$. We have the following lemma on the singular values and vectors of $B^{(2)}$.

Lemma 1. *Let v_i and v_j denote two singular vectors of B with singular values μ_i and μ_j. Then $v_i \otimes v_j$ is a singular vector of $B^{(2)}$ and its singular value is $\mu_i \mu_j$.*

Now, recall that the largest singular value of B is $\mu_0 = 1$, with the singular vector $v_0 = \left[\sqrt{P_X}, x \in \mathcal{X}\right]^T$, which corresponds to the direction orthogonal to the distribution simplex. This implies that the largest singular value of $B^{(2)}$ is also 1, again corresponds to the direction that is orthogonal to all valid choices of the perturbation vectors.

The second largest singular value of $B^{(2)}$ is a tie between $\mu_0\mu_1$ and $\mu_1\mu_0$, with singular vectors $v_0 \otimes v_1$ and $v_1 \otimes v_0$. The optimal solution of (4) is thus to set the perturbation vectors to be along these two vectors. This can be written as

$$\begin{aligned} &P_{X^2|U=u} \\ =&P_X \otimes P_X + \left[\sqrt{P_X \otimes P_X}\right] \cdot (\epsilon v_0 \otimes v_1 + \epsilon' v_1 \otimes v_0) \\ =&\left(P_X + \epsilon' \left[\sqrt{P_X}\right] v_1\right) \otimes \left(P_X + \epsilon \left[\sqrt{P_X}\right] v_1\right) + O(\epsilon^2). \end{aligned}$$

Here, we use the fact that $v_0 = \left[\sqrt{P_X}, x \in \mathcal{X}\right]^T$. This means that the optimal conditional distribution $P_{X^2|U=u}$ for any u has the product form, up to the first order approximation. With a simple time-sharing argument, it is easy to see that we can indeed set $\epsilon = \epsilon'$, that is, pick this conditional distribution to be i.i.d. over the two symbols, to achieve the optimum.

The simplicity of this proof of the optimality of the single letter solutions is astonishing. All we have used is the fact that the singular vector of $B^{(2)}$ corresponding to the second largest singular value has a special form. A distribution in the neighborhood of $P_X \otimes P_X$ is a product distribution if and only if it can be written as a perturbation from $P_X \otimes P_X$, along the subspace spanned by vectors $v_0 \otimes v_i$ and $v_j \otimes v_0$, in the form of $v_0 \otimes v + v' \otimes v_0$, for some v and v'. Thus, all we need to do is to find the eigen-structure of the B-matrix, and verify if the optimal solutions have this form. This procedure is used in more general problems.

One way to explain why the local approximation is useful is as follows. In general, tradeoff between multiple K-L divergence (mutual information) is a non-convex problem. Thus, finding global optimum for such problems is in general intrinsically intractable and extremely hard. In contrast, with our local approximation, the K-L divergence becomes a quadratic function. Now, the tradeoff between quadratic functions remains quadratic. Effectively, our approach focus on verifying the local optimality of the quadratic solutions, which is a natural thing to do, since the overall problem is not convex.

IV. The General Broadcast Channel

Let us now apply our local approximation approach to the general broadcast channel. A general broadcast channel with input $X \in \mathcal{X}$, and outputs $Y_1 \in \mathcal{Y}_1$, $Y_2 \in \mathcal{Y}_2$, is specified by the memoryless channel matrices W_1 and W_2. These channel matrices specify the conditional distributions of the output signals at two users, 1 and 2 as $W_i(y_i|x) = P_{Y_i|X}(y_i|x)$, for $i = 1, 2$. Let M_1, M_2, and M_0 be the two private messages and the common message, with rate R_1, and R_2, and R_0, respectively. The multi-letter capacity region can be written as

$$\begin{cases} R_0 & \leq \frac{1}{n}\min\{I(U;Y_1^n), I(U;Y_2^n)\}, \\ R_1 & \leq \frac{1}{n} I(V_1;Y_1^n), \\ R_2 & \leq \frac{1}{n} I(V_2;Y_2^n), \end{cases}$$

for some mutually independent random variables U, V_1, and V_2, such that $(U, V_1, V_2) \to X^n \to (Y_1^n, Y_2^n)$ forms a Markov chains.

The linear information coupling problems of the private messages, given that the common message is decoded, are essentially the same as the point-to-point channel case. Thus, we only need to focus on the linear information coupling problem of the common message:

$$\max . \ \frac{1}{n}\min\left\{I(U;Y_1^n), I(U;Y_2^n)\right\}, \tag{5}$$
$$\text{subject to: } U \to X^n \to (Y_1^n, Y_2^n) : \frac{1}{n} I(U;X^n) \leq \frac{1}{2}\epsilon^2$$

The core problem we want to address here is that whether or not the single-letter solutions are optimal for (5). To do this, suppose that $P_{X^n|U=u} = P_X^{(n)} + \epsilon \cdot J_u$. Define the DTMs $B_i \triangleq \left[\sqrt{P_{Y_i}}^{-1}\right] W_i \left[\sqrt{P_X}\right]$, for $i = 1, 2$, and the scaled perturbation $L_u = \left[\sqrt{P_X^{(n)}}^{-1}\right] J_u$, the problem then becomes

$$\max_{L_u:\|L_u\|^2=1} \min\left\{\|B_1^{(n)} L_u\|^2, \|B_2^{(n)} L_u\|^2\right\}, \tag{6}$$

where $B_i^{(n)}$ is the n^{th} Kronecker product of the single-letter DTM B_i, for $i = 1, 2$.

Different from the point-to-point problem, we need to choose the perturbation vectors L_u to have large images simultaneously through two different linear systems. In general, the tradeoff between two SVD structures can be rather messy problems. However, in this problem, for both $i = 1, 2$, $B_i^{(n)}$ have the special structure of being the Kronecker product of the single letter DTMs. Furthermore, both B_1 and B_2 have the largest singular value of 1, corresponding to the same singular vector $\phi_0 = \left[\sqrt{P_X}, x \in \mathcal{X}\right]^T$, although the rest of their SVD structures are not specified. The following theory characterizes the optimality of single-letter and finite-letter solutions for the general cases.

Theorem 1. *Suppose that B_i are DTM's for some DMC and input/output distributions, for $i = 1, 2, \ldots, k$, then the linear information coupling problem*

$$\max_{L_u:\|L_u\|^2=1} \min_{1\leq i\leq k} \left\{\|B_i^{(n)} L_u\|^2\right\}, \tag{7}$$

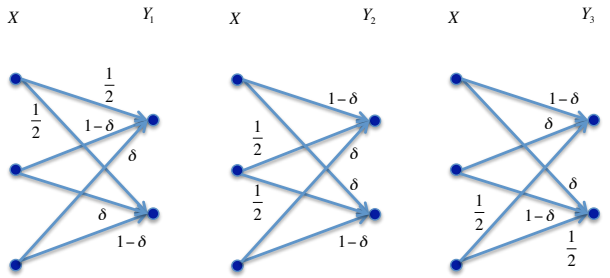
has optimal single letter solutions for the case with 2 receivers. In general, when there are $k > 2$ receivers, single letter solutions can not be optimal, when the cardinality $|\mathcal{U}|$ is bounded by some function of $|\mathcal{X}|$. However, there still exists k-letter solutions that are optimal.

While we will not present the full proof of this result in this paper, it worth pointing out how conceptually straightforward it is. We can write the right singular vectors of the two DTMs, B_1 and B_2, as $\phi_0, \phi_1, \ldots, \phi_{n-1}$ and $\varphi_0, \varphi_1, \ldots, \varphi_{n-1}$. The only structure we have is that $\phi_0 = \varphi_0 = \left[\sqrt{P_X(x)}, x \in X\right]^T$, both correspond to the largest singular value of 1. For other vectors, the relation between the two bases can be written as a unitary matrix Ψ, with $\phi_i = \sum_j \Psi_{ij}\varphi_j$. Now, we can define an orthonormal basis for the space of multi-letter distributions on X^n. For example, with 2-letter distributions, we can use $\phi_i \otimes \phi_j, i, j \in \{0, 1, ..., n-1\}$ and $(i, j) \neq (0, 0)$. Note that any L_u can be written as $L_u = \sum_{i,j\neq(0,0)} \alpha_{ij}\phi_i \otimes \phi_j$. If a perturbation vector L_u has any non-zero component along $\phi_i \otimes \phi_j$, with $i, j \neq 0$, we can always move this component to either $\phi_i \otimes \phi_0$ or $\phi_0 \otimes \phi_j$ to have, say, $\phi_i \otimes \phi_0 = \sum_j \Psi_{ij}\varphi_j \otimes \varphi_0$. This results in larger norms of the output vectors through both channels. As a result, the optimizer of (7) can only have components on the vectors $\phi_0 \otimes \phi_j$ and $\phi_i \otimes \phi_0$. This means that the resulting conditional distribution must be product distributions, i.e. $P_{X^n|U=u} = P_{X_1|U=u} \cdot P_{X_2|U=u} \ldots$. This simple observation greatly simplifies the multi-letter optimization problem: instead of searching for general joint distributions, now we have the further constraint of conditional independence. This directly gives rise to the proof of the optimality of i.i.d. distributions and hence single letter solutions for the 2 user case, which is the first definitive answer on the general broadcast channels.

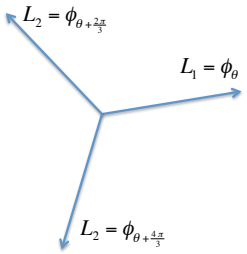
The more interesting case is when there are more than 2 receivers. In such cases, i.i.d. distributions simply do not have enough degrees of freedom to be optimal in the tradeoff of more than 2 linear systems. Instead, one has to design multi-letter product distributions to achieve the optimal. The following example, constructed with the geometric method, illustrate the key ideas.

Example:

We consider a 3-user broadcast channel. Let the input alphabet $\mathcal{X}$ be ternary, so that the perturbation vectors have



(a)



(b)

Fig. 2. (a) A ternary input broadcast channel (b) The optimal perturbations over 3 time slots.

2 dimensions and can be easily visualized. Suppose that the three DTMs are rotations of 0, $2\pi/3$, $4\pi/3$, respectively, followed (left multiplied) by the projection to the horizontal axis. This corresponds to the ternary input channels as shown in Figure 2(a). Now if we use single-letter inputs, it can be seen that for any L_u with $\|L_u\|^2 = 1$, $\min\left\{\|B_1 L_u\|^2, \|B_2 L_u\|^2, \|B_3 L_u\|^2\right\} \leq 1/4$. The problem here is that no matter what direction L_{u_0} takes, the three output norms are unequal, and the minimum one always limits the performance. Now, if we use 3-letter input, and denote $\phi_\theta = [\cos\theta, \sin\theta]^T$, then we can take

$$P_{X^3|U=u} = (P_X + \epsilon\phi_\theta) \otimes (P_X + \epsilon\phi_{\theta+\frac{2\pi}{3}}) \otimes (P_X + \epsilon\phi_{\theta+\frac{4\pi}{3}})$$

for any value of θ, as shown in Figure 2(b). Intuitively, this input equalizes the three channels, and gives for all $i = 1, 2, 3$, $\|B_i^{(3)} L_u^{(3)}\|^2 = 1/2$, which doubles the information coupling rate. Translating this solution to the coding language, it means that we take turns to feed the common information to each individual user. Note that the solution is not a standard time-sharing input, and hence the performance is strictly out of the convex hull of i.i.d. solutions. One can interpret this input as a repetition of the common message over three time-slots, where the information is modulated along equally rotated vectors. For this reason, we call this example the "windmill" channel. Additionally, it is easy to see that the construction of the windmill channel can be generalized to the cases of $k > 3$ receivers, where k-letter solutions is necessary.

Note that in this example, we let U be a binary random variable, and in this case, while there are optimal 3-letter solutions, the optimal single-letters do not exist. However, one can in fact take U to be non-binary. For example, let $\mathcal{U} = \{0, 1, 2, 3, 4, 5\}$ with $P_U(u) = 1/6$ for all u, and let $L_{U=0} = -L_{U=1} = \phi_\theta$, $L_{U=2} = -L_{U=3} = \phi_{\theta+\frac{2\pi}{3}}$, and $L_{U=4} = -L_{U=5} = \phi_{\theta+\frac{4\pi}{3}}$, then we can still achieve the information coupling rate 1/2. Thus, there actually exits an optimal single-letter solution with cardinality $|\mathcal{U}| = 6$. However, when there are k receivers, it requires cardinality $|\mathcal{U}| = 2k$ for obtaining optimal single-letter solutions. Essentially, this example shows that finding a single perturbation vector with a large image at the outputs of all 3 channels is difficult. The tension between these 3 linear systems requires more degrees of freedom in choosing the perturbations, or in other words, the way that common information is modulated. Such more degrees of freedom can be provided either by using multi-letter solutions or have larger cardinality bounds. This effect is not captured by the conventional single-letterization approach.

Theorem 1 reduces most of the difficulty of solving the multi-letter optimization problem (5). The remaining is to find the optimal scaled perturbation L_u for the single-letter version of (5), if the number of receivers $k = 2$, or the k-letter version, if $k > 2$. These are finite dimensional convex optimization problems [8], which can be readily solved.

We can see that all these information theory problems are solved with essentially the same procedure, and all we need in solving these problems is simple linear algebra. This again, demonstrates the simplicity and uniformity of our approach in dealing with information theory problems.

V. Conclusion

In this paper, we present the local approximation approach, and show that with this approach, we can handle the issue of single-letterization in information theory by just solving simple linear algebra problems. Moreover, we demonstrate that our approach can be applied to different communication problems with the same procedure, which is a very attractive property. Finally, we provide the geometric insight of the optimal finite-letter solutions in sending the common message to $k > 2$ receivers, which also explains why optimal single-letter solutions fail to be existed in these cases.

References

[1] T. Cover and J. Thomas, *Elementary of Information Theory,*, Wiley Interscience, 1991.

[2] Shun-ichi Amari and Hiroshi Nagaoka, *Methods of Information Geometry,*, Oxford University Press, 2000.

[3] S. Borade and L. Zheng, "Euclidean Information Theory," *IEEE International Zurich Seminars on Communications*, March, 2008.

[4] T. Cover, "An Achievable Rate Region for the Broadcast Channel," *IEEE Transactions on Information Theory*, Vol. IT-21, pp. 399-404, July, 1975

[5] E. van der Meulen, "Random Coding Theorems for the General Discrete Memoryless Broadcast Channel," *IEEE Transactions on Information Theory,*, Vol. IT-21, pp. 180-190, March, 1975

[6] B. E. Hajek and M. B. Pursley, "Evaluation of an Achievable Rate Region for the Broadcast Channel," *IEEE Transactions on Information Theory,*, Vol. IT-25, pp. 36-46, Jan, 1979

[7] K. Marton, "A Coding Theorem for the Discrete Memoryless Broadcast Channel," *IEEE Transactions on Information Theory*, Vol. IT-25, pp. 306-311, May, 1979

[8] S. Boyd and L. Vandenberghe, *Convex Optimization,*, Cambridge University Press, 2004 .