Geometric Programming Duals of Channel Capacity and Rate Distortion

Mung Chiang, *Member, IEEE,* and Stephen Boyd, *Fellow, IEEE*

***Abstract*—We show that the Lagrange dual problems of the channel capacity problem with input cost and the rate distortion problem are simple geometric programs. Upper bounds on channel capacity and lower bounds on rate distortion can be efficiently generated from their duals. For channel capacity, the geometric programming dual characterization is shown to be equivalent to the minmax Kullback–Leibler (KL) characterization in [10], [14]. For rate distortion, the geometric programming dual is extended to rate distortion with two-sided state information.**

A "duality by mapping" is then given between the Lagrange dual problems of channel capacity with input cost and rate distortion, which resolves several apparent asymmetries between their primal problems in the familiar form of mutual information optimization problems. Both the primal and dual problems can be interpreted in a common framework of free energy optimization from statistical physics.

***Index Terms*—Channel capacity, convex optimization, duality, free energy, geometric programming, rate distortion.**

I. Introduction

A. Overview

SHANNON made the following well-known remarks in [27]:

> *There is a curious and provocative duality between the properties of a source with a distortion measure and those of a channel. This duality is enhanced if we consider channels in which there is a cost associated with the different input letters…Solving this problem corresponds, in a sense, to finding a source that is right for the channel and the desired cost…In a somewhat dual way, evaluating the rate distortion function for a source…corresponds to finding a channel that is just right for the source and allowed distortion level.*

Thus, the two fundamental limits of data transmission and data compression are "somewhat dual" for the basic discrete memoryless systems in [26], [27]. This paper gives an exact and detailed characterization of this Shannon duality between data transmission and compression through Lagrange duality in convex optimization.

In Sections II and III, we show that the Lagrange dual problems of channel capacity with input cost and of rate distortion are simple geometric programs, a special type of nonlinear optimization problems to be introduced in Section I-B. This implies that any Lagrange dual variable satisfying the Lagrange dual constraints gives an upper bound on channel capacity or a lower bound on rate distortion, and that the optimized value of the Lagrange dual problem equals channel capacity or rate distortion. For channel capacity, the geometric programming characterization is shown to be equivalent to the minmax Kullback–Leibler (KL) characterization in [10], [14], which obtains channel capacity by minimizing over output distributions. For rate distortion, the geometric programming dual is extended to rate distortion with state information [8], [31]. Section IV-A shows a "duality by mapping" relationship between the Lagrange dual problems of channel capacity and rate distortion, thus characterizing Shannon duality through Lagrange duality. Section IV-B interprets channel capacity as a free energy optimization problem in statistical physics, complementing the free energy interpretation of rate distortion in [1].

It is not too surprising that channel capacity and rate distortion can be obtained by geometric programs. Section I-B shows that geometric programs can easily be turned into convex optimization problems using the same inequality that proves the convexity of KL divergence, and the appendix connects the special forms of the geometric program duals with typicality arguments. In [6], it is shown that large deviations bounds for independent and identically distributed (i.i.d.) random variables and Markov chains can be obtained through geometric programs, and that generalized free energy optimization problems and various lossless source coding relaxations are also geometric programs.

We will use the following notation. Probability distributions are represented as row vectors. Both column and row vectors are denoted in boldface, and matrices in capital letter boldface. Given two column vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ of length n, we express the sum $\sum_{i=1}^{n} x_i y_i$ as an inner product $\boldsymbol{x}^T\boldsymbol{y}$. If $\boldsymbol{x}$ is a row vector instead, the inner product is denoted as $\boldsymbol{xy}$. Componentwise inequalities on a vector $\boldsymbol{x}$ with n entries are expressed using the $\succeq$ symbol: $\boldsymbol{x} \succeq 0$ denotes $x_i \geq 0, i = 1, 2, \ldots, n$. A column vector with all entries being 1 is denoted as $\mathbf{1}$, and the length of $\mathbf{1}$ will be clear from the context.

B. Geometric Programming

All problems treated in this paper are convex optimization problems: minimizing a convex objective function subject to upper bound inequality constraints on other convex functions. It is well known that for a convex optimization problem, a local

Manuscript received January 6, 2003; revised October 6, 2003. This work was supported by the Hertz Foundation Fellowship and the Stanford Graduate Fellowship. The material in this paper was presented in part at the 40th Allerton Conference on Communication, Control and Computing, Monticello, IL, October 2002, and the IEEE International Symposium on Information Theory, Yokohama, Japan, June/July 2003.

M. Chiang is with the Electrical Engineering Department, Princeton University, Princeton, NJ 08544 USA (e-mail: chiangm@princeton.edu).

S. Boyd is with the Electrical Engineering Department, Stanford University, Stanford, CA 94305 USA (e-mail: boyd@stanford.edu).

Communicated by R. W. Yeung, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2003.822581

minimum is also a global minimum. Lagrange duality theory is also well developed for convex optimization. For example, the duality gap is zero under mild technical conditions such as Slater's condition [4] that requires the existence of a strictly feasible point. When put in an appropriate form with the right data structure, a convex optimization problem is also easy to solve numerically by efficient algorithms such as the primal-dual interior-point methods [4], [22]. In this paper, we will use geometric programming [11], a type of nonlinear problems that can be turned into convex optimization. Geometric programs have been used for various engineering problems, including recently for resource allocation in communication networks [5]–[7], [16], [17].

We first define a monomial as a function $f : \boldsymbol{R}_+^n \rightarrow \boldsymbol{R}$

$$f(\boldsymbol{x}) = d x_1^{a^{(1)}} x_2^{a^{(2)}} \ldots x_n^{a^{(n)}}$$

where the multiplicative constant $d \geq 0$ and the exponential constants $a^{(j)} \in \boldsymbol{R}$, $j = 1, 2, \ldots, n$. A sum of monomials is called a posynomial

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} d_k x_1^{a_k^{(1)}} x_2^{a_k^{(2)}} \ldots x_n^{a_k^{(n)}}$$

where $d_k \geq 0$, $k = 1, 2, \ldots, K$, and $a_k^{(j)} \in \boldsymbol{R}$, $j = 1, 2, \ldots, n$, $k = 1, 2, \ldots, K$.

Minimizing a posynomial subject to posynomial inequality constraints and monomial equality constraints is called a geometric program in standard form

$$\begin{array}{lll} \text{minimize} & f_0(\boldsymbol{x}) & \\ \text{subject to} & f_i(\boldsymbol{x}) \leq 1, & i = 1, 2, \ldots, m \\ & h_l(\boldsymbol{x}) = 1, & l = 1, 2, \ldots, M \end{array} \qquad (1)$$

where f_i, $i = 0, 1, \ldots, m$ are posynomials

$$f_i(\boldsymbol{x}) = \sum_{k=1}^{K_i} d_{ik} x_1^{a_{ik}^{(1)}} x_2^{a_{ik}^{(2)}} \cdots x_n^{a_{ik}^{(n)}}$$

and h_l, $l = 1, 2, \ldots, M$ are monomials

$$h_l(\boldsymbol{x}) = d_l x_1^{a_l^{(1)}} x_2^{a_l^{(2)}} \cdots x_n^{a_l^{(n)}}.$$

Note that the objective of a geometric program can also be the maximization of a monomial, since this is equivalent to the minimization of the reciprocal of a monomial, which is again a monomial. Given a geometric program in standard form, we can form a matrix $\boldsymbol{A}$ where each row consists of the exponential constants associated with each term of the posynomials and monomials, and a vector $\boldsymbol{d}$ consisting of all the multiplicative constants.

Geometric programming in the above standard form is not a convex optimization problem. However, with a logarithmic change of variables and constants: $y_j = \log x_j$, $b_{ik} = \log d_{ik}$, $b_l = \log d_l$, we can turn it into the following geometric program in convex form with variables $\boldsymbol{y}$:

$$\begin{array}{ll} \text{minimize} & p_0(\boldsymbol{y}) = \log \sum_{k=1}^{K_0} \exp\left(\boldsymbol{a}_{0k}^T \boldsymbol{y} + b_{0k}\right) \\ \text{subject to} & p_i(\boldsymbol{y}) = \log \sum_{k=1}^{K_i} \exp\left(\boldsymbol{a}_{ik}^T \boldsymbol{y} + b_{ik}\right) \leq 0, \\ & \qquad\qquad\qquad\qquad i = 1, 2, \ldots, m \\ & q_l(\boldsymbol{y}) = \boldsymbol{a}_l^T \boldsymbol{y} + b_l = 0, \qquad l = 1, 2, \ldots, M. \end{array} \qquad (2)$$

To show that (2) is indeed a convex optimization problem, consider the following log sum inequality (readily proved by the convexity of $f(t) = t \log t$, $t \geq 0$, [9], [10]):

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \qquad (3)$$

where $a_i, b_i \in \boldsymbol{R}_+$, $i = 1, 2, \ldots, n$. This inequality easily leads to the convexity of KL divergence

$$D(\boldsymbol{p} \| \boldsymbol{q}) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}$$

in $(\boldsymbol{p}, \boldsymbol{q})$ [9], which in turn shows that channel capacity and rate distortion problems are convex optimization problems.

Now consider the convexity of the LogSumExp function $f(\boldsymbol{x}) = \log \sum_{i=1}^{n} e^{x_i}$, which can be shown from the following conjugacy relationship. Given a function $f : \boldsymbol{R}^n \rightarrow \boldsymbol{R}$, the function $f^\star : \boldsymbol{R}^n \rightarrow \boldsymbol{R}$, defined as

$$f^\star(\boldsymbol{y}) = \sup_{\boldsymbol{x} \in \mathrm{dom} f} \left(\boldsymbol{y}^T \boldsymbol{x} - f(\boldsymbol{x})\right) \qquad (4)$$

is called the conjugate function of f. Since $f^\star$ is the pointwise supremum of a family of affine functions of $\boldsymbol{y}$, it is always a convex function. It is easy to verify that, if $f^\star(\boldsymbol{y})$ is the conjugate of $f(\boldsymbol{x})$, then for a given $T > 0$, the perspective function $T f^\star(\frac{\boldsymbol{y}}{T})$ is the conjugate of the scaled function $T f(\boldsymbol{x})$.

Let $\hat{b}_i = \log b_i$ and $\sum_{i=1}^{n} a_i = 1$ in the log sum inequality (3). We obtain

$$\log\left(\sum_{i=1}^{n} e^{\hat{b}_i}\right) \geq \boldsymbol{a}^T \hat{\boldsymbol{b}} - \sum_{i=1}^{n} a_i \log a_i$$

which by definition shows that LogSumExp is the conjugate function of negative entropy. Since all conjugate functions are convex, LogSumExp is convex. Therefore, geometric programs in convex form are indeed convex optimization problems.

As a small example, it is easily seen that the following optimization problem is a geometric program in standard form:

$$\begin{array}{ll} \text{minimize} & x(y + z) \\ \text{subject to} & x^2 \sqrt{yz} \geq 0.8, \ \sqrt{xy} \geq 0.5, \ x \geq 1. \end{array}$$

We will show in Section II-B that the above geometric program is, in fact, computing a channel capacity with an input cost constraint.

II. Lagrange Dual of Channel Capacity

A. The Channel Capacity Problem With Input Cost

First consider the problem of data transmission over a discrete memoryless channel with input $X \in \mathcal{X} = \{1,2,\dots,N\}$, output $Y \in \mathcal{Y} = \{1,2,\dots,M\}$, and channel law

$$P_{ij} = \mathbf{Prob}\{Y=j|X=i\},\ i=1,2,\dots,N,\ j=1,2,\dots,M.$$

The channel law forms a channel matrix $\boldsymbol{P} \in \boldsymbol{R}^{N\times M}$, where the (i,j) entry of $\boldsymbol{P}$ is $P_{ij} \geq 0$ with $\boldsymbol{P}\mathbf{1} = \mathbf{1}$. A distribution $\boldsymbol{p} \in \boldsymbol{R}^{1\times N}$ on the input, together with a given channel matrix $\boldsymbol{P}$, induces a distribution $\boldsymbol{q} \in \boldsymbol{R}^{1\times M}$ on the output by $\boldsymbol{q} = \boldsymbol{pP}$, and a joint distribution $\boldsymbol{Q}$ on the input output pair by $Q_{ij} = p_i P_{ij}$. We also associate with each input alphabet symbol i an input cost $s_i \geq 0$, forming a column vector $\boldsymbol{s}$.

It is a key result in information theory [20] that the capacity $C(S)$ of a discrete memoryless channel under the input cost constraint $\boldsymbol{E_p}[\boldsymbol{s}] = \boldsymbol{ps} \leq S$ is

$$C(S) = \max_{\boldsymbol{p}:\boldsymbol{ps}\leq S} I(X;Y) \qquad (5)$$

where the mutual information between input X and output Y is defined as

$$\begin{aligned} I(X;Y) &= \sum_{i=1}^{N}\sum_{j=1}^{M} Q_{ij}\log\frac{Q_{ij}}{p_i q_j} \\ &= H(Y) - H(Y|X) \\ &= -\sum_{j=1}^{M} q_j \log q_j - \boldsymbol{pr} \end{aligned}$$

where $\boldsymbol{r} \in \boldsymbol{R}^{N\times 1}$ and $r_i = -\sum_{j=1}^{M} P_{ij}\log P_{ij}$ is the conditional entropy of Y given $X = i$.

Therefore, we view channel capacity as the optimal objective value of the following maximization problem, referred to as the channel capacity problem with input cost:

$$\begin{aligned} \text{maximize} \quad & -\boldsymbol{pr} - \sum\nolimits_{j=1}^{M} q_j \log q_j \\ \text{subject to} \quad & \boldsymbol{pP} = \boldsymbol{q}, \quad \boldsymbol{ps} \leq S \\ & \boldsymbol{p}\mathbf{1} = 1, \quad \boldsymbol{p} \succeq 0 \end{aligned} \qquad (6)$$

where the optimization variables are $\boldsymbol{p}$ and $\boldsymbol{q}$. The constant parameters are $\boldsymbol{P}$ the channel matrix and

$$r_i = -\sum_{j=1}^{M} P_{ij}\log P_{ij}.$$

In the special case of no input cost constraint, the channel capacity problem becomes

$$\begin{aligned} \text{maximize} \quad & -\boldsymbol{pr} - \sum\nolimits_{j=1}^{M} q_j \log q_j \\ \text{subject to} \quad & \boldsymbol{pP} = \boldsymbol{q}, \\ & \boldsymbol{p}\mathbf{1} = 1, \quad \boldsymbol{p} \succeq 0. \end{aligned} \qquad (7)$$

If we substituted $\boldsymbol{q} = \boldsymbol{pP}$ in the objective function of (6), we would have found that the Lagrange dual problem can only be implicitly expressed through the solution of a system of linear equations. Keeping two sets of optimization variables $\boldsymbol{p}$ and $\boldsymbol{q}$, and introducing the equality constraint $\boldsymbol{pP} = \boldsymbol{q}$ in the primal problem is a key step to derive an explicit and simple Lagrange dual problem of (6).

B. Geometric Programming Dual

Proposition 1: The Lagrange dual of the channel capacity problem with input cost (6) is the following geometric program (in convex form):

$$\begin{aligned} \text{minimize} \quad & \log \sum\nolimits_{j=1}^{M} \exp(\alpha_j + \gamma S) \\ \text{subject to} \quad & \boldsymbol{P\alpha} + \gamma\boldsymbol{s} \succeq -\boldsymbol{r}, \\ & \gamma \geq 0 \end{aligned} \qquad (8)$$

where the optimization variables are $\boldsymbol{\alpha}$ and γ, and the constant parameters are $\boldsymbol{P}$, $\boldsymbol{s}$, and S.

An equivalent version of the Lagrange dual problem is the following geometric program (in standard form):

$$\begin{aligned} \text{minimize} \quad & w^S \sum\nolimits_{j=1}^{M} z_j \\ \text{subject to} \quad & w^{s_i} \prod\nolimits_{j=1}^{M} z_j^{P_{ij}} \geq e^{-H(\boldsymbol{P}^{(i)})},\ i=1,2,\dots,N, \\ & w \geq 1,\ z_j \geq 0,\ j=1,2,\dots,M \end{aligned} \qquad (9)$$

where the optimization variables are $\boldsymbol{z}$ and w, and $\boldsymbol{P}^{(i)}$ is the ith row of $\boldsymbol{P}$.

Lagrange duality between problems (6) and (8) means the following.

- *Weak duality.* Any feasible $(\boldsymbol{\alpha},\gamma)$ of the Lagrange dual problem (8) produce an upper bound on channel capacity with input cost: $\log\sum_{j=1}^{M}\exp(\alpha_j + \gamma S) \geq C(S)$.
- *Strong duality.* The optimal value of the Lagrange dual problem (8) is $C(S)$.

Proof: In order to find the Lagrange dual of problem (6), we first form the Lagrangian L as

$$\begin{aligned} L(\boldsymbol{p},\boldsymbol{q},\boldsymbol{\nu},\mu,\boldsymbol{\lambda},\gamma) = &-\boldsymbol{pr} - \sum_j q_j \log q_j + (\boldsymbol{q} - \boldsymbol{pP})\boldsymbol{\nu} \\ &+ \mu(1 - \boldsymbol{p}\mathbf{1}) + \boldsymbol{p\lambda} + \gamma(S - \boldsymbol{ps}) \end{aligned} \qquad (10)$$

with Lagrange multiplier vector $\boldsymbol{\nu} \in \boldsymbol{R}^{M\times 1}$, Lagrange multiplier $\mu, \gamma \in \boldsymbol{R}$, and Lagrange multiplier vector $\boldsymbol{\lambda} \in \boldsymbol{R}^{N\times 1}$. Since $\boldsymbol{\lambda}$ and γ correspond to the inequality constraints, we have $\boldsymbol{\lambda} \succeq 0$ and $\gamma \geq 0$.

We then find the Lagrange dual function $g(\boldsymbol{\nu},\mu,\boldsymbol{\lambda},\gamma) = \sup_{\boldsymbol{p},\boldsymbol{q}} L(\boldsymbol{p},\boldsymbol{q},\boldsymbol{\nu},\mu,\boldsymbol{\lambda},\gamma)$ by finding the $\boldsymbol{p}$ and $\boldsymbol{q}$ that maximize L, which is a concave function of $(\boldsymbol{p},\boldsymbol{q})$. First, note that L is a linear function of $\boldsymbol{p}$, thus bounded from above only when it is identically zero. As a result, $g(\boldsymbol{\nu},\mu,\boldsymbol{\lambda},\gamma) = \infty$ unless $\boldsymbol{r} + \boldsymbol{P\nu} + \mu\mathbf{1} - \boldsymbol{\lambda} + \gamma\boldsymbol{s} = 0$, which is equivalent to $\boldsymbol{r} + \boldsymbol{P\nu} + \mu\mathbf{1} + \gamma\boldsymbol{s} \succeq 0$ since $\boldsymbol{\lambda} \succeq 0$.

Assuming $\boldsymbol{r} + \boldsymbol{P\nu} + \mu\mathbf{1} + \gamma\boldsymbol{s} \succeq 0$, the Lagrangian becomes

$$-\sum_j (q_j \log q_j - \nu_j q_j) + \mu + \gamma S$$

which we must now maximize over $\boldsymbol{q}$. To find the maximum of $-q_j \log q_j + \nu_j q_j$ over q_j, we set the derivative with respect to q_j equal to zero: $\log q_j + 1 - \nu_j = 0$. Thus, $q_j = \exp(\nu_j - 1)$ is the maximizer, with the associated maximum value

$$-q_j \log q_j + \nu_j q_j = -e^{\nu_j - 1}(\nu_j - 1) + \nu_j e^{\nu_j - 1} = e^{\nu_j - 1}.$$

Therefore, the Lagrange dual function is

$$g(\boldsymbol{\nu}, \mu, \gamma) = \begin{cases} \sum_j \exp(\nu_j - 1) + \mu + \gamma S, & \boldsymbol{r} + \boldsymbol{P\nu} + \mu\boldsymbol{1} + \gamma\boldsymbol{s} \succeq 0 \\ \infty, & \text{otherwise.} \end{cases} \tag{11}$$

By making the constraint $\boldsymbol{r} + \boldsymbol{P\nu} + \mu\boldsymbol{1} + \gamma\boldsymbol{s} \succeq 0$ explicit, we obtain the Lagrange dual problem

$$\begin{array}{ll} \text{minimize} & \sum_j \exp(\nu_j - 1) + \mu + \gamma S \\ \text{subject to} & \boldsymbol{r} + \boldsymbol{P\nu} + \mu\boldsymbol{1} + \gamma\boldsymbol{s} \succeq 0, \qquad \gamma \geq 0. \end{array}$$

The optimization variables are $\boldsymbol{\nu}$, μ, and γ, and the constant parameters are $\boldsymbol{P}$.

Letting $\boldsymbol{\alpha} = \boldsymbol{\nu} + \mu\boldsymbol{1}$, and using the fact $\boldsymbol{P1} = \boldsymbol{1}$, we rewrite the dual problem as

$$\begin{array}{ll} \text{minimize} & \exp(-1 - \mu) \sum_j e^{\alpha_j} + \mu + \gamma S \\ \text{subject to} & \boldsymbol{r} + \boldsymbol{P\alpha} + \gamma\boldsymbol{s} \succeq 0, \qquad \gamma \geq 0 \end{array}$$

where the optimization variables are $\boldsymbol{\alpha}$, μ, and γ. Since the dual variable μ, which is the Lagrange multiplier corresponding to the primal constraint $\boldsymbol{p1} = 1$, is unconstrained in the dual problem, we can minimize the dual objective function over μ analytically, and obtain the minimizing $\mu = \log \sum_j e^{\alpha_j} - 1$. The resulting dual objective function is $\log \sum_j e^{\alpha_j} + \gamma S$. The Lagrange dual problem is simplified to the following geometric program in convex form:

$$\begin{array}{ll} \text{minimize} & \log \sum_j e^{\alpha_j} + \gamma S \\ \text{subject to} & \boldsymbol{P\alpha} + \gamma\boldsymbol{s} \succeq -\boldsymbol{r}, \qquad \gamma \geq 0 \end{array}$$

where the optimization variables are $\boldsymbol{\alpha}$ and γ, and the constant parameters are $\boldsymbol{P}$.

We can turn this geometric program into standard form, through an exponential change of the variables $z_j = e^{\alpha_j}$ and the dual objective function

$$\begin{array}{ll} \text{minimize} & w^S \sum_j z_j \\ \text{subject to} & w^{s_i} \prod_j z_j^{P_{ij}} \geq e^{-H(\boldsymbol{P}^{(i)})}, \quad i = 1, 2, \dots, N, \\ & z_j \geq 0, \ j = 1, 2, \dots, M, \ w \geq 1 \end{array}$$

where the optimization variables are $\boldsymbol{z}$ and w, the constant parameters are $\boldsymbol{P}$, and $\boldsymbol{P}^{(i)}$ is the ith row of $\boldsymbol{P}$.

The weak duality part of the proposition follows directly from a standard fact in Lagrange duality theory [4]: the Lagrange dual function is always an upper bound on the primal maximization problem.

It is well known that the objective function to be maximized in the primal problem (6) is concave in $(\boldsymbol{p}, \boldsymbol{q})$ and the constraint functions are affine. The strong duality part of the proposition holds because the primal problem (5) is a convex optimization satisfying Slater's condition [4]. □

Corollary 1: The Lagrange dual of the channel capacity problem without input cost (7) is the following geometric program (in convex form):

$$\begin{array}{ll} \text{minimize} & \log \sum_{j=1}^{M} e^{\alpha_j} \\ \text{subject to} & \boldsymbol{P\alpha} \succeq -\boldsymbol{r} \end{array} \tag{12}$$

where the optimization variables are $\boldsymbol{\alpha}$, and the constant parameters are $\boldsymbol{P}$.

An equivalent version of the Lagrange dual problem is the following geometric program (in standard form):

$$\begin{array}{ll} \text{minimize} & \sum_{j=1}^{M} z_j \\ \text{subject to} & \prod_{j=1}^{M} z_j^{P_{ij}} \geq e^{-H(\boldsymbol{P}^{(i)})}, \quad i = 1, 2, \dots, N, \\ & z_j \geq 0, \ j = 1, 2, \dots, M \end{array} \tag{13}$$

where the optimization variables are $\boldsymbol{z}$, the constant parameters are $\boldsymbol{P}$, and $\boldsymbol{P}^{(i)}$ is the ith row of $\boldsymbol{P}$.

Lagrange duality between problems (7) and (12) means the following.

- *Weak duality.* $\log\left(\sum_{j=1}^{M} e^{\alpha_j}\right) \geq C$, for all $\boldsymbol{\alpha}$ that satisfy $\boldsymbol{P\alpha} + \boldsymbol{r} \succeq 0$.
- *Strong duality.* $\log\left(\sum_{j=1}^{M} e^{\alpha_j^*}\right) = C$, where $\boldsymbol{\alpha}^*$ are the optimal dual variables.

We can also prove the weak duality result in Corollary 1 on channel capacity upper bound in a simple way without using the machinery of Lagrange duality.

Proof: We are given

$$\boldsymbol{\alpha} \in \boldsymbol{R}^{M \times 1} : \sum_j P_{ij}\alpha_j \geq \sum_j P_{ij} \log P_{ij}, \quad i = 1, 2, \dots, N$$

and $\boldsymbol{p} \in \boldsymbol{R}^{1 \times N}$, $\boldsymbol{q} \in \boldsymbol{R}^{1 \times M} : \boldsymbol{p} \succeq 0, \boldsymbol{p1} = 1, \boldsymbol{pP} = \boldsymbol{q}$. Through second derivative test, $f(t) = t \log t$, $t \geq 0$, is readily verified to be convex, i.e.,

$$\sum_j \theta_j f(t_j) \geq f\left(\sum_j \theta_j t_j\right)$$

with $\boldsymbol{\theta} \succeq 0$, $\boldsymbol{1}^T\boldsymbol{\theta} = 1$. Letting $t_j = \frac{q_j}{e^{\alpha_j}}$ and $\theta_j = \frac{e^{\alpha_j}}{\sum_k e^{\alpha_k}}$ and using $\boldsymbol{q1} = \boldsymbol{pP1} = \boldsymbol{p1} = 1$ gives

$$\log \sum_j e^{\alpha_j} \geq \sum_j \alpha_j q_j - \sum_j q_j \log q_j.$$

Since

$$\begin{aligned} \sum_j \alpha_j q_j &= \sum_j \alpha_j \sum_i p_i P_{ij} \\ &= \sum_i p_i \sum_j P_{ij}\alpha_j \\ &\geq \sum_i p_i \sum_j P_{ij} \log P_{ij} \end{aligned}$$

we have

$$\log \sum_j e^{\alpha_j} \geq \sum_{i,j} p_i P_{ij} \log P_{ij} - \sum_j q_j \log q_j = I(X;Y)$$

i.e., any feasible dual objective value is an upper bound on channel capacity. □

Note that the Lagrange dual (13) of the channel capacity problem is a simple geometric program with a linear objective function and only monomial inequality constraints. Also, dual problem (9) is a generalized version of dual problem (13), weighing the objective function by w^S and each constraint by w^{s_i}, where w is the Lagrange multiplier associated with the input cost constraint. If the costs for all alphabet symbols are 0, we can analytically minimize the objective function over w by simply letting $w=0$, indeed recovering the dual problem (13) for channels without the input cost constraint.

We can interpret the Lagrange dual problem (12) as follows. Let $\Lambda : \{1,\ldots,M\} \rightarrow \boldsymbol{R}$ be a real-valued function on the output space, with $\Lambda(j)=\alpha_j$. We can think of the variables $\boldsymbol{\alpha}$ as parameterizing all real-valued functions on the output space, so the dual problem is one over all real-valued functions on the output space. Since

$$(\boldsymbol{P\alpha})_i = \sum_{j=1}^{M} \alpha_j P_{ij} = \boldsymbol{E}(\Lambda | X=i)$$

the inequality constraint in the dual states that for each i, $\boldsymbol{E}(\Lambda|X=i)$ exceeds $-r_i = -H(Y|X=i)$. Since

$$\max_j \alpha_j \leq \log\left(\sum_{j=1}^{M} e^{\alpha_j}\right) \leq \max_j \alpha_j + \log M$$

the objective function in the dual is a smooth approximation of the maximum function. Thus, the Lagrange dual problem asks us to consider all real-valued functions Λ on the output space, for which $\boldsymbol{E}(\Lambda|X=i)$ exceeds $-H(Y|X=i)$ for each i. Among all such Λ, we are to find the one that minimizes a smoothed approximation of the maximum value of Λ.

Suppose we have solved the geometric program dual of channel capacity. By strong duality, we obtain $C(S)$. We can also recover the optimal primal variables, i.e., the capacity-achieving input distribution, from the optimal dual variables. For example, we can recover a least norm capacity-achieving input distribution for a channel without an input cost constraint as follows. First, the optimal output distribution $\boldsymbol{q}^*$ can be recovered from the optimal dual variable $\boldsymbol{\alpha}^*$

$$q_j^* = \exp\left(\alpha_j^* - C\right), \qquad j=1,2,\ldots,M \tag{14}$$

where $C = \log \sum_{j=1}^{M} e^{\alpha_j^*}$, and the optimal input distribution $\boldsymbol{p}^*$ is a vector that satisfies the linear equations

$$\begin{aligned} -\boldsymbol{pr} &= C + e^{-C}\left(\sum_{j=1}^{M}\alpha_j^* e^{\alpha_j^*} - C\sum_{j=1}^{M} e^{\alpha_j^*}\right)\\ \boldsymbol{pP} &= \boldsymbol{q}^*\\ \boldsymbol{p1} &= 1. \end{aligned}$$

In fact, both the primal and dual problems of $C(S)$ can be simultaneously and efficiently solved through a primal–dual interior point algorithm [4], [22], which scales smoothly for different channels and alphabet sizes. Utilizing the structure and sparsity of the exponent constant matrix $\boldsymbol{A}$ of the geometric program dual for channel capacity, standard convex optimization algorithms can be further accelerated in this case.

Proposition 1 can be easily extended when there are K input cost constraints indexed by $k : \sum_i p_i s_{ik} \leq S_k$, $k=1,2,\ldots,K$. The Lagrange dual problem becomes

$$\begin{aligned} \text{minimize} \quad & \log \sum_{j=1}^{M} \exp\left(\alpha_j + \sum_{k=1}^{K}\gamma_k S_k\right)\\ \text{subject to} \quad & \sum_{j=1}^{M} P_{ij}(\alpha_j - \log P_{ij}) + \sum_{k=1}^{M}\gamma_k s_{ik} \geq 0,\\ & \qquad i=1,2,\ldots,N,\\ & \boldsymbol{\gamma} \succeq 0 \end{aligned} \tag{15}$$

where the optimization variables are $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$.

C. Complementary Slackness

From the complementary slackness property [4], if $p_i^* > 0$, $\forall\, i$, i.e., every mass point in the capacity-achieving input distribution is positive, then solving a system of linear equations $\boldsymbol{P\alpha}+\gamma\boldsymbol{s}+\boldsymbol{r}=0$ obtains $\boldsymbol{\alpha}^*$ and γ^*, hence, the channel capacity with input cost

$$C(S) = \log \sum_{j=1}^{M} \exp(\alpha_j^* + \gamma^* S).$$

In the case of no input cost constraint, this recovers the observation made by Gallager in [13].

A dual argument based on complementary slackness shows that $p_i^*=0$ if $r_i + (\boldsymbol{P\alpha}^*)_i + \gamma^* s_i > 0$ in the Lagrange dual of channel capacity. Therefore, from the optimal dual variable $\boldsymbol{\alpha}^*$, γ^*, we immediately obtain the support of the capacity-achieving input distribution as

$$\{i | r_i + (\boldsymbol{P\alpha}^*)_i + \gamma^* s_i = 0\}.$$

From the primal and dual problems of channel capacity, we also obtain various optimality conditions. For example, if there are $\boldsymbol{\lambda} \in \boldsymbol{R}^{N\times 1}$ and $\boldsymbol{\alpha} \in \boldsymbol{R}^{M\times 1}$ satisfying the following Karush–Kuhn–Tucker (KKT) conditions [4] for a given $\boldsymbol{P}$

$$\begin{aligned} \boldsymbol{\lambda} &\succeq 0\\ \boldsymbol{r} + \boldsymbol{P\alpha} &\succeq 0\\ \frac{e^{\alpha_j}}{\sum_{j'=1}^{M} e^{\alpha_{j'}}} + \sum_{i=1}^{N}\lambda_i P_{ij} &= 0, \qquad j=1,2,\ldots,M\\ \lambda_i\left(r_i + \sum_{j=1}^{M} P_{ij}\alpha_j\right) &= 0, \qquad i=1,2,\ldots,N \end{aligned}$$

then the resulting $\log \sum_{j=1}^{M} e^{\alpha_j}$ is the channel capacity C.

There is a minmax KL divergence (minmaxKL) characterization of discrete memoryless channel capacity with input cost in [10], [14]

$$C(S) = \min_{\boldsymbol{q}} \min_{\gamma \geq 0} \max_i \left[D\left(\boldsymbol{P}^{(i)} \| \boldsymbol{q}\right) + \gamma(S - s_i)\right] \tag{18}$$

where the minimization over $\boldsymbol{q}$ is over all possible output distributions.

Since the Lagrange dual (8) and minmaxKL (18) characterizations both give $C(S)$, they must be equivalent. This equivalence can also be established directly. Let the dual variables $z_j = e^{\alpha_j} = \beta q_j$ where $\beta > 0$ and $\boldsymbol{q}$ is any distribution. Then the dual constraints become

$$\sum_{j=1}^{M} P_{ij} \log \frac{P_{ij}}{q_j} - \gamma s_i \leq \log \beta, \qquad i=1,2,\ldots,N.$$

Since the case of $C(S)=0$ is trivial, assume $C(S)>0$. By complementary slackness, if at optimality all the dual constraints are satisfied with strict inequalities, then the optimal Lagrange multipliers (readily seen to be the optimal input distribution) of this geometric program must all be zero, contradicting our assumption that $C(S) > 0$. Therefore,

$$\max_i \left[\sum_{j=1}^M P_{ij} \log \frac{P_{ij}}{q_j^*} - \gamma^* s_i\right] = \log \beta^*.$$

By the strong duality part of Proposition 1

$$\begin{aligned} C(S) &= \log \sum_{j=1}^M \beta^* q_j^* + \gamma^* S \\ &= \log \beta^* + \gamma^* S = \max_i \left(\sum_{j=1}^M P_{ij} \log \frac{P_{ij}}{q_j^*} - \gamma^* s_i\right) + \gamma^* S. \end{aligned}$$

Since at optimality, $\boldsymbol{q}^*$ must correspond to the output distribution induced by an optimal input distribution, restricting the minimization of dual variables $\boldsymbol{z}$ to a scaled version of an output distribution incurs no loss of generality. Thus, the minmaxKL characterization (18) is recovered.

D. Bounding From the Dual

Because the inequality constraints in the dual problem (8) are affine, it is easy to obtain a dual feasible $\boldsymbol{\alpha}$ by finding any solution to a system of linear inequalities, and the resulting value of the dual objective function provides an easily derivable upper bound on channel capacity. Many channel matrices also exhibit sparsity patterns: special patterns of a small number of nonzero entries. Based on the sparsity pattern of the given channel matrix, tight analytic bounds may also be obtained from an appropriate selection of dual variables.

As a simple example, it is easy to verify that $\alpha_j = \max_i -H(\boldsymbol{P}^{(i)}), \forall\, j$, satisfy the dual constraints and generate an upper bound on channel capacity

$$C \le \log M - \min_i H\left(\boldsymbol{P}^{(i)}\right).$$

Similarly, it is easy to verify that $\alpha_j = \log \max_i P_{ij}$ satisfy the dual constraints and give the following.

Corollary 2: Channel capacity is upper-bounded in terms of a maximum-likelihood receiver selecting $\mathrm{argmax}_i\, P_{ij}$ for each output alphabet symbol j

$$C \le \log \sum_{j=1}^M \max_i P_{ij} \tag{19}$$

which is tight if and only if the optimal output distribution q^* is

$$q_j^* = \frac{\max_i P_{ij}}{\sum_{k=1}^M \max_i P_{ik}}, \qquad j = 1, 2, \dots, M.$$

When there is an input cost constraint $\boldsymbol{ps} \le S$, the above upper bound becomes

$$C(S) \le \log \sum_{j=1}^M \max_i (e^{-s_i} P_{ij}) + S \tag{20}$$

where each maximum-likelihood decision is modified by the cost vector $\boldsymbol{s}$.

Of course, it is also easy to find a primal feasible point satisfying the linear inequality constraints of the primal problem (7), which gives a lower bound on channel capacity. $C(S)$ must be within the range determined by this pair of upper and lower bounds.

Sometimes dual feasible variables alone give an estimate of the channel capacity with a bounded error. For example, for channel capacity without input cost, find any $\boldsymbol{\alpha}$ such that $\boldsymbol{P\alpha} + \boldsymbol{r} \succeq 0$, from which we generate

$$q_j = \frac{e^{\alpha_j}}{\sum_{k=1}^M e^{\alpha_k}}, \qquad j = 1, 2, \dots, M.$$

If there is a $\boldsymbol{p} \succeq 0$ such that $\boldsymbol{pP} = \boldsymbol{q}$, then the estimated channel capacity $\tilde{C}(\boldsymbol{\alpha}) = \log \sum_{j=1}^M e^{\alpha_j}$ can only be Γ away from the true capacity C, where

$$\Gamma = \boldsymbol{pr} + \frac{\sum_{j=1}^M \alpha_j e^{\alpha_j}}{\sum_{j=1}^M e^{\alpha_j}}.$$

Note that the minmaxKL characterization of C (18) obviously leads to the following known class of upper bounds on channel capacity: for any output distribution $\boldsymbol{q}$

$$C \le \max_i \sum_{j=1}^M P_{ij} \log \frac{P_{ij}}{q_j} \tag{21}$$

which is shown in [10], [13], and has recently been used for simulating finite-state channels in [29] and bounding capacity of noncoherent multiple-antenna fading channels in [18].

An argument similar to that in Section II-C shows that the geometric program Lagrange dual (8) generates a broader class of upper bounds on $C(S)$, including the class of bounds from (21) as a special case. Specifically, the following bounds, readily extended from (21) and parameterized by output distributions $\boldsymbol{q}$ and $\gamma \ge 0$:

$$C(S) \le \max_i \left[\sum_{j=1}^M P_{ij} \log \frac{P_{ij}}{q_j} - \gamma s_i\right] + \gamma S$$

can be obtained from the geometric program dual by restricting the dual variables $(\boldsymbol{z}, \gamma)$ to be such that

$$\max_i \left[\sum_{j=1}^M P_{ij} \log \frac{P_{ij}}{z_j} - \gamma s_i\right] = 0$$

and by restricting $\boldsymbol{z}$ to be a scaled output distribution.

For a memoryless channel with continuous alphabets, where the channel is a family of conditional distributions $P(y|x)$ and the input cost constraint is $\int p(x)s(x)dx \le S$, a derivation similar to Proof 1 shows that the Lagrange dual of the channel capacity problem is the following continuous analog of geometric program with variables $z(y)$ and γ:

$$\begin{aligned} &\text{minimize} \quad \log \int z(y)dy + \gamma S \\ &\text{subject to} \quad \int P(y|x) \log \frac{z(y)}{P(y|x)} dy + \gamma s(x) \ge 0, \quad \forall\, x, \\ &\qquad\qquad z(y) \ge 0,\ \forall\, y, \quad \gamma \ge 0. \end{aligned} \tag{22}$$

Weak duality and the Lagrange dual problem (22) readily lead to the following class of bounds, which is also shown in [18]: for any distribution $q(y)$ and $\gamma \geq 0$

$$C(S) \leq \max_x \left[\int P(y|x) \log \frac{P(y|x)}{q(y)} dy - \gamma s(x) \right] + \gamma S.$$

III. Lagrange Dual of Rate Distortion

A. The Rate Distortion Problem

Consider the following standard problem of data compression. Assume a source that produces a sequence of i.i.d. random variables $X_1, X_2, \ldots, X_n \sim \boldsymbol{p}$, where the state space of X_i is a discrete source alphabet $\mathcal{X}$ with N alphabet symbols and $\boldsymbol{p} \in \boldsymbol{R}^{1\times N}$ is the source distribution. The encoder describes the source sequence X^n by an index $f_n(x^n) \in \{1, 2, \ldots, 2^{nR}\}$, where x^n is a realization of X^n. The decoder reconstructs X^n by an estimate $\hat{X}^n = g_n(f_n(X^n))$ in a finite reconstruction alphabet $\hat{\mathcal{X}}$. Given a bounded distortion measure $d : \mathcal{X} \times \hat{\mathcal{X}} \to \boldsymbol{R}_+$, the distortion $d(x^n, \hat{x}^n)$ between sequences x^n and $\hat{x}^n$ is the average distortion of these two n-letter blocks.

It is another key result in information theory that the rate distortion function $R(D)$ of a discrete source can be evaluated as the minimum mutual information $I(X; \hat{X})$ between the source and the reconstruction under the distortion constraint

$$R(D) = \min_{\boldsymbol{P}: \boldsymbol{E}[d(X,\hat{X})] \leq D} I(X; \hat{X}) \tag{23}$$

where

$$P_{ij} = \mathbf{Prob}\{\hat{X} = j | X = i\}, \quad i = 1, 2, \ldots, N, \; j = 1, 2, \ldots, M.$$

In Section III-B, we will focus on the following rate distortion problem:

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{N} \sum_{j=1}^{M} p_i P_{ij} \log \frac{P_{ij}}{\sum_k P_{kj} p_k} \\ \text{subject to} \quad & \sum_{i=1}^{N} \sum_{j=1}^{M} p_i P_{ij} d_{ij} \leq D, \\ & \sum_{j=1}^{M} P_{ij} = 1, \; i = 1, 2, \ldots, N \\ & P_{ij} \geq 0, \; i = 1, 2, \ldots, N, \; j = 1, 2, \ldots, M \end{aligned} \tag{24}$$

where the variables are the reconstruction probabilities P_{ij}. The constant parameters are the source distribution $\boldsymbol{p}$, the distortion measures $d_{ij} = d(X = i, \hat{X} = j)$, and the distortion constraint D.

B. Geometric Programming Dual

Proposition 2: The Lagrange dual of the rate distortion problem (24) is the following geometric program (in convex form):

$$\begin{aligned} \text{maximize} \quad & \boldsymbol{p\alpha} - \gamma D \\ \text{subject to} \quad & \log \sum_{i=1}^{N} \exp(\log p_i + \alpha_i - \gamma d_{ij}) \leq 0, \\ & \qquad j = 1, 2, \ldots, M, \\ & \gamma \geq 0 \end{aligned} \tag{25}$$

where the optimization variables are $\boldsymbol{\alpha} \in \boldsymbol{R}^{N\times 1}$ and γ, and the constant parameters are $\boldsymbol{p} \in \boldsymbol{R}^{1\times N}$, d_{ij}, and D.

An equivalent version of the Lagrange dual problem is the following geometric program (in standard form):

$$\begin{aligned} \text{maximize} \quad & w^{-D} \prod_{i=1}^{N} z_i^{p_i} \\ \text{subject to} \quad & \sum_{i=1}^{N} p_i z_i w^{-d_{ij}} \leq 1, \quad j = 1, 2, \ldots, M, \\ & w \geq 1, \; z_i \geq 0, \; i = 1, 2, \ldots, N \end{aligned} \tag{26}$$

where the optimization variables are $\boldsymbol{z}$ and w, and the constant parameters are $\boldsymbol{p}$, d_{ij}, and D.

Lagrange duality between problems (23) and (25) means the following.

- *Weak duality.* Any feasible $(\boldsymbol{\alpha}, \gamma)$ of the Lagrange dual problem (25) produce a lower bound on the rate distortion function
$$\boldsymbol{p\alpha} - \gamma D \leq R(D).$$
- *Strong duality.* The optimal value of the Lagrange dual problem (25) is $R(D)$.

Note that in [1], Berger proved an equivalent formulation as (25). The proof here is simpler by directly using the Lagrange duality approach.

Proof: In order to find the Lagrange dual of problem (24), we first form the Lagrangian

$$\begin{aligned} L(\boldsymbol{P}, \boldsymbol{\mu}, \gamma, \boldsymbol{\Lambda}) = & \sum_{i,j} p_i P_{ij} \log \frac{P_{ij}}{\sum_k P_{kj} p_k} + \sum_i \mu_i \sum_j P_{ij} \\ & - \sum_i \mu_i + \gamma \sum_{i,j} p_i P_{ij} d_{ij} - \gamma D - \sum_{i,j} \lambda_{ij} P_{ij} \end{aligned} \tag{27}$$

with Lagrange multiplier vector $\boldsymbol{\mu} \in \boldsymbol{R}^{N\times 1}$, Lagrange multiplier $\gamma \in \boldsymbol{R}$, and Lagrange multiplier matrix $\boldsymbol{\Lambda} \in \boldsymbol{R}^{M\times N}$, with (i,j) entry of $\boldsymbol{\Lambda}$ denoted as λ_{ij}. Since γ and $\boldsymbol{\Lambda}$ correspond to the inequality constraints, we have $\gamma \geq 0$ and $\lambda_{ij} \geq 0$, $i = 1, 2, \ldots, N$, $j = 1, 2 \ldots, M$.

We then find the Lagrange dual function

$$g(\boldsymbol{\mu}, \gamma, \boldsymbol{\Lambda}) = \inf_{\boldsymbol{P}} L(\boldsymbol{P}, \boldsymbol{\mu}, \gamma, \boldsymbol{\Lambda})$$

by finding the $\boldsymbol{P}$ that minimizes L, which is a convex function of P_{ij}. We let the derivatives of L with respect to P_{ij} be equal to 0

$$p_i \left[\log \left(\frac{P_{ij}}{z_i \sum_k P_{kj} p_k} \right) + \gamma d_{ij} - \frac{\lambda_{ij}}{p_i} \right] = 0$$

where $z_i = \exp(-\frac{\mu_i}{p_i})$. This gives the following condition on the minimizer $\boldsymbol{P}$ of L:

$$P_{ij} = z_i q_j \exp\left(\frac{\lambda_{ij}}{p_i} - \gamma d_{ij} \right) \tag{28}$$

where $q_j = \sum_k P_{kj} p_k$. Now multiply both sides of (28) by p_i, sum over i, and cancel q_j on both sides; we obtain the following condition:

$$\sum_i z_i p_i \exp\left(\frac{\lambda_{ij}}{p_i} - \gamma d_{ij} \right) = 1, \qquad j = 1, 2, \ldots, M$$

which, by the definition of z_i and the condition $\lambda_{ij} \geq 0$, is equivalent to

$$\sum_i p_i \exp\left(-\frac{\mu_i}{p_i} - \gamma d_{ij}\right) \leq 1, \qquad j = 1, 2, \ldots, M. \quad (29)$$

Substituting the minimizer (28) and the condition (29) into L (27), we obtain the Lagrange dual function

$$g(\boldsymbol{\mu}, \gamma) = \begin{cases} -\sum_i \mu_i - \gamma D, & \sum_i p_i \exp\left(-\frac{\mu_i}{p_i} - \gamma d_{ij}\right) \leq 1 \\ -\infty, & \text{otherwise.} \end{cases} \quad (30)$$

By making the constraints explicit, we obtain the Lagrange dual problem

$$\begin{array}{ll} \text{maximize} & -\sum_i \mu_i - \gamma D \\ \text{subject to} & \sum_i p_i \exp\left(-\frac{\mu_i}{p_i} - \gamma d_{ij}\right) \leq 1, \ j = 1, 2, \ldots, M, \\ & \gamma \geq 0 \end{array}$$

where the variables are $\boldsymbol{\mu}$ and γ, and the constant parameters are $\boldsymbol{p}$, d_{ij}, and D.

Now we change the dual variables from $\boldsymbol{\mu}$ to $\boldsymbol{\alpha}$: $\alpha_i = -\frac{\mu_i}{p_i}$, and rewrite the dual problem as

$$\begin{array}{ll} \text{maximize} & \sum_i p_i \alpha_i - \gamma D \\ \text{subject to} & \log \sum_i \exp(\log p_i + \alpha_i - \gamma d_{ij}) \leq 0, \\ & \qquad j = 1, 2, \ldots, M, \\ & \gamma \geq 0 \end{array}$$

where the variables are $\boldsymbol{\alpha}$ and γ, and the constant parameters are $\boldsymbol{p}$, d_{ij}, and D.

In order to bring the dual problem (25) to the standard form of geometric programming, we use an exponential change of the variables $w = e^{\gamma}$, $z_i = e^{\alpha_i}$ to rewrite the dual problem as

$$\begin{array}{ll} \text{maximize} & w^{-D} \prod_i z_i^{p_i} \\ \text{subject to} & \sum_i p_i z_i w^{-d_{ij}} \leq 1, \quad j = 1, 2, \ldots, M, \\ & \qquad w \geq 1, \ z_i \geq 0, \ i = 1, 2, \ldots, N \end{array}$$

where the variables are $\boldsymbol{z}$ and w, and the constant parameters are $\boldsymbol{p}$, d_{ij}, and D.

The weak duality part of the proposition follows directly from a standard fact in Lagrange duality theory [4]: the Lagrange dual function is always a lower bound on the primal minimization problem.

It is well known that the objective function in the primal problem (24) is convex in P_{ij}, and the constraints are affine. The strong duality part of the proposition holds because the primal problem (5) is a convex optimization satisfying Slater's condition [4]. □

The Lagrange dual (26) of the rate distortion problem (24) is a simple geometric program: maximizing a monomial over posynomial constraints, in the form of maximizing a geometric mean $\prod_{i=1}^{N} z_i^{p_i}$ weighted by w^{-D}, under constraints on arithmetic means $\sum_{i=1}^{N} p_i z_i$ weighted by $w^{-d_{ij}}$. A smaller shadow price w would reduce the objective value but also loosen each constraint, allowing larger dual variables z_i and possibly higher objective value.

Proposition 2 can be easily generalized to rate distortion with K distortion constraints indexed by k : $E[d^k(X, \hat{X})] \leq D_k$, $k = 1, 2, \ldots, K$, where the Lagrange dual problem becomes

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^{N} p_i \alpha_i - \sum_{k=1}^{K} \gamma_k D_k \\ \text{subject to} & \log \sum_{i=1}^{N} \exp\left(\log p_i + \alpha_i - \sum_{k=1}^{K} \gamma_k d_{ij}^k\right) \leq 0, \\ & \qquad j = 1, 2, \ldots, M, \\ & \boldsymbol{\gamma} \succeq 0 \end{array} \quad (31)$$

where the optimization variables are $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$.

Similar to the case for channel capacity, we can now efficiently lower-bound the rate distortion function from the geometric programming dual. In particular, due to the structure of the constraints in (26), for any given w, finding a dual feasible $\boldsymbol{z}$ reduces to the easy task of solving a system of linear inequalities. For example, with Hamming distortion measure, it is easy to verify that

$$\alpha_i = \log\left(\frac{1-D}{p_i}\right) \quad \text{and} \quad \gamma = \log\left(\frac{(1-D)(N-1)}{D}\right)$$

satisfy the Lagrange dual constraints in (25), and give the following lower bound:

$$R(D) \geq H(X) - H_0(D) - D \log(N-1) \quad (32)$$

where $H_0(x) = -x \log x - (1-x) \log(1-x)$ is the binary entropy function.

Now consider guessing a random variable X based on another random variable $\hat{X}$. If we replace D by the probability of estimation error P_e and use the fact that

$$R(D) = \min I(X; \hat{X}) \leq H(X) - H(X|\hat{X})$$

then the lower bound (32) recovers Fano's inequality

$$H(X|\hat{X}) \leq H_0(P_e) + P_e \log(N-1) \quad (33)$$

that readily proves the converse theorem for channel capacity [9].

More results on bounding and computing the rate distortion function can be found in [1] and [2].

C. Rate Distortion With State Information

As an extension of the basic data compression problem described in Section III-A, rate distortion with state information has been studied (e.g., in [31]) and applied to data compression systems such as distributed source coding in a sensor web (e.g., in [24]).

The following general problem of rate distortion with two-sided state information was considered in [8]. Assume correlated random variables (X, S_1, S_2) i.i.d. $\sim p(x, s_1, s_2)$ with finite alphabets $\mathcal{X}, \mathcal{S}_1, \mathcal{S}_2$, respectively. Let $\{(X_i, S_{1,i}, S_{2,i})\}$ be a sequence of independent drawings of (X, S_1, S_2), and let $S_1^n = \{S_{1,i}, \ i = 1, 2, \ldots, n\}$ be the state information available to the encoder and S_2^n to the decoder. We wish to describe $\{X_k\}$ at rate R bits per symbol and reconstruct $\{\hat{X}_k\}$ with distortion smaller than or equal to D. The sequence $\{X_k\}$ is encoded in blocks of length n into a binary stream of rate R, which will in turn be decoded as a sequence $\{\hat{X}_k\}$ in the reproduction alphabet. For a given bounded distortion measure $d(x, \hat{x})$, the average distortion is $\frac{1}{n} \sum_{k=1}^{n} \boldsymbol{E}[d(X_k, \hat{X}_k)]$. We say that rate R is achievable at distortion level D if there exists a sequence of $(2^{nR}, n)$ codes $i : \mathcal{X}^n \times \mathcal{S}_1^n \rightarrow \{1, 2, \ldots, 2^{nR}\}$, $\hat{X}^n : \{1, 2, \ldots, 2^{nR}\} \times \mathcal{S}_2^n \rightarrow \hat{\mathcal{X}}^n$ such that

$$\boldsymbol{E}[d(X^n, \hat{X}^n(i(X^n, S_1^n), S_2^n))] \leq D.$$

The rate distortion function $R_{S_1,S_2}(D)$ is the infimum of the achievable rates with distortion D.

The rate distortion function with two-sided state information is [8]

$$R_{S_1,S_2}(D) = \min_{p(u|x,s_1)p(\hat{x}|u,s_2)} [I(U; S_1, X) - I(U; S_2)] \quad (34)$$

where U is an auxiliary random variable and the minimization is under the distortion constraint

$$\sum_{x,u,s_1,s_2,\hat{x}} p(x, s_1, s_2)p(u|x, s_1)p(\hat{x}|u, s_2)d(x, \hat{x}) \leq D.$$

The problem of rate distortion with state information at the decoder, i.e., when $S_1 = \phi$ and $S_2 = S$, was studied by Wyner and Ziv [31], who proved the rate distortion function to be

$$R_{\phi,S}(D) = \min_{p(u|x)p(\hat{x}|u,s)} [I(U; X) - I(U; S)] \quad (35)$$

where the minimum is under the distortion constraint

$$\sum_{x,\hat{x},u,s} p(x, s)p(u|x)p(\hat{x}|u, s)d(x, \hat{x}) \leq D$$

and $|\mathcal{U}| \leq |\mathcal{X}| + 1$.

Although the above source coding theorems have been proved, bounding or computing the rate distortion function with state information have not appeared in the literature. In Section III-D, we will show that the Lagrange dual of the rate distortion problem with state information is again a geometric program, which allows us to lower-bound $R_{S_1,S_2}(D)$ from the dual.

D. Geometric Programming Dual

We can view (34) as the primal optimization problem. However, deriving the Lagrange dual in an explicit form directly from (34) turns out to be difficult. Therefore, we first simplify the primal problem by using the Markovity structure of the problem and then Shannon's signaling strategy [28].

Lemma 1: The rate distortion function with two-sided state information can be written as

$$R_{S_1,S_2}(D) = \min_{p(u|x,s_1), \hat{x}=f(u,s_2)} I(U; S_1, X|S_2) \quad (36)$$

where $f : \mathcal{U} \times \mathcal{S}_2 \rightarrow \hat{\mathcal{X}}$ is a deterministic function, and the minimization is under the constraint

$$\sum_{x,u,s_1,s_2} d(x, f(u, s_2))p(x, s_1, s_2)p(u|x, s_1) \leq D.$$

Proof: It was shown in [8], [31] that restricting the minimization over $p(\hat{x}|u, s_2)$ to deterministic functions $\hat{x} = f(u, s_2)$ incurs no loss of generality. Using the conditional independence of U and S_2 given (X, S_1), we have

$$\begin{aligned} I(U; X, S_1) - I(U; S_2) &= H(U|S_2) - H(U|X, S_1) \\ &= H(U|S_2) - H(U|X, S_1, S_2) \\ &= I(U; X, S_1|S_2). \end{aligned} \qquad \square$$

Now, similar to Shannon's signaling strategy for channels with state information at the encoder [28], consider all strategies $v(\bullet)$ that map S_2 into $\hat{X}$. Each strategy can be represented as a vector $(v(1), v(2), \ldots, v(|\mathcal{S}_2|))$. Therefore, there are a total of $|\hat{\mathcal{X}}|^{|\mathcal{S}_2|}$ strategies. Following the arguments in [8], [28], [30], [31], both the achievability and converse proofs of the following lemma can be readily verified.

Lemma 2: The rate distortion function with two-sided state information can be written as

$$R_{S_1,S_2}(D) = \min_{p(v|x,s_1)} I(V; S_1, X|S_2) \quad (37)$$

where the minimization is over all conditional distributions of strategies v and under the distortion constraint

$$\sum_{x,v,s_1,s_2} d(x, v(s_2))p(x, s_1, s_2)p(v|x, s_1) \leq D.$$

Note that for a given S_2, once a strategy v is determined by $p(v|x, s_1)$, the reconstruction $\hat{x}$ is fixed. Therefore, we only need to optimize over $p(v|x, s_1)$. By exponentially expanding the size of the domain of the reconstruction function (from $\hat{x} = f(u, s_2)$ to $\hat{x} = v(s_2)$), we have reduced the primal problem from (34) to the convex optimization of a conditional mutual information over a set of conditional probabilities (37).

To facilitate a more compact expression of the primal and dual problems, we will use the following simplified notation. We use i, j, l, k to index X, V, S_1, S_2, respectively, and suppress ranges of the summations. We are given distributions $q_k = \mathbf{Prob}\{S_2 = k\}$ and $Q_{kil} = \mathbf{Prob}\{X = i, S_1 = l | S_2 = k\}$, and minimize over the conditional distribution of strategies $P_{ilj} = \mathbf{Prob}\{V = j | X = i, S_1 = l\}$. Using Lemma 2, the rate distortion function with two-sided state information $R_{S_1,S_2}(D)$ becomes the optimal value of the following convex optimization with variables P_{ilj}:

$$\begin{aligned} \text{minimize} \quad & \textstyle\sum_{i,j,l,k} q_k Q_{kil} P_{ilj} \log \frac{P_{ilj}}{\sum_{i',l'} Q_{ki'l'} P_{i'l'j}} \\ \text{subject to} \quad & \textstyle\sum_{i,j,l,k} q_k Q_{kil} P_{ilj} d_{ij} \leq D, \\ & \textstyle\sum_j P_{ilj} = 1, \quad \forall\, i, l, \\ & P_{ilj} \geq 0, \quad \forall\, i, l, j. \end{aligned} \quad (38)$$

Following the proof for Proposition 2, we derive and simplify the Lagrange dual problem for (38) in the following.

Proposition 3: The Lagrange dual of the problem of rate distortion with two-sided state information (38) is the following geometric program (in convex form):

$$\begin{aligned} \text{maximize} \quad & \textstyle\sum_{i,l} \mu_{il} - \gamma D \\ \text{subject to} \quad & \log \textstyle\sum_{i,l} Q_{kil} \exp\left(\frac{\mu_{il}}{\sum_{k'} q_{k'} Q_{k'il}} - \gamma d_{ij} \right) \leq 0, \\ & \qquad\qquad\qquad\qquad\qquad \forall\, j, k, \\ & \gamma \geq 0 \end{aligned} \quad (39)$$

where the optimization variables are μ_{il} and γ, and the constant parameters are the given distributions q_k, Q_{kil}, distortion measures d_{ij}, and constraint D.

An equivalent version of the Lagrange dual problem is the following geometric program (in standard form):

$$\begin{array}{ll} \text{maximize} & w^{-D}\prod_{i,l} z_{il}^{(\sum_{k'} q_{k'}Q_{k'il})} \\ \text{subject to} & \sum_{i,l} Q_{kil} z_{il} w^{-d_{ij}} \leq 1, \quad \forall\, j, k, \\ & w \geq 1,\ z_{il} \geq 0,\ \forall\, i, l \end{array} \tag{40}$$

where the optimization variables are z_{il} and w.

Lagrange duality between problems (38) and (39) means the following.

- *Weak duality.* Any feasible (μ_{il}, γ) of the Lagrange dual problem (39) produce a lower bound on the rate distortion function
$$R_{S_1,S_2}(D) \geq \sum_{i,l} \mu_{il} - \gamma D.$$
- *Strong duality.* The optimal value of the Lagrange dual problem (39) is $R_{S_1,S_2}(D)$.

The geometric program Lagrange dual (40) for rate distortion with two-sided state information indeed extends the following Lagrange dual problem for rate distortion without state (26):

$$\begin{array}{ll} \text{maximize} & w^{-D}\prod_{i=1}^{N} z_i^{p_i} \\ \text{subject to} & \sum_{i=1}^{N} p_i z_i w^{-d_{ij}} \leq 1, \quad j = 1, 2, \ldots, M, \\ & w \geq 1,\ z_i \geq 0,\ i = 1, 2, \ldots, N \end{array}$$

where the optimization variables are z_i and w.

The Lagrange dual gives a class of lower bounds for rate distortion function with two-sided state information: any dual feasible (μ_{il}, γ) in (39) lower bounds $R_{S_1,S_2}(D)$. Due to strong duality, this class of bounds can be made arbitrarily tight by choosing appropriate dual variables. At the same time, it is trivial to generate an upper bound: any P_{ilj} in the primal problem (38) gives one. This pair of bounds gives an estimate of $R_{S_1,S_2}(D)$ and at optimality, they coincide.

Assuming Hamming distortion measure, we now give an example lower bound that is valid for any joint distribution $p(x, s_1, s_2)$. It is readily verified that

$$\mu_{il} = \left(\sum_{k'} q_{k'} Q_{k'il}\right) \log \frac{1-D}{\max_k Q_{kil}}$$

and

$$\gamma = \log\left(\frac{(1-D)(N-1)}{D}\right)$$

where N is the size of source alphabet, satisfy the dual constraints in (39), and lead to the following.

Corollary 3: The rate distortion function with two-sided state information and Hamming distortion measure is lower-bounded by

$$\begin{aligned} R_{S_1,S_2}(D) \geq & \sum_{i,l} \mathbf{Prob}\{X = i, S_1 = l\} \\ & \times \left(-\max_k \log \mathbf{Prob}\{X = i, S_1 = l | S_2 = k\}\right) \\ & - H_0(D) - D\log(N-1) \end{aligned} \tag{41}$$

where H_0 is the binary entropy function.

Consider the following problem in the special case of rate distortion with state information S_2 at the decoder only. We guess random variable X based on $(\hat{X}, S_2)$ and would like to upper-bound $H(X|\hat{X}, S_2)$ similar to Fano's inequality (33) that bounds $H(X|\hat{X})$ in terms of the probability of guessing error P_e. Corollary 3 gives the following extension of Fano's inequality:

$$\begin{aligned} H(X|\hat{X}, S_2) \leq & H_0(P_e) + P_e \log(N-1) \\ & + H(X|S_2) + \sum_i \mathbf{Prob}\{X = i\} \\ & \times \max_k \log \mathbf{Prob}\{X = i | S_2 = k\}. \end{aligned} \tag{42}$$

Since the Lagrange dual problem (39) is a simple geometric program, it can be very efficiently solved through a primal–dual interior-point algorithm [4], [22] for large alphabet sizes. By strong duality, the resulting optimal dual value is $R_{S_1,S_2}(D)$. Using a similar technique for rate distortion without state information in [1], we can further recover the optimal $\mathbf{Prob}\{V = j|X = i, , S_1 = l\}$.

There have been discussions of the duality between channel capacity and rate distortion with state information, e.g., in [3], [8], [23]. A different Markovity structure for channel capacity with state information makes it difficult to convert the problem into the convex optimization form of minimizing a conditional mutual information over conditional distribution. This was the key step (Lemmas 1 and 2) to decouple the primal variables and derive an explicit Lagrange dual problem for rate distortion with state information.

IV. Shannon Duality Through Lagrange Duality

A. "Duality By Mapping" Between the Lagrange Duals

Lagrange duality is usually stated as follows. Given an optimization problem called the primal problem, the objective and constraint functions in the dual problem can be obtained from those in the primal problem by some simple mappings of signs, variables, constant parameters, and mathematical operations. This "duality by mapping" that turns one optimization problem into its Lagrange dual is also used in other duality relationships [19], such as that between controllability and observability. However, as can be easily verified, "duality by mapping" does not hold between the primal problems of channel capacity (6) and rate distortion (24). It turns out that their Lagrange dual problems exhibit a precise "duality by mapping." Due to strong duality, this induces a "duality by mapping" between the primal problems through the geometric programming duals, as shown in Fig. 1. Note that Lagrange duality is different from Shannon duality. Indeed, while channel capacity and rate distortion are "somewhat dual," as commented by Shannon, their Lagrange dual problems are both geometric programs.

We first summarize two versions of the Lagrange dual problems for $C(S)$ and $R(D)$ in Table I, where the geometric program dual problems in standard form better illustrate the "duality by mapping" relationships. The objective functions, constraint functions, variables, and constant parameters in the Lagrange dual problems of $C(S)$ and $R(D)$ can be obtained from one another through the following simple mappings.

TABLE I
LAGRANGE DUAL PROBLEMS OF CHANNEL CAPACITY WITH INPUT COST AND RATE DISTORTION

	Lagrange Dual of Channel Capacity	*Lagrange Dual of Rate Distortion*
Convex Form	minimize $\log \sum_{j=1}^{M} e^{\alpha_j+\gamma S}$ subject to $\sum_{j=1}^{M} P_{ij}(\alpha_j - \log P_{ij}) + \gamma s_i \geq 0,$ $\gamma \geq 0,$ variables: α_j, γ constants: channel P_{ij}, cost s_i, S	maximize $\mathbf{p}\boldsymbol{\alpha} - \gamma D$ subject to $\log \sum_{i=1}^{N} e^{\log p_i + \alpha_i - \gamma d_{ij}} \leq 0$ $\gamma \geq 0$ variables: α_i, γ constants: source p_i, distortion d_{ij}, D
Standard Form	minimize $w^S \sum_{j=1}^{M} z_j$ subject to $\prod_{j=1}^{M} e^{H(\mathbf{P}^{(i)})} w^{s_i} z_j^{P_{ij}} \geq 1,$ $w \geq 1, \quad z_j \geq 0$ variables: z_j, w constants: channel P_{ij}, cost s_i, S	maximize $w^{-D} \prod_{i=1}^{N} z_i^{p_i}$ subject to $\sum_i p_i z_i w^{-d_{ij}} \leq 1,$ $w \geq 1, \quad z_i \geq 0$ variables: z_i, w constants: source p_i, distortion d_{ij}, D

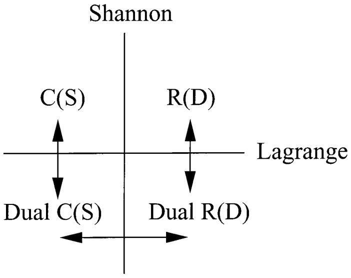

Fig. 1. Shannon duality characterized through the Lagrange dual problems of channel capacity and rate distortion.

Shannon Duality Correspondence

Dual of channel capacity		Dual of rate distortion
monomial	$\leftrightarrow$	posynomial
posynomial	$\leftrightarrow$	monomial
minimization	$\leftrightarrow$	maximization
$\geq$ constraints	$\leftrightarrow$	$\leq$ constraints
j (receiver side index)	$\leftrightarrow$	i (sender side index)
i (sender side index)	$\leftrightarrow$	j (receiver side index)
w^S	$\leftrightarrow$	w^{-D}
w^{s_i}	$\leftrightarrow$	$w^{-d_{ij}}$
z_j	$\leftrightarrow$	$z_i^{p_i}$
$z_j^{P_{ij}}$	$\leftrightarrow$	z_i
$H(\boldsymbol{P}^{(i)})$	$\leftrightarrow$	$-\log \dfrac{1}{p_i}$

Lagrange duality gives an exact and detailed analysis of the structures in Shannon duality.

- It resolves the apparent asymmetry between maximizing over a vector $\boldsymbol{p}$ in the channel capacity problem and minimizing over a matrix $\boldsymbol{P}$ in the rate distortion problem. In the Lagrange dual of $C(S)$, there are as many optimization variables as output alphabet symbols, plus a shadow price for the cost constraint. In the Lagrange dual of $R(D)$, there are as many optimization variables as input alphabet symbols, plus a shadow price for the distortion constraint.
- It answers the following question: since a vector $\boldsymbol{p}$ (the source distribution) is given in the rate distortion problem, and a matrix $\boldsymbol{P}$ (the channel matrix) is given in the channel capacity problem, what is the proper analog of p_i (the ith entry in $\boldsymbol{p}$) in the channel capacity problem? The last pair in the Shannon duality correspondence shows that the proper analog of $\log \frac{1}{p_i}$ in rate distortion is $H(\boldsymbol{P}^{(i)})$ in channel capacity: $\log \frac{1}{p_i}$ is the number of bits to describe an alphabet symbol with probability p_i in the Shannon code for lossless compression, and $H(\boldsymbol{P}^{(i)})$ is the number of bits needed to describe without loss the ith row of channel matrix. This correspondence can be interpreted in the context of universal source coding, where each row $\boldsymbol{P}^{(i)}$ of the channel matrix represents a possible distribution of the source.
- It confirms Shannon's remark [27] on introducing input costs to enhance duality. From the geometric programming Lagrange duals in standard form, it is easy to see that input costs $\boldsymbol{s}$ and cost constraint S in the channel capacity dual problem are complementary to distortion measures d_{ij} and distortion constraint D in the rate distortion dual problem.
- The dual variable $w \geq 1$ can be interpreted as the shadow price associated with the input cost constraint S and with the reconstruction distortion constraint D, respectively. From local sensitivity analysis [4], the optimal $-w^*$ tells us approximately how much increase in capacity $C(S)$ or reduction in rate $R(D)$ would result if the cost or distortion constraint could be relaxed by a small amount. From global sensitivity analysis [4], if w^* is large, then tightening the cost or distortion constraint will greatly decrease capacity, or increase the rate required to describe the source for a given distortion constraint. If w^* is small, then loosening the cost or distortion constraint will not significantly increase capacity or decrease rate.

The above characterization of Shannon duality through the Lagrange dual problems is different from the functional duality characterization in [23]. However, the two characterizations together imply that solving one geometric program in the form of (8) induces a set of problem parameters for another geometric program in the form of (25), whose optimal value equals that of the first geometric program.

B. Free Energy Interpretations

Various parallels between physics and information theory have been drawn, for example, in [9], [15]. In recent years, free energy concepts in statistical physics have also been used to identify threshold behaviors in error exponents [12], to understand iterative decoding through sum product algorithms on graphs [32], and to design low-density parity-check codes [21]. In this subsection, we interpret both the primal and dual problems of channel capacity as free energy optimization problems from statistical physics, complementing the analogy made by Berger in [1] between rate distortion and free energy optimization.

Consider a system with n states at temperature T, where each state i has energy e_i and probability p_i of occurring. Given an energy column vector $\boldsymbol{e}$ and a probability row vector $\boldsymbol{p}$, average energy is $U(\boldsymbol{p},\boldsymbol{e}) = \boldsymbol{pe}$ and entropy is $H(\boldsymbol{p}) = -\sum_{i=1}^n p_i \log p_i$. The Gibbs free energy is defined as

$$G(\boldsymbol{p},\boldsymbol{e}) = U(\boldsymbol{p},\boldsymbol{e}) - TH(\boldsymbol{p}) = \boldsymbol{pe} + T\sum_{i=1}^n p_i \log p_i.$$

Solving the problem of the Gibbs free energy minimization

$$\begin{array}{ll} \text{minimize} & \boldsymbol{pe} + T\sum_{i=1}^n p_i \log p_i \\ \text{subject to} & \boldsymbol{p1} = 1, \qquad \boldsymbol{p} \succeq 0 \end{array} \tag{43}$$

where the optimization variables are $\boldsymbol{p}$ and the constant parameters are $\boldsymbol{e}$, is important in statistical physics with several interpretations, such as striking a balance between energy minimization and entropy maximization.

Following the same argument in the discussion of conjugacy in Section I-B, it is easy to see that the Boltzmann distribution $\tilde{\boldsymbol{b}}$ minimizes $G(\boldsymbol{p},\boldsymbol{e})$ over $\boldsymbol{p}$ for a given energy vector $\boldsymbol{e}$, where $\tilde{b}_i$ is proportional to $\exp(-\frac{e_i}{T})$. The proportionality constant needed for normalization is called the partition function

$$Z(\boldsymbol{e}) = \sum_{i=1}^n \exp\left(\frac{-e_i}{T}\right).$$

The Gibbs free energy $G(\boldsymbol{p},\boldsymbol{e})$ induced by the Boltzmann distribution $\boldsymbol{p} = \tilde{\boldsymbol{b}}$ is called the Helmholtz free energy $F(\boldsymbol{e})$, which is the negative logarithm of $Z(\boldsymbol{e})$ scaled by T

$$F(\boldsymbol{e}) = G(\tilde{\boldsymbol{b}},\boldsymbol{e}) = -T\log\sum_{i=1}^n \exp\left(-\frac{e_i}{T}\right).$$

Due to convexity of the Gibbs free energy in $\boldsymbol{p}$ and concavity of the Helmholtz free energy in $\boldsymbol{e}$,

$$\max_{\boldsymbol{e}}\min_{\boldsymbol{p}} G(\boldsymbol{p},\boldsymbol{e}) = \min_{\boldsymbol{p}}\max_{\boldsymbol{e}} G(\boldsymbol{p},\boldsymbol{e}).$$

Therefore, maximizing the Helmholtz free energy is equivalent to finding the minimum Gibbs free energy for the worst case energy vector.

Now suppose the distribution on the states is not the Boltzmann distribution $\tilde{\boldsymbol{b}}$ but some general distribution $\boldsymbol{q}$. In this case, we will get a corresponding value for the Gibbs free energy $G(\boldsymbol{q},\boldsymbol{e})$, and the difference between this value and the Helmholtz free energy, normalized by the temperature T, is given by

$$\begin{aligned}\frac{1}{T}\left(G(\boldsymbol{q},\boldsymbol{e}) - G(\tilde{\boldsymbol{b}},\boldsymbol{e})\right) &= \frac{\boldsymbol{qe}}{T} - H(\boldsymbol{q}) - \frac{F(\boldsymbol{e})}{T} \\ &= \sum_{i=1}^n q_i \log\frac{q_i}{\tilde{b}_i} = D(\boldsymbol{q}\|\tilde{\boldsymbol{b}}).\end{aligned} \tag{44}$$

Therefore, another way to derive the Boltzmann distribution is through minimizing the difference between a general Gibbs free energy and the Helmholtz free energy, expressed as a KL divergence, over the probability simplex. Minimizing a KL divergence $D(\boldsymbol{q}\|\tilde{\boldsymbol{b}})$ over $\boldsymbol{q}$ is precisely the dual objective function of an unconstrained geometric program as shown in the Appendix.

For the primal problem of rate distortion (23), Berger [1] shows that minimizing the Lagrangian of rate distortion is a Gibbs free energy minimization problem. Furthermore, as can be readily verified, the Lagrange dual problem (25) of $R(D)$ is minimizing an average energy under the Helmholtz free energy constraints, where the energy of state i in the jth constraint is $e_{ij} = \gamma d_{ij} - \alpha_i - \log p_i$.

Turning to channel capacity problems, we note that the primal problem of channel capacity without input cost is a generalized Gibbs free energy minimization, where each state i has energy r_i, temperature is unity, average energy $U = \boldsymbol{Pr}$ is on the input distribution, but entropy is on the output distribution $\boldsymbol{q}$ induced by the input distribution and the channel. Minimizing the Lagrangian $\boldsymbol{p}(\boldsymbol{r}+\boldsymbol{s}) + \sum_{j=1}^M q_j \log q_j$ of channel capacity with input cost is still a Gibbs free energy minimization problem, with the energy for each state increased by the input cost s_i. Furthermore, the Lagrange dual problem (12) of $C(S)$ is a Helmholtz free energy maximization problem under average energy constraints: energy for each state i is $-\alpha_i$, the objective is to maximize the Helmholtz free energy $F(-\boldsymbol{\alpha})$, and the average energy constraints are $\sum_{j=1}^M P_{ij}(-\alpha_j) \le r_i$, $i = 1,2,\ldots,N$.

When channel capacity is written as $C = \max_{\boldsymbol{p}}[H(Y) - H(Y|X)]$, there is a standard typicality interpretation of C: the number of typical output sequences is about $e^{nH(Y)}$, but $e^{nH(Y|X)}$ typical output sequences may be associated with the same input sequence due to the unreliability of the channel. When channel capacity is written in the minmaxKL form (18), [10] gives a geometric interpretation of C: it is the radius of the smallest sphere containing all distributions $\boldsymbol{P}^{(i)}$, where the distance between two distributions is measured by KL divergence. When channel capacity is written in the geometric programming form (8), (9), we have a new physical interpretation that C is a maximized Helmholtz free energy.

More discussion on the free energy interpretation of error exponents, lossless source codes, and general geometric programs can be found in [6] and [12].

V. Conclusion

The Lagrange dual problems of channel capacity with input cost and rate distortion are geometric programs. The structures of these geometric programs allow us to efficiently generate upper bounds on channel capacity and lower bounds on rate distortion by solving systems of linear inequalities, to characterize

Shannon duality between transmission and compression in the discrete memoryless cases, and to interpret channel capacity and rate distortion problems as free energy optimization in statistical physics.

Appendix
Dual Proof of Corollary 1

Conjugate functions can be used to derive the Lagrange dual of an optimization problem [4]. It is readily verified that for the following geometric program over $\boldsymbol{y}$:

$$\begin{array}{ll}\text{minimize} & \log \sum_{j=1}^{k_0} \exp\left(\boldsymbol{a}_{0j}^T \boldsymbol{y} + b_{0j}\right) \\ \text{subject to} & \log \sum_{j=1}^{k_i} \exp\left(\boldsymbol{a}_{ij}^T \boldsymbol{y} + b_{ij}\right) \leq 0, \quad i = 1, \ldots, m\end{array}$$

the dual problem is

$$\begin{array}{ll}\text{maximize} & \boldsymbol{b}_0^T \boldsymbol{\nu}_0 - \sum_{j=1}^{k_0} \nu_{0j} \log \nu_{0j} \\ & \quad + \sum_{i=1}^{m} \left(\boldsymbol{b}_i^T \boldsymbol{\nu}_i - \sum_{j=1}^{k_i} \nu_{ij} \log \frac{\nu_{ij}}{\boldsymbol{1}^T \boldsymbol{\nu}_i} \right) \\ \text{subject to} & \boldsymbol{\nu}_i \succeq 0, \qquad i = 0, \ldots, m, \\ & \boldsymbol{1}^T \boldsymbol{\nu}_0 = 1, \\ & \sum_{i=0}^{m} \boldsymbol{A}_i^T \boldsymbol{\nu}_i = 0 \end{array} \tag{45}$$

where the optimization variables are $\boldsymbol{\nu}_i$, $i = 0, 1, \ldots, m$. Note that $\boldsymbol{A}_0$ is the matrix of the exponential constants in the primal objective function, and $\boldsymbol{A}_i$, $i = 1, 2, \ldots, m$, are the matrices of the exponential constants in the primal constraint functions.

A special case is an unconstrained geometric program

$$\text{minimize}_{\boldsymbol{y}} \ \log \sum_{i=1}^{N} \exp(\boldsymbol{a}_i^T \boldsymbol{y} + b_i).$$

From (45), the Lagrange dual problem of unconstrained geometric program is

$$\begin{array}{ll}\text{maximize} & \boldsymbol{b}^T \boldsymbol{\nu} - \sum_{i=1}^{N} \nu_i \log \nu_i \\ \text{subject to} & \boldsymbol{1}^T \boldsymbol{\nu} = 1, \\ & \boldsymbol{\nu} \succeq 0, \\ & \boldsymbol{A}^T \boldsymbol{\nu} = 0 \end{array} \tag{46}$$

where the optimization variables are $\boldsymbol{\nu}$. Notice that we can rewrite the dual objective in (46) as a KL divergence minimization: minimize$_{\boldsymbol{\nu}} D(\boldsymbol{\nu} \| \boldsymbol{d})$, or in exponentiated form: maximize$_{\boldsymbol{\nu}} \prod_i \left(\frac{d_i}{\nu_i}\right)^{\nu_i}$, where $b_i = \log d_i$. In general, a KL divergence $D(\boldsymbol{q}_1 \| \boldsymbol{q}_2)$ is the Gibbs free energy $G(\boldsymbol{q}_1, \boldsymbol{e})$ where the energy vector is the negative log likelihood of $\boldsymbol{q}_2$: $e_i = -\log q_{2,i}$. Therefore, the dual objective of an unconstrained geometric program (46) is equivalent to minimizing Gibbs free energy $G(\boldsymbol{\nu}, -\boldsymbol{b})$ at unit temperature.

Now recall the perspective suggested by Shannon [26]: Consider communication as a problem of reproducing an i.i.d. stochastic source (with N alphabet symbols) at the destination. If each alphabet symbol in the source has probability p_i of appearing, and the string is n symbols long, then for large n, the probability of a typical string is approximately

$$\prod_{i=1}^{N} p_i^{np_i} = e^{-nH(\boldsymbol{p})}$$

and to the first order in the exponent, the number of typical sequences is

$$K = e^{nH(\boldsymbol{p})} = \prod_{i=1}^{N} p_i^{-np_i}.$$

It turns out that K is the exponentiated objective function of the dual problem of an unconstrained geometric program: if we let the constants $d_i = n, \forall\, i$, and the variables $\nu_i = np_i$, then

$$K = \prod_{i=1}^{N} \left(\frac{d_i}{\nu_i}\right)^{\nu_i}.$$

Therefore, maximizing the number of typical sequences K, which is equivalent to maximizing $H(\boldsymbol{p})$, over linear constraints on $\boldsymbol{p}$, is Lagrange dual to an unconstrained geometric program.

It is natural to ask if maximizing a mutual information is the Lagrange dual of a properly constrained geometric program. This intuition, together with the Lagrange dual geometric program (45) quickly shows (through simple identification of the appropriate terms) that

$$\begin{array}{ll}\text{maximize} & I(X;Y) = \sum_{i=1}^{N} p_i \sum_{j=1}^{M} P_{ij} \log P_{ij} \\ & \quad - \sum_{j=1}^{M} q_j \log q_j \\ \text{subject to} & \boldsymbol{pP} = \boldsymbol{q}, \\ & \boldsymbol{p1} = 1, \qquad \boldsymbol{p} \succeq 0 \end{array}$$

is indeed the Lagrange dual of a linearly constrained geometric program with variables $\boldsymbol{\alpha}$

$$\begin{array}{ll}\text{minimize} & \log \sum_{j=1}^{M} e^{\alpha_j} \\ \text{subject to} & \sum_{j=1}^{M} P_{ij} \alpha_j - \sum_{j=1}^{M} P_{ij} \log P_{ij} \geq 0, \\ & \qquad\qquad i = 1, 2, \ldots, N. \end{array}$$

This geometric program is indeed the Lagrange dual of channel capacity as derived in Section II-B, and the above argument is a "dual" and shorter, proof of Corollary 1.

Acknowledgment

We would like to thank Tom Cover and David Forney for many stimulating interactions, Lin Xiao and Wei Yu for their helpful comments in the derivation of the duals, Kannan Ramchandran and Sandeep Pradhan for discussions on Shannon duality, Pramod Viswanath and Amos Lapidoth for pointing out the class of bounds in [13], and other reviewers for their suggestions.

REFERENCES

[1] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, NJ: Prentice Hall, 1971.

[2] R. Blahut, "Computation of channel capacity and rate distortion function," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 460–473, July 1972.

[3] R. J. Barron, B. Chen, and G. W. Wornell, "The duality between information embedding and source coding with side information, and some appplications," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1159–1180, May 2003.

[4] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[5] M. Chiang, "To layer or not to layer: Balancing transport and physical layers in wireless multihop networks," in *Proc. IEEE INFOCOM*, Hong Kong, China, Mar. 2004, to be published.

[6] ——, "Solving nonlinear problems in communication systems using geometric programming and dualities," Ph.D. dissertation, Stanford Univ., Stanford, CA, 2003.

[7] M. Chiang and A. Sutivong, "Efficient optimization of constrained nonlinear resource allocation," in *Proc. IEEE GLOBECOM*, San Francisco, CA, Dec. 2003.

[8] T. M. Cover and M. Chiang, "Duality between channel capacity and rate distortion with state information," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1629–1638, June 2002.

[9] T. M. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[10] I. Csiszár and J. Körner, *Information Theory: coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981.

[11] R. J. Duffin, E. L. Peterson, and C. Zener, *Geometric Programming: Theory and Applications*. New York: Wiley, 1967.

[12] G. D. Forney Jr, "On exponential error bounds for random codes on the BSC," unpublished manuscript, 2001.

[13] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[14] ——, "Source coding with side information and universal coding," in *Proc. IEEE Int. Symp. Information Theory*, 1976.

[15] P. Harremoes and F. Topsoe, "Unified approach to optimization techniques in Shannon theory," in *Proc. IEEE Int. Symp. Information Theory*, Lausanne, Switzerland, June/July 2002, p. 238.

[16] D. Julian, M. Chiang, D. O'Neill, and S. Boyd, "QoS and fairness constrained convex optimization of resource allocation for wireless cellular and ad hoc networks," in *Proc. IEEE INFOCOM*, New York, June 2002.

[17] S. Kandukuri and S. Boyd, "Optimal power control in interference limited fading wireless channels with outage probability specifications," *IEEE Trans. Wireless Commun.*, vol. 1, pp. 46–55, Jan. 2002.

[18] A. Lapidoth and S. M. Moser, "Capacity bounds via duality with applications to multiple-antenna systems on flat fading channels," *IEEE Trans. Inform. Theory*, submitted for publication.

[19] D. Luenberger, "A double look at duality," *IEEE Trans. Automat. Contr.*, vol. 37, pp. 1474–1482, Oct. 1992.

[20] R. McEliece, *Information Theory and Coding*. New York: Wiley, 1976.

[21] J. van Mourik, D. Saad, and Y. Kabashima, "Magnetization enumerator for LDPC codes–A statistical physics approach," in *Proc. IEEE Int. Symp. Information Theory*, Lausanne, Switzerland, June/July 2002, p. 256.

[22] Y. Nesterov and A. Nemirovsky, *Interior Point Polynomial Method in Convex Programming*. Philadelphia, PA: SIAM, 1994.

[23] S. Pradhan, J. Chou, and K. Ramchandran, "A characterization of functional duality between source and channel coding," in *Proc. IEEE Int. Symp. Information Theory*, Lausanne, Switzerland, June/July 2002, p. 224.

[24] S. Pradhan and K. Ramchandran, "Distributed source coding: symmetric rates and applications to sensor networks," in *Proc. IEEE Data Compression Conf*, Snowbird, UT, Mar. 2000.

[25] F. Reif, *Fundamentals of Statistical and Thermal Physics*. New York: McGraw-Hill, 1965.

[26] C. E. Shannon, "A mathematical theory of communications," *Bell. Syst. Tech. J.*, vol. 27, pp. 379–423, 1948.

[27] ——, "Coding theorems for a discrete source with a fidelity criterion," in *IRE Nat. Conv. Rec.*, Mar. 1959, pp. 142–163.

[28] ——, "Channels with side information at the transmitter," *IBM J. Res. Develop.*, vol. 2, pp. 289–293, 1958.

[29] P. O. Vontobel and D. M. Arnold, "An upper bound on the capacity of channels with memory and constraint input," in *Proc. Information Theory Workshop*, Cairns, Australia, Sept. 2001.

[30] F. M. J. Willems, "Computation of the Wyner-Ziv rate distortion funciton," Eindhoven Univ. Technol., Eindhoven, The Netherlands, Tech. Rep. 83-E-140, 1983.

[31] A. D. Wyner and J. Ziv, "The rate distortion function for source coding with side information at the decoder," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 1–10, Jan. 1976.

[32] J. Yedidia, W. Freeman, and Y. Weiss, "Generalized belief propagation," in *Proc. Neural Information Processing Systems Conf.*, Dec. 2001, pp. 689–695.