

Optimization in Information Theory

Dawei Shen

November 11, 2005

Abstract

This tutorial introduces the application of optimization techniques in information theory. We revisit channel capacity problem from the convex optimization perspective and show that the Lagrange dual problem of the channel capacity problem with input cost is a simple geometric problem, which can be efficiently solved using numerical algorithms such as primal-dual interior point method. Upper bound on the channel capacity can be efficiently generated. This tutorial assumes the reader has some background in optimization theory, especially in convex optimization and geometric optimization techniques.

1 The Channel Capacity Problem with Input Cost

First we are going to formulate the channel capacity problem with input cost into a standard optimization problem. 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 the channel law $P_{ij} = \text{Prob}\{Y = j | X = i\}$, $i = 1, 2, \dots, N, j = 1, 2, \dots, M$. The channel law forms channel matrix $\mathbf{P} \in \mathbf{R}^{N \times M}$, where the (i, j) entry of \mathbf{P} is $P_{ij} \geq 0$ with $\mathbf{P}\mathbf{1} = \mathbf{1}$. Here $\mathbf{1}$ represents for the column vector with all entries equal to 1. A distribution $\mathbf{p} \in \mathbf{R}^{1 \times N}$ on the input, together with a given channel matrix \mathbf{P} , induces a distribution $\mathbf{q} \in \mathbf{R}^{1 \times M}$ on the output by $\mathbf{q} = \mathbf{p}\mathbf{P}$, and a joint distribution \mathbf{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 \mathbf{s} .

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

$$C(S) = \max_{\mathbf{p}: \mathbf{p}\mathbf{s} \leq S} I(X; Y) \quad (1)$$

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

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

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

Therefore, the channel capacity problem can be formulated as the following maximization problem, referred to as the channel capacity problem with input cost:

$$\begin{aligned} \text{maximize} \quad & -\mathbf{p}\mathbf{r} - \sum_{j=1}^M q_j \log q_j \\ \text{subject to} \quad & \mathbf{p}\mathbf{P} = \mathbf{q}, \quad \mathbf{p}\mathbf{s} \leq S \\ & \mathbf{p}\mathbf{1} = 1, \quad \mathbf{p} \succeq 0 \end{aligned} \tag{2}$$

where the optimization variables are \mathbf{p} and \mathbf{q} . ‘ \succeq ’ means componentwise inequalities on a vector. The constant parameters are P , the channel matrix and

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

If there are no input cost constraint, the channel capacity problem becomes

$$\begin{aligned} \text{maximize} \quad & -\mathbf{p}\mathbf{r} - \sum_{j=1}^M q_j \log q_j \\ \text{subject to} \quad & \mathbf{p}\mathbf{P} = \mathbf{q} \\ & \mathbf{p}\mathbf{1} = 1, \quad \mathbf{p} \succeq 0 \end{aligned} \tag{3}$$

The primal optimization problem for the channel capacity has been formulated. Note that keeping two sets of optimization variables \mathbf{p} and \mathbf{q} , and introducing the equality constraint $\mathbf{p}\mathbf{P} = \mathbf{q}$ in the primal problem is a key step to derive an explicit and simple lagrange dual problem of (2).

2 Geometric Programming Dual

We are mainly interested in the Lagrange dual problem rather than the primal problem itself. After some manipulation as done in [2], it can be shown that the Lagrange dual problem of the channel capacity problem with input cost (2) is the following geometric program in convex form:

$$\begin{aligned} \text{minimize} \quad & \log \sum_{j=1}^M \exp(\alpha_j + \gamma S) \\ \text{subject to} \quad & \mathbf{P}\alpha + \gamma\mathbf{s} \succeq -\mathbf{r}, \\ & \gamma \geq 0 \end{aligned} \tag{4}$$

where the optimization variables are α and γ , and the constant parameters are \mathbf{P} , \mathbf{s} , and S .

An equivalent version of the lagrange dual problem is the following geometric program, in

standard form:

$$\begin{aligned}
& \text{minimize} && w^S \sum_j z_j \\
& \text{subjectto} && w^{s_i} \prod_{j=1}^M z_j^{P_{ij}} \geq e^{-H(\mathbf{P}^{(i)})}, i = 1, 2, \dots, N, \\
& && w \geq 1, z_j \geq 0, j = 1, 2, \dots, M
\end{aligned} \tag{5}$$

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

Lagrange duality between problems (2) and (4) means the following properties:

- Weak Duality. Any feasible (α, γ) of the Lagrange dual problem (4) 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 (4) is $C(S)$.

The weak duality part follows directly from the duality theory that the Lagrange dual function is always an upper bound on the primal maximization problem. The strong duality part of the proposition holds because the primal problem (1) is a convex optimization satisfying Slater's condition.

An immediate corollary could be obtained. The Lagrange dual of the channel capacity problem without input cost (3) is the following geometric program in convex form:

$$\begin{aligned}
& \text{minimize} && \log \sum_{j=1}^M e^{\alpha_j} \\
& \text{subjectto} && \mathbf{P}\alpha \succeq -\mathbf{r}
\end{aligned} \tag{6}$$

where the optimization variables are α , and the constant parameters are \mathbf{P} .

The equivalent version of the Lagrange dual problem is the following geometric program in standard form:

$$\begin{aligned}
& \text{minimize} && \sum_{j=1}^M z_j \\
& \text{subjectto} && \prod_{j=1}^M z_j^{P_{ij}} \geq e^{-H(\mathbf{P}^{(i)})}, i = 1, 2, \dots, N, \\
& && z_j \geq 0, j = 1, 2, \dots, M
\end{aligned} \tag{7}$$

Lagrange duality between problems (3) and (6) means the following:

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

The Lagrange dual (7) of the channel capacity problem is a simple geometric program with a linear objective function and only monomial inequality constraints. Also, dual problem (5)

is a generalized version of the dual problem (7), 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 $w = 0$, so that we indeed recover the dual problem (7) for channels without the input cost constraint.

Both the primal and the dual problems of $C(S)$ can be simultaneously and efficiently solved through a primal-dual interior point method, which scales smoothly for different channels and alphabet sizes. Utilizing the structure and sparsity of the exponent constant matrix \mathbf{A} of the geometric program dual for channel capacity, standard convex optimization algorithms can be further accelerated in this case.

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 \mathbf{q} can be recovered from the optimal dual variable α^*

$$q_j^* = \exp(\alpha_j^* - C), \quad j = 1, 2, \dots, M \quad (8)$$

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

$$\begin{aligned} -\mathbf{p}\mathbf{r} &= C + e^{-C} \left(\sum_{j=1}^M \alpha_j^* e^{\alpha_j^*} - C \sum_{j=1}^M e^{\alpha_j^*} \right) \\ \mathbf{p}\mathbf{P} &= \mathbf{q}^* \\ \mathbf{p}\mathbf{1} &= 1. \end{aligned}$$

At this point, we have shown that we can formulate the channel capacity problem into a geometric programming problem. Note that in this tutorial, we assume the reader has some familiarities with optimization theories, especially in convex optimization and geometric optimization. A convex optimization problem is easy to solve numerically by efficient algorithms such as the primal-dual interior point methods. Geometric programs have been used for various engineering problems. Geometric programming is a type of nonlinear problems that can be turned into convex optimization. Interested readers could refer to [1][2] for more complete introduction to geometric programming.

Thus, by the convex optimization techniques, we could compute the channel capacity and the capacity-achieving optimal input distributions efficiently.

3 Some observations from the formulation

From the complementary slackness property, 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 $\mathbf{P}\alpha + \gamma\mathbf{s} \succeq +\mathbf{r} = 0$ obtains α^* 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 [4].

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

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

From the primal and dual problems of channel capacity, we also obtain various optimality conditions. For example, if there are $\lambda \in \mathbf{R}^{N \times 1}$ and $\alpha \in \mathbf{R}^{M \times 1}$ satisfying the following KKT conditions for a given \mathbf{P}

$$\begin{aligned} \lambda &\succeq 0 \\ \mathbf{r} + \mathbf{P}\alpha &\succeq 0 \\ \frac{e^{\alpha_j}}{\sum_{j'=1}^M e^{\alpha_{j'}}} + \sum_{i=1}^N \lambda_i P_{ij} &= 0, \quad j = 1, 2, \dots, M \\ \lambda_i \left(r_i + \sum_{j=1}^M P_{ij} \alpha_j \right) &= 0, \quad i = 1, 2, \dots, N \end{aligned} \quad (9)$$

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

There is a minmax KL divergence (minmax KL) characterization of discrete memoryless channel capacity with input cost in [5][6]

$$C(S) = \min_{\mathbf{q}} \min_{\gamma \geq 0} \max_i \left[D(\mathbf{P}^{(i)} || \mathbf{q}) + \gamma(S - s_i) \right] \quad (10)$$

where the minimization over \mathbf{q} is over all possible output distributions. It is proved in [2] that the minmaxKL characterization (10) can be recovered from the complementary slackness property.

4 Bounding From the Dual

Because the inequality constraints in the dual problem (4) are affine, it is easy to obtain a dual feasible α 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.

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

$$C \leq \log M - \min_i H(\mathbf{P}^{(i)}).$$

Similarly, it is easy to verify that $\alpha_j = \log \max_i P_{ij}$ satisfy the dual constraints and give the following corollary: Corollary: Channel capacity is upper-bounded in terms of a maximum-likelihood receiver selecting $\arg \max_i P_{ij}$ for each output alphabet symbol j

$$C \leq \log \sum_{j=1}^M \max_i P_{ij} \quad (11)$$

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

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

When there is an input cost constraint $\mathbf{ps} \leq S$, the above upper bound would become

$$C(S) \leq \log \sum_{j=1}^M \max_i (e^{-s_i} P_{ij}) + S \tag{12}$$

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

5 Conclusions and Remarks

In this tutorial, we introduce the application of optimization theory in the information theory research. We revisit the classic channel capacity problem and show that finding the channel capacity and capacity-achieving optimal input distribution can be cast into a geometric programming problem in both standard form and convex form, which can be efficiently solved using primal-dual interior point method.

We also discuss some observations from the formulation. The geometric programming dual characterization is shown to be equivalent to the minmax KL characterization in [5][6]. Upper bound for the channel capacity can also be efficiently generated from the dual problem.

Note that in this tutorial, we focus on the channel capacity problem. In fact, the rate distortion problem can also be formulated into a geometric program and it has many similar forms, properties and observations. More interestingly, the dual of the channel capacity problem and the dual of the rate distortion problem can be connected beautifully by the Shannon duality correspondence. However, it's pretty mathematically heavy and has been beyond the scope of this tutorial. Interested readers may refer to [2] for a more detailed introduction.

References

- [1] S. Boyd, S. J. Kim, L. Vandenberghe, and A. Hassibi. "A Tutorial on Geometric Programming". *Optimization and Engineering*, 2005. to appear
- [2] M. Chiang and S. Boyd, "Geometric Programming Duals of Channel Capacity and Rate Distortion". *IEEE Trans. on Information Theory*, Vol.50, NO.2, pp. 245-258, February 2004
- [3] T.M.Cover and M. Chiang, "Duality between channel capacity and rate distortion with state information," , *IEEE Trans. on Information Theory*, Vol.48, pp.1629-1638, June 2002
- [4] R.G.Gallager, *Information Theory and Reliable Communication*, New York: Wiley, 1968
- [5] R.G.Gallager, "Source coding with side information and universal coding", in *Proc. IEEE Int. Symp. Information Theory*, 1976

- [6] I.Csiszar and J.Korner, *Information Thoery: Coding Theorems for Discrete Memoryless Systems*. New York: Academic, 1981
- [7] M. Chiang and A.Sutivong, "Efficient Optimization of constrained nonlinear resource allocation," in *Proc. IEEE GLOBECOM*, San Francisco, CA, Dec. 2003