

Optimality Conditions for Channel Capacity Problems

Supplemental Material for Information Theory

Henry D. Pfister

November 8th, 2024

1 Introduction

This note provides a description of optimality conditions for the channel capacity problem with and without cost constraints. The target audience is graduate students in communications and signal processing and the goal is to provide supplemental material for a graduate class on information theory which is based on the book by Cover and Thomas [1].

2 Channel Capacity

Consider a memoryless channel with a finite input and output alphabets \mathcal{X} and \mathcal{Y} . Let $W(y|x)$ be the probability of observing $y \in \mathcal{Y}$ given that $x \in \mathcal{X}$ was transmitted. The channel capacity is defined to be

$$C = \max_{p(x) \in \mathcal{P}(\mathcal{X})} I(X; Y),$$

where $\mathcal{P}(\mathcal{X})$ is the set of probability distributions on \mathcal{X} . This maximum is well-defined because $\mathcal{P}(\mathcal{X})$ is compact and $I(X; Y)$ is continuous. The achievability and converse for this rate is proven in [1].

Now, we consider some conditions that can be easily checked to verify the optimality of an input distribution. Let the divergence $D(Q(\cdot)||R(\cdot))$ between $Q(\cdot)$ and $R(\cdot)$ be

$$D(Q(\cdot)||R(\cdot)) \triangleq \sum_{y \in \mathcal{Y}} Q(y) \log \frac{Q(y)}{R(y)}.$$

Let the per-input mutual information be

$$I(X = x; Y) \triangleq D(W(\cdot|x)||pW(\cdot)) = \sum_{y \in \mathcal{Y}} W(y|x) \log \frac{W(y|x)}{pW(y)},$$

where $pW(y) \triangleq \sum_{x' \in \mathcal{X}} p(x')W(y|x')$ and thus $I(X; Y) = \sum_x p(x)I(X = x; Y)$. The following arguments are based on [2].

Lemma 2.1. *Let $p(x)$ and $q(x)$ be two capacity achieving input distributions. Then, $pW(y) = qW(y)$ for all $y \in \mathcal{Y}$. Thus, there is a unique output distribution generated by all capacity achieving input distributions.*

Proof. To see this, we write

$$\begin{aligned}
C &= \frac{1}{2}C + \frac{1}{2}C \\
&= \frac{1}{2} \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} W(y|x) \log \frac{W(y|x)}{pW(y)} + \frac{1}{2} \sum_{x \in \mathcal{X}} q(x) \sum_{y \in \mathcal{Y}} W(y|x) \log \frac{W(y|x)}{qW(y)} \\
&= \frac{1}{2} \sum_{y \in \mathcal{Y}} pW(y) \log \frac{1}{pW(y)} + \frac{1}{2} \sum_{y \in \mathcal{Y}} qW(y) \log \frac{1}{qW(y)} - \sum_{x \in \mathcal{X}} \frac{p(x) + q(x)}{2} \sum_{y \in \mathcal{Y}} W(y|x) \log \frac{1}{W(y|x)} \\
&\leq \sum_{y \in \mathcal{Y}} \frac{pW(y) + qW(y)}{2} \log \frac{2}{pW(y) + qW(y)} - \sum_{x \in \mathcal{X}} \frac{p(x) + q(x)}{2} \sum_{y \in \mathcal{Y}} W(y|x) \log \frac{1}{W(y|x)},
\end{aligned}$$

where the last step follows from the strict concavity of entropy and, thus, equality occurs iff $pW(y) = qW(y)$ for all $y \in \mathcal{Y}$. Next, we observe that equality must occur because the final expression is the mutual information associated with the input distribution $\frac{1}{2}(p(x) + q(x))$ and is thus upper bounded by C . Hence, the two output distributions must be equal. \square

Lemma 2.2. *For any distribution $R(y)$ on \mathcal{Y} ,*

$$C \leq \max_{x \in \mathcal{X}} D(W(\cdot|x)||R(\cdot))$$

with equality iff $R(y)$ is the unique output distribution generated by all capacity achieving input distributions.

Proof. To see this, let $p(x)$ be any capacity achieving input distribution and observe that

$$\begin{aligned}
&\max_{x \in \mathcal{X}} D(W(\cdot|x)||R(\cdot)) - C \\
&= \max_{x \in \mathcal{X}} D(W(\cdot|x)||R(\cdot)) - \sum_{x \in \mathcal{X}} p(x) D(W(\cdot|x)||pW(\cdot)) \\
&\geq \sum_{x \in \mathcal{X}} p(x) D(W(\cdot|x)||R(\cdot)) - \sum_{x \in \mathcal{X}} p(x) D(W(\cdot|x)||pW(\cdot)) \\
&= \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} W(y|x) \log \frac{W(y|x)}{R(y)} - \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} W(y|x) \log \frac{W(y|x)}{pW(y)} \\
&= \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} W(y|x) \log \frac{pW(y)}{R(y)} \\
&= D(pW(\cdot)||R(\cdot)) \\
&\geq 0,
\end{aligned}$$

with equality iff there is a constant γ such that $D(W(\cdot|x)||R(\cdot)) = \gamma$ whenever $p(x) > 0$ (i.e., first inequality holds with equality) and $R(y) = pW(y)$ for all $y \in \mathcal{Y}$ (i.e., second inequality holds with equality). \square

Theorem 2.3 (Optimality Conditions). *An input distribution $p(x)$ achieves capacity iff*

$$I(X = x; Y) \leq I(X; Y)$$

for all $x \in \mathcal{X}$, with equality whenever $p(x) > 0$. If this holds, then $C = I(X; Y)$. Thus, the capacity can be defined equivalently as

$$C = \min_{R(\cdot)} \max_{x \in \mathcal{X}} D(W(\cdot|x)||R(\cdot)).$$

Proof. If $I(X = x; Y) \leq I(X; Y)$ for all $x \in \mathcal{X}$ with $p(x) > 0$, then

$$I(X; Y) = \sum_{x \in \mathcal{X}} p(x) I(X = x; Y) \leq I(X; Y)$$

implies that $I(X = x; Y) = I(X; Y)$ for all $x \in \mathcal{X}$ with $p(x) > 0$. Combining this with Lemma 2.2, we find that $C \leq \max_{x \in \mathcal{X}} I(X = x; Y) = I(X; Y)$. Since, by definition, $C \geq I(X; Y)$ for all input distributions, we conclude that $C = I(X; Y)$ and $p(x)$ achieves capacity.

Conversely, if $p(x)$ achieves capacity, then the stationary condition for the constrained optimization over $p(x)$ shows that

$$0 = \frac{d}{dp(z)} \left(I(X; Y) + \eta \left(1 - \sum_{x \in \mathcal{X}} p(x) \right) \right) = I(X = z; Y) - \eta$$

for some η whenever $p(z) > 0$. Averaging this expression over $p(z)$ shows that $\eta = C = I(X = z; Y)$ for all z such that $p(z) > 0$.

For the second statement, we can minimize the capacity upper bound in Lemma 2.2 over $R(\cdot)$. Let $R^*(\cdot)$ be the minimizer so that the lemma implies

$$C \leq \max_{x \in \mathcal{X}} D(W(\cdot|x) || R^*(\cdot)),$$

with equality iff $R^*(y) = pW(y)$ for all $y \in \mathcal{Y}$ (i.e., $R^*(y)$ equals the capacity-achieving output distribution) and $D(W(\cdot|x) || R^*(\cdot)) = \gamma$ for all $x \in \mathcal{X}$ with $p(x) > 0$. The first condition implies that $D(W(\cdot|x) || R^*(\cdot)) = I(X = x; Y)$ for some input distribution $p(x)$ and the second condition implies the first-order optimality condition for the concave capacity optimization. Thus, we find that $\gamma = I(X = x; Y) = C$. □

3 Capacity with Cost Constraints

Consider a memoryless channel with a finite input alphabet \mathcal{X} and output alphabet \mathcal{Y} . Let $c(x)$ be the non-negative real cost of transmitting $x \in \mathcal{X}$ and $W(y|x)$ be the probability (density if \mathcal{Y} is not discrete) of observing $y \in \mathcal{Y}$ given that $x \in \mathcal{X}$ was transmitted. For a fixed cost constraint $\theta \geq 0$, the constrained capacity is given by

$$C_\theta = \max_{p(x): \sum_x p(x)c(x) \leq \theta} I(X; Y).$$

The achievability and converse for this rate can be proven using the channel coding theorem for DMCs and the techniques for handling costs developed for the coding theorem for the Gaussian channel. In fact, writing this proof clearly would be a good exercise for students. The average cost associated with an input distribution is $\theta = \sum_x p(x)c(x)$.

Theorem 3.1 (Optimality Conditions). *An input distribution achieves the unconstrained capacity C_∞ iff*

$$I(X = x; Y) \leq I(X; Y)$$

for all $x \in \mathcal{X}$ and equality occurs iff $p(x) > 0$. An input distribution achieves the constrained capacity C_θ iff, for some $\lambda \geq 0$,

$$\frac{I(X = x; Y) - I(X; Y)}{c(x) - \theta} \leq \lambda$$

for all $x \in \mathcal{X}$ and equality occurs iff $p(x) > 0$.

Proof. Since $I(X; Y)$ is a smooth concave function of the input distribution and the constraints (i.e., cost and normalization) are linear, the Karush-Kuhn-Tucker (KKT) conditions are necessary and sufficient for a global optimum value. The Lagrangian is given by

$$L(\mathbf{p}, \boldsymbol{\mu}, \nu, \lambda) = \sum_{x \in \mathcal{X}} p(x) I(X = x; Y) + \sum_{x \in \mathcal{X}} \mu(x) p(x) + \nu \left(1 - \sum_{x \in \mathcal{X}} p(x) \right) + \lambda \left(\theta - \sum_{x \in \mathcal{X}} p(x) c(x) \right),$$

where μ enforces the positivity constraint, ν enforces the normalization constraint, and $\lambda \geq 0$ enforces the cost constraint. Taking the derivative¹ w.r.t. $p(z)$ gives the first order necessary conditions

$$\frac{d}{dp(z)} L(\mathbf{p}, \boldsymbol{\mu}, \nu, \lambda) = I(X = z; Y) - \log(\mathbf{e}) + \mu(z) - \nu - \lambda c(z) = 0.$$

Since $\mu(z) \geq 0$ and complementary slackness implies $\mu(z) = 0$ iff $p(z) > 0$, one can rewrite this as

$$I(X = z; Y) - \log(\mathbf{e}) - \nu - \lambda c(z) \leq 0,$$

where equality holds iff $p(z) > 0$. Since equality holds if $p(z) > 0$, we can sum $p(z)$ times the RHS to get

$$\sum_{z \in \mathcal{X}} p(z) (I(X = z; Y) - \log(\mathbf{e}) - \nu - \lambda c(z)) = I(X; Y) - \log(\mathbf{e}) - \nu - \lambda \theta = 0.$$

Using this to eliminate ν allows one to simplify the condition to

$$I(X = z; Y) - I(X; Y) - \lambda c(z) + \lambda \theta \leq 0.$$

If there is no cost constraint (e.g., $\theta = \infty$), then we can choose $\lambda = 0$ and the condition simplifies to

$$I(X = z; Y) \leq I(X; Y).$$

Otherwise, we find that

$$\frac{I(X = z; Y) - I(X; Y)}{c(z) - \theta} \leq \lambda.$$

□

Remark 3.2. One can interpret these conditions from a marginal utility point of view. Without costs, one has “If there is any input, x , which provides more mutual information $I(X = x; Y)$ than the average $I(X; Y)$, then using this input more often can increase the mutual information.” With costs, one has “If there is any input, x , which provides more than its fair share of mutual information, then using it more often can increase the mutual information without increasing the cost.”

Example 3.3. Consider a costly Z-channel where $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{0, 1\}$, $c(x) = 2 - x$, and

$$W(y|x) = \begin{cases} 0 & \text{if } x = 0 \text{ and } y = 1 \\ \frac{1}{2} & \text{if } x = 1 \\ 1 & \text{if } x = y = 0. \end{cases}$$

If we let $\Pr(X = 0) = p$, then we have

$$I(X = x; Y) = \begin{cases} \log \frac{2}{1+p} & \text{if } x = 0 \\ \frac{1}{2} \log \frac{1}{1-p^2} & \text{if } x = 1, \end{cases}$$

which implies that

$$I(X; Y) = p \log \frac{2}{1+p} - \frac{1-p}{2} \log(1-p^2).$$

From this, we can solve the Lagrangian to get

$$p^*(\lambda) = \arg \max_p I(X; Y) - \lambda(2p + (1-p)) = \frac{4 - 2^{2\lambda}}{4 + 2^{2\lambda}},$$

which implies the cost is given by

$$\theta(\lambda) = 1 + \frac{4 - 2^{2\lambda}}{4 + 2^{2\lambda}} = \frac{8}{4 + 2^{2\lambda}}.$$

¹Though straightforward, it is a good exercise to compute the derivative of $I(X = x; Y)$ w.r.t. $p(z)$.

Finally, we verify the result for $\lambda = \frac{1}{2}$ and find that $p^*(\frac{1}{2}) = \frac{1}{3}$, $\theta(\frac{1}{2}) = \frac{4}{3}$, and

$$\begin{aligned}\frac{I(X=0;Y) - I(X;Y)}{2-\theta} &= \frac{\log(\frac{3}{2}) - \frac{1}{3}\log(\frac{3}{2}) - \frac{1}{3}\log(\frac{9}{8})}{2-\frac{4}{3}} = \frac{1}{2} \\ \frac{I(X=0;Y) - I(X;Y)}{1-\theta} &= \frac{\frac{1}{2}\log(\frac{9}{8}) - \frac{1}{3}\log(\frac{3}{2}) - \frac{1}{3}\log(\frac{9}{8})}{1-\frac{4}{3}} = \frac{1}{2}.\end{aligned}$$

Therefore, Theorem 3.1 implies this input distribution achieves the constrained capacity with cost $4/3$.

Remark 3.4. A careful reader might observe that the example is somewhat trivial because there is only one distribution with cost $4/3$. Still, the example follows a path that works for arbitrary channels.

While the method described above extends naturally to discrete-input continuous-output channels (via integration over \mathcal{Y}), the extension to continuous-input channels would require techniques from functional analysis and the calculus of variations. Fortunately, there is an information-theoretic proof for a slightly weaker result that extends easily to continuous-input channels with cost constraints.

Theorem 3.5. *Let $W(y|x)$ be a pdf on \mathcal{Y} for each $x \in \mathcal{X}$, $c(x)$ be the cost of transmitting $x \in \mathcal{X}$, and $R(y)$ be an arbitrary pdf on \mathcal{Y} . For any $\lambda \geq 0$, we have*

$$C_\theta \leq \lambda\theta + \sup_{x \in \mathcal{X}} (D(W(\cdot|x)||R(\cdot)) - \lambda c(x)), \quad (1)$$

where the inequality is strict unless $R(y)$ is induced by some capacity-achieving input distribution.

Proof. Let $p(x)$ be any capacity-achieving input distribution and let $pW(\cdot)$ be the pdf defined by $pW(y) = \int_{\mathcal{X}} p(x)W(y|x) dx$. Then, we have

$$\begin{aligned}& \sup_{x \in \mathcal{X}} (D(W(\cdot|x)||R(\cdot)) - \lambda c(x)) - (C - \lambda\theta) \\ &= \sup_{x \in \mathcal{X}} (D(W(\cdot|x)||R(\cdot)) - \lambda c(x)) - \int_{\mathcal{X}} p(x) (D(W(\cdot|x)||pW(\cdot)) - \lambda c(x)) dx \\ &\geq \int_{\mathcal{X}} p(x) (D(W(\cdot|x)||R(\cdot)) - \lambda c(x)) dx - \int_{\mathcal{X}} p(x) (D(W(\cdot|x)||pW(\cdot)) - \lambda c(x)) dx \\ &= \int_{\mathcal{X}} p(x) \int_{\mathcal{Y}} W(y|x) \log \frac{W(y|x)}{R(y)} dy dx - \int_{\mathcal{X}} p(x) \int_{\mathcal{Y}} W(y|x) \log \frac{W(y|x)}{pW(y)} dy dx \\ &= \int_{\mathcal{X}} p(x) \int_{\mathcal{Y}} W(y|x) \log \frac{pW(y)}{R(y)} dy dx \\ &= D(pW(\cdot)||R(\cdot)) \\ &\geq 0,\end{aligned}$$

where the inequality is strict unless $pW(y) = R(y)$ almost everywhere. \square

Corollary 3.6. *The following dual expression for capacity can also be useful*

$$C_\theta \leq \inf_{\lambda \geq 0} \left(\lambda\theta + \inf_{R(\cdot)} \sup_{x \in \mathcal{X}} (D(W(\cdot|x)||R(\cdot)) - \lambda c(x)) \right).$$

Proof. Using Theorem 3.5, we can take the infimum of (1) first over $R(\cdot)$ and then over $\lambda \geq 0$ to get the smallest possible upper bound on C_θ . By carefully considering the Lagrangian dual problem implied by Theorem 3.1, one can show that this bound is indeed tight. \square

4 Channel Symmetry

Consider a memoryless channel with a finite input alphabet \mathcal{X} and output alphabet \mathcal{Y} . Let $c(x)$ be the non-negative real cost of transmitting $x \in \mathcal{X}$ and $W(y|x)$ be the probability (density if \mathcal{Y} is not discrete)

of observing $y \in \mathcal{Y}$ given that $x \in \mathcal{X}$ was transmitted. Let $\pi : \mathcal{X} \rightarrow \mathcal{X}$ and $\sigma : \mathcal{Y} \rightarrow \mathcal{Y}$ be one-to-one mappings. We say that the channel has a (π, σ) symmetry if $c(x) = c(\pi(x))$ and

$$W(y|x) = W(\sigma(y)|\pi(x))$$

for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. More generally, one can consider the set of all channel symmetries

$$S = \{(\pi, \sigma) | c(x) = c(\pi(x)), W(y|x) = W(\sigma(y)|\pi(x)) \forall (x, y) \in \mathcal{X} \times \mathcal{Y}\}.$$

This set can be equipped with the structure of a group whose operation is based on the composition of permutations: $(\pi', \sigma') \circ (\pi, \sigma) = (\pi' \circ \pi, \sigma' \circ \sigma)$. Since this operation is closed on the set of all permutations, one can verify that S is a group by using the finite input alphabet property to show that S is finite.

Let the channel input/output pair X, Y be distributed according to $p(x)W(y|x)$, then

$$I_p(X; Y) \triangleq \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x) W(y|x) \log \frac{W(y|x)}{\sum_z p(z) W(y|z)},$$

where the sum over \mathcal{Y} becomes an integral if \mathcal{Y} is not discrete. The cost of the input distribution $p(x)$ is given by

$$\theta_p \triangleq \sum_{x \in \mathcal{X}} p(x) c(x).$$

Lemma 4.1. *If there exists an input distribution $p(x)$ with mutual information I_p and cost θ_p , then there exists an input distribution $q(x)$ with mutual information $I_q \geq I_p$ and cost $\theta_q = \theta_p$ that satisfies $q(x) = q(\pi(x))$ for all $x \in \mathcal{X}$ and $(\pi, \sigma) \in S$. Therefore, if $p(x)$ is capacity-achieving under this cost constraint, then so is $q(x)$. Moreover, an input distribution $p(x)$ cannot be capacity achieving if $\sum_x p(x)W(y|x) \neq \sum_x p(\pi(x))W(y|x)$ for some $(\pi, \sigma) \in S$.*

Proof. Let $p(\cdot)$ be an input distribution with mutual information I_p and cost E_p . Then, for any $(\pi, \sigma) \in S$, we have

$$\begin{aligned} I_p &= I_p(X; Y) \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x) W(y|x) \log \frac{W(y|x)}{\sum_z p(z) W(y|z)} \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(\pi(x)) W(\sigma(y)|\pi(x)) \log \frac{W(\sigma(y)|\pi(x))}{\sum_z p(\pi(z)) W(\sigma(y)|\pi(z))} \\ &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(\pi(x)) W(y|x) \log \frac{W(y|x)}{\sum_z p(\pi(z)) W(y|z)} \\ &= I_{\tilde{p}}(X; Y), \end{aligned}$$

where $\tilde{p}(x) = p(\pi(x))$ is a permuted input distribution. Since $c(x) = c(\pi(x))$, the permuted input distribution has cost $\theta_{\tilde{p}} = \theta_p$. Next, we define

$$q(x) = \frac{1}{|S|} \sum_{(\pi, \sigma) \in S} p(\pi(x)) \quad (2)$$

and use the concavity of mutual information to see that $I_q = I_q(X; Y) \geq I_p$. Therefore, $q(x)$ is an input distribution with mutual information $I_q \geq I_p$ and cost $\theta_q = \theta_p$ that satisfies $q(\pi(x)) = q(x)$ for all $x \in \mathcal{X}$ and $(\pi, \sigma) \in S$. Since every capacity-achieving input distribution induces the same output distribution, this also implies that the output distribution must be invariant under input relabeling $x \rightarrow \pi(x)$ for all $(\pi, \sigma) \in S$. \square

Example 4.2. For any m , consider the m^2 -QAM constellation

$$\mathcal{X} = \{x \in \mathbb{C} | x = (2a - m + 1) + (2b - m + 1)i, a, b \in \mathbb{Z}_m\}$$

with input cost $c(x) = |x|^2$. Let $\mathcal{Y} = \mathbb{C}$ and $W(y|x) = P_Z(|y - x|)$ where $P_Z(z)$ is the pdf of the magnitude of some complex circularly-symmetric additive noise Z . The symmetry group of this channel is isomorphic to the dihedral group D_4 which is well-known as the symmetry group of the square. It is generated by: (i) the 90-degree rotation $\pi(x) = e^{i\pi/2}x$ with $\sigma(y) = e^{i\pi/2}y$ and (ii) the reflection $\pi(x) = -x$ with $\sigma(y) = y$. Applying Lemma 4.1 shows that there exist capacity-achieving input distributions for m^2 -QAM where $p(x)$ takes on at most $\frac{1}{2} \lfloor \frac{m+1}{2} \rfloor \lfloor \frac{m+3}{2} \rfloor$ distinct values. For example, 16-QAM (i.e., $m = 4$) has at most 3 distinct values.

References

- [1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley Series in Telecommunications, Wiley, 2nd. ed., 2006.
- [2] F. Topsøe, “A new proof of a result concerning computation of the capacity for a discrete channel,” *Probability Theory and Related Fields*, vol. 22, no. 2, pp. 166–168, 1972.