I232 Information Theory
Chapter 13: Optimization in Information Theory

Brian Kurkoski

Japan Advanced Institute of Science and Technology

2023 May

Do Pop Quiz 13 on the LMS.

13.1 Convexity of Information Measures

Jensen's Inequality

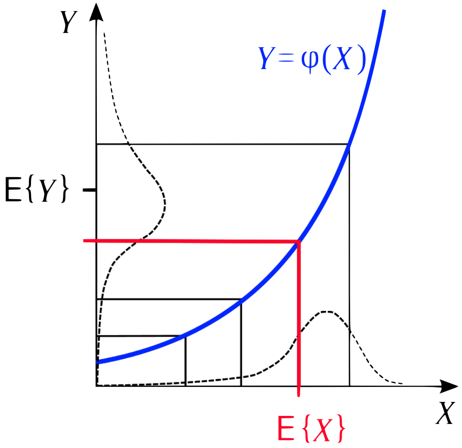
Jensen's Inequality if $\phi(x)$ is a convex function and X is a random variable then:

$$E[\phi(\mathsf{X})] \geq \phi\left(E[\mathsf{X}]\right)$$

$$\sum_{x \in \mathcal{X}} \phi(x) p_{\mathsf{X}}(x) \geq \phi\Big(\sum_{x \in \mathcal{X}} x p_{\mathsf{X}}(x)\Big)$$

- Equality holds if and only if ϕ is linear.
- The inequality $\geq$ changes to $\leq$ if ϕ is a concave function.

Jensen's Inequality

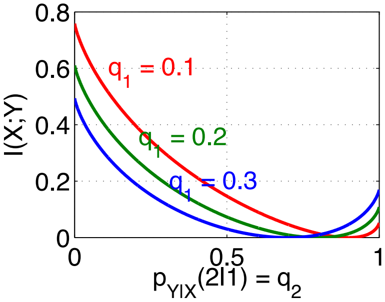


$$E[\phi(\mathsf{X})] \geq \phi\,(E[\mathsf{X}])$$

Image from Wikipedia "Jensen's Inequality"

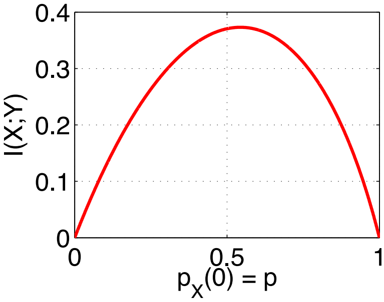
Importance of Convexity

Rate-distortion $R(D)$:
Minimization of a convex function



$$R(D) = \min I(\mathsf{X}; \widehat{\mathsf{X}})$$

Capacity C:
Maximization of a concave function



$$C = \max I(\mathsf{X}; \mathsf{Y})$$

If the objective function is convex, there are many mathematical tools to find solution.

13.2 Convexity of KL Divergence

Convexity of KL Divergence

Mutual Information

Entropy is Concave

KL Divergence is Convex

Log-Sum Inequality

Jensen's Inequality

Log-Sum Inequality

Proposition

Log-Sum Inequality For non-negative numbers, $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$:

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \big(\sum_{i=1}^{n} a_i\big) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

with equality if and only if $a_i/b_i =$ constant for any i.

Log-Sum Inequality

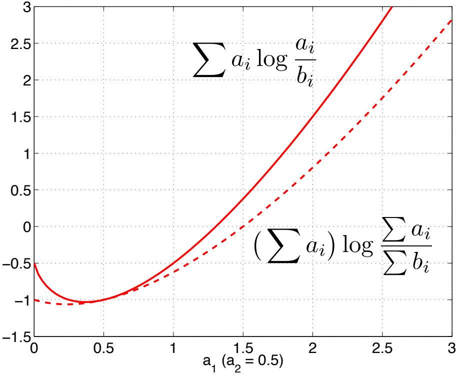


Figure shows specific values: $n = 2$, $a_2 = \frac{1}{2}, b_1 = b_2 = 1$

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

★SSQ: Log-Sum Inequality

Convexity of KL Divergence

Proposition

Convexity of KL divergence. $D(\mathbf{p}||\mathbf{q})$ is a convex in the pair of distributions $(\mathbf{p}, \mathbf{q})$.

★1

13.3 Computation of Channel Capacity

We know that the channel capacity C is:

$$C = \max_{p_{\mathsf{X}}(x)} I(\mathsf{X};\mathsf{Y})$$

We solved this optimization problem for a few special cases:

$$\text{Binary symmetric channel } C = 1 - h(p)$$
$$\text{Binary erasure channel } C = 1 - p$$

How to find the capacity of an arbitrary DMC?

- Must use numerical methods
- Arimoto-Blahut algorithm is one such method

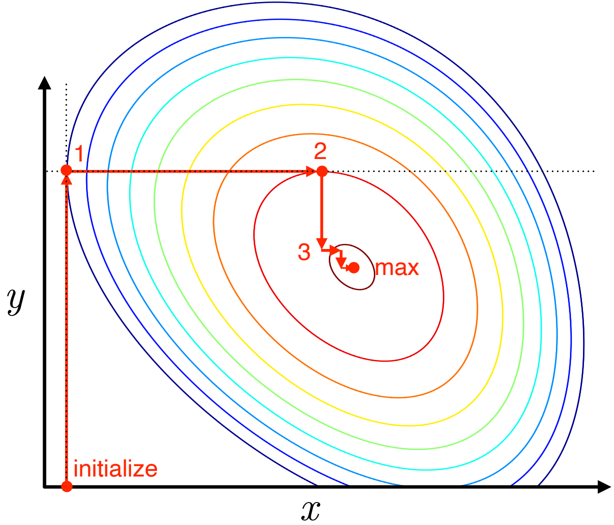
Principle of Alternating Maximization

Consider maximizing a concave function $f(x, y)$ as shown in the figure:

$$\max_{x,y} f(x, y)$$

One approach is to alternate between x and y:

$$\max_{x} \max_{y} f(x, y)$$



Two Papers on Solving this Problem

Two papers were published independently in 1972 that solved the problem of computing the channel capacity:

- Suguru Arimoto, "An Algorithm for Computing the Capacity of Arbitrary Discrete Memoryless Channels," *IEEE Transactions on Information Theory*, January 1972.
- Richard Blahut, "Computation of Channel Capacity and Rate-Distortion Functions," *IEEE Transactions on Information Theory*, July 1972.

大会企画（AT-1）　**情報理論研究会**

Arimoto-Blahut アルゴリズムの50年

参加費無料
(要事前登録)

会場：２号館
２３０２教室

- Arimoto-Blahut アルゴリズムにおける速い収束と遅い収束　中川健治（長岡技科大）
- 量子 AB アルゴリズム再訪　長岡浩司（電通大）
- 有本 -Blahut アルゴリズムの多端子モデルへの拡張とその収束性について　松嶋敏泰（早大）
- 多端子通信路に対する容量域計算アルゴリズムについて　大濱靖匡（電通大）
- Arimoto の指数計算アルゴリズム　實松　豊（東工大）





Webinarも無料

日時：2023年3月10日（金）13：00～16：30

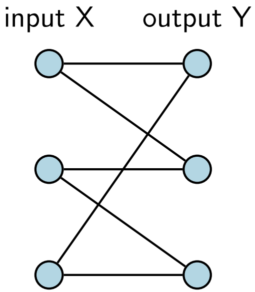
是非ご参加ください！

2023年電子情報通信学会総合大会

Arimoto-Blahut Algorithm: Definitions

Define the following for convenience:

$$\begin{aligned} r(x) &= p_{\mathsf{X}}(x) && \text{Input distribution} \\ p(y|x) &= p_{\mathsf{Y|X}}(y|x) && \text{Channel} \\ q(x|y) &= p_{\mathsf{X|Y}}(x|y) && \text{``Backward'' Channel} \end{aligned}$$



$r(x)$ depends on $q(x|y)$ and likewise $q(x|y)$ depends on $r(x)$:

$$p(y|x)r(x) = q(x|y) \sum_{x \in \mathcal{X}} p(y|x)r(x)$$

Recall that channel $p(y|x)$ is given and we must find $r(x)$

Capacity Computation as Alternating Maximization

Write the capacity as an alternating maximization problem:

$$C = \max_{q(x|y)} \max_{r(x)} I(\mathsf{X};\mathsf{Y})$$

Explicitly, the optimization problem is:

$$C = \max_{q(x|y)} \max_{r(x)} \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} r(x)p(y|x) \log \frac{q(x|y)}{r(x)}$$

Capacity Computation Part 1: Fix $r(x)$, max $q(x|y)$

The first step is to fix $r(x)$ and find $q^*(x|y)$ which maximizes:

$$C = \max_{q(x|y)} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} r(x) p(y|x) \log \frac{q(x|y)}{r(x)}.$$

This is an optimization problem, and the solution is given by:

$$q^*(x|y) = \frac{r(x)p(y|x)}{\sum_{x' \in \mathcal{X}} r(x')p(y|x')}$$

Capacity Computation Part 2: Fix $q(x|y)$, max $r(x)$

The second step is to fix $q(x|y)$ and find $r^*(x)$ which maximizes:

$$C = \max_{r(x)} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} r(x) p(y|x) \log \frac{q(x|y)}{r(x)}.$$

This is an optimization problem, and the solution is given by:

$$r^*(x) = \frac{\prod_y \left(q(x|y)\right)^{p_{\mathsf{Y}|\mathsf{X}}(y|x)}}{\sum_{x' \in \mathcal{X}} \prod_y \left(q(x'|y)\right)^{p(y|x')}}.$$

Arimoto-Blahut Algorithm for Capacity Computation

Require: A discrete memoryless channel $p_{\mathsf{Y}|\mathsf{X}}(y|x)$.

Ensure: Channel capacity C, capacity-achieving input distribution $p^*_{\mathsf{X}}(x)$

(a) Initialize $r(x)$ with a random distribution

(b) Fix $r(x)$, maximize over $q(x|y)$. For all $x \in \mathcal{X}$, $y \in \mathcal{Y}$:

$$q(x|y) = \frac{r(x) p_{\mathsf{Y}|\mathsf{X}}(y|x)}{\sum_{x' \in \mathcal{X}} r(x') p(y|x')}$$

(c) Fix $q(x|y)$, maximize over $r(x)$. For all $\mathbf{x} \in \mathcal{X}$:

$$r(x) = \frac{\prod_y \left(q(x|y)\right)^{p_{\mathsf{Y}|\mathsf{X}}(y|x)}}{\sum_{x' \in \mathcal{X}} \prod_y \left(q(x'|y)\right)^{p(y|x')}}$$

(d) Go to step (b) until the solution $r(x)$ stabilizes.

(e) Capacity C is $I(\mathsf{X};\mathsf{Y})$ computed using using $r(x)$ and $p_{\mathsf{Y}|\mathsf{X}}(y|x)$

Example: Computation of Channel Capacity

Consider the DMC given by:

$$p_{\mathsf{Y}|\mathsf{X}}(y|x) = \begin{bmatrix} \frac{3}{4} & \frac{1}{4} & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

Initialize A-B algorithm with:

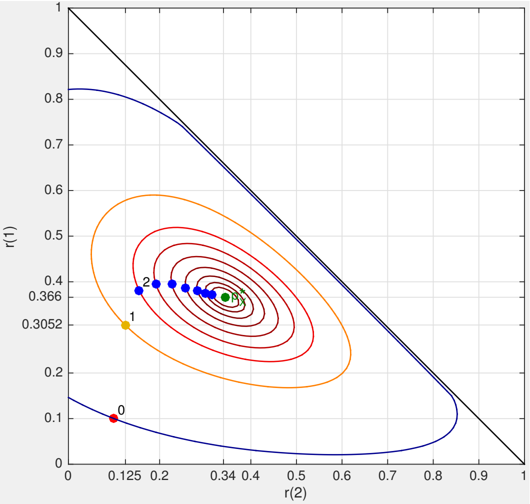
$$\text{initial} \quad r(x) = \begin{bmatrix} 0.1 & 0.1 & 0.8 \end{bmatrix}$$

After several iterations, stabilizes at:

$$p^*_{\mathsf{X}}(x) = [0.3657, 0.3440, 0.2903],$$

This $p^*_{\mathsf{X}}(x)$ is used to compute the capacity:

$$C = 0.7845 \text{ bits/channel use.}$$



13.4 Computation of Rate-Distortion Function $R(D)$

Recall the rate-distortion function $R(D)$:

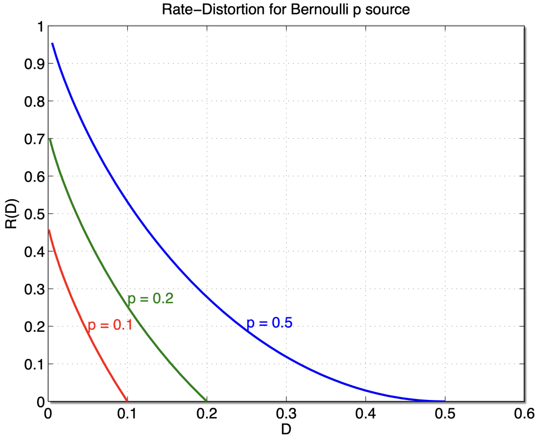
$$R(D) = \min_{p_{\widehat{\mathsf{X}}|\mathsf{X}}(\widehat{x}|x)} I(\mathsf{X};\widehat{\mathsf{X}})$$

We solved this optimization problem for a few special cases:

$$\text{Binary i.i.d. source } C = h(p) - h(D)$$

How to find $R(D)$ for an arbitrary source?

- Similar to channel capacity, use Arimoto-Blahut algorithm
- with suitable modifications



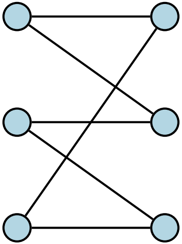
$R(D)$ Arimoto-Blahut Algorithm: Definitions

Define the following for convenience:

$$
\begin{aligned}
p(x) &= p_{\mathsf{X}}(x) && \text{Source distribution} \\
q(\widehat{x}|x) &= p_{\widehat{\mathsf{X}}|\mathsf{X}}(\widehat{x}|x) && \text{Quantizer} \\
r(\widehat{x}) &= p_{\widehat{\mathsf{X}}}(x) && \text{Reconstruction dist.}
\end{aligned}
$$

$r(\widehat{x})$ depends on $q(\widehat{x}|x)$ and likewise $q(\widehat{x}|x)$ depends on $r(\widehat{x})$.

Recall that channel $p(y|x)$ is given and we must find $r(x)$



Capacity$R(D)$ Computation as Alternating Maximization

Write the capacity as an alternating maximization problem:

$$R(D) = \min_{r(\widehat{x})} \min_{q(\widehat{x}|x)} I(\mathsf{X}; \widehat{\mathsf{X}}).$$

Explicitly, the optimization problem is:

$$R(D) = \min_{r(\widehat{x})} \min_{q(\widehat{x}|x)} \sum_{x \in \mathcal{X}} \sum_{\widehat{x} \in \widehat{\mathcal{X}}} q(\widehat{x}|x) p_{\mathsf{X}}(x) \log \frac{q(\widehat{x}|x)}{r(\widehat{x})}$$

Capacity Computation Part 1: Fix $r(\widehat{x})$, max $q(\widehat{x}|x)$

The first step is to fix $r(\widehat{x})$ and to find $q^*(x|y)$ which minimizes the mutual information:

$$q^*(x|y) = \arg\min_{q(\widehat{x}|x)} \sum_{x\in\mathcal{X}} \sum_{\widehat{x}\in\widehat{\mathcal{X}}} q(\widehat{x}|x) p_{\mathsf{X}}(x) \log \frac{q(\widehat{x}|x)}{r(\widehat{x})}$$

This is an optimization problem, the solution is given by:

$$q^*(x|y) = \frac{r(\widehat{x}) e^{-\lambda d(x,\widehat{x})}}{\sum_{\widehat{x}'\in\widehat{\mathcal{X}}} r(x') e^{-1} e^{-\lambda d(x,\widehat{x}')}}$$

Found using method of Lagrange multipliers. λ is the Lagrange multiplier.

Capacity Computation Part 2: Fix $q(\widehat{x}|x)$, max $r(\widehat{x})$

The second step is to fix $q(\widehat{x}|x)$ and find the $r^*(x)$ which minimizes the mutual information:

$$r^*(x) = \arg\min_{r(\widehat{x})} D\big(q(\widehat{x}|x)p_{\mathsf{X}}(x)||r(\widehat{x})p_{\mathsf{X}}(x)\big).$$

This is an optimization problem, and the solution is given by:

$$r^*(x) = \sum_{x \in \mathcal{X}} p_{\mathsf{X}}(x) q(\widehat{x}|x)$$

Arimoto-Blahut Algorithm for Rate-Distortion

Require: A discrete probability distribution $p_{\mathsf{X}}(x)$, an output alphabet $\widehat{\mathcal{X}}$, a distortion measure $d(x, \widehat{x})$, a parameter λ.

Ensure: Using optimized $q^*(x|y)$ and $r^*(x)$ from the last iteration, output R and D.

(a) Initialize with any choice of $r(\widehat{x})$, for example a random distribution.

(b) Fix $r(\widehat{x})$, minimize over $q(\widehat{x}|x)$:

$$q(\widehat{x}|x) = \frac{r(\widehat{x})e^{-\lambda d(x,\widehat{x})}}{\sum_{\widehat{x}' \in \widehat{\mathcal{X}}} r(x')e^{-\lambda d(x,\widehat{x}')}}$$

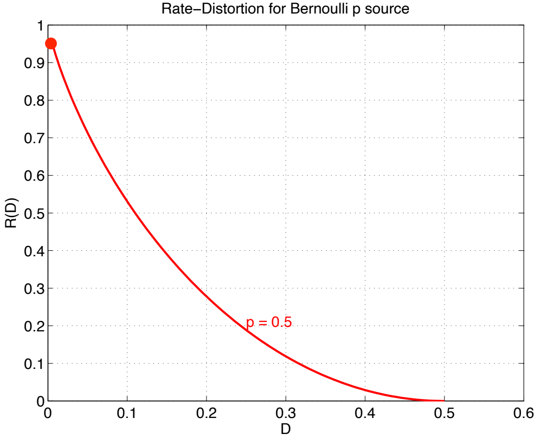
(c) Fix $q(\widehat{x}|x)$, minimize over $r(\widehat{x})$:

$$r(\widehat{x}) = \sum_{x \in \mathcal{X}} p_{\mathsf{X}}(x) q(\widehat{x}|x)$$

(d) Go to step (b) until the solution stabilizes.

Sweeping the Rate-Distortion Curve

1. Choose some value of λ
2. Perform A-B algorithm
3. Obtain some pair $R(D), D$.
4. Change value of λ, go to Step 2.



Class Info

- Next lecture: Monday, May 29. Network Information Theory.
- Tutorial Hours: Monday, May 29 at 13:30. Ask questions about homework.
- Last lecture: Wednesday, May 31. Review of Information Theory. No Pop Quiz.
- Final exam: Monday, June 5.