Elements of Information Theory

George Voutsadakis[1]

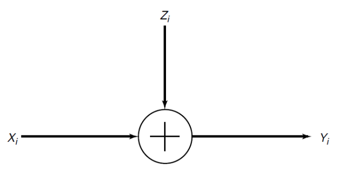
[1] Mathematics and Computer Science
Lake Superior State University

LSSU Math 500

1. Gaussian Channel
   - Gaussian Channel: Definitions
   - Converse to the Coding Theorem for Gaussian Channels
   - Parallel Gaussian Channels
   - Channels with Colored Gaussian Noise

The Gaussian Channel

- The most important continuous alphabet channel is the Gaussian channel.
- This is a time-discrete channel with output Y_i at time i, where Y_i is the sum of the input X_i and the noise Z_i.



- Z_i is drawn i.i.d. from a Gaussian distribution with variance N.
- Thus, $Y_i = X_i + Z_i$, $Z_i \sim \mathcal{N}(0, N)$.
- The noise Z_i is assumed to be independent of the signal X_i.
- Without further conditions, the capacity of this channel may be infinite.

Case of Infinite Capacity

- If the noise variance is zero, the receiver receives the transmitted symbol perfectly.

  Since X can take on any real value, the channel can transmit an arbitrary real number with no error.
- Suppose, next, that:
  - The noise variance is nonzero;
  - There is no constraint on the input.

  Then we can choose an infinite subset of inputs arbitrarily far apart, so that they are distinguishable at the output with arbitrarily small probability of error.

  Such a scheme has an infinite capacity as well.
- Thus, if the noise variance is zero or the input is unconstrained, the capacity of the channel is infinite.

Energy Constraint and Gaussian Noise

- The most common limitation on the input is an energy or power constraint.
- We assume an average power constraint.
- For any codeword $(x_1, x_2, \ldots, x_n)$ transmitted over the channel, we require that
$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 \leq P.$$
- The additive noise may be due to a variety of causes.
- However, by the Central Limit Theorem, the cumulative effect of a large number of small random effects will be approximately normal.
- So the Gaussian assumption is valid in a large number of situations.

Use as a Discrete Binary Channel

- We want to send 1 bit over the channel in one use of the channel.
- Given the power constraint, we send one of two levels,
  - $+\sqrt{P}$;
  - $-\sqrt{P}$.
- The receiver:
  - Looks at the corresponding Y received;
  - Tries to decide which of the two levels was sent.
- Assuming that both levels are equally likely (this would be the case if we wish to send exactly 1 bit of information), the optimum decoding rule is to decide that:
  - $+\sqrt{P}$ was sent, if $Y > 0$;
  - $-\sqrt{P}$ was sent, if $Y < 0$.

Use as a Discrete Binary Channel (Cont'd)

- The probability of error with such a decoding scheme is

$$\begin{array}{rcl} P_e & = & \frac{1}{2}\Pr(Y < 0|X = +\sqrt{P}) + \frac{1}{2}\Pr(Y > 0|X = -\sqrt{P}) \\ & = & \frac{1}{2}\Pr(Z < -\sqrt{P}|X = +\sqrt{P}) + \frac{1}{2}\Pr(Z > \sqrt{P}|X = -\sqrt{P}) \\ & = & \Pr(Z > \sqrt{P}) \\ & = & 1 - \Phi\left(\sqrt{\frac{P}{N}}\right). \end{array}$$

  Here

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt.$$

- Using such a scheme, we have converted the Gaussian channel into a discrete binary symmetric channel with crossover probability P_e.

Subsection 1

Gaussian Channel: Definitions

Information Capacity

Definition

The **information capacity** of the Gaussian channel with power constraint P is

$$C = \max_{f(x): EX^2 \leq P} I(X;Y).$$

- We can calculate the information capacity as follows:

$$\begin{array}{rcl} I(X;Y) & = & h(Y) - h(Y|X) \\ & = & h(Y) - h(X+Z|X) \\ & = & h(Y) - h(Z|X) \\ & \overset{\text{indep.}}{=} & h(Y) - h(Z). \end{array}$$

Information Capacity (Cont'd)

We have

$$h(Z) = \frac{1}{2} \log 2\pi e N.$$

Moreover, recall that:

- X and Z are independent;
- $EZ = 0$.

Therefore,

$$EY^2 = E(X+Z)^2 = EX^2 + 2EXEZ + EZ^2 = P + N.$$

The normal distribution maximizes the entropy for a given variance.

It follows that the entropy of Y is bounded by

$$\frac{1}{2} \log 2\pi e(P+N).$$

Bound for the Mutual Information

- We bound the mutual information:

$$\begin{array}{rcl} I(X;Y) & = & h(Y)-h(Z) \\ & \leq & \frac{1}{2}\log 2\pi e(P+N)-\frac{1}{2}\log 2\pi eN \\ & = & \frac{1}{2}\log\left(1+\frac{P}{N}\right). \end{array}$$

Hence, the information capacity of the Gaussian channel is

$$C=\max_{EX^2\leq P} I(X;Y)=\frac{1}{2}\log\left(1+\frac{P}{N}\right).$$

The maximum is attained when $X\sim\mathcal{N}(0,P)$.

We will show that this capacity is also the supremum of the rates achievable for the channel.

Codes for the Gaussian Channel

Definition

An (M, n) **code** for the Gaussian channel with power constraint P consists of the following:

1. An index set $\{1, 2, \ldots, M\}$;
2. An encoding function $x : \{1, 2, \ldots, M\} \to \mathcal{X}^n$, yielding codewords $x^n(1), x^n(2), \ldots, x^n(M)$, satisfying the power constraint P, i.e., for every codeword
$$\sum_{i=1}^{n} x_i^2(w) \leq nP, \quad w = 1, 2, \ldots, M;$$
3. A decoding function $g : \mathcal{Y}^n \to \{1, 2, \ldots, M\}$.

Capacity of a Gaussian Channel

- The **rate** is given by $R = \frac{\log M}{n}$.
- The **probability of error** is $\lambda_i = \Pr(g(Y^n) \neq i | X^n = x^n(i))$.
- The **arithmetic average of the probability of error** is defined by

$$P_e^{(n)} = \frac{1}{2^{nR}} \sum \lambda_i .$$

Definition

A rate R is said to be **achievable** for a Gaussian channel with a power constraint P if there exists a sequence of $(2^{nR}, n)$ codes, with codewords satisfying the power constraint, such that the maximal probability of error $\lambda^{(n)}$ tends to zero.

The **capacity** of the channel is the supremum of the achievable rates.

Capacity Theorem for a Gaussian Channel

Theorem (Capacity Theorem for a Gaussian Channel)

The capacity of a Gaussian channel with power constraint P and noise variance N is

$$C = \frac{1}{2} \log \left(1 + \frac{P}{N} \right) \text{ bits per transmission.}$$

Capacity Theorem (Plausibility Argument)

- We first present a plausibility argument as to why we may be able to construct $(2^{nC}, n)$ codes with a low probability of error.
- Consider any codeword of length n.
- The received vector is normally distributed with mean equal to the true codeword and variance equal to the noise variance.
- So, with high probability, the received vector is contained in a sphere of radius $\sqrt{n(N+\epsilon)}$ around the true codeword.
- We can assign everything within this sphere to the given codeword.
- When this codeword is sent, there will be an error only if the received vector falls outside the sphere, which happens with low probability.
- Similarly, we can choose other codewords and their corresponding decoding spheres.

Capacity Theorem (Plausibility Argument Cont'd)

- The volume of an n-dimensional sphere is of the form $C_n r^n$, where r is the radius of the sphere.
- In this case, each decoding sphere has radius $\sqrt{nN}$.
- The spheres are scattered throughout the space of received vectors.
- The received vectors have energy no greater than $n(P+N)$.
- So they lie in a sphere of radius $\sqrt{n(P+N)}$.
- The maximum number of nonintersecting decoding spheres in this volume is no more than

$$\frac{C_n(n(P+N))^{\frac{n}{2}}}{C_n(nN)^{\frac{n}{2}}} = \left(1+\frac{P}{N}\right)^{\frac{n}{2}}.$$

- The rate of the code is $\frac{\log{(1+\frac{P}{N})^{\frac{n}{2}}}}{n} = \frac{1}{2}\log{(1+\frac{P}{N})}$.

Capacity Theorem (Achievability 1)

- **Achievability**: We will use the same ideas as in the proof of the channel coding theorem in the case of discrete channels.

  Namely, we use random codes and joint typicality decoding.

  However, we must make some modifications to take into account the power constraint and the continuity of the variables.

1. **Generation of the Codebook**: We wish to generate a codebook in which all the codewords satisfy the power constraint.

   To ensure this, we generate the codewords with each element i.i.d. according to a normal distribution with variance $P - \epsilon$.

   Since for large n, $\frac{1}{n}\sum X_i^2 \to P - \epsilon$, the probability that a codeword does not satisfy the power constraint will be small.

   Let $X_i(w)$, $i = 1, 2, \ldots, n$, $w = 1, 2, \ldots, 2^{nR}$ be i.i.d. $\sim \mathcal{N}(0, P - \epsilon)$, forming codewords $X^n(1), X^n(2), \ldots, X^n(2^{nR}) \in \mathbb{R}^n$.

Capacity Theorem (Achievability 2-3)

2. **Encoding**: After the generation of the codebook, the codebook is revealed to both the sender and the receiver.

   To send the message index w, the transmitter sends the w-th codeword $X^n(w)$ in the codebook.

3. **Decoding**: The receiver looks down the list of codewords $\{X^n(w)\}$ and searches for one that is jointly typical with the received vector.
   - If there is one and only one such codeword $X^n(w)$, the receiver declares $\widehat{W} = w$ to be the transmitted codeword.
   - Otherwise, the receiver declares an error.

   The receiver also declares an error if the chosen codeword does not satisfy the power constraint.

Capacity Theorem (Achievability 4)

4. **Probability of Error**: Without loss of generality, assume that codeword 1 was sent. Thus, $Y^n = X^n(1) + Z^n$.
   Define the following events:
   $$\begin{aligned} E_0 &= \left\{ \frac{1}{n}\sum_{j=1}^{n} X_j^2(1) > P \right\}; \\ E_i &= \{(X^n(i), Y^n) \text{ is in } A_\epsilon^{(n)}\}. \end{aligned}$$
   Then an error occurs if one of the following happens:
   - E_0 occurs (the power constraint is violated);
   - E_1^c occurs (the transmitted codeword and the received sequence are not jointly typical);
   - $E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}$ occurs (some wrong codeword is jointly typical with the received sequence).

   Let $\mathcal{E}$ denote the event $\widehat{W} \neq W$.

   Let P denote the conditional probability given that $W = 1$.

Capacity Theorem (Achievability 4 Cont'd)

- We have

$$\begin{array}{rcl}\Pr(\mathcal{E}|W=1) & = & P(\mathcal{E}) \\ & = & P(E_0 \cup E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}) \\ & \overset{\text{union}}{\leq} & P(E_0) + P(E_1^c) + \sum_{i=2}^{2^{nR}} P(E_i).\end{array}$$

  - By the Law of Large Numbers, $P(E_0) \to 0$ as $n \to \infty$.
  - By the joint AEP (which can be proved using the same argument as that used in the discrete case), $P(E_1^c) \to 0$. Hence $P(E_1^c) \leq \epsilon$, for n sufficiently large.
  - By the code generation process, $X^n(1)$ and $X^n(i)$ are independent. Thus, so are Y^n and $X^n(i)$. Hence, the probability that $X^n(i)$ and Y^n will be jointly typical is $\leq 2^{-n(I(X;Y)-3\epsilon)}$ by the joint AEP.

Capacity Theorem (Achievability 4 Cont'd)

- Now let W be uniformly distributed over $\{1, 2, \ldots, 2^{nR}\}$. Then

$$\Pr(\mathcal{E}) = \frac{1}{2^{nR}} \sum \lambda_i = P_e^{(n)}.$$

So, for n sufficiently large and $R < I(X;Y) - 3\epsilon$, we have

$$\begin{array}{rcl} P_e^{(n)} & = & \Pr(\mathcal{E}) = \Pr(\mathcal{E}|W = 1) \\ & \leq & P(E_0) + P(E_1^c) + \sum_{i=2}^{2^{nR}} P(E_i) \\ & \leq & \epsilon + \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)} \\ & = & 2\epsilon + (2^{nR} - 1) 2^{-n(I(X;Y)-3\epsilon)} \\ & \leq & 2\epsilon + 2^{3n\epsilon} 2^{-n(I(X;Y)-R)} \\ & \leq & 3\epsilon. \end{array}$$

This proves the existence of a good $(2^{nR}, n)$ code.

Capacity Theorem (Achievability Conclusion)

- Choosing a good codebook and deleting the worst half of the codewords, we obtain a code with low maximal probability of error.

  The codewords that do not satisfy the power constraint have probability of error 1.

  So they must belong to the worst half of the codewords.

  So the power constraint is satisfied by each of the non-deleted codewords.

  Hence we have constructed a code that achieves a rate arbitrarily close to capacity.
- The next task is to show that the achievable rate cannot exceed the capacity.

Subsection 2

Converse to the Coding Theorem for Gaussian Channels

Converse to the Coding Theorem

- We complete the proof that the capacity of a Gaussian channel is $C = \frac{1}{2}\log(1+\frac{P}{N})$ by proving that rates $R > C$ are not achievable.
- We show, if $P_e^{(n)} \to 0$ for a sequence of $(2^{nR}, n)$ codes for a Gaussian channel with power constraint P, then $R \leq C = \frac{1}{2}\log(1+\frac{P}{N})$.

  Consider any $(2^{nR}, n)$ code that satisfies the power constraint

  $$\frac{1}{n}\sum_{i=1}^{n} x_i^2(w) \leq P, \quad w = 1, 2, \ldots, 2^{nR}.$$

  Let W be distributed uniformly over $\{1, 2, \ldots, 2^{nR}\}$.

  The uniform distribution over the index set $W \in \{1, 2, \ldots, 2^{nR}\}$ induces a distribution on the input codewords.

  This, in turn, induces a distribution over the input alphabet.

  We get a joint distribution on $W \to X^n(W) \to Y^n \to \widehat{W}$.

Converse to the Coding Theorem (Error)

- To relate probability of error and mutual information, we can apply Fano's Inequality to obtain $H(W|\widehat{W}) \leq 1 + nRP_\epsilon^{(n)} = n\epsilon_n$, where $\epsilon_n \to 0$ as $P_\epsilon^{(n)} \to 0$. Hence,

$$\begin{array}{rcl} nR & = & H(W) = I(W;\widehat{W}) + H(W|\widehat{W}) \\ & \leq & I(W;\widehat{W}) + n\epsilon_n \\ & \leq & I(X^n;Y^n) + n\epsilon_n \\ & = & h(Y^n) - h(Y^n|X^n) + n\epsilon_n \\ & = & h(Y^n) - h(Z^n) + n\epsilon_n \\ & \leq & \sum_{i=1}^n h(Y_i) - h(Z^n) + n\epsilon_n \\ & = & \sum_{i=1}^n h(Y_i) - \sum_{i=1}^n h(Z_i) + n\epsilon_n \\ & = & \sum_{i=1}^n I(X_i;Y_i) + n\epsilon_n. \end{array}$$

Here $X_i = x_i(W)$, where W is drawn according to the uniform distribution on $\{1, 2, \ldots, 2^{nR}\}$.

Converse to the Coding Theorem (Capacity)

- Let
$$P_i = \frac{1}{2^{nR}} \sum_w x_i^2(w).$$
Now we have:
  - $Y_i = X_i + Z_i$;
  - X_i and Z_i are independent.

  So the average power EY_i^2 of Y_i is $P_i + N$.

  Hence, since entropy is maximized by the normal distribution,
$$h(Y_i) \leq \frac{1}{2} \log 2\pi e(P_i + N).$$
Continuing with the inequalities of the converse, we obtain
$$\begin{array}{rcl} nR & \leq & \sum(h(Y_i) - h(Z_i)) + n\epsilon_n \\ & \leq & \sum(\frac{1}{2} \log (2\pi e(P_i + N)) - \frac{1}{2} \log 2\pi eN) + n\epsilon_n \\ & = & \sum \frac{1}{2} \log (1 + \frac{P_i}{N}) + n\epsilon_n. \end{array}$$

Converse to the Coding Theorem (Capacity)

- Each of the codewords satisfies the power constraint.

  Thus, so does their average,

  $$\frac{1}{n}\sum_i P_i \leq P.$$

  Now $f(x) = \frac{1}{2}\log(1+x)$ is a concave function of x.

  So we can apply Jensen's inequality to obtain

  $$\begin{aligned}\tfrac{1}{n}\textstyle\sum_{i=1}^n \tfrac{1}{2}\log(1+\tfrac{P_i}{N}) &\leq \tfrac{1}{2}\log(1+\tfrac{1}{n}\textstyle\sum_{i=1}^n \tfrac{P_i}{N})\\ &\leq \tfrac{1}{2}\log(1+\tfrac{P}{N}).\end{aligned}$$

  Thus, we get

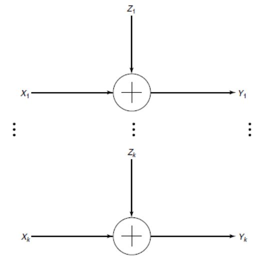
  $$R \leq \frac{1}{2}\log\left(1+\frac{P}{N}\right)+\epsilon_n, \quad \epsilon_n \to 0.$$

Subsection 3

Parallel Gaussian Channels

Parallel Gaussian Channels

- In this section we consider k independent Gaussian channels in parallel with a common power constraint.
- The objective is to distribute the total power among the channels so as to maximize the capacity.
- Assume that we have a set of Gaussian channels in parallel.
- The output of each channel is the sum of the input and Gaussian noise.
- For channel j, $Y_j = X_j + Z_j$, $j = 1, 2, \ldots, k$, with $Z_j \sim \mathcal{N}(0, N_j)$, and the noise is assumed to be independent from channel to channel.



- We assume that there is a common power constraint on the total power used, that is, $E \sum_{j=1}^{k} X_j^2 \leq P$.

Information Capacity of Parallel Gaussian Channels

- The information capacity of the channel C is

$$C = \max_{f(x_1,x_2,\ldots,x_k):\sum EX_i^2 \leq P} I(X_1, X_2, \ldots, X_k; Y_1, Y_2, \ldots, Y_k).$$

We calculate the distribution that achieves the information capacity. Since $Z_1, Z_2, \ldots, Z_k$ are independent,

$$\begin{aligned} & I(X_1, X_2, \ldots, X_k; Y_1, Y_2, \ldots, Y_k) \\ & = h(Y_1, Y_2, \ldots, Y_k) - h(Y_1, Y_2, \ldots, Y_k | X_1, X_2, \ldots, X_k) \\ & = h(Y_1, Y_2, \ldots, Y_k) - h(Z_1, Z_2, \ldots, Z_k | X_1, X_2, \ldots, X_k) \\ & = h(Y_1, Y_2, \ldots, Y_k) - h(Z_1, Z_2, \ldots, Z_k) \\ & = h(Y_1, Y_2, \ldots, Y_k) - \textstyle\sum_i h(Z_i) \\ & \leq \textstyle\sum_i (h(Y_i) - h(Z_i)) \\ & \leq \textstyle\sum_i \frac{1}{2} \log\left(1 + \frac{P_i}{N_i}\right), \end{aligned}$$

where $P_i = EX_i^2$ and $\sum P_i = P$.

The Case of Equality

- Equality is achieved by

$$(X_1, X_2, \ldots, X_k) \sim \mathcal{N}\left(0, \begin{bmatrix} P_1 & 0 & \cdots & 0 \\ 0 & P_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & P_k \end{bmatrix}\right).$$

Maximization of Information Capacity

- The problem is reduced to finding the power allotment that maximizes the capacity subject to the constraint that $\sum P_i = P$.

  This is a standard optimization problem and can be solved using Lagrange multipliers.

  We write

  $$J(P_1, \ldots, P_k) = \sum \frac{1}{2} \log \left(1 + \frac{P_i}{N_i} \right) + \lambda \left(\sum P_i \right).$$

  Differentiate with respect to P_i:

  $$\frac{1}{2} \frac{1}{P_i + N_i} + \lambda = 0.$$

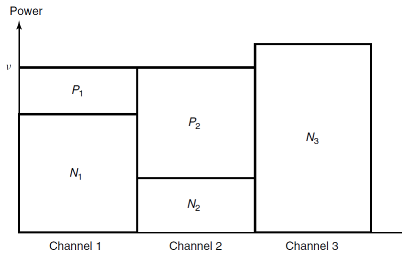
  Equivalently, $P_i = \nu - N_i$, where $\nu = -\frac{1}{2\lambda}$.

Maximization of Information Capacity (Cont'd)

- We found
$$P_i = \nu - N_i, \quad \nu = -\frac{1}{2\lambda}.$$
- Since the P_i's must be nonnegative, it may not always be possible to find a solution of this form.
- In this case, we use the Kuhn-Tucker conditions to verify that the solution
$$P_i = (\nu - N_i)^+$$
is the assignment that maximizes capacity, where ν is chosen so that
$$\sum (\nu - N_i)^+ = P$$
and we have used the notation
$$(x)^+ = \begin{cases} x, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases}.$$

The Solution

- This solution is illustrated on the right.
- The vertical levels indicate the noise levels in the various channels.
- As the signal power is increased from zero, we allot the power to the channels with the lowest noise.



- When the available power is increased still further, some of the power is put into noisier channels.
- The process by which the power is distributed among the various bins is identical to the way in which water distributes itself in a vessel.
- So this process is sometimes referred to as **water-filling**.

Subsection 4

Channels with Colored Gaussian Noise

Channels with Colored Gaussian Noise

- We consider the case of a set of parallel Gaussian channels in which the noise samples from different channels is dependent.
- This represents not only the case of parallel channels, but also the case when the channel has Gaussian noise with memory.
- For channels with memory, we can consider a block of n consecutive uses of the channel as n channels in parallel with dependent noise.
- We will calculate only the information capacity for this channel.

Parameters

- Let K_Z be the covariance matrix of the noise.
- Let K_X be the input covariance matrix.
- The power constraint on the input can be written as

$$\frac{1}{n}\sum_i EX_i^2 \leq P.$$

- Equivalently,

$$\frac{1}{n}\mathrm{tr}(K_X) \leq P.$$

- The power constraint here depends on n.
- So the capacity has to be calculated for each n.

Maximizing the Entropy

- Just as in the case of independent channels, we can write

$$I(X_1, X_2, \ldots, X_n; Y_1, Y_2, \ldots, Y_n) = h(Y_1, Y_2, \ldots, Y_n) - h(Z_1, Z_2, \ldots, Z_n).$$

- Here $h(Z_1, Z_2, \ldots, Z_n)$ is determined only by the distribution of the noise and is not dependent on the choice of input distribution.
- So finding the capacity amounts to maximizing $h(Y_1, Y_2, \ldots, Y_n)$.
- The entropy of the output is maximized when Y is normal.
- This is achieved when the input is normal.
- We are working under the hypotheses that:
  - The input and the noise are independent;
  - The covariance of the output Y is $K_Y = K_X + K_Z$.
- So the entropy is

$$h(Y_1, Y_2, \ldots, Y_n) = \frac{1}{2} \log\left((2\pi e)^n |K_X + K_Z|\right).$$

- Now the problem is reduced to choosing K_X so as to maximize $|K_X + K_Z|$, subject to a trace constraint on K_X.

Maximizing $|K_X + K_Z|$

- To maximize $|K_X + K_Z|$, we decompose K_Z into its diagonal form, $K_Z = Q\Lambda Q^t$, where $QQ^t = I$.

  Then
  $$\begin{array}{rcl} |K_X + K_Z| & = & |K_X + Q\Lambda Q^t| \\ & = & |Q||Q^t K_X Q + \Lambda||Q^t| \\ & = & |Q^t K_X Q + \Lambda| \\ & = & |A + \Lambda|, \end{array}$$

  where $A = Q^t K_X Q$.

  Since for any matrices B and C, $\text{tr}(BC) = \text{tr}(CB)$, we have
  $$\text{tr}(A) = \text{tr}(Q^t K_X Q) = \text{tr}(QQ^t K_X) = \text{tr}(K_X).$$

  Now the problem is reduced to maximizing $|A + \Lambda|$ subject to a trace constraint $\text{tr}(A) \leq nP$.

Maximizing $|A+\Lambda|$ Subject to $\mathrm{tr}(A) \leq nP$

- We apply Hadamard's inequality that states that the determinant of any positive definite matrix K is less than the product of its diagonal elements, i.e., $|K| \leq \prod_i K_{ii}$, with equality iff the matrix is diagonal.

  Thus,
  $$|A+\Lambda| \leq \prod_i (A_{ii} + \lambda_i),$$
  with equality iff A is diagonal.

  Now $A_{ii} \geq 0$ and A is subject to a trace constraint, $\frac{1}{n}\sum_i A_{ii} \leq P$.

  So the maximum value of $\prod_i (A_{ii} + \lambda_i)$ is attained when
  $$A_{ii} + \lambda_i = \nu.$$

Maximizing $|A+\Lambda|$ Subject to $\mathrm{tr}(A) \leq nP$ (Cont'd)

- Given the constraints, it may not always be possible to satisfy this equation with positive A_{ii}.
- In such cases, we can show by the standard Kuhn-Tucker conditions that the optimum solution corresponds to setting
$$A_{ii} = (\nu - \lambda_i)^+,$$
where the water level ν is chosen so that $\sum A_{ii} = nP$.
- This value of A maximizes the entropy of Y.
- Hence, it also maximizes the mutual information.