8 Capacity of the Gaussian Channel

Consider the following additive noise channel

\[ Y_i = X_i + Z_i, \]

where the \( Z_i \) are i.i.d., zero-mean, variance \( N \) Gaussian random variables representing the noise and the \( X_i \) represent the signal. It is assumed that the noise and signal are independent. We are interested in the capacity of the above channel under the following average power constraint:

\[ \frac{1}{n} \sum_{i=1}^{n} x_i^2 \leq P, \]

for any codeword \( \mathbf{x} = (x_1, x_2, \cdots, x_n) \) transmitted through the channel.

**Definition 1** The information capacity of the Gaussian channel with power constraint \( P \) is

\[ C = \max_{p(x): E(X^2) \leq P} I(X;Y). \]

We have:

\[ I(X;Y) = h(Y) - h(Y|X) \]
\[ = h(Y) - h(X + Z|X) \]
\[ = h(Y) - h(Z|X) \]
\[ = h(Y) - h(Z), \quad (Z \text{ is independent of } X) \]

Form the previous chapter,

\[ h(Z) = \frac{1}{2} \log(2\pi eN). \]
Also,


Now, given that \( E[Y^2] \leq P + N \), from a previous theorem, we can bound the entropy of \( Y \) by

\[ h(Y) \leq \frac{1}{2} \log(2\pi e(P + N)), \]

where equality is when \( Y \) is Gaussian and the maximum available average power \( P \) is used. Since \( Z \) is Gaussian, \( Y \) is Gaussian when \( X \) is Gaussian. Thus,

\[
I(X; Y) = h(Y) - h(Z) \\
\leq \frac{1}{2} \log(2\pi e(P + N)) - \frac{1}{2} \log(2\pi eN) \\
= \frac{1}{2} \log \left( 1 + \frac{P}{N} \right).
\]

Thus, the information capacity of the additive Gaussian channel is

\[ C = \max_{E[X^2] \leq P} I(X; Y) = \frac{1}{2} \log \left( 1 + \frac{P}{N} \right), \]

attained by a Gaussian \( X \) utilizing all available average power \( P \).

We still need to show that, as for the discrete case, this information capacity is the supremum of all achievable rates for the channel.

**Definition 2** An \((M, n)\) code for the Gaussian channel with power constraint \( P \) consists of:

1. An index set \( \{1, 2, \cdots, M\} \)
2. An encoding function that maps indices into $n$-symbol codewords $x^n(1), x^n(2), \ldots, x^n(M)$. Each codeword satisfies the average power constraint

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

3. A decoding function $g$ that maps the outputs of the channel corresponding to transmitted codewords into message indices.

The rate of the code, as for the discrete case, is

$$R = \frac{\log M}{n}.$$ 

The probability of error is the probability the decoded message index not being the same as the transmitted message index.

**Definition 3** A rate $R$ is said to be achievable for a Gaussian channel with 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 as $n \to \infty$. The capacity of the channel is the supremum of all achievable rates.

8.1 The Channel Coding Theorem

**Theorem 1** The capacity of the Gaussian channel with power constraint $P$ and noise variance $N$ is

$$C = \frac{1}{2} \log \left(1 + \frac{P}{N}\right).$$
Proof: (Achievability) We must show that all rates $R$ below capacity are achievable, i.e. there is a sequence of $(2^{nR}, n)$ codes for which the maximal probability of error can be made arbitrarily small for all $R$ below capacity. As for the discrete case we have the following encoding decoding process:

1. We generate a codebook with i.i.d Gaussian codewords that satisfy the power constraint. For this, the variance of each component of the codeword is chosen as $(P - \epsilon)$. For large $n$,

$$\frac{1}{n} \sum_{i=1}^{n} X_i^2 \rightarrow P - \epsilon$$

and with high probability the power constraint is satisfied. Thus, we have \{ $X_i(w), i = 1, 2, \cdots, n, w = 1, 2, \cdots, 2^{nR}$\} i.i.d. zero-mean Gaussian random variables with variance $P - \epsilon$.

2. Given the codebook, it is made known to both the transmitter and receiver. Given a message index $w$, the encoder produces the corresponding codeword $X^n(w)$.

3. Given the received data $Y^n$, the decoder looks for the codeword that is jointly typical with it. If there is exactly one such codeword, the decoder produces the message corresponding to it as its decision. Otherwise an error is declared. The receiver also declares an error if the chosen codeword does not satisfy the power constraint.

4. Because of the symmetry of the (random) encoding process, we can assume without loss of generality that codeword 1 is transmitted.
Thus:

\[ Y^n = X^n(1) + Z^n. \]

Let

\[ E_0 = \left\{ \frac{1}{n} \sum_{i=1}^{n} X_i^2(1) > P \right\}, \]

and

\[ E_i = \{(X^n(i), Y^n) \in A_i^{(n)} \}, \quad i = 1, 2, \ldots, 2^{nR}. \]

Then, an error occurs if the power constraint is violated, i.e. \( E_0 \) occurs, or when \( X^n(1) \) is not jointly typical with the data (i.e. \( E_1^c \) occurs), or \( E_2 \cup E_3 \cdots \cup E_{2^{nR}} \) occurs.

Let \( \mathcal{E} \) denote the event \( \hat{W} \neq W \) and \( P(\mathcal{E}) = \Pr(\mathcal{E}|W = 1) \). Then

\[
P(\mathcal{E}) = P(E_0 \cup E_1^c \cup E_2, \cdots \cup E_{2^{nR}}) \leq P(E_0) + P(E_1^c) + \sum_{i=2}^{2^{nR}} P(E_i), \quad \text{(union bound)}.
\]

By the law of large numbers, \( P(E_0) \to 0 \) as \( n \to \infty \). Thus, for large enough \( n \) and a given \( \epsilon \), we have \( P(E_0) \leq \epsilon \). Also, by the joint AEP (as for the discrete case),

\[ P(E_1^c) \leq \epsilon, \quad \text{for sufficiently large } n. \]

Since the code is randomly generated, \( X^n(1) \) is independent of \( X^n(i) \) for \( i = 2, 3, \ldots, 2^{nR} \). Thus, \( Y^n \) is independent of \( X^n(i), \quad i \neq 1 \). Thus, the probability \( Y^n \) and \( X^n(i) \) will be jointly typical is upperbounded by
$$2^{-n(I(X;Y)-3\epsilon)}$$ (see proof for discrete case). Then,

$$P_e^{(n)} = P(E)$$

$$\leq P(E_0) + P(E_1^c) + \sum_{i=2}^{2^{nR}} P(E_i)$$

$$\leq \epsilon + \epsilon + \sum_{i=2}^{2^{nR}} 2^{-n(I(X;Y)-3\epsilon)}$$

$$= 2\epsilon + (2^{nR} - 1)2^{-n(I(X;Y)-3\epsilon)}$$

$$\leq 2\epsilon + 2^{3\epsilon n} 2^{-n(I(X;Y)-R)}$$

$$\leq 3\epsilon$$

where the last inequality is for $n$ sufficiently large and $R < I(X;Y) - 3\epsilon$.

Now, as for the discrete case, choosing a good code and deleting half of the codewords with the worst conditional error probability, we obtain a code with low maximal error probability and negligible rate loss. Also, the power constraint should be satisfied since any codewords that violate it must belong to the worst half.

Proof of Converse: We must show that if $P_e^{(n)} \to 0$ for a sequence of $(2^{nR}, n)$ codes for the Gaussian channel with power constraint $P$, it must be

$$R \leq C = \frac{1}{2} \log \left(1 + \frac{P}{N}\right).$$

Consider a $(2^{nR}, n)$ code that satisfies the power constraint, i.e.,

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

We have (Fano’s inequality)

$$H(W|Y^n) \leq H(P_e^{(n)}) + \log(2^{nR} - 1)P_e^{(n)} \leq 1 + nRP_e^{(n)} = n\epsilon_n,$$
where $\epsilon_n \to 0$ as $n \to \infty$ ($P_e^{(n)} \to 0$.)

We have:

\[
R = H(W) = I(W; Y^n) + H(W|Y^n)
\]

\[
\leq I(W; Y^n) + n\epsilon_n \quad \text{(Fano’s inequality)}
\]

\[
\leq I(X^n; Y^n) + n\epsilon_n \quad \text{(data processing theorem)}
\]

\[
= h(Y^n) - h(Y^n|X^n) + n\epsilon_n
\]

\[
= h(Y^n) - h(Z^n) + n\epsilon_n
\]

\[
\leq \sum_{i=1}^{n} h(Y_i) - h(Z^n) + n\epsilon_n
\]

\[
= \sum_{i=1}^{n} h(Y_i) - \sum_{i=1}^{n} h(Z_i) + n\epsilon_n
\]

Now let $P_i$ be the average power of the $i$-th element of a codeword, i.e.

\[
P_i = \frac{1}{2^{nR}} \sum_{w=1}^{2^n} x_i^2(w).
\]

Since $Y_i = X_i + Z_i$ and $X_i$ is independent of $Z_i$, we have

\[
E[Y_i^2] = E[X_i^2] + E[Z_i^2] = P_i + N.
\]

Thus, the entropy of $Y_i$ is bounded by

\[
h(Y_i) \leq \frac{1}{2} \log(2\pi e(P_i + N)).
\]
Thus,

\[
\begin{align*}
    nR & \leq \sum_{i=1}^{n} h(Y_i) - \sum_{i=1}^{n} h(Z_i) + n\epsilon_n \\
    & \leq \sum_{i=1}^{n} \left( \frac{1}{2} \log(2\pi e (P_i + N)) - \frac{1}{2} \log(2\pi e N) \right) + n\epsilon_n \\
    & = \sum_{i=1}^{n} \frac{1}{2} \log \left( 1 + \frac{P_i}{N} \right) + n\epsilon_n.
\end{align*}
\]

Now, since each codeword satisfies the power constraint, we have:

\[
\frac{1}{n} \sum_{i=1}^{n} x_i^2(w) \leq P.
\]

Summing on both sides of the inequality above with respect to \( w \) and dividing by \( 2^{nR} \) we obtain

\[
\begin{align*}
    \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \left[ \frac{1}{n} \sum_{i=1}^{n} x_i^2(w) \right] &= \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} x_i^2(w) \right] \\
    &= \frac{1}{n} \sum_{i=1}^{n} P_i \leq \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} P \\
    &= P.
\end{align*}
\]

Now, since \( \log(1 + x) \) is a concave function of \( x \), Jensen’s inequality implies

\[
\begin{align*}
    \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \log \left( 1 + \frac{P_i}{N} \right) & \leq \frac{1}{2} \log \left( 1 + \frac{1}{n} \sum_{i=1}^{n} \frac{P_i}{N} \right) \\
    & \leq \frac{1}{2} \log \left( 1 + \frac{P}{N} \right).
\end{align*}
\]

Thus,

\[
\begin{align*}
nR & \leq \frac{n}{2} \log \left( 1 + \frac{P}{N} \right) + n\epsilon_n,
\end{align*}
\]
or
\[ R \leq \frac{1}{2} \log \left( 1 + \frac{P}{N} \right) + \epsilon_n. \]

As \( n \to \infty, \epsilon_n \to 0 \) and we have
\[ R \leq \frac{1}{2} \log \left( 1 + \frac{P}{N} \right). \]
\( \square \)

### 8.2 The Capacity of Bandlimited Channels

So far we have looked at discrete channels (not waveform) and the capacity expression above is in bits per channel use. In practice, we are interested in the maximum bit-rate in bits/s that can be communicated through a channel. In order for this to be meaningful, we must impose a bandwidth constraint on the transmitted waveforms. Thus, the channel becomes:

\[ Y(t) = (X(t) + Z(t)) * h(t), \]

where \( Z(t) \) is additive white Gaussian noise and \( h(t) \) is the impulse response of a channel ideally bandlimited to \( W \) Hz. We know from the sampling theorem that any signal bandlimited to \( W \) Hz can be perfectly reconstructed from its samples taken at a rate of \( 2W \) samples per second. Strictly speaking, the sampling theorem holds for strictly bandlimited signals, which cannot be time-limited. However, it can be approximately applied to time-limited (and thus not strictly bandlimited signals) by employing a bandwidth definition that states that most energy of a signal be within \([-W,W]\), but not all. In this case, if one asks the question how many orthogonal signals there are in a time
interval $T$ which are bandlimited to $WHz$, the so-called dimensionality theorem of Slepian, Landau and Pollak states that there are approximately $2WT$ such signals. Thus, for a $WHz$ bandlimited signal there are approximately $2WT$ independent samples in a $T$ second interval.

If the noise spectral density is $N_0/2$, then the noise power is $\frac{N_0}{2}2W = N_0W$. Thus, each of the $2WT$ samples has variance $N_0WT/2WT = N_0/2$. Also, the noise samples are i.i.d. Gaussian.

The signal energy per sample is then $PT/2WT = P/2W$. Thus,

$$C = \frac{1}{2} \log \left( 1 + \frac{P}{\frac{2W}{N_0}} \right) = \frac{1}{2} \log \left( 1 + \frac{P}{N_0W} \right) \text{ bits per sample.}$$

Since $2W$ samples per second suffice, the capacity in bits per second is

$$C = W \log \left( 1 + \frac{P}{N_0W} \right) \text{ bits per second.}$$

For a fixed signal power $P/N_0$, the capacity approaches

$$C \to \frac{P}{N_0} \log(e) \text{ bits/s}$$

as $W \to \infty$.

For a fixed signal-to-noise ratio (SNR) per symbol defined as

$$\text{SNR} = \frac{P}{N_0W}$$

the capacity increases linearly with bandwidth:

$$C = W \log (1 + \text{SNR}).$$
8.3 Maximum Achievable Spectral Efficiency

In terms of the spectral efficiency, defined as the bit rate per unit bandwidth, the maximum spectral efficiency is

\[ \frac{C}{W} = \log \left( 1 + \frac{P}{N_0 W} \right) \text{ bits/s/Hz.} \]

If we are communicating at the maximum bit-rate \( C \), then the energy per bit is

\[ E_b = \frac{P}{C} \text{ J/bit.} \]

Thus, the capacity in bits/s/Hz becomes

\[ \frac{C}{W} = \log \left( 1 + \frac{E_b C}{N_0 W} \right) \text{ bits/s/Hz.} \]

Thus,

\[ \frac{E_b}{N_0} = \frac{2^{C/W} - 1}{C/W}, \]

where \( E_b/N_0 \) is the SNR per information bit. Note that as \( C/W \to 0 \),

\[ \frac{E_b}{N_0} \to \ln(2) = -1.6 \text{dB}. \]

Thus, it is possible theoretically to achieve practically zero error probability with an SNR per information bit as low as -1.6dB at very low spectral efficiencies.

In general, a plot of maximum achievable spectral efficiency as a function of SNR per bit is presented in Figure 1.
8.4 Input-Constrained Capacity

We have shown that the capacity of the continuous input continuous output, average power constrained additive white Gaussian channel is achieved when the inputs are i.i.d Gaussian. In practice, the transmitted symbols come from a discrete set (constellation), and thus cannot possibly be Gaussian distributed. In this case we have a discrete input, continuous output channel:

\[ Y = X + Z \]

where the \( X \) now come from some discrete set \( X \in \{x_1, x_2, \cdots, x_M\} \) with probabilities \( P(x_i) \) and \( Z \) is zero-mean Gaussian noise of variance
N. The capacity of this channel is

\[ C = \max_{\{P(x_i)\}} I(X; Y), \]

where

\[
I(X; Y) = \sum_{i=1}^{M} I(Y; X = x_i) P(x_i)
\]
\[
= \sum_{i=1}^{M} \int_{-\infty}^{\infty} p(y|x_i) P(x_i) \ln \frac{p(y|x_i)}{p(y)} dy
\]
\[
= h(Y) - h(Y|X)
\]
\[
= h(Y) - h(Z)
\]
\[
= h(Y) - \frac{1}{2} \ln(2\pi eN). \tag{1}
\]

Now,

\[
h(Y) = -\int_{-\infty}^{\infty} p(y) \ln p(y) dy
\]
where

\[
p(y) = \sum_{i=1}^{M} p(y|x_i) P(x_i).
\]

For the binary antipodal signaling case, \( X \in \{\sqrt{E}, -\sqrt{E}\} \). In this case, it can be shown (symmetry) that a uniform prior distribution achieves capacity. Then,

\[
p(y) = \frac{1}{2} p(y|X = \sqrt{E}) + \frac{1}{2} p(y|X = -\sqrt{E})
\]
\[
= \frac{1}{2\sqrt{2\pi N}} \left[ e^{-\frac{(y-\sqrt{E})^2}{2N}} + e^{-\frac{(y+\sqrt{E})^2}{2N}} \right] \tag{2}
\]
\[
= \frac{1}{\sqrt{2\pi N}} e^{-\frac{(y^2+E)}{2N}} \cosh \left( \frac{y\sqrt{E}}{N} \right) \tag{3}
\]
\[ h(Y) = -E \left[ \ln p(y) \right] \]
\[ = -E \left[ \ln \cosh \left( \frac{y\sqrt{E}}{N} \right) - \frac{y^2 + E}{2N} - \frac{1}{2} \ln(2\pi N) \right] \]
\[ = -E \left[ \ln \cosh \left( \frac{y\sqrt{E}}{N} \right) \right] + \frac{N + 2E}{2N} + \frac{1}{2} \ln(2\pi N) \]
\[ = -E \left[ \ln \cosh \left( \frac{y\sqrt{E}}{N} \right) \right] + \frac{E}{N} + \frac{1}{2} \ln(2\pi eN) \] (4)

Now,
\[ E \left[ \ln \cosh \left( \frac{y\sqrt{E}}{N} \right) \right] = \frac{1}{2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi e N}} e^{-\frac{(y-\sqrt{E})^2}{2N}} \ln \cosh \left( \frac{y\sqrt{E}}{N} \right) dy \]
\[ + \frac{1}{2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi e N}} e^{-\frac{(y+\sqrt{E})^2}{2N}} \ln \cosh \left( \frac{y\sqrt{E}}{N} \right) dy \]
\[ = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi e N}} e^{-\frac{(y-\sqrt{E})^2}{2N}} \ln \cosh \left( \frac{y\sqrt{E}}{N} \right) dy \]
\[ = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi e N}} e^{-\frac{y^2}{2}} \ln \cosh \left( \frac{y\sqrt{E}}{N} + \frac{E}{N} \right) dy \] (5)

Combining (1), (4) and (5), we have finally the capacity in nats per channel use:

\[ C = \frac{E}{N} - \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \ln \cosh \left( \frac{y\sqrt{E}}{N} + \frac{E}{N} \right) dy \] (6)
\[ = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi e N}} \left[ \frac{E}{N} - \ln \cosh \left( \frac{y\sqrt{E}}{N} + \frac{E}{N} \right) \right] dy \] (7)
\[ = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi e N}} \ln \left[ \frac{e^{\frac{E}{N}}}{\cosh \left( \frac{y\sqrt{E}}{N} + \frac{E}{N} \right)} \right] dy \] (8)
Figure 2: The capacity of binary antipodal signaling.

Figure 2 plots the unconstrained capacity and the capacity of binary antipodal signaling as a function of SNR per symbol. As can be seen, for low SNR, the two capacities are close.

8.5 Parallel Gaussian Channels

Consider now $k$ independent Gaussian channels tied together by a common power constraint. In other words, the available power is to be distributed among the $k$ channels so as to maximize the maximum achievable rate. We have

$$Y_j = X_i + Z_j, \quad j = 1, 2, \cdots, k,$$
where as before $Z_j$ is zero-mean Gaussian with variance $N_j$. The noise is independent from one channel to another. The power constraint is

$$\sum_{j=1}^{k} E[X_j^2] \leq P.$$  

The information capacity of the channel is

$$C = \max_{f(x_1, x_2, \cdots, x_k): \sum E(X_j^2) \leq P} I(X^k; Y^k).$$

It can be shown that the information capacity defined above is in fact the supremum of all achievable rates for the parallel Gaussian channel using derivations similar to the single channel case.

We have

$$I(X^k; Y^k) = h(Y^k) - h(Y^k | X^k)$$

$$= h(Y^k) - h(Z^k)$$

$$= h(Y^k) - \sum_{i=1}^{k} h(Z_i)$$

$$\leq \sum_{i=1}^{k} [h(Y_i) - h(Z_i)]$$

$$\leq \sum_{i=1}^{k} \frac{1}{2} \log \left( 1 + \frac{P_i}{N_i} \right),$$

where

$$P_i = E[X_i^2],$$

and to satisfy the power constraint we must have

$$\sum_{i=1}^{k} P_i = P.$$
Note that we have equality in the power constraint as this will further upper bound the mutual information. Equality in the first inequality above is when the $Y_i$ are independent and in the last inequality when all available power is used and the $Y_i$ are Gaussian. This happens when the $X_i$ are independent Gaussian with variance $P_i$.

We now have a standard optimization problem to allocate the power over the $k$ channels so as to maximize the bound to the mutual information. Using Lagrange multipliers, we have

$$J(P_1, P_2, \cdots, P_k) = \frac{1}{2} \sum_{i=1}^{k} \log \left( 1 + \frac{P_i}{N_i} \right) + \lambda \left( \sum_{i=1}^{k} P_i \right).$$

Taking derivatives w.r.t. $P_i$ we have

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

from which

$$P_i = \nu - N_i,$$

for some constant $\nu$ chosen to satisfy the power constraint. Clearly, the $P_i$ cannot be negative. It can be verified through the Kuhn-Tucker conditions that the solution, taking into account the non-negativity of the $P_i$, is to set

$$P_i = (\nu - N_i)^+ = \begin{cases} (\nu - N_i), & (\nu - N_i) \geq 0; \\ 0, & \text{otherwise}. \end{cases}$$

The value of $\nu$ is chosen to satisfy the power constraint

$$\sum_{i=1}^{k} (\nu - P_i)^+ = P.$$
The process by which we allocate power to the various channels is referred to as “water-filling”, as illustrated in Figure 3.

8.6 The Correlated Gaussian Noise Channel

Let the covariance matrix of the noise be $K_Z$ and that for the signal $K_X$. It is still assumed that $X$ and $Z$ are independent. Consider the power constraint

$$\frac{1}{n} \sum_{i=1}^{n} E[X_i^2] \leq P.$$

Equivalently, we can express the power constraint as

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

Note that in contrast to the power constraint in the parallel channels case, the power constraint now is a function of time $n$. This is nec-
essential in order to account for the temporal correlation of the additive noise process. Thus, the capacity must now be computed based on transmitted sequences.

We have

\[ I(X^n; Y^n) = h(Y^n) - h(Z^n), \]

as previously. Thus, the mutual information is maximized by maximizing \( h(Y^n) \), which is achieved when \( Y^n \) is Gaussian. This in turn is achieved when the input is Gaussian. Since \( X \) and \( Z \) are independent, we have

\[ K_Y = K_X + K_Z. \]

Thus,

\[ h(Y^n) = \frac{1}{2} \log((2\pi e)^n|K_X + K_Z|), \]

for Gaussian \( X \). The problem now is to choose \( K_X \) to maximize the above entropy while satisfying the power constraint. For this, we express the noise covariance matrix as

\[ K_Z = Q\Lambda Q^t, \]

where \( Q \) is unitary, i.e., \( QQ^t = I \), and \( \Lambda \) is diagonal with elements the eigenvalues of \( K_Z \). Then

\[ |K_X + K_Z| = |K_X + QAQ^t| \]

\[ = |Q (Q^tK_XQ + \Lambda) Q^t| \]

\[ = |Q||Q^tK_XQ + \Lambda||Q^t| \]

\[ = |Q^tK_XQ + \Lambda| \]

\[ = |A + \Lambda|, \]
where $A = Q^t K_X Q$. Since $A$ and $K_X$ are unitarily similar, we have

$$\text{tr}(A) = \text{tr}(K_X),$$

which can also be seen from $\text{tr}(BC) = \text{tr}(CB)$, i.e.

$$\text{tr}(A) = \text{tr}(Q^t K_X Q) = \text{tr}(K_X Q Q^t) = \text{tr}(K_X).$$

Thus, we now need to maximize $|A + \Lambda|$ subject to the power constraint $\text{tr}(A) \leq nP$.

From Hadamard’s inequality, we have

$$|A + \Lambda| \leq \prod_{i=1}^{n} (A_{ii} + \lambda_i),$$

with equality iff $A$ is diagonal. Thus, we have the following constrained maximization problem:

$$\max \prod_{i=1}^{n} (A_{ii} + \lambda_i),$$

subject to the constraint

$$\frac{1}{n} \sum_{i=1}^{n} A_{ii} \leq P.$$

Taking derivatives we obtain that the optimal $A_{ii}$ must satisfy

$$A_{ii} + \lambda_i = \nu$$

for some constant $\nu$. To satisfy the fact that the $A_{ii}$ must be non-negative, the Kuhn-Tucker conditions show that the optimum solution
is

\[ A_{ii} = (\nu - \lambda_i)^+ \quad (9) \]

where \( \nu \) is chosen again to satisfy the power constraint. Note that the corresponding capacity expression becomes

\[
C = \frac{1}{2} \log \left( \frac{|K_X + K_Z|}{|K_Z|} \right) \\
= \frac{1}{2} \log \left( \frac{\prod_i (A_{ii} + \lambda_i)}{\prod_i \lambda_i} \right) \\
= \frac{1}{2} \sum_i \log \left( 1 + \frac{A_{ii}}{\lambda_i} \right),
\]

where again the \( A_{ii} \) are chosen according to (9).

In the limit as \( n \to \infty \), it turns out that the density of the noise eigenvalues tends to the spectral density of the noise. Thus,

\[
C = \int \frac{1}{2} \log \left( 1 + \frac{\nu - N(f)^+}{N(f)} \right) df,
\]

where \( \nu \) is chosen to satisfy

\[
\int (\nu - N(F))^+ df = P.
\]