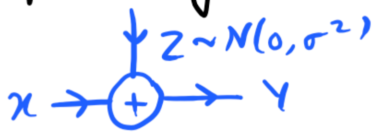
Unit 14: Channel Coding-3

(Gaussian Channels)

A Discrete-time Gaussian channel

Towards analysing a real communication system, we consider now a discrete-time channel with continuous alphabet. Specifically, consider a channel $W: \mathbb{R} \to \mathbb{R}$ which for input x yields an output $Y \sim N(x, \sigma^2)$.



Additive Gaussian noise channel

We use this memoryless channel n times and send a codeword $\underline{x} = (x_1, \dots, x_n)$ over it to receive $Y^n \sim N(\underline{x}, \sigma^2 I)$.

It can be seen that, even with $n=1$, we can send $m\Delta$, $m \in \{-M, \dots, M\}$, to get the prob. of error less than $\approx e^{-\Delta^2}$. Thus, by choosing Δ arbitrary large we can get a vanishing error for every M giving infinite capacity. However, this scheme requires infinite "power" since the transmitted codewords are unbounded.

Closer to practise, we impose power constraints for codewords: $\frac{1}{n}\sum_{i=1}^{n} x_i^2 \le P$.

We define the capacity of an Additive Gaussian noise channel as the max. (sup.) over achievable rates of codes with vanishing error and such that each codeword satisfying power constraint P as before, and denote it by $C(P, \sigma^2)$.

The following result is among the most famous in information theory.

Theorem (Additive Gaussian Noise Channel)

$$C(P, \sigma^2) = \frac{1}{2} \log \left(1 + \frac{P}{\sigma^2}\right)$$

To prove this result, we need to learn some more maths!

[B] Mutual Information and Differential Entropy

* The first thing we need is to extend the definition of KL-divergence and mutual information to general prob. measures (beyond just the discrete measures)

  The main idea is that the ratio of pmfs has a very general counterpart.

  Recall that for pmfs P and Q such that $\mathrm{supp}(P) \subseteq \mathrm{supp}(Q)$, for every set A we have

$$(\#)\left\{\begin{aligned} P(A) &= \sum_{x\in A} P(x) = \sum_{x\in A} \underbrace{\frac{P(x)}{Q(x)}}_{g(x)} \cdot Q(x) \\ &= \sum_{x\in A} g(x)\, Q(x). \end{aligned}\right.$$

We used such expressions several times; in particular, we used it to control the cardinality of sets $A = \{x : g(x) \geq 2^{-\lambda}\}$. In fact, the counterpart of the function $g(x)$ exists for very general prob. measures.

Radon-Nikodym Theorem

Consider two probability measures P and Q such that $Q(A)=0 \Rightarrow P(A)=0$. We say P is absolutely continuous w.r.t. Q, denoted $P \ll Q$. (e.g. for discrete P and Q, $P \ll Q$ iff $\mathrm{supp}(P) \subseteq \mathrm{supp}(Q)$)

If $P \ll Q$, there exists a "random variable" $g(X)$ such that for every f

$$(\#\#)\left\{ \mathbb{E}_P[f(X)] = \mathbb{E}_Q[f(X)\, g(X)]. \right.$$

→ Equation (#) and (##) are similar. For $f(x) = \mathbb{1}_{\{x\in A\}}$ (##) gives $P(A) = \int_A g(x)\, dQ(x) = \mathbb{E}_Q[f(X) g(X)]$.

→ The quantity $g(X)$ is called the Radon-Nikodym derivative and is denoted by $\frac{dP}{dQ}$.

For discrete P, Q, $\frac{dP}{dQ}(x) = \frac{P(x)}{Q(x)}$.

For P, Q with densities $f_P(x)$ and $f_Q(x)$,

$$\frac{dP}{dQ}(x) = \frac{f_P(x)}{f_Q(x)}.$$

→ Why this notation? Some voodoo maths:

$$\mathbb{E}_P[f(X)] = \int_x f(x)\, dP(x) = \int_x f(x) \frac{dP}{\cancel{dQ}}(x)\, \cancel{dQ}(x)$$

$$= \int_x f(x)\, g(x)\, dQ(x).$$

* Radon-Nikodym derivative allows us to consider log-likelihood ratios for arbitrary measures.

→ For $P \ll Q$,

$$D(P\|Q) = \mathbb{E}_Q\left[\frac{dP}{dQ}(X) \log \frac{dP}{dQ}(X)\right]$$

e.g. P, Q discrete:

$$D(P\|Q) = \sum_x \cancel{Q(x)} \cdot \frac{P(x)}{\cancel{Q(x)}} \log \frac{P(x)}{Q(x)}$$

$$= \sum_x P(x) \log \frac{P(x)}{Q(x)}.$$

P, Q with densities:

$$D(P\|Q) = \int_x \cancel{f_Q(x)} \cdot \frac{f_P(x)}{\cancel{f_Q(x)}} \log \frac{f_P(x)}{f_Q(x)}\, dx$$

$$= \int f_P(x) \log \frac{f_P(x)}{f_Q(x)} \, dx .$$

→ For $P_{XY} \ll P_X P_Y$,

$$I(X \wedge Y) =: D(P_{XY} \| P_X P_Y).$$

A reader can go back to the course notes and verify that many of our results for hypothesis testing hold with $\frac{P(x)}{Q(x)}$ replaced with $\frac{dP}{dQ}(x)$.

More importantly for this unit, the single-shot achievability of Unit 13D applies to general channels and input distributions.

* Differential Entropy

Looking for the counterpart of Shannon entropy in $D(P\|Q)$ above, note that for P and Q with densities

$$D(P\|Q) = \int_x f_P(x) \log \frac{1}{f_Q(x)} \, dx \; \underbrace{- \int_x f_P(x) \log \frac{1}{f_P(x)} \, dx}_{\text{"differential entropy of } P \text{"}}$$

denoted $h(P)$.

Note that $h(P)$ is only defined prob. with densities.
Its operational definition is a bit complicated and requires the notion of "dimension of a large prob. set."
But roughly $h(P)$ is the log of volume of a large prob. set (measured in appropriate dimension).

* Some basic properties

(1) Data-processing inequality :

$$D(PW \| QW) \le D(P \| Q) \qquad \text{(Jensen's inequality)}$$

For $U - X - Y - V$ forming a Markov chain,

$$I(U \wedge V) \le I(X \wedge Y).$$

(2) Mutual Information and Differential Entropy

Consider a channel $W : \mathcal{X} \to \mathcal{Y}$ such that for every $x \in \mathcal{X}$ the output distribution has density. Further, let P_X be a discrete distribution.

(This is the situation when a random message is encoded and sent over a Gaussian channel).

$$I(X \wedge Y) = \underbrace{h(P_Y)}_{h(Y)} - \underbrace{\mathbb{E}[h(W_X)]}_{h(Y|X)}$$

density given x

Note that Y has density $\sum_x P_X(x)\, g_W(y|x)$.

(3) Gaussian Maximizes Differential Entropy

Consider a random variable $X^n \in \mathbb{R}^n$ satisfying

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i^2\right] \le nP.$$

Further, assume that X^n has a density. Then,

$$h(X^n) \le \underbrace{\frac{n}{2}\log 2\pi e P}_{\longrightarrow\, h(\mathcal{N}(0, PI))}.$$

Proof. Let $P_{X^n} \equiv P$ and $Q \equiv \mathcal{N}(0, PI)$. Then,

$$0 \le D(P\|Q) = \mathbb{E}_P\left[\log \frac{1}{\frac{1}{(2\pi P)^{n/2}} \exp\left[-\frac{1}{2P}\sum_{i=1}^{n} X_i^2\right]}\right] - h(P)$$

$$= \frac{n}{2}\log 2\pi P + \frac{1}{2P}\sum_{i=1}^{n} \mathbb{E}_P[X_i^2] \log e$$

$$\le \frac{n}{2}\log 2\pi e P. \qquad \blacksquare$$

C Proof of converse

Consider a code for the Additive Gaussian Noise channel of rate R and prob. of error less than ε, and such that each codeword $\underline{x}$ satisfies:

$$\sum_{i=1}^{n} x_i^2 \le nP.$$

Then, for $U \sim \text{unif}\{1, \dots, M\}$ and $\hat{U}$ the decoded message, we get

$$
\begin{aligned}
nR = H(U) &= I(U \wedge \hat{U}) + H(U|\hat{U}) \\
&\leq I(U \wedge \hat{U}) + \varepsilon nR + 1 \qquad \text{(Fano's inequality)} \\
&\leq I(X^n \wedge Y^n) + \varepsilon nR + 1 \qquad \text{(Data-processing ineq.)} \\
&= h(Y^n) - \underbrace{h(Y^n|X^n)}_{\frac{n}{2}\log 2\pi e \sigma^2} + \varepsilon nR + 1
\end{aligned}
$$

Note that

$$
\mathbb{E}\left[\sum_{i=1}^{n} Y_i^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[(X_i + Z_i)^2\right]
= \mathbb{E}\left[\sum_{i=1}^{n} X_i^2\right] + n\sigma^2 \leq n(P + \sigma^2).
$$

Thus, since Gaussian maximizes entropy,

$$
h(Y^n) \leq \frac{n}{2} \log 2\pi e (P + \sigma^2).
$$

Therefore,

$$
R(1-\varepsilon) \leq \frac{1}{2} \log\left(1 + \frac{P}{\sigma^2}\right) + \frac{1}{n}
$$

which gives $C(P;\sigma^2) \leq \frac{1}{2} \log\left(1 + \frac{P}{\sigma^2}\right)$. ∎

→ An alternative converse proof can be given using a sphere-packing argument similar to the one we saw for BSC.

D Proof of achievability

We already mentioned that our single-shot achievability holds. The only modification we need is to generate codewords using $P_X \equiv P_{X^n}$ such that the codewords satisfy power constraints. A first attempt can be to use iid P_{X^n} s.t. each X_i satisfies $\mathbb{E}[X_i^2] \le P-\eta$. Then, with large prob., X^n will satisfy $\mathbb{E}\left[\sum_{i=1}^n X_i^2\right] \le nP$. But we generate M codewords and they may not all satisfy power constraints simultaneously. (A union bound will not give the desired performance).

Instead, we use a slightly different, non-iid P_{X^n}.

Consider $\tilde{X}_1, \ldots, \tilde{X}_n \sim$ iid $N(0, P-\eta)$ and let

$$P_{X^n} = P_{\tilde{X}^n \mid \underbrace{\sum_{i=1}^n \tilde{X}_i^2 \le nP}_{\mathcal{E} \equiv \{\tilde{X}^n \text{ satisfies the power const.}\}}} \equiv \text{Conditional dist. given } \tilde{X}^n \text{ satisfies the power constraint}$$

We make a simple observation. For any $P_{\tilde{X}\tilde{Y}}$, let $P_{Y|X} = P_{\tilde{Y}|\tilde{X}}$ and $P_X = P_{\tilde{X} \mid \tilde{X} \in \mathcal{E}}$. Then,

$$\frac{f_{\tilde{Y}}(y)}{f_Y(y)} = \frac{\int f_{\tilde{Y}|\tilde{X}}(y|x)\, f_{\tilde{X}}(x)\, dx}{\int_{\mathcal{E}} f_{Y|X}(y|x)\, f_X(x)\, dx}$$

$$\ge \frac{\int_{\mathcal{E}} f_{Y|X}(y|x)\, f_{\tilde{X}}(x)\, dx}{\int_{\mathcal{E}} f_{Y|X}(y|x)\, f_X(x)\, dx} = P(\tilde{X} \in \mathcal{E})$$

(since for $x \in \mathcal{E}$ $f_X(x) = \dfrac{f_{\tilde{X}}(x)}{\mathbb{P}(\tilde{X} \in \mathcal{E})}$)

Therefore,

$$\log \frac{f_{Y|X}(y|x)}{f_Y(y)} \geq \log \frac{\overbrace{f_{Y|X}(y|x)}^{= f_{\tilde{Y}|\tilde{X}}(y|x)}}{f_{\tilde{Y}}(y)} - \log \frac{1}{\mathbb{P}(\tilde{X} \in \mathcal{E})},$$

for every $x \in \mathcal{E}$. This further yields,

$$P_{XY}\left(\left\{(x,y) : \log \frac{f_{XY}(x,y)}{f_X f_Y(x,y)} \geq \lambda\right\}\right)$$

$$\geq P_{XY}\left(\left\{(x,y) : \log \frac{f_{\tilde{Y}|\tilde{X}}(y|x)}{f_{\tilde{Y}}(y)} \geq \lambda - \log \frac{1}{P(\tilde{X} \in \mathcal{E})}\right\}\right)$$

$$\geq 1 - \frac{1}{\mu} P_{\tilde{X}\tilde{Y}} \quad : \log \frac{f_{\tilde{X}\tilde{Y}}(x,y)}{f_{\tilde{X}} f_{\tilde{Y}}(x,y)} < \lambda - \log \frac{1}{P(\tilde{X} \in \mathcal{E})}\})$$

(why?)

In particular, for $P_{\tilde{X}}$ s that $\mathbb{P}(\tilde{X} \in \mathcal{E}) \geq \frac{1}{2}$ and λ such that

$$P_{\tilde{X}\tilde{Y}}\left(\left\{(x,y) : \log \frac{f_{\tilde{Y}|\tilde{X}}(y|x)}{f_{\tilde{Y}}(y)} > \lambda - 1\right\}\right) \geq 1 - \varepsilon,$$

we get

$$P_{XY}\left(\left\{(x,y) : \log \frac{f_{XY}(x,y)}{f_X f_Y(x,y)} > \lambda\right\}\right) \geq 1 - 2\varepsilon.$$

Therefore, by our single-shot achievability result, we can find $\lfloor 2\varepsilon 2^{\lambda} \rfloor$ codewords with prob. of error less than 4ε.

The key point is that all the codewords are now, with prob. 1, generated from $\mathcal{E} = \text{supp}(P_X)$.

Returning to our problem, since $P_{\tilde{X}_i} \equiv N(0, P-\eta)$ and $\mathcal{E} = \{ \underline{x} : \sum_{i=1}^{n} x_i^2 \le nP \}$, for all n sufficiently large $\mathbb{P}(\tilde{X}^n \in \mathcal{E}) \ge \frac{1}{2}$. Furthermore, by the law of large numbers, a good choice of λ is

$$\lambda \simeq \mathbb{E}\left[\log \frac{f_{\tilde{Y}^n|\tilde{X}^n}(\tilde{Y}^n|\tilde{X}^n)}{f_{\tilde{Y}^n}(\tilde{Y}^n)} \right]$$

$$= I(\tilde{X}^n \wedge \tilde{Y}^n) \quad = \frac{n}{2} \log\left(1 + \frac{P-\eta}{\sigma^2}\right)$$

Therefore, $\frac{1}{2}\log\left(1 + \frac{P-\eta}{\sigma^2}\right)$ is an achievable rate for every $\eta > 0 \Rightarrow C(P, \sigma^2) \ge \frac{1}{2}\log\left(1 + \frac{P}{\sigma^2}\right)$.

$\boxed{\text{E}}$ Parallel Gaussian channels and water-filling

In communication systems, often we need to encode for multiple Gaussian channels in parallel while maintaining joint power constraints. Specifically, we have K parallel additive Gaussian noise channel with channel i having noise variance σ_i^2, $1 \le i \le K$. Further, we seek codewords

$\underline{x}_1, \ldots, \underline{x}_K$ (choosen to be of the same length for simplicity) such that

$$\sum_{k=1}^{K} \sum_{t=1}^{n} x_{k,t}^2 \leq P.$$

Equivalently, for each k

$$\sum_{t=1}^{n} x_{k,t}^2 \leq P_k,$$

and

$$\sum_{k=1}^{K} P_k \leq P.$$

It is easy to see that the capacity of the parallel channels is given by

$$\max_{\substack{P_1, \ldots, P_K: \\ \sum_k P_k \leq P \\ P_k \geq 0}} \frac{1}{2} \sum_{k=1}^{K} \log\left(1 + \frac{P_i}{\sigma_i^2}\right).$$

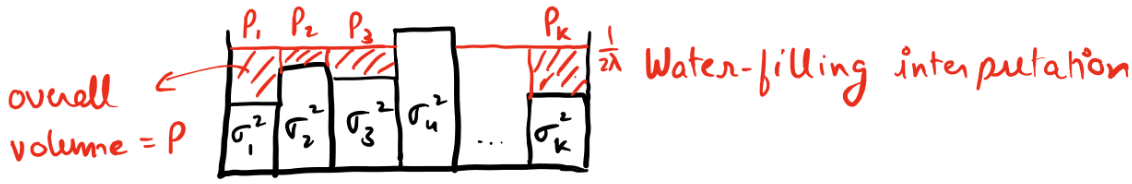
Note that for every $\lambda \geq 0$, the maximum above is less than

$$\max_{\substack{P_1, \ldots, P_K \\ P_k \geq 0}} \underbrace{\frac{1}{2} \sum_{k=1}^{K} \log\left(1 + \frac{P_i}{\sigma_i^2}\right) + \lambda\left(P - \sum_{k=1}^{K} P_k\right)}_{=: f_\lambda(P_1, \ldots, P_K)}$$

$$\frac{d}{dP_k} f_\lambda(P_1, \ldots, P_K) = \frac{1}{2} \frac{1}{\sigma_k^2 + P_k} - \lambda, \quad 1 \leq k \leq K.$$

Thus, f_λ is increasing in P_k till $P_k = \frac{1}{2\lambda} - \sigma_k^2$ and decreasing after that. If $\frac{1}{2\lambda} - \sigma_k^2 < 0$, the function is decreasing in P_k for all $P_k \geq 0$. Thus, the optimal choice of P_k is given by $P_k = \left(\frac{1}{2\lambda} - \sigma_k^2\right)_+$, $1 \leq k \leq K$, where $(x)_+ = \max\{x, 0\}$. Note that the optimal value attained exceeds our original maximum for every $\lambda \geq 0$. In particular, the two will be equal if λ is set so that $P = \sum_{k=1}^{K} \left(\frac{1}{2\lambda} - \sigma_k^2\right)_+$, $1 \leq k \leq K$.

The figure below depicts the optimal P_ks.



The capacity is given by

$$\frac{1}{2} \sum_{k:\ B \geq \sigma_k^2} \log \frac{B}{\sigma_k^2},$$

where $$\sum_{k:\ B \geq \sigma_k^2} \left(B - \sigma_k^2\right) = P.$$