CS3236 Lecture Notes #5:
Continuous-Alphabet Channels

Jonathan Scarlett

March 16, 2020

Useful references:

- Cover/Thomas Chapters 8 and 9
- MacKay Chapter 11

1 Differential Entropy

Introduction.

- So far, we have considered channels with finite input and output alphabets, and accordingly used probability mass functions (PMFs) P_X and conditional PMFs $P_{Y|X}$.
- In this lecture, we will consider *continuous* (real-valued) inputs and outputs, and accordingly consider probability density functions (PDFs) f_X and conditional PDFs $f_{Y|X}$.
- First, we need to revise the main definitions of information measures (entropy, mutual information, KL divergence)

Differential entropy.

- The *differential entropy* of a continuous random variable X with PDF f_X is seemingly natural given the regular version:

$$\begin{aligned} h(X) &= \mathbb{E}_{f_X}\left[\log_2 \frac{1}{f_X(X)}\right] \\ &= \int_{\mathbb{R}} f_X(x) \log_2 \frac{1}{f_X(x)} dx. \end{aligned}$$

  However, in contrast with the discrete version, this should generally *not* be directly viewed as a measure of information/uncertainty (in particular, see the properties that *no longer hold* below)

  - As usual, we can also consider the joint version

$$h(X,Y) = \mathbb{E}_{(X,Y)\sim f_{XY}}\left[\log_2 \frac{1}{f_{XY}(X,Y)}\right],$$

and the conditional version

$$
\begin{aligned}
h(Y|X) &= \mathbb{E}_{(X,Y)\sim f_{XY}}\left[\log_2 \frac{1}{f_{Y|X}(Y|X)}\right] \\
&= \int_{\mathbb{R}} f_X(x) H(Y|X=x) dx
\end{aligned}
$$

when (X,Y) have a joint density function $f_{XY}(x,y) = f_X(x) f_{Y|X}(y|x)$.

- Properties of entropy that <u>still hold</u> for differential entropy:
  - Chain rule: $h(X_1,\dots,X_n) = \sum_{i=1}^n h(X_i|X_1,\dots,X_{i-1})$
  - Conditioning reduces entropy: $h(X|Y) \le h(X)$
  - Sub-additivity: $h(X_1,\dots,X_n) \le \sum_{i=1}^n h(X_i)$
  - $h(X) = h(X+c)$ for constant c
- Properties that <u>no longer hold</u>:
  - Non-negativity
  - Invariance under one-to-one transformations

  Counter-examples to both of these can be deduced as follows: If $Y = cX$ for some constant c, then a standard formula for the density of a function gives $f_Y(y) = \frac{1}{|c|} f_X\left(\frac{y}{c}\right)$, and substitution into the formula for differential entropy gives $h(Y) = h(X) + \log_2|c|$. As $c \to 0$, we have $\log_2|c| \to -\infty$, meaning $h(Y)$ may be arbitrarily negative.

Examples.

- **Claim.** For a uniform random variable $X \sim \text{Uniform}(a,b)$ with $a < b$, we have

$$
h(X) = \log_2(b-a).
$$

  - <u>Proof</u>: By definition $f_X(x) = \frac{1}{b-a}$ for $a < x < b$, and $f_X(x) = 0$ elsewhere. Substitute this into the expression for $h(X)$.

- **Claim.** For a univariate Gaussian $X \sim N(\mu, \sigma^2)$, we have

$$
h(X) = \frac{1}{2}\log_2\left(2\pi e \sigma^2\right).
$$

  - <u>Proof</u>: We give the proof for the case $\mu = 0$; the general case is very similar. The PDF of X is given by $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/(2\sigma^2)}$, and hence

$$
\begin{aligned}
h(X) &= \mathbb{E}\left[\log_2 \frac{1}{f_X(x)}\right] \\
&= \mathbb{E}\left[\log_2\left(\sqrt{2\pi\sigma^2}\cdot e^{X^2/(2\sigma^2)}\right)\right] \\
&= \frac{1}{2}\log_2(2\pi\sigma^2) + \frac{\log_2 e}{2\sigma^2}\mathbb{E}[X^2],
\end{aligned}
$$

where we have used $\log_2(ab) = \log_2(a) + \log_2(b)$ and $\log_2(a^c) = c\log_2 a$. But by definition $\mathbb{E}[X^2] = \sigma^2$, and so we get

$$h(X) = \frac{1}{2}\log_2(2\pi\sigma^2) + \frac{\log_2 e}{2} = \frac{1}{2}\log_2\left(2\pi e\sigma^2\right).$$

Mutual information and KL divergence.

- The definitions of KL divergence and mutual information also extend naturally:

$$D(f\|g) = \int_{\mathbb{R}} f(x)\log_2\frac{f(x)}{g(x)}dx$$

  and

$$\begin{aligned} I(X;Y) &= D(f_{XY}\|f_X \times f_Y) \\ &= \mathbb{E}_{f_{XY}}\left[\log_2\frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}\right] \\ &= h(Y) - h(Y|X) \\ &= h(X) - h(X|Y). \end{aligned}$$

- In contrast with differential entropy, it is uncontroversial to consider $I(X;Y)$ as a measure of how much information Y reveals about X (or vice versa). Indeed, both mutual information and KL divergence retain all of their key properties, including non-negativity.
  - It can also be shown that $I(X;Y) = I(\phi(X);\psi(Y))$ for *invertible* functions $\phi(\cdot)$ and $\psi(\cdot)$.

2 Gaussian Random Variables

Univariate case.

- As mentioned above, for $X \sim N(\mu,\sigma^2)$, we have $h(X) = \frac{1}{2}\log_2\left(2\pi e\sigma^2\right)$.
- **Maximum entropy property (univariate case).** For any random variable X having a density f_X and variance $\mathrm{Var}[X]$, we have

$$h(X) \le \frac{1}{2}\log_2\left(2\pi e\,\mathrm{Var}[X]\right)$$

  with equality if and only if X is Gaussian.
  - Proof: Let f be the density function of X, and let g be the Gaussian density with the same mean and variance as X. For brevity, denote this mean and variance by μ and σ^2, so that

$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$. Then observe that

$$\begin{aligned}
D(f\|g) &= \mathbb{E}_f\left[\log_2 \frac{f(X)}{g(X)}\right] \\
&\stackrel{(a)}{=} \mathbb{E}_f\left[\log_2 \frac{1}{g(X)}\right] + \mathbb{E}_f\left[\log_2 f(X)\right] \\
&\stackrel{(b)}{=} \mathbb{E}_f\left[\log_2 \frac{1}{g(X)}\right] - h(X) \\
&\stackrel{(c)}{=} \mathbb{E}_f\left[\log_2 \left(\sqrt{2\pi\sigma^2} \cdot e^{(X-\mu)^2/(2\sigma^2)}\right)\right] - h(X) \\
&\stackrel{(d)}{=} \frac{1}{2}\log_2(2\pi\sigma^2) + \frac{\log_2 e}{2\sigma^2}\mathbb{E}_f[(X-\mu)^2] - h(X) \\
&\stackrel{(e)}{=} \frac{1}{2}\log_2(2\pi e\sigma^2) - h(X),
\end{aligned}$$

where (a) and (d) simply expand the logarithms, (b) uses the definition of $h(X)$, (c) substitutes the definition of g, and (e) uses $\mathbb{E}_f[(X-\mu)^2] = \sigma^2$. The maximum entropy property now follows from the fact that $D(f\|g) \ge 0$ with equality if and only if $f = g$.

(Optional) Multivariate case.

- The following are written without proof, mainly for the sake of completeness (we will only make use of the univariate result).
- **Claim.** For a multivariate Gaussian $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we have

$$h(\mathbf{X}) = \frac{1}{2}\log_2 \det\big(2\pi e \boldsymbol{\Sigma}\big).$$

- **Maximum entropy property (multivariate case).** For any random vector $\mathbf{X}$ having a joint density $f_{\mathbf{X}}$ and covariance matrix $\mathrm{Cov}[\mathbf{X}]$, we have

$$h(\mathbf{X}) \le \frac{1}{2}\log_2 \det\big(2\pi e\, \mathrm{Cov}[\mathbf{X}]\big)$$

  with equality if and only if $\mathbf{X}$ is a multivariate Gaussian.

3 Gaussian Channel

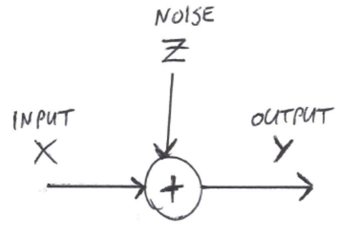
Model.

- In general, a continuous channel can be described by a conditional PDF $f_{Y|X}$. However, we will focus on a more specific class of *additive noise* channels:

$$Y = X + Z,$$

  where Z is a noise term independent of the input X. This means that $f_{Y|X}(y|x) = f_Z(y-x)$.

  - In particular, when $Z \sim N(0, \sigma^2)$ for some noise variance $\sigma^2 > 0$, this is called the *additive white Gaussian noise (AWGN) channel.*



  - Well-motivated in many applications where a large number of tiny disturbances impact the output; these combine to give approximately Gaussian noise (by the central limit theorem).
  - Also very convenient to analyze mathematically!

- If X is unconstrained, then we can transmit arbitrarily many bits arbitrarily reliably in a single channel use: Just send different messages using the inputs $0, \pm\Delta, \pm 2\Delta, \dots$ for a huge value of Δ (e.g., a million times larger than the noise variance).

- However, in practice, the energy consumed by transmitting X is proportional to X^2, and we need to satisfy a *power constraint* of the form
$$\mathbb{E}[X^2] \le P.$$
Sometimes, *peak power constraints* of the form $X^2 \le P_{\max}$ also arise, but we will not consider those.

- The symbol $\mathbb{E}[\cdot]$ above is somewhat ambiguous. If we have a codebook $\mathcal{C} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(M)}\}$ of length-n codewords $\mathbf{x}^{(m)} = (x_1^{(m)}, \dots, x_n^{(m)})$, then we could require that every codeword has power at most P averaged over the block length,
$$\frac{1}{n}\sum_{i=1}^{n} \big(x_i^{(m)}\big)^2 \le P, \qquad \forall m \in \{1, \dots, M\},$$
or we could require a less stringent constraint that averages over both the message and block length:
$$\frac{1}{M}\sum_{m=1}^{M} \frac{1}{n}\sum_{i=1}^{n} \big(x_i^{(m)}\big)^2 \le P.$$
In fact, either requirement leads to the same channel capacity.

Channel capacity.

- In the following, the channel capacity $C(P)$ is defined in the same way as discrete memoryless channels, but with codebooks constrained to satisfy the average power constraint.

- **Theorem.** For general noise models, the channel capacity with power constraint P is given by
$$C(P) = \max_{f_X \,:\, \mathbb{E}_{f_X}[X^2] \le P} I(X;Y).$$
The proof is outlined below.

- **Corollary.** For the AWGN channel with power constraint P and noise variance σ^2, the channel capacity is
$$C(P) = \frac{1}{2}\log_2\left(1 + \frac{P}{\sigma^2}\right),$$

and the capacity-achieving f_X is Gaussian, namely $N(0, P)$.

- Proof: For fixed f_X such that $\mathbb{E}[X^2] \le P$, we expand the mutual information as follows:

$$\begin{aligned}I(X;Y) &\overset{(a)}{=} h(Y) - h(Y|X)\\ &\overset{(b)}{=} h(Y) - h(X+Z|X)\\ &\overset{(c)}{=} h(Y) - h(Z|X)\\ &\overset{(d)}{=} h(Y) - h(Z)\end{aligned}$$

  where (a) is by definition of mutual information, (b) is by $Y = X + Z$, (c) is since shifting by a constant doesn't change entropy (and X is a constant conditioned on X), and (d) holds since X and Z are independent.

  Now, since Z is Gaussian, we have $h(Z) = \frac{1}{2}\log_2(2\pi e\sigma^2)$. Moreover, since $Y = X + Z$ with X and Z being independent, we have

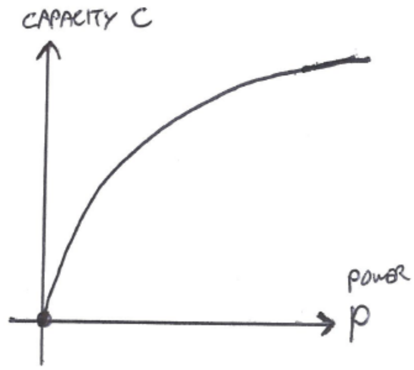
$$\begin{aligned}\mathrm{Var}[Y] &= \mathrm{Var}[X] + \mathrm{Var}[Z]\\ &\le P + \sigma^2,\end{aligned}$$

  where the first term uses $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \le \mathbb{E}[X^2] \le P$, and the second term uses $\mathrm{Var}[Z] = \sigma^2$. By the maximum entropy property of Gaussians, we deduce that $h(Y) \le \frac{1}{2}\log_2\big(2\pi e(P+\sigma^2)\big)$. Substituting this and the expression for $h(Z)$ into $I(X;Y) = h(Y) - h(Z)$, we obtain

$$\begin{aligned}I(X;Y) &\le \frac{1}{2}\log_2\big(2\pi e(P+\sigma^2)\big) - \frac{1}{2}\log_2(2\pi e\sigma^2)\\ &= \frac{1}{2}\log_2\frac{2\pi e(P+\sigma^2)}{2\pi e\sigma^2}\\ &= \frac{1}{2}\log_2\left(1 + \frac{P}{\sigma^2}\right).\end{aligned}$$

  Finally, both the inequalities used ($\mathrm{Var}[Y] \le P + \sigma^2$ and $h(Y) \le \frac{1}{2}\log\big(2\pi e(P+\sigma^2)\big)$) hold with equality when $X \sim N(0, P)$, and so we deduce that the upper bound $I(X;Y) \le \frac{1}{2}\log_2\big(1 + \frac{P}{\sigma^2}\big)$ is achieved with equality by such Gaussian f_X.

- Properties of the Gaussian channel capacity:
  - Depends on P and σ^2 only through the *signal-to-noise ratio* $\frac{P}{\sigma^2}$.
  - Equals zero when $P = 0$.
  - When $\frac{P}{\sigma^2}$ is very small, we have $C(P) \approx \frac{P}{2\sigma^2}$, so doubling P may (nearly) double the capacity.
  - When $\frac{P}{\sigma^2}$ is very large, we have $C(P) \approx \frac{1}{2}\log_2\frac{P}{\sigma^2}$, so doubling P only (roughly) adds a constant to the capacity (diminishing returns).
  - An illustration:



(Optional) Outline of proofs.

- Achievability:
  - Again random coding is used – generate each symbol of each codeword independently according to some f_X such that $\mathbb{E}[X^2] < P$.[1] Under this condition, most (but not all) of the codewords satisfy the power constraint, with high probability.
  - To prove vanishing error probability, we follow similar arguments to the previous lecture with suitable modifications:
    * Extend the joint typicality definition and properties to the continuous setting (a tutorial question makes a start on this);
    * Follow the "joint typicality decoding" analysis from the discrete case to deduce that vanishing average error probability still holds for rates below the mutual information.
  - The desired result is then obtained by a fairly simple expurgation argument in which any codewords violating the power constraint are discarded (there are so few such codewords that this has a negligible effect on the rate and average error probability).
- Converse:
  - An argument based on Fano's inequality can still be used, but a bit of extra effort is required to handle the power constraint $\mathbb{E}[X^2] \le P$. See Chapter 9 of Cover/Thomas for details.

4 (Optional) Geometric Intuition: Sphere Packing

- At least for the converse part, we can get some intuition on the AWGN capacity formula $C = \frac{1}{2}\log_2\left(1+\frac{P}{\sigma^2}\right)$ by considering geometric arguments in the space of all output sequences $\mathbf{y}$.
- To satisfy the power constraint, assume that every codeword $\mathbf{x}^{(m)}$ lies in the sphere of radius $\sqrt{nP}$ centered at zero:

$$\|\mathbf{x}^{(m)}\|^2 \le nP, \quad \forall m = 1, \dots, M.$$

[1] The need for strict inequality here is a minor technical issue.

- Since the noise vector $\mathbf{Z}$ is independent of $\mathbf{x}$, a "Pythagoras-type" argument gives

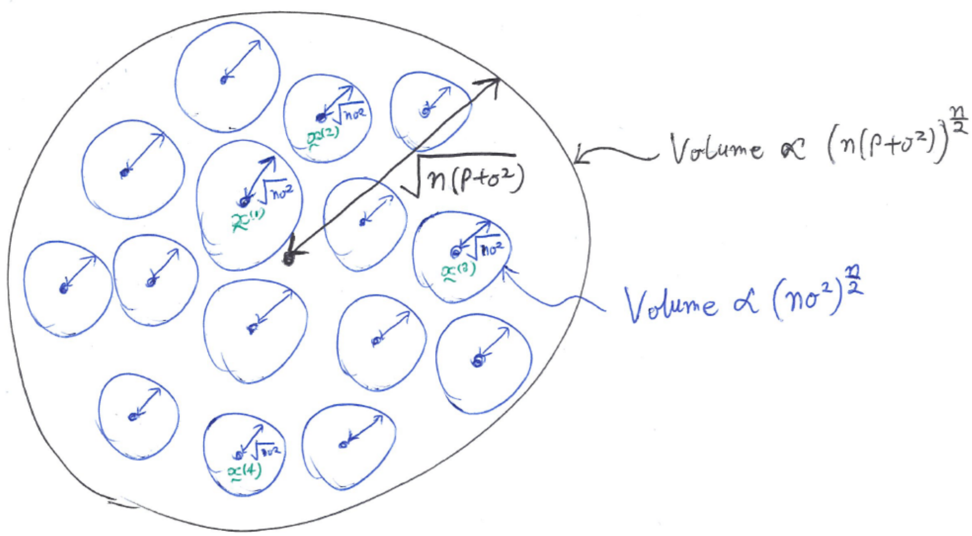
$$\begin{aligned}\|\mathbf{Y}\|^2 &\approx \|\mathbf{x}\|^2 + \|\mathbf{Z}\|^2 \\ &\leq nP + \|\mathbf{Z}\|^2 \\ &\approx n(P+\sigma^2),\end{aligned}$$

  where the last line uses the fact that $\|\mathbf{Z}\|^2 \approx n\sigma^2$ with high probability by the law of large numbers.

  - Hence, $\mathbf{Y}$ typically lies within the sphere of radius $\sqrt{n(P+\sigma^2)}$.

- Now, for a specific transmitted codeword $\mathbf{x}^{(m)}$, using a similar argument to the one just shown, transmitting it will produce an output sequence $\mathbf{Y}$ such that $\|\mathbf{Y}-\mathbf{x}^{(m)}\|^2 \lesssim n\sigma^2$ with high probability. That is, the output will roughly be in a sphere of radius $\sqrt{n\sigma^2}$ centered at the transmitted codeword.

- <u>Intuition</u>: For successful decoding, these "high-probability spheres" of radius $\sqrt{n\sigma^2}$ should be *non-overlapping*. An illustration:



- But there are only so many non-overlapping spheres of radius $\sqrt{n\sigma^2}$ we can fit inside the overall sphere of radius $\sqrt{n(P+\sigma^2)}$! Specifically, since the volume of a sphere of radius r in n dimensions is $\alpha_n \cdot r^n$ for some constant α_n, we have

$$\#\text{spheres} \lesssim \frac{\left(\sqrt{n(P+\sigma^2)}\right)^n}{\left(\sqrt{n\sigma^2}\right)^n} = \left(\frac{P+\sigma^2}{\sigma^2}\right)^{n/2}. \tag{1}$$

- But the number of spheres is simply the number of codewords M; hence, and taking logs in the previous equation, we obtain $\frac{1}{n}\log_2 M \lesssim \frac{1}{2}\log_2\left(1+\frac{P}{\sigma^2}\right)$.