

ECE 587 / STA 563: Lecture 10 – Review and Applications

Information Theory
Duke University, Fall 2023

Author: Galen Reeves

Last Modified: November 16, 2023

Outline of lecture:

10.1 Point-to-Point Communication	1
10.1.1 Recap of Main Theorems	1
10.1.2 Source Channel Separation Theorem	2
10.2 Application to Statistical inference	3
10.2.1 Parameter Estimation	3
10.2.2 Linear Model	4

10.1 Point-to-Point Communication

The fundamental problem of communication is that of reproducing at one point, either exactly or approximately, a message selected at another point

— Claude Shannon

10.1.1 Recap of Main Theorems

- **Lossless Source Coding:** For a discrete iid source with pmf $p(x)$, the expected length $\mathbb{E}[\ell(X)]$ of the optimal uniquely decodable D -ary source code satisfies

$$\frac{H(X)}{\log D} \leq \mathbb{E}[\ell(X)] < \frac{H(X)}{\log D} + 1$$

By coding over blocks of length n , the expected number of code symbols per source symbol of the optimal uniquely decodable D -ary source code satisfies

$$\frac{H(X)}{\log D} \leq \frac{1}{n} \mathbb{E}[\ell(X^n)] < \frac{H(X)}{\log D} + \frac{1}{n}$$

More generally, for a stationary ergodic source X_1, X_2, \dots , the fundamental limit of compression, measured in bits per source symbol, is given by the entropy rate $H(\mathcal{X})$, computed using the base 2 logarithm.

- **Channel Coding:** For a discrete memoryless channel $p(y | x)$, there exists a sequence of rate R block-length n coding schemes with error probability tending to zero provided that R is less than the capacity C , which is given by

$$C = \max_{p(x)} I(X; Y).$$

Conversely, if $R > C$ then the probability of error is bounded away from zero.

- **Gaussian Channel:** For an additive white gaussian noise channel with power constraint P and noise variance N , the capacity is given by

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

- **Lossy Source Coding:** For a discrete iid source with pmf $p(x)$ and bounded distortion measure $d(x, \hat{x})$ there exists a sequence of rate R block-length n coding scheme with distortion satisfying

$$\limsup_{n \rightarrow \infty} \mathbb{E} \left[d(X^n, \hat{X}^n) \right] \leq D$$

provided that R is greater than the rate-distortion function $R(D)$, which is given by

$$R(D) = \min_{p(\hat{x}|x) : \mathbb{E}[d(X, \hat{X})] \leq D} I(X; \hat{X})$$

- **Gaussian Source:** For an iid Gaussian source $\mathcal{N}(0, \sigma^2)$ and squared-error distortion $d(x, \hat{x}) = (x - \hat{x})^2$, the rate distortion function is given by

$$R(D) = \begin{cases} \frac{1}{2} \log \left(\frac{\sigma^2}{D} \right), & 0 \leq D \leq \sigma^2 \\ 0, & D > \sigma^2 \end{cases}$$

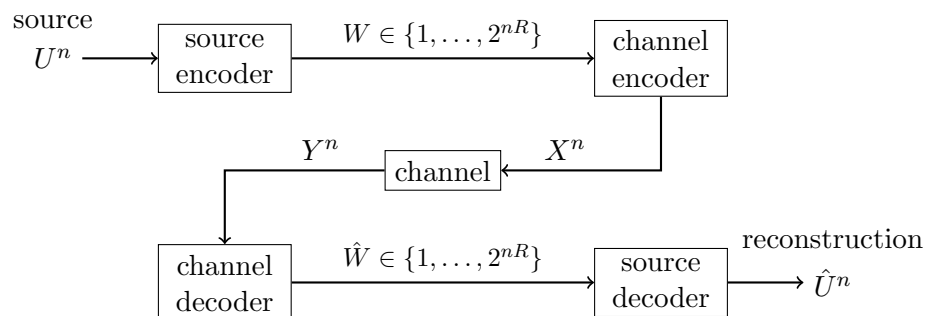
10.1.2 Source Channel Separation Theorem

- Suppose we want to communicate an iid source U_1, U_2, \dots, U_n with n uses of a memoryless channel with capacity C while incurring a distortion no greater than D .
- **Joint source and channel coding:**



- Encoder: maps source U^n into channel input X^n
- Decoder: maps channel output Y^n into reconstruction \hat{U}^n

- **Separate source and channel coding:**



- Source encoder: maps source U^n into message $W \in \{1, 2, \dots, 2^{nR}\}$
- Channel encoder: maps message W into channel input X^n

- Channel decoder: maps channel output Y^n into message estimate $\hat{W} \in \{1, 2, \dots, 2^{nR}\}$
- Source decoder: maps message estimate into reconstruction \hat{U}^n
- **Theorem:** (Source-Channel Separation) Suppose we want to send an iid source with with rate distortion function $R(D)$ across a discrete memoryless channel with capacity C . A distortion D is achievable if and only if

$$C > R(D)$$

Furthermore, there is no loss in using separate source and channel coding.

- **Example with uncoded transmission:** Let U_1, U_2, \dots be a sequence of iid Gaussian variables with mean zero and variance σ^2 . Suppose these values are transmitted over a Gaussian noise channel with noise power N and power constraint P according to the encoding scheme

$$X_i = \frac{\sqrt{P}}{\sigma} U_i$$

such that

$$Y_i = \frac{\sqrt{P}}{\sigma} U_i + Z_i$$

Suppose that the reconstruction of U_i is given by the conditional expectation:

$$\hat{U}_i = \mathbb{E}[U_i | Y_i] = \frac{\sqrt{P}\sigma}{P + N} Y_i$$

Then, the squared error distortion of this coding scheme is

$$D = \mathbb{E}[(U_i - \mathbb{E}[U_i | Y])^2] = \frac{\sigma^2}{1 + P/N}$$

Equivalently,

$$\frac{\sigma^2}{D} = 1 + \frac{P}{N} \iff \underbrace{\frac{1}{2} \log\left(\frac{\sigma^2}{D}\right)}_{R(D)} = \underbrace{\frac{1}{2} \log\left(1 + \frac{P}{N}\right)}_C$$

Hence, the distortion is precisely the distortion-rate function $D(R)$ of the Gaussian source evaluated at the the capacity of the Gaussian channel.

10.2 Application to Statistical inference

10.2.1 Paramter Estimation

- Suppose that data X_1, \dots, X_n are drawn i.i.d. from a family of probability measure P_θ indexed by a parameter $\theta \in \Theta$.
- The **minimax risk** associated with a loss function $L(\theta, \hat{\theta})$ is defined by

$$M(\Theta) = \inf_{\delta} \max_{\theta \in \Theta} \mathbb{E}_{X^n \sim P_\theta} [L(\theta, \delta(X^n))]$$

where the infimum is over all estimators $\delta(\cdot)$.

- For any prior distribution π supported on Θ , the minimax risk is bounded from below by the **Bayes risk**, which is defined by

$$B(\pi) = \inf_{\delta} \mathbb{E}_{\theta \sim \pi, X^n \stackrel{iid}{\sim} P_{\theta}} [L(\theta, \delta(X^n))]$$

An estimator that achieves the infimum is called the Bayes rule.

- Suppose that:
 - the “source” defined by π has rate-distortion function $R(D)$ with respect to the loss function $L(\theta, \hat{\theta})$.
 - the mutual information defined by π and $P_{\theta}^{\otimes n}$ is given by $I(\theta; X^n)$.

Then, the Bayes risk is bounded from below by the distortion-rate function $D(R)$ evaluated at $I(\theta; X^n)$, i.e.,

$$\text{Bayes risk at } \pi \geq D(I(\theta; X^n))$$

Note that if $I(\theta; X^n)$ can be bounded from above by the capacity of the “channel” defined by $P_{\theta}^{\otimes n}$.

10.2.2 Linear Model

- Suppose that data $(X, Y) \in \mathbb{R}^{n \times p} \times \mathbb{R}^n$ depend on an unknown parameter $\beta \in \mathbb{R}^p$ according to the model

$$Y = X\beta + Z, \quad Z \sim \mathcal{N}(0, \sigma^2 I_n)$$

- Suppose that β is drawn according to a prior distribution π with mean zero and identity covariance.
- The mutual information satisfies

$$\begin{aligned} I(\beta; X, Y) &= I(\beta; Y | X) \\ &= h(Y | X) - h(Y | X, \beta) \\ &= h(Y | X) - \frac{n}{2} \log(2\pi e \sigma^2) \end{aligned}$$

Furthermore, the entropy of $Y | X$ is bounded from above by the Gaussian distribution of the same mean and variance, and so

$$h(Y | X) \leq \frac{n}{2} \log(2\pi e) + \log \det(\sigma^2 I_n + XX^{\top})$$

Whence,

$$I(\beta; X, Y) \leq \frac{1}{2} \log \det(I_n + \sigma^{-2} XX^{\top})$$