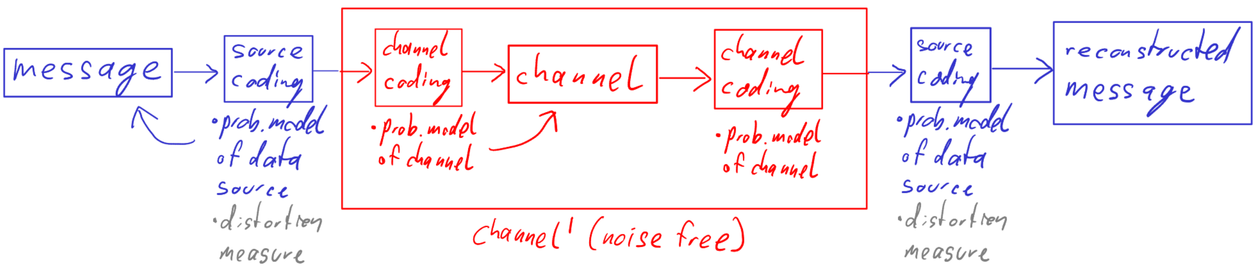
The (Noisy) Channel Coding Theorem

Robert Bamler · 7 July 2022

This lecture constitutes part 10 of the Course “Data Compression With and Without Deep Probabilistic Models” at University of Tübingen.

More course materials (lecture notes, problem sets, solutions, and videos) are available at:
https://robamler.github.io/teaching/compress22/

Recall from very first lecture:



- so far: focus on source coding (blue)
- (only) today: channel coding (following closely MacKay, “Information Theory, Inference, and Learning Algorithms”)
- next week: use “inverse channel coding” to derive theory of lossy compression

Motivating Example

$$\underset{\{0,1\}^k}{\overset{\Cap}{\mathbf{S}}} \xrightarrow[P(\mathbf{X}\,|\,\mathbf{S})]{\text{channel encoder}} \underset{\{0,1\}^n}{\overset{\Cap}{\mathbf{X}}} \xrightarrow[P(\mathbf{Y}\,|\,\mathbf{X})]{\text{channel}} \underset{\{0,1\}^n}{\overset{\Cap}{\mathbf{Y}}} \xrightarrow[P(\hat{\mathbf{S}}\,|\,\mathbf{Y})]{\text{channel decoder}} \underset{\{0,1\}^k}{\overset{\Cap}{\hat{\mathbf{S}}}}$$

- $\mathbf{S}$ is uniformly random distributed over $\{0,1\}^k$ and $n \geq k$.
- The channel transmits each bit independently but it introduces random bit flips:

$$P(\mathbf{Y}\,|\,\mathbf{X}) = \prod_{i=1}^{n} P(Y_i\,|\,X_i) \quad \text{with} \quad P(Y_i{=}y_i\,|\,X_i{=}x_i) = \begin{cases} 1-f & \text{if } y_i = x_i; \\ f & \text{if } y_i \neq x_i. \end{cases} \quad (0 \leq f \leq 1)$$

1. Assume there's no channel coding (i.e., $n = k$, $P(\mathbf{X}\,|\,\mathbf{S}) = \delta_{\mathbf{X},\mathbf{S}}$, $P(\hat{\mathbf{S}}\,|\,\mathbf{Y}) = \delta_{\hat{\mathbf{S}},\mathbf{Y}}$):
   - How many bits are flipped in expectation? $\mathbb{E}_P\big[\sum_{i=1}^k (1-\delta_{S_i,\hat{S}_i})\big] = k\,\mathbb{E}_P[1-\delta_{s_i,\hat{s}_i}] = kf$
   - What is the probability that no bits are flipped? $P(\hat{\mathbf{S}}{=}\mathbf{S}) = \mathbb{E}_P\left[\prod_{i=1}^{k} \delta_{s_i,\hat{s}_i}\right] = (1-f)^k$ (example: $f = 0.01$, $k = 10$ kbit $\Rightarrow (1-f)^k \approx 10^{-44}$)

Motivating Example

$$\begin{array}{ccccccc} \mathbf{S} & \xrightarrow[P(\mathbf{X}\,|\,\mathbf{S})]{\text{channel encoder}} & \mathbf{X} & \xrightarrow[P(\mathbf{Y}\,|\,\mathbf{X})]{\text{channel}} & \mathbf{Y} & \xrightarrow[P(\hat{\mathbf{S}}\,|\,\mathbf{Y})]{\text{channel decoder}} & \hat{\mathbf{S}} \\ \cap & & \cap & & \cap & & \cap \\ \{0,1\}^k & & \{0,1\}^n & & \{0,1\}^n & & \{0,1\}^k \end{array}$$

- $\mathbf{S}$ is uniformly random distributed over $\{0,1\}^k$ and $n \geq k$.
- $P(\mathbf{Y}\,|\,\mathbf{X}) = \prod_{i=1}^{n} P(Y_i\,|\,X_i)$ with $P(Y_i{=}y_i\,|\,X_i{=}x_i) = \begin{cases} 1-f & \text{if } y_i = x_i \\ f & \text{if } y_i \neq x_i \end{cases}$ $(0 \leq f \leq 1)$

Transmit 3 copies of each bit, receiver takes majority vote.

2. Come up with a simple encoding/decoding scheme to transmit **S** more reliably.
   - What is the ratio of transmitted bits k per channel invocations: $\frac{k}{n} = \frac{1}{3}$
   - What is the expected number of bit errors: $\mathbb{E}_P\big[\sum_{i=1}^{k}(1-\delta_{S_i,\hat{S}_i})\big] = k\left(3(1-f)f^2 + f^3\right) \approx k\left(3f^2 + \mathcal{O}(f^3)\right)$
   - What is the probability of having no error: $P(\hat{\mathbf{S}}{=}\mathbf{S}) \approx (1-3f^2)^k$

(same example as on last slide: $f = 0.01$; $k = 10$ kbit
$\Rightarrow (1-3f^2)^k \approx 0.05$
still really bad despite 3x reduction in transfer rate)

(Noisy) Channel Coding Theorem

Claim: we can do a lot better than replicating each bit three times:

- For a memoryless channel $P(\mathbf{Y}\,|\,\mathbf{X}) = \prod_{i=1}^{n} P(Y_i\,|\,X_i)$ (where $X_i \in \mathbb{X}$ and $Y_i \in \mathbb{Y}$ are not necessarily binary), let the *channel capacity* C be:

$$C := \max_{P(X_i)} I_P(X_i;\,Y_i).$$

→ examples on problem set (Problem 10.2)

- Then: in the limit of long messages (i.e., large n) there exists a channel coding scheme that satisfies both of the following:
  - the ratio $\frac{k}{n}$ can be made arbitrarily close to C; and
  - the error probability $P(\hat{\mathbf{S}}{\neq}\mathbf{s}\,|\,\mathbf{S}{=}\mathbf{s})$ can be made arbitrarily small for all $\mathbf{s} \in \{0,1\}^k$.
- More formally: $\forall \varepsilon > 0$ and $R < C$, there exists an $n_0 \in \mathbb{N}$ such that $\forall n \geq n_0$: there exists a code with $k \geq Rn$ and $P(\hat{\mathbf{S}}{\neq}\mathbf{s}\,|\,\mathbf{S}{=}\mathbf{s}) < \varepsilon$ for all $\mathbf{s} \in \{0,1\}^k$.

Intuition: block error correction

- We only care whether the *entire* bit string **S** gets transmitted without error. Thus:
  - make it as probable as possible that *no* bit is transmitted incorrectly;
  - if *one* bit S_i is transmitted incorrectly then we don't care if the other bits are also incorrect.
- E.g., split $\mathbf{S} \in \{0,1\}^k$ into blocks of 2 bits:

(S_{2i}, S_{2i+1})	3x replication	shorter code
$(0,0)$	000 000	00000
$(0,1)$	000 111	00111
$(1,0)$	111 000	11100
$(1,1)$	111 111	11011
k/n	$1/3 = 2/6$	$2/5 > 2/6$

In both codes, any two code words differ in at least 3 bits.
$\Rightarrow$ both codes can correct errors as long as at most one bit per block is corrupted.
But the shorter code achieves this property at higher ratio $\frac{k}{n}$

- The proof of the channel coding theorem scales up this idea to giant blocks.

Prerequisits (1 of 2): Chebychev's Inequality

- Let X be a nonnegative (discrete or continuous) scalar random variable with a finite expectation $\mathbb{E}_P[X]$. Then:

$$\boxed{P(X \geq \beta) \;\leq\; \frac{\mathbb{E}_P[X]}{\beta} \qquad \forall \beta > 0.}$$

- Proof:

$$\mathbb{1}_{X\geq\beta} = \begin{cases} 1 & \text{if } X \geq \beta \\ 0 & \text{otherwise} \end{cases}$$

$$P(X \geq \beta) = \mathbb{E}_P\left[\mathbb{1}_{X\geq\beta}\right] \leq \mathbb{E}_P\Big[\underbrace{\frac{X}{\beta}}_{\substack{\geq 1 \text{ for all} \\ \text{contributing terms}}} \mathbb{1}_{X\geq\beta}\Big] = \frac{1}{\beta}\mathbb{E}_P\Big[X \underbrace{\mathbb{1}_{X\geq\beta}}_{\leq 1}\Big] \leq \frac{1}{\beta}\mathbb{E}_P[X]$$

Prerequisits (2 of 2): Weak Law of Large Numbers

- Let $X_1, \ldots, X_n$ be independent random variables, all with the same expectation value $\mu := \mathbb{E}_P[X_i]$ and with the same (finite) variance $\sigma^2 := \mathbb{E}_P\big[(X_i - \mu)^2\big] < \infty$.
- Denote the *empirical mean* of all X_i as $\langle X_i \rangle_i := \frac{1}{n}\sum_{i=1}^n X_i$ (thus, $\langle X_i \rangle_i$ is itself a random variable).
- Then: $\boxed{P\big(\big|\langle X_i\rangle_i - \mu\big| \geq \beta\big) \;\leq\; \dfrac{\sigma^2}{n\,\beta^2} \qquad \forall \beta > 0.}$
- Proof: $P(|\langle x_i\rangle_i - \mu| \geq \beta) = P((\langle x_i\rangle - \mu)^2 \geq \beta^2) \overset{\text{chebychev}}{\leq} \dfrac{\mathbb{E}_P[(\langle x_i\rangle - \mu)^2]}{\beta} \overset{(*)}{=} \dfrac{\sigma^2}{n\beta}$

where $(*)$: $\mathbb{E}_P[(\langle x_i\rangle_i - \mu)^2] = \mathbb{E}_P\Big[\Big(\frac{1}{n}\sum_{i=1}^n x_i - \mu\Big)^2\Big] = \frac{1}{n^2}\mathbb{E}_P\Big[\Big(\sum_{i=1}^n (x_i - \mu)\Big)^2\Big]$

$$= \frac{1}{n^2}\mathbb{E}_P\Big[\sum_{i=1}^n\sum_{j=1}^n (x_i-\mu)(x_j-\mu)\Big] = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \underbrace{\mathbb{E}_P[(x_i-\mu)(x_j-\mu)]}_{= \begin{cases} 0 & \text{if } i \neq j \text{ (since } x_i, x_j \text{ indep)} \\ \sigma^2 & \text{if } i = j \end{cases}} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Apply Weak Law of Large Numbers to Information Content

Consider a data source P of messages $\mathbf{X} \equiv (X_1, \ldots, X_n) \in \mathbb{X}^n$ where all X_i are i.i.d.

Thus, the information content of a symbol X_i is a random variable: $-\log P(X_i)$.

- Its *expectation* is the entropy of a symbol: $\mathbb{E}_P\big[-\log_2 P(X_i)\big] = H_P[X_i]$
- Its *empirical mean* is: $\langle -\log_2 P(X_i)\rangle_i = -\frac{1}{n}\sum_{i=1}^n \log_2 P(X_i) \overset{\text{(i.i.d.)}}{=} -\frac{1}{n}\log_2 P(\mathbf{X})$
- Apply weak law of large numbers: for long messages (i.e., large n), large deviations β of the empirical mean from the expectation value are improbable:

$$\boxed{P\left(\left|\frac{-\log_2 P(\mathbf{X})}{n} - H_P[X_i]\right| \geq \beta\right) \leq \frac{\sigma^2}{n\,\beta^2} \qquad \forall \beta > 0.}$$

(where σ^2 is the variance of $-\log P(X_i)$) ← (assume $\sigma^2 < \infty$ as, e.g., for a finite alphabet)

What are "typical" messages?

Last slide: $P\left(\left|\frac{-\log_2 P(\mathbf{X})}{n} - H_P[X_i]\right| \geq \beta\right) \leq O\left(\frac{1}{n\,\beta^2}\right) \qquad \forall \beta > 0.$

- Thus, for "most" long random messages, the information content per symbol is close to the entropy of a symbol.
- Define the *typical set* $T_{P(X_i),n,\beta}$ as the set of messages of length n whose information content per symbol deviates from the entropy of a symbol by less than some given threshold β:

$$T_{P(X_i),n,\beta} := \left\{ \mathbf{x} \in \mathbb{X}^n \quad \text{that satisfy:} \quad \left|\frac{-\log_2 P(\mathbf{X}{=}\mathbf{x})}{n} - H_P[X_i]\right| < \beta \right\}$$

- Thus: $P\big(\mathbf{X} \in T_{P(X_i),n,\beta}\big) \geq 1 - \frac{\sigma^2}{n\,\beta^2} \xrightarrow{n\to\infty} 1 \quad \forall \beta > 0$

(weak law of large numbers)

Examples of Typical Sets

Consider sequences of binary symbols, $\mathbf{X} \in \{0,1\}^n$, with $\begin{cases} P(X_i{=}1) = \alpha \\ P(X_i{=}0) = 1-\alpha \end{cases}$. $(0 \leq \alpha \leq 1)$

- Entropy per symbol: $H_P[X_i] = H_2(\alpha) = -\alpha \log_2 \alpha - (1-\alpha)\log_2(1-\alpha) \in [0,1]$
- Size of full message space: $\big|\{0,1\}^n\big| = 2^n$
- If $\alpha = \frac{1}{2}$ then all messages $\mathbf{x} \in \{0,1\}^n$ have the same information content, and thus all messages are typical: $T_{P(X_i),n,\beta} = \{0,1\}^n \; \forall n, \beta > 0$.
- But if $\alpha \neq \frac{1}{2}$ then, for long messages, *significantly* (exponentially) fewer messages are typical: $\big|T_{P(X_i),n,\beta}\big| \approx 2^{nH_2(\alpha)} \ll 2^n$ ← (see next slide)
  - fraction of typical messages: $\frac{|T_{P(X_i),n,\beta}|}{|\{0,1\}^n|} \approx 2^{-n(1-H_2(\alpha))} \xrightarrow{n\to\infty} 0$ (exponentially fast)

Size of the Typical Set

$$T_{P(X_i),n,\beta} := \left\{ \mathbf{x} \in \mathbb{X}^n \quad \text{that satisfy:} \quad \left|\frac{-\log_2 P(\mathbf{X}{=}\mathbf{x})}{n} - H_P[X_i]\right| < \beta \right\}$$

- **Claim:** $\big|T_{P(X_i),n,\beta}\big| < 2^{n(H_P[X_i]+\beta)}$
- **Proof:** $\forall \underline{x} \in T_{P(X_i),n,\beta}: -\frac{1}{n}\log_2 P(\underline{X}=\underline{x}) - H_P[X_i] < \beta$

$\Rightarrow P(\underline{X}=\underline{x}) > 2^{-n(H_P[X_i]+\beta)}$

$\Rightarrow$ There can be at most $\frac{1}{2^{-n(H_P[X_i]+\beta)}} = 2^{n(H_P[X_i]+\beta)}$ elements in $T_{P(X_i),n,\beta}$.