Science Department

Course “Data Compression With and Without Deep Probabilistic Models” · Department of Computer Science

The (Noisy) Channel Coding Theorem

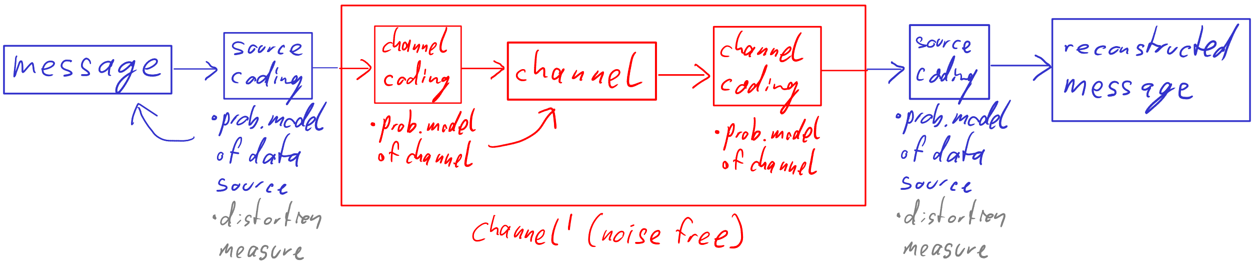
Robert Bamler · 7 July 2022

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

$$\begin{tikzcd}
\mathbf{S} \arrow[r, "\text{channel encoder}", "P(\mathbf{X}\,|\,\mathbf{S})"'] & \mathbf{X} \arrow[r, "\text{channel}", "P(\mathbf{Y}\,|\,\mathbf{X})"'] & \mathbf{Y} \arrow[r, "\text{channel decoder}", "P(\hat{\mathbf{S}}\,|\,\mathbf{Y})"'] & \hat{\mathbf{S}} \\
\{0,1\}^k \arrow[u, phantom, sloped, "\ni"] & \{0,1\}^n \arrow[u, phantom, sloped, "\ni"] & \{0,1\}^n \arrow[u, phantom, sloped, "\ni"] & \{0,1\}^k \arrow[u, phantom, sloped, "\ni"]
\end{tikzcd}$$

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

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

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?
   - What is the probability that no bits are flipped? $P(\hat{\mathbf{S}}{=}\mathbf{S}) =$

Motivating Example

$$\begin{tikzcd}[column sep=huge]
\mathbf{S} \arrow[r, "\text{channel encoder}", "P(\mathbf{X}\,|\,\mathbf{S})"'] \arrow[d, phantom, sloped, "\in"] & \mathbf{X} \arrow[r, "\text{channel}", "P(\mathbf{Y}\,|\,\mathbf{X})"'] \arrow[d, phantom, sloped, "\in"] & \mathbf{Y} \arrow[r, "\text{channel decoder}", "P(\hat{\mathbf{S}}\,|\,\mathbf{Y})"'] \arrow[d, phantom, sloped, "\in"] & \hat{\mathbf{S}} \arrow[d, phantom, sloped, "\in"] \\
\{0,1\}^k & \{0,1\}^n & \{0,1\}^n & \{0,1\}^k
\end{tikzcd}$$

- **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 $P(Y_i{=}y_i\,|\,X_i{=}x_i) = \begin{cases} 1-\alpha & \text{if } y_i = x_i \\ \alpha & \text{if } y_i \neq x_i \end{cases}$

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: $\frac{k}{n} =$
   - What is the expected number of bit errors: $\mathbb{E}_P\big[\sum_{i=1}^{k}(1-\delta_{S_i,\hat{S}_i})\big] =$
   - What is the probability of having no error: $P(\hat{\mathbf{S}}{=}\mathbf{S}) =$

(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:

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

- 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
$(0,0)$		
$(0,1)$		
$(1,0)$		
$(1,1)$		
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:

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 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: $\boxed{P\big(\big|\langle X_i \rangle_i - \mu\big| \geq \beta\big) \;\leq\; \frac{\sigma^2}{n\,\beta^2} \qquad \forall \beta > 0.}$
- Proof:

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: $\langle -\log_2 P(X_i) \rangle_i = -\frac{1}{n}\sum_{i=1}^n \log_2 P(X_i) \overset{(i.i.d.)}{=} -\frac{1}{n}\log_2 P(\mathbf{X})$
- 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 $-\log P(X_i)$)

What are “typical” messages?

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

- 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:

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

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

Examples of Typical Sets

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

- Entropy per symbol: $H_P[X_i] = H_2(\alpha)$
- 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: $\big|T_{P(X_i),n,\beta}\big| \approx 2^{nH_2(\alpha)} \ll 2^n$
  - fraction of typical messages: $\frac{\left|T_{P(X_i),n,\beta}\right|}{\left|\{0,1\}^n\right|} \approx$

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:**

Back to Channel Coding: Transmitting “Typical” Messages

$$\begin{tikzcd}
\mathbf{S} \arrow[r, "\text{channel encoder}", "P(\mathbf{X}\,|\,\mathbf{S})"'] \arrow[d, phantom, "\rotatebox{-90}{$\in$}"] & \mathbf{X} \arrow[r, "\text{channel}", "P(\mathbf{Y}\,|\,\mathbf{X})"'] \arrow[d, phantom, "\rotatebox{-90}{$\in$}"] & \mathbf{Y} \arrow[r, "\text{channel decoder}", "P(\hat{\mathbf{S}}\,|\,\mathbf{Y})"'] \arrow[d, phantom, "\rotatebox{-90}{$\in$}"] & \hat{\mathbf{S}} \arrow[d, phantom, "\rotatebox{-90}{$\in$}"] \\
\{0,1\}^k & \mathbb{X}^n & \mathbb{Y}^n & \{0,1\}^k
\end{tikzcd}$$

- Draw a message $\mathbf{x} \in \mathbb{X}^n$ from some input distribution $P(\mathbf{X}) = \prod_{i=1}^n P(X_i)$.
- Transmit $\mathbf{x}$ over the channel $\Rightarrow$ receive $\mathbf{y} \sim P(\mathbf{Y}\,|\,\mathbf{X}{=}\mathbf{x})$.
- Thus:
  - $\mathbf{x} \sim P(\mathbf{X})$ and therefore $P(\mathbf{x} \in T_{P(X_i),n,\beta}) \xrightarrow{n\to\infty} 1 \quad \forall\beta > 0$
  - $\mathbf{y} \sim P(\mathbf{Y})$ and therefore $P(\mathbf{y} \in T_{P(Y_i),n,\beta}) \xrightarrow{n\to\infty} 1 \quad \forall\beta > 0$
  - $(\mathbf{x},\mathbf{y}) \sim P(\mathbf{X},\mathbf{Y}) = \prod_{i=1}^n P(X_i)\,P(Y_i\,|\,X_i)$ and thus $P\big((\mathbf{x},\mathbf{y}) \in T_{P(X_i,Y_i),n,\beta}\big) \xrightarrow{n\to\infty} 1 \quad \forall\beta > 0$
- We say that $\mathbf{x}$ and $\mathbf{y}$ are *jointly typical*: $P\big((\mathbf{x},\mathbf{y}) \in J_{P(X_i,Y_i),n,\beta}\big) \xrightarrow{n\to\infty} 1 \quad \forall\beta > 0$

Understanding Joint Typicality

Compare the example on the last slide to a situation where $\mathbf{x}$ and $\mathbf{y}$ are drawn *independently* from their respective marginal distributions, i.e.,

- $\mathbf{x} \sim P(\mathbf{X})$;and
- $\mathbf{y} \sim P(\mathbf{Y})$ where $P(\mathbf{Y}) = \sum_{\mathbf{x}' \in \mathbb{X}^n} P(\mathbf{X}{=}\mathbf{x}')\,P(\mathbf{Y}{=}\mathbf{y}\,|\,\mathbf{X}{=}\mathbf{x}')$

Question: What is the probability that $\mathbf{x}$ and $\mathbf{y}$ are jointly typical?

Answer: $P\big((\mathbf{x},\mathbf{y}) \in J_{P(X_i,Y_i),n,\beta}\big) =$

Insight: *Randomly Designed* Channel Codes Work Surprisingly Well

$$\begin{tikzcd}
\mathbf{S} \arrow[r, "\text{channel encoder}", "P(\mathbf{X}\,|\,\mathbf{S})"'] \arrow[d, phantom, "\rotatebox{-90}{$\in$}"] & \mathbf{X} \arrow[r, "\text{channel}", "P(\mathbf{Y}\,|\,\mathbf{X})"'] \arrow[d, phantom, "\rotatebox{-90}{$\in$}"] & \mathbf{Y} \arrow[r, "\text{channel decoder}", "P(\hat{\mathbf{S}}\,|\,\mathbf{Y})"'] \arrow[d, phantom, "\rotatebox{-90}{$\in$}"] & \hat{\mathbf{S}} \arrow[d, phantom, "\rotatebox{-90}{$\in$}"] \\
\{0,1\}^k & \mathbb{X}^n & \mathbb{Y}^n & \{0,1\}^k
\end{tikzcd}$$

For given n, k, β, $P(X_i)$ and channel $P(Y_i\,|\,X_i)$, construct a random channel code $\mathcal{C}$:

- For each $\mathbf{s} \in \{0,1\}^k$, draw a code word $\mathcal{C}(\mathbf{s}) \in \mathbb{X}^k$ from $P(\mathbf{X})$.
- Define a deterministic encoder: $P(\mathbf{X}{=}\mathbf{x}\,|\,\mathbf{S}{=}\mathbf{s}) := \delta_{\mathbf{x},\mathcal{C}(\mathbf{s})}$
- Decoder: map $\mathbf{y}$ to $\hat{\mathbf{s}}$ if $(\mathcal{C}(\hat{\mathbf{s}}),\mathbf{y}) \in J_{P(X_i,Y_i),n,\beta}$ for exactly one $\hat{\mathbf{s}}$. Otherwise, fail.

Claim: In expectation over all random codes $\mathcal{C}$ that are constructed in this way, the error probability for long messages goes to zero as long as $\frac{k}{n} < I_P(X_i, Y_i) - 3\beta$:

$$\boxed{\mathbb{E}_{P(\mathcal{C})}\big[P(\hat{\mathbf{S}} \neq \mathbf{S})\big] \xrightarrow{n\to\infty} 0 \quad \text{if} \quad \frac{k}{n} < I_P(X_i, Y_i) - 3\beta.}$$

Proof of $\mathbb{E}_{P(\mathcal{C})}\big[P(\hat{\mathbf{S}}\neq\mathbf{S})\big] \xrightarrow{n\to\infty} 0$ if $\frac{k}{n} < I_P(X_i, Y_i) - 3\beta$

2 possibilities for errors:

- $(\mathcal{C}(\mathbf{s}), \mathbf{y}) \notin J_{P(X_i,Y_i),n,\beta}$:
- $(\mathcal{C}(\mathbf{s}'), \mathbf{y}) \in J_{P(X_i,Y_i),n,\beta}$ for some $\mathbf{s}' \neq \mathbf{s}$:

Total error probability:

Proof of the Noisy Channel Coding Theorem

Theorem (reminder): $\forall\varepsilon > 0$ 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 $\mathbf{s} \in \{0,1\}^k$.

- Set $P(X_i) := \arg\max_{P(X_i)} I_P(X_i; Y_i)$. Thus, $I_P(X; Y) = C$.
- Assume $\frac{k}{n} < C - 3\beta$. Thus, $\mathbb{E}_{P(\mathcal{C})}\big[P(\hat{\mathbf{S}}\neq\mathbf{S})\big] \xrightarrow{n\to\infty} 0$.
- This means that $\forall\varepsilon$: $\exists n_0$ such that $\mathbb{E}_{P(\mathcal{C})}\big[P(\hat{\mathbf{S}}\neq\mathbf{S})\big] < \frac{\varepsilon}{2}\ \forall n > n_0$.
  $\Rightarrow$ For all $n > n_0$, there must exist at least one code $\mathcal{C}$ that satisfies $P(\hat{\mathbf{S}}\neq\mathbf{S}) < \frac{\varepsilon}{2}$.
- $P(\hat{\mathbf{S}}\neq\mathbf{S}) = \mathbb{E}_{\mathbf{s}\sim P(\mathbf{S})}\big[P(\hat{\mathbf{S}}\neq\mathbf{s}\,|\,\mathbf{S}=\mathbf{s})\big] < \frac{\varepsilon}{2}$ where $P(\mathbf{S})$ is the uniform distribution.
  $\Rightarrow$ The $\big|\{0,1\}^k\big|/2$ code words $\mathcal{C}(\mathbf{s})$ with lowest $P(\hat{\mathbf{S}}\neq\mathbf{s}\,|\,\mathbf{S}=\mathbf{s})$ must all satisfy $P(\hat{\mathbf{S}}\neq\mathbf{s}\,|\,\mathbf{S}=\mathbf{s}) < \varepsilon$. $\Rightarrow$ Defines a code with ratio $\frac{k-1}{n}$ ($\approx \frac{k}{n}$ for $n\to\infty$).
- We can make $\frac{k}{n}$ and therefore R arbitrarily close to C by letting $\beta \to 0$.

Outlook

Problem Set:

- Implement a VAE for lossy compression.
- Calculate the channel capacity for a few toy channels models (both discrete and continuous).
- Derive a fundamental property of mutual information along a Markov chain: the information processing inequality (we'll need this next week).

Next Week:

- Use the noisy channel coding theorem to prove a lower bound on the bit rate of *lossy* compression ("rate/distortion-theory").