EBERHARD KARLS UNIVERSITÄT TÜBINGEN
Faculty of Science · Department of Computer Science · Group of Prof. Robert Bamler

Lecture 12, Part 2:

Channel Coding and Source/Channel Separation

Robert Bamler · Summer Term of 2023

These slides are part of the course *"Data Compression With and Without Deep Probabilistic Models"* taught at University of Tübingen. More course materials—including video recordings, lecture notes, and problem sets with solutions—are publicly available at `https://robamler.github.io/teaching/compress23/`.

Recall: Lower Bound on the Rate/Distortion Curve

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- Encoder/decoder form a *Markov chain:*

$$\text{message } \mathbf{X} \xrightarrow[P(\mathbf{S}|\mathbf{X})]{\text{encoder}} \text{bit string } \mathbf{S} \xrightarrow[P(\mathbf{X}'|\mathbf{S})]{\text{decoder}} \text{reconstruction } \mathbf{X}'$$

$\Longrightarrow$ By data processing inequality:
$I_P(\mathbf{X};\mathbf{X}') \leq I_P(\mathbf{X};\mathbf{S}) = HP(\mathbf{S}) - H_P(\mathbf{S}\,|\,\mathbf{X}) \leq H_P(\mathbf{S}) \leq \text{bit rate}$

- Typical formulation in the literature:
  - Consider distortion metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ and distortion threshold $\mathcal{D}$
  - Then, all lossy compression codes that satisfy $\mathbb{E}_P\big[d(\mathbf{X},\mathbf{X}')\big] \leq D$ have:

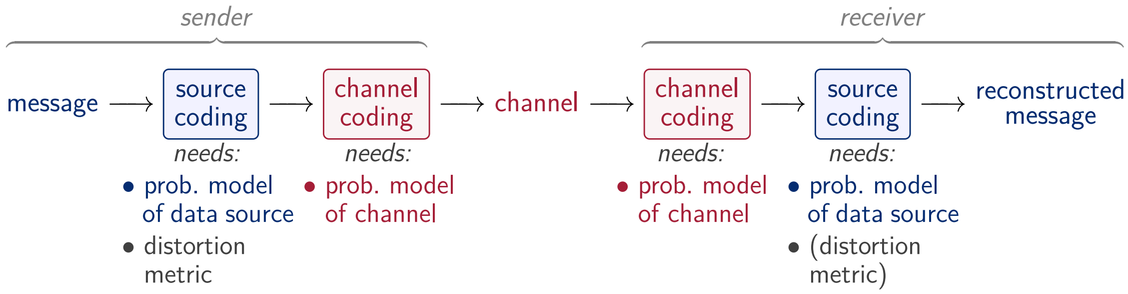
$$\text{bit rate} \geq \mathcal{R}(\mathcal{D}) \qquad \text{with the rate/distortion curve:} \qquad \mathcal{R}(\mathcal{D}) \geq \inf_{\substack{P(\mathbf{X},\mathbf{X}'):\\ \mathbb{E}_P[d(\mathbf{X},\mathbf{X}')]\leq\mathcal{D}}} I_P(\mathbf{X};\mathbf{X}')$$

- **Today:** finish proof that this lower bound is (almost) achievable

Channel Coding

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- **Recap from very first lecture:**



Intuition: Block Error Correction

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

$$\mathbf{S} \in \{0,1\}^n \xrightarrow[P(\mathbf{Y}|\mathbf{X})]{\text{channel encoder}} \mathbf{X} \in \mathcal{X}^k \xrightarrow[\prod_{i=1}^k P(Y_i|X_i)]{\text{memoryless channel}} \mathbf{Y}' \in \mathcal{Y}^k \xrightarrow[P(\mathbf{X}'|\mathbf{Y})]{\text{channel decoder}} \mathbf{S}' \in \{0,1\}^n$$

Examples: $(\mathcal{X} = \mathcal{Y} = \{0,1\})$

S	Y
"0"	"000"
"1"	"111"

$\Longrightarrow$

S	Y
"00"	"000000"
"01"	"000111"
"10"	"111000"
"11"	"111111"

better:

S	Y
"00"	"00000"
"01"	"00111"
"10"	"11100"
"11"	"11111"

- Assume that the channel flips symbols with probability $f \ll 1$.
- Both codes can recover a bit string of length N if at most 1 symbol per K-block is flipped.
  $\Longrightarrow$ for $|\mathbf{S}| = N$: $P(\mathbf{S}'=\mathbf{S}) = (1-f)^K + Kf(1-f)^{K-1} \approx 1 - \binom{K}{2} f^2 + O(f^3)$
  $\Longrightarrow$ for a sequence of $n \gg N$ bits:
  $$P(\mathbf{S}'=\mathbf{S}) \approx \left(1 - \binom{K}{2} f^2\right)^{n/N} = \exp\left[\tfrac{n}{N} \ln\left(1 - \binom{K}{2} f^2\right)\right] \approx \exp\left[-\binom{K}{2} \tfrac{f^2}{N}\, n\right]$$

(Noisy) Channel Coding Theorem

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

$$\mathbf{S} \in \{0,1\}^n \xrightarrow[P(\mathbf{Y}|\mathbf{X})]{\text{channel encoder}} \mathbf{X} \in \mathcal{X}^k \xrightarrow[\prod_{i=1}^k P(Y_i|X_i)]{\text{memoryless channel}} \mathbf{Y}' \in \mathcal{Y}^k \xrightarrow[P(\mathbf{X}'|\mathbf{Y})]{\text{channel decoder}} \mathbf{S}' \in \{0,1\}^n$$

- **Goal:** transmit a bit string $\mathbf{S}$ that is *as long as possible* using the channel *as little as possible* and recover original bit string with *high probability*.
  $\Longrightarrow$ we want: large n, small k, and high $P(\mathbf{S}=\mathbf{S}')$
- For a memoryless channel $P(\mathbf{Y}\,|\,\mathbf{X}) = \prod_{i=1}^{k} P(Y_i\,|\,X_i)$, let the *channel capacity* be:
  $$C := \sup_{P(X_i)} I_P(X_i; Y_i).$$
- **Theorem:** in the limit of long messages ($n \gg 1$), there exists a channel coding scheme that satisfies both of the following:
  - the ratio $\frac{n}{k}$ can be made arbitrarily close to the channel capacity C; and
  - the error probability $P(\mathbf{S}' \neq \mathbf{s}\,|\,\mathbf{S}=\mathbf{s})$ can be made arbitrarily small for all $\mathbf{s} \in \{0,1\}^n$.

Prerequisite 1 of 2: Chebychev's Inequality

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

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

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

- **Proof:**

Prerequisite 2 of 2: Weak Law of Large Numbers

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- Let $X_1, \ldots, X_k$ 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 $\langle X_i \rangle_i := \frac{1}{n} \sum_{i=1}^{k} X_i$
  (thus, $\langle X_i \rangle_i$ is itself a random variable).
- Then: $P\Big(\big|\langle X_i \rangle_i - \mu\big| \geq \beta\Big) \leq \frac{\sigma^2}{k\beta^2} \qquad \forall \beta > 0.$
- **Proof:**

Implications on Information Content

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- Consider a data source of messages $\mathbf{X} = (X_1, \ldots, X_k)$ where all X_i are i.i.d.
- The information content $-\log_2 P(X_i)$ of a symbol is a random variable.
  - Its *expectation* is the entropy of a symbol: $\mathbb{E}_P[-\log_2 P(X_i)] = H_P(X_i)$
  - Its *empirical mean* is: $\langle -\log_2 P(X_i) \rangle_i =$
- Apply weak law of large numbers:

$$P\left(\left|\frac{-\log_2 P(\mathbf{X})}{k} - H_P(X_i)\right| \geq \beta\right) \leq O\left(\frac{\sigma^2}{k\beta^2}\right) \qquad \forall \beta > 0$$

(where σ^2 is the variance of $-\log_2 P(X_i)$)

"For long messages (i.e., $k \gg 1$), large deviations β between the mean information content and the entropy per symbol are improbable."

What are "Typical" Messages?

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- **Last Slide:** $P\left(\left|\frac{-\log_2 P(\mathbf{X})}{k} - H_P(X_i)\right| \geq \beta\right) \leq O\left(\frac{\sigma^2}{k\beta^2}\right) \qquad \forall \beta > 0$
  "For most long random messages, the information content per symbol is close to $H_P(X_i)$."
- Define the *typical set* $T_{P(X_i),k,\beta}$ as the set of messages of length k whose information content per symbol deviates from $H_P(X_i)$ by less than some given threshold β:

$$T_{P(X_i),k,\beta} := \left\{\mathbf{x} \in \mathcal{X}^k \quad \text{that satisfty:} \quad \left|\frac{-\log_2 P(\mathbf{X}=\mathbf{x})}{k} - H_P(X_i)\right| < \beta\right\}$$

- Thus, by weak law of large numbers: $P(\mathbf{X} \in T_{P(X_i),k,\beta}) \geq$

Examples of Typical Sets

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

- Entropy per symbol: $H_P(X_i) \equiv H_2(\alpha)$
- Size of the full message space: $\left|\{0,1\}^k\right| = 2^k$
- If $\alpha = \frac{1}{2}$ then all messages $\mathbf{x} \in \{0,1\}^k$ have the same information content.
  $\Longrightarrow$ All messages are typical: $T_{P(X_i),k,\beta} = \{0,1\}^k \quad \forall k, \beta > 0.$
- But if $\alpha \neq \frac{1}{2}$ then, for long messages, *significantly* (exponentially) fewer messages are typical: $\left|T_{P(X_i),k,\beta}\right| \approx 2^{nH_2(\alpha)}$ (see next slide)

  $\Longrightarrow$ fraction of typical messages: $\dfrac{|T_{P(X_i),k,\beta}|}{|\{0,1\}^k|}$

Size of the Typical Set

$$T_{P(X_i),k,\beta} := \left\{ \mathbf{x} \in \mathcal{X}^k \quad \text{that satisfty:} \quad \left| \frac{-\log_2 P(\mathbf{X}=\mathbf{x})}{k} - H_P(X_i) \right| < \beta \right\}$$

- **Claim:** $\left|T_{P(X_i),k,\beta}\right| < 2^{n(H_P(X_i)+\beta)}$
- **Proof:**

Application to Channel Coding

$$\mathbf{S} \in \{0,1\}^n \xrightarrow[P(\mathbf{Y}|\mathbf{X})]{\text{channel encoder}} \mathbf{X} \in \mathcal{X}^k \xrightarrow[\prod_{i=1}^k P(Y_i|X_i)]{\text{memoryless channel}} \mathbf{Y}' \in \mathcal{Y}^k \xrightarrow[P(\mathbf{X}'|\mathbf{Y})]{\text{channel decoder}} \mathbf{S}' \in \{0,1\}^n$$

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

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
  - $\mathbf{y} \sim P(\mathbf{Y})$.
- **Question:** what is the probability that $\mathbf{x}$ and $\mathbf{y}$ are jointly tyipcal?

Random Channel Codes

$$\mathbf{S} \in \{0,1\}^n \xrightarrow[P(\mathbf{Y}|\mathbf{X})]{\text{channel encoder}} \mathbf{X} \in \mathcal{X}^k \xrightarrow[\prod_{i=1}^k P(Y_i|X_i)]{\text{memoryless channel}} \mathbf{Y}' \in \mathcal{Y}^k \xrightarrow[P(\mathbf{X}'|\mathbf{Y})]{\text{channel decoder}} \mathbf{S}' \in \{0,1\}^n$$

(Crazy) idea: assign *random* code words to bit strings:

- For each $\mathbf{s} \in \{0,1\}^n$, draw a code word $\mathcal{C}(\mathbf{s}) \in \mathcal{X}^k$ from $P(\mathbf{X})$.
- Define a (deterministic) channel encoder: $P(\mathbf{X}=\mathbf{x} \,|\, \mathbf{S}=\mathbf{s}, \mathcal{C}) = \delta_{\mathbf{x},\mathcal{C}(\mathbf{s})}$.
- Channel decoder: map $\mathbf{y}$ to $\hat{\mathbf{s}}$ if $(\mathcal{C}(\hat{\mathbf{s}}), \mathbf{y}) \in J_{P(X_i,Y_i),k,\beta}$ for exactly one $\hat{\mathbf{s}}$. Otherwise fail.
- **Claim** (Problem Set): In expectation over all random codes $\mathcal{C}$ that are constructed in this way, and over all input strings $\mathbf{s} \sim P(\mathbf{S}) := \text{Uniform}(\{0,1\}^k)$, the error probability for long messages goes to zero as long as $\frac{k}{n} < I_P(X_i, Y_i) - 3\beta$.

Proof of the Noisy Channel Coding Theorem

Theorem (reminder): for long messages ($n \gg 1$), there exists a channel coding scheme such that $\frac{n}{k}$ can be made arbitrarily close to the channel capacity C and the error probability $P(\mathbf{S}' \neq \mathbf{s} \,|\, \mathbf{S}=\mathbf{s})$ can be made arbitrarily small for all $\mathbf{s} \in \{0,1\}^n$.

Proof (cont'd)

Application to Lossy Data Compression