Lecture 13, Part 1:

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: Two Theorems That Are Awaiting Proofs

- **Rate/Distortion Theorem:** $\mathbf{X} \longrightarrow \mathbf{S} \in \{0,1\}^n \longrightarrow \mathbf{X}'$
  all lossy compression codes that satisfy $\mathbb{E}_P[d(\mathbf{X},\mathbf{X}')] \leq \mathcal{D}$ have $\mathbb{E}_P[\text{bit rate}] \geq \mathcal{R}(\mathcal{D})$ with the rate/distortion curve:

$$\mathcal{R}(\mathcal{D}) := \inf_{\substack{P(\mathbf{X}'|\mathbf{X}):\\ \mathbb{E}_P[d(\mathbf{X},\mathbf{X}')]\leq\mathcal{D}}} I_P(\mathbf{X};\mathbf{X}')$$

- **Channel Coding Theorem:** $\mathbf{S} \in \{0,1\}^n \longrightarrow \mathbf{X} \in \mathcal{X}^k \longrightarrow \mathbf{Y}' \in \mathcal{Y}^k \longrightarrow \mathbf{S}' \in \{0,1\}^n$
  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 $C := \sup_{P(X_i)} I_P(X_i;Y_i)$; and
  - the error probability $P(\mathbf{S}' \neq \mathbf{s} \,|\, \mathbf{S}=\mathbf{s})$ can be made arbitrarily small for all $\mathbf{s} \in \{0,1\}^n$.

Recall: Typicality and Joint Typicality

- **Def. typical set:** $T_{P(X_i),k,\beta} := \Big\{\mathbf{x} \in \mathcal{X}^k \quad \text{that satisfty:} \quad \Big|\frac{-\log_2 P(\mathbf{X}=\mathbf{x})}{k} - H_P(X_i)\Big| < \beta\Big\}$
  - most (long) messages are *not* typical: $\big|T_{P(X_i),k,\beta}\big| < 2^{k(H_P(X_i)+\beta)} \implies \frac{|T_{P(X_i),k,\beta}|}{|\mathcal{X}^k|} < 2^{k(H_P(X_i)-|\mathcal{X}|+\beta)}$
  - But: most (long) *random* messages are typical: $P(\mathbf{X} \in T_{P(X_i),k,\beta}) \geq 1 - \frac{\sigma^2}{k\beta^2} \xrightarrow{k\to\infty} 1$
- **Def. joint typicality:**
  $(\mathbf{x},\mathbf{y}) \in J_{P(X_i,Y_i),k,\beta}$ iff: $\mathbf{x} \in T_{P(X_i),k,\beta}$, $\mathbf{y} \in T_{P(Y_i),k,\beta}$, and $(\mathbf{x},\mathbf{y}) \in T_{P(X_i,Y_i),k,\beta}$.
  - Again, most random samples $(\mathbf{x},\mathbf{y}) \sim P(\mathbf{X},\mathbf{Y})$ are jointly typical.
  - Thus, if we draw $\mathbf{x} \sim P(\mathbf{X})$ and then transmit it over the noisy channel to get $\mathbf{y} \sim P(\mathbf{Y}\,|\,\mathbf{X}=\mathbf{x})$, the resulting pair $(\mathbf{x},\mathbf{y})$ is jointly typical with high probability.
  - But: drawing $\mathbf{x} \sim P(\mathbf{X})$ and $\mathbf{y} \sim P(\mathbf{Y})$ independently from their marginal distributions usually does *not* lead to joint typicality:

Random Channel Codes

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

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

(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 $\mathbf{s}'$ if $(\mathcal{C}(\mathbf{s}'), \mathbf{y}) \in J_{P(X_i,Y_i),k,\beta}$ for exactly one $\mathbf{s}'$. 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 $\frac{n}{k} < I_P(X_i, Y_i) - 3\beta$.

Proof of *Expected* Performance of Random Codes

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Claim: $\mathbb{E}_{P(\mathcal{C})P(\mathbf{S})}\big[P(\mathbf{S}' \neq \mathbf{S} \,|\, \mathbf{S}, \mathcal{C})\big] \xrightarrow{k\to\infty} 0$ if $\frac{n}{k} < I_P(X_i, Y_i) - 3\beta$ $\quad (P(\mathbf{S}) = \text{Uniform}(\{0,1\}^n))$

- 2 possibilities for errors:
  - $(\mathcal{C}(\mathbf{s}), \mathbf{y}) \notin J_{P(X_i,Y_i),k,\beta}$:
  - $(\mathcal{C}(\mathbf{s}'), \mathbf{y}) \in J_{P(X_i,Y_i),k,\beta}$ for some $\mathbf{s}' \neq \mathbf{s}$:
- Total error probability:

Proof of the Noisy Channel Coding Theorem

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

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 $\mathbf{s} \in \{0,1\}^n$.

- Set $P(X_i) := \arg\max_{P(X_i)} I_P(X_i; Y_i)$. Thus, $I_P(X_i; Y_i) = C$
- Assume $\frac{n}{k} < C - 3\beta$. Thus, $\mathbb{E}_{P(\mathcal{C})P(\mathbf{S})}\big[P(\mathbf{S}' \neq \mathbf{S} \,|\, \mathbf{S}, \mathcal{C})\big] \xrightarrow{n\to\infty} 0$.
- This means that $\forall \varepsilon > 0 : \exists n_0$ such that $\mathbb{E}_{P(\mathcal{C})P(\mathbf{S})}\big[P(\mathbf{S}' \neq \mathbf{S} \,|\, \mathbf{S}, \mathcal{C})\big] < \frac{\varepsilon}{2} \; \forall n > n_0$.
  $\Longrightarrow$ For all $n > n_0$, there exists at least one code $\mathcal{C}$ with $\mathbb{E}_{P(\mathbf{S})}\big[P(\mathbf{S}' \neq \mathbf{S} \,|\, \mathbf{S}, \mathcal{C})\big] < \frac{\varepsilon}{2}$.
  $\Longrightarrow$ Since $P(\mathbf{S})$ is a uniform distribution over 2^n bit strings, the $2^n/2 = 2^{n-1}$ bit strings $\mathbf{s}$ with lowest $P(\mathbf{S}' \neq \mathbf{s} \,|\, \mathbf{S}=\mathbf{s}, \mathcal{C})$ must all satisfy $P(\mathbf{S}' \neq \mathbf{s} \,|\, \mathbf{S}=\mathbf{s}, \mathcal{C}) < \varepsilon$.
  $\Longrightarrow$ Use their 2^{n-1} code words $\mathcal{C}(\mathbf{s})$ to define a code with ratio $\frac{n-1}{k}$ ($\approx \frac{n}{k}$ for $n \to \infty$)
- Thus, we can make $\frac{n}{k}$ arbitrarily close to the capacity C by letting $\beta \to 0$.

Recall: Agenda

- **Rate/Distortion Theorem:** $\mathbf{X} \longrightarrow \mathbf{S} \in \{0,1\}^n \longrightarrow \mathbf{X}'$

  all lossy compression codes that satisfy $\mathbb{E}_P[d(\mathbf{X},\mathbf{X}')] \leq \mathcal{D}$ have $\mathbb{E}_P[\text{bit rate}] \geq \mathcal{R}(\mathcal{D})$ with the rate/distortion curve:

$$\mathcal{R}(\mathcal{D}) := \inf_{\substack{P(\mathbf{X}'|\mathbf{X}): \\ \mathbb{E}_P[d(\mathbf{X},\mathbf{X}')] \leq \mathcal{D}}} I_P(\mathbf{X};\mathbf{X}')$$

- **Channel Coding Theorem:** $\mathbf{S} \in \{0,1\}^n \longrightarrow \mathbf{X} \in \mathcal{X}^k \longrightarrow \mathbf{Y}' \in \mathcal{Y}^k \longrightarrow \mathbf{S}' \in \{0,1\}^n$

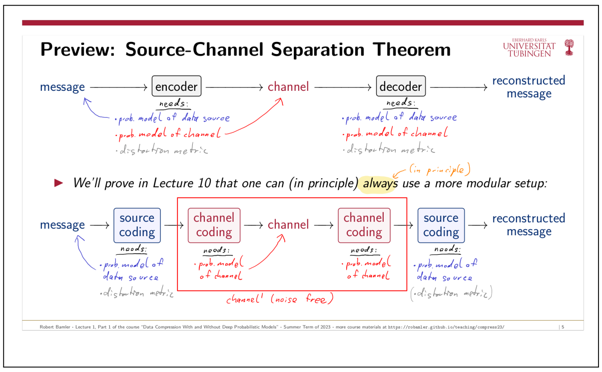
  In the limit of long messages ($n \gg 1$), there exists a channel coding scheme that satisfies both of the following:
  - the ratio $\frac{n}{k}$ can be made arbitrarily close to the channel capacity $C := \sup_{P(X_i)} I_P(X_i; Y_i)$; and
  - the error probability $P(\mathbf{S}' \neq \mathbf{s} \,|\, \mathbf{S} = \mathbf{s})$ can be made arbitrarily small for all $\mathbf{s} \in \{0,1\}^n$.

Proof of Rate/Distortion Theorem

- Channel coding: $\mathbf{S} \in \{0,1\}^n \xrightarrow[P(\mathbf{Y}|\mathbf{X})]{\text{encoder}} \mathbf{X} \in \mathcal{X}^k \xrightarrow[\prod_{i=1}^k P(Y_i|X_i)]{\text{channel}} \mathbf{Y}' \in \mathcal{Y}^k \xrightarrow[P(\mathbf{X}'|\mathbf{Y})]{\text{decoder}} \mathbf{S}' \in \{0,1\}^n$
- (Lossy) source coding: $\mathbf{X} \xrightarrow[P(\mathbf{S}|\mathbf{X})]{\text{encoder}} \mathbf{S} \in \{0,1\}^n \xrightarrow[P(\mathbf{X}'|\mathbf{S})]{\text{decoder}} \mathbf{X}'$
  - Assume data source $P(\mathbf{X})$ and mapping $P(\mathbf{X}'|\mathbf{X})$ are both given.
  - **Idea:** consider *inference channel* $P(\mathbf{X}|\mathbf{X}') = \frac{P(\mathbf{X})\,P(\mathbf{X}'|\mathbf{X})}{\sum_{\mathbf{x}} P(\mathbf{X})\,P(\mathbf{X}'|\mathbf{X})}$

Source/Channel Separation Theorem

- Recall from very first lecture:



- **Claim:** a joint source and channel coder cannot have a better rate/distortion performance than an optimal source coder combined with an optimal channel coder.