Theory of Lossy Compression (Rate/Distortion Theory)

Robert Bamler · 21 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: Noisy Channel Coding Theorem

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

- **Memoryless channel:** $P(\mathbf{Y}|\mathbf{X}) = \prod_{i=1}^n P(Y_i|X_i)$
- **Channel capacity:** $C := \max_{P(X_i)} I_P(X_i; Y_i)$
- **Proved so far:** error-free communication is possible as long as $\frac{k}{n} < C$.
- **Problem 10.3 (e):** prove that error-free communication is *not* possible if $\frac{k}{n} > C$.
  (follows from *data processing inequality*: $I_P(\mathbf{S}; \hat{\mathbf{S}}) \leq I_P(\mathbf{X}; \mathbf{Y})$)
- **But:** communication with $\frac{k}{n} > C$ *is* possible if we accept errors.
  - How many errors do we have to accept for a given $\frac{k}{n} > C$?

Poll

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

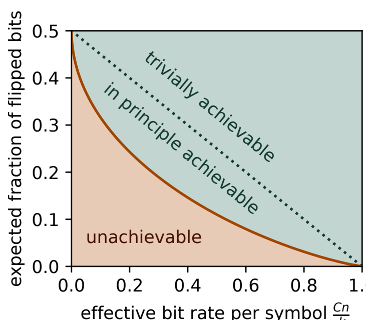
Assume you want to transmit $k > Cn$ uniformly distributed random bits using n invocations of a channel with capacity C. How many bit flips should you expect?

satisfy additional constraint: receiver knows which bits might be flipped.

(a) about $k - Cn$; ← transmit only first Cn bits

(b) about $(k - Cn)/2$; ← transmit only first Cn bits and let decoder guess the remaining $k - Cn$ bits → will get half of them right, in expectation

receiver can't tell which bits might be flipped → (c) fewer than $(k - Cn)/2$.



Problem 11.5: $\frac{k}{n}$ can go up to $\frac{C}{1 - H_2(D)}$

Application of Channel Coding Theorem:

Theoretical bound for *lossy* compression

Theoretical Bound for Lossy Compression

Consider a lossy compression code:

$$\text{message } \mathbf{X} \xrightarrow[P(\mathbf{S}\,|\,\mathbf{X})]{\text{source encoder}} \underset{\{0,1\}^*}{\overset{\mathbf{S}}{\cap}} \xrightarrow[P(\hat{\mathbf{X}}\,|\,\mathbf{S})]{\text{source decoder}} \text{reconstruction } \hat{\mathbf{X}}$$

- Assume the data distribution $P(\mathbf{X})$ and the mapping from $\mathbf{X}$ to its reconstruction $\hat{\mathbf{X}}$ is given and we want to find a suitable encoder/decoder pair.
- **Theorem:** optimal $\mathbb{E}_P\big[\text{amortized bit rate}\big] = I_P(\mathbf{X};\hat{\mathbf{X}})$.
  - Below: prove that $\exists$ code with $\mathbb{E}_P\big[\text{amortized bit rate}\big]$ arbitrarily close to $I_P(\mathbf{X};\hat{\mathbf{X}})$
  - Problem 11.2: prove that $\nexists$ code with $\mathbb{E}_P\big[\text{amortized bit rate}\big] < I_P(\mathbf{X};\hat{\mathbf{X}})$

Proof of Theoretical Bound for Lossy Compression

$$\text{message } \mathbf{X} \xrightarrow[P(\mathbf{S}\,|\,\mathbf{X})]{\text{source encoder}} \mathbf{S} \in \{0,1\}^* \xrightarrow[P(\hat{\mathbf{X}}\,|\,\mathbf{S})]{\text{source decoder}} \text{reconstruction } \hat{\mathbf{X}}$$

- **Given:** $P(\mathbf{X})$ and $P(\hat{\mathbf{X}}|\mathbf{X})$; **we seek:** source encoder $P(\mathbf{S}|\mathbf{X})$ and decoder $P(\hat{\mathbf{X}}|\mathbf{S})$.

• Consider posterior distribution $P(\underline{X}|\hat{\underline{X}}) = \dfrac{P(\underline{x})\,P(\hat{\underline{x}}|\underline{x})}{\int P(\underline{x})\,P(\hat{\underline{x}}|\underline{x})\,d\underline{x}}$

↳ interpret $P(\underline{X}|\hat{\underline{X}})$ as a model for a communication channel & find optimal channel coder:

$$\underline{S} \in \{0,1\}^k \xrightarrow[P(\hat{\underline{X}}|\underline{S})]{\text{chan. encoder}} \hat{\underline{X}} \in \mathfrak{X}^n \xrightarrow[\prod_{i=1}^n P(X_i|\hat{X}_i)]{\text{channel (for multiple } \hat{X}_i\text{'s)}} \underline{X} \in \mathfrak{X}^n \xrightarrow[P(\hat{\underline{S}}|\underline{X})]{\text{chan. decoder}} \hat{\underline{S}}$$

($= \underline{S}$ since error-free communication)

• Now re-interpret channel decoder as source encoder and channel encoder as source decoder:

$$\underline{X} \in \mathfrak{X}^n \xrightarrow[P(\hat{\underline{S}}|\underline{X})]{\text{source encoder}} \hat{\underline{S}} \in \{0,1\}^k \xrightarrow{\text{identity}} \underline{S} \in \{0,1\}^k \xrightarrow[P(\hat{\underline{X}}|\underline{S})]{\text{source decoder}} \hat{\underline{X}} \in \mathfrak{X}^n$$

⇒ Bitrate for n messages $X_i \in \mathfrak{X}$: $k = Cn + \varepsilon$
⇒ Amortized bit rate for single message: $C \le I_P(X;\hat{X}) + \varepsilon$

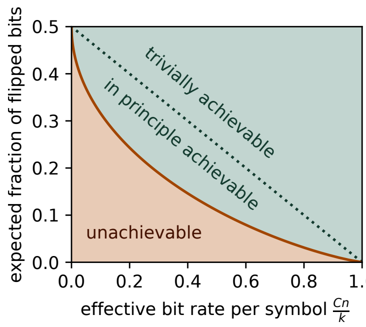
Rate/Distortion Theorem

Recap: For given $P(\mathbf{X})$ and $P(\hat{\mathbf{X}}|\mathbf{X})$: optimal $\mathbb{E}_P\big[\text{amortized bit rate}\big] = I_P(\mathbf{X};\hat{\mathbf{X}})$.

Corollary: ("rate/distortion theorem")

- consider a distortion metric $d(\mathbf{X},\hat{\mathbf{X}})$ between messages and their reconstructions, and a distortion threshold $\mathcal{D} \geq 0$.
- Then: optimal $\mathbb{E}_P\big[\text{amortized bit rate}\big]$ of code that satisfies $\mathbb{E}_P[d(\mathbf{X},\hat{\mathbf{X}})] \leq \mathcal{D}$ is:

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



see Problem 11.5

Example: Lossy Compression of Analog Messages

- Message $\mathbf{X} \in \mathbb{R}^k$, from normal-distributed data source: $P(\mathbf{X}) = \mathcal{N}(0, \sigma^2 I)$
- Distortion metric: mean squared error $d(\mathbf{X},\hat{\mathbf{X}}) = \frac{1}{k}||\mathbf{X} - \hat{\mathbf{X}}||_2^2 = \frac{1}{k}\sum_{i=1}^{k}(X_i - \hat{X}_i)^2$
- Goal: find (lower bound on) optimal rate $\mathcal{R}(\mathcal{D}) = \inf_{P(\hat{\mathbf{X}}|\mathbf{X}):\,\mathbb{E}_P[d(\mathbf{X},\hat{\mathbf{X}})]\leq\mathcal{D}} I_P(\mathbf{X};\hat{\mathbf{X}})$

→ Problem 12.5

Source-Channel Separation Theorem

Practical communication pipelines include both source and channel coding:

$$\mathbf{X} \xrightarrow[P(\mathbf{S}|\mathbf{X})]{\text{source encoder}} \mathbf{S} \xrightarrow[P(\mathcal{X}|\mathbf{S})]{\text{channel encoder}} \mathcal{X} \xrightarrow[P(\mathcal{Y}|\mathcal{X})=\prod_{i=1}^{n}P(\mathcal{Y}_i|\mathcal{X}_i)]{\text{(memoryless) channel}} \mathcal{Y} \xrightarrow[P(\hat{\mathbf{S}}|\mathcal{Y})]{\text{channel decoder}} \hat{\mathbf{S}} \xrightarrow[P(\hat{\mathbf{X}}|\hat{\mathbf{S}})]{\text{source decoder}} \hat{\mathbf{X}}$$

- Source coder: compress $\mathbf{X}$ to $\mathcal{R}(\mathcal{D})$ bits (in expectation)
- Channel coder: transmit these $\mathcal{R}(\mathcal{D})$ bits in $n_{\text{min}} := \mathcal{R}(\mathcal{D})/C$ channel invocations.

Question: can we get away with fewer than n_{min} channel invocations if we don't separate source coding from channel coding?

$$\mathbf{X} \xrightarrow[P(\mathcal{X}|\mathbf{X})]{\text{encoder}} \mathcal{X} \xrightarrow[P(\mathcal{Y}|\mathcal{X})=\prod_{i=1}^{n}P(\mathcal{Y}_i|\mathcal{X}_i)]{\text{(memoryless) channel}} \mathcal{Y} \xrightarrow[P(\hat{\mathbf{X}}|\mathcal{Y})]{\text{decoder}} \hat{\mathbf{X}}$$

- This is not possible (source-channel separation theorem); Proof: Problem 11.4

Outlook:

Recent advances in machine-learning based data compression

Overview: Research in ML-based Data Compressoin

- Main focus of community so far: improve rate/distortion performance for images and videos above classical codecs.
- Some work on medical data
- less focus on:
  - computational efficiency
  - real-time applications
  - other data types
  - strengths of ML-based methods (e.g., deliberate overfitting to specific kinds of videos, etc.)
  - dangers (especially due to semantically meaningful distortions)

Overview: Open Research Questions

- Connection between data compression & inference methods
- Quantization methods (e.g., learn good quantization grid, perform inference directly over discrete representation space)
- Models that don't require quantization (e.g., integer discrete flows)
- Dealing with distortions that change semantics of the data in Ml-based lossy compression (crucial for applications in medical and security relevant areas)
- Effective model architectures for data compression (e.g., for video data)