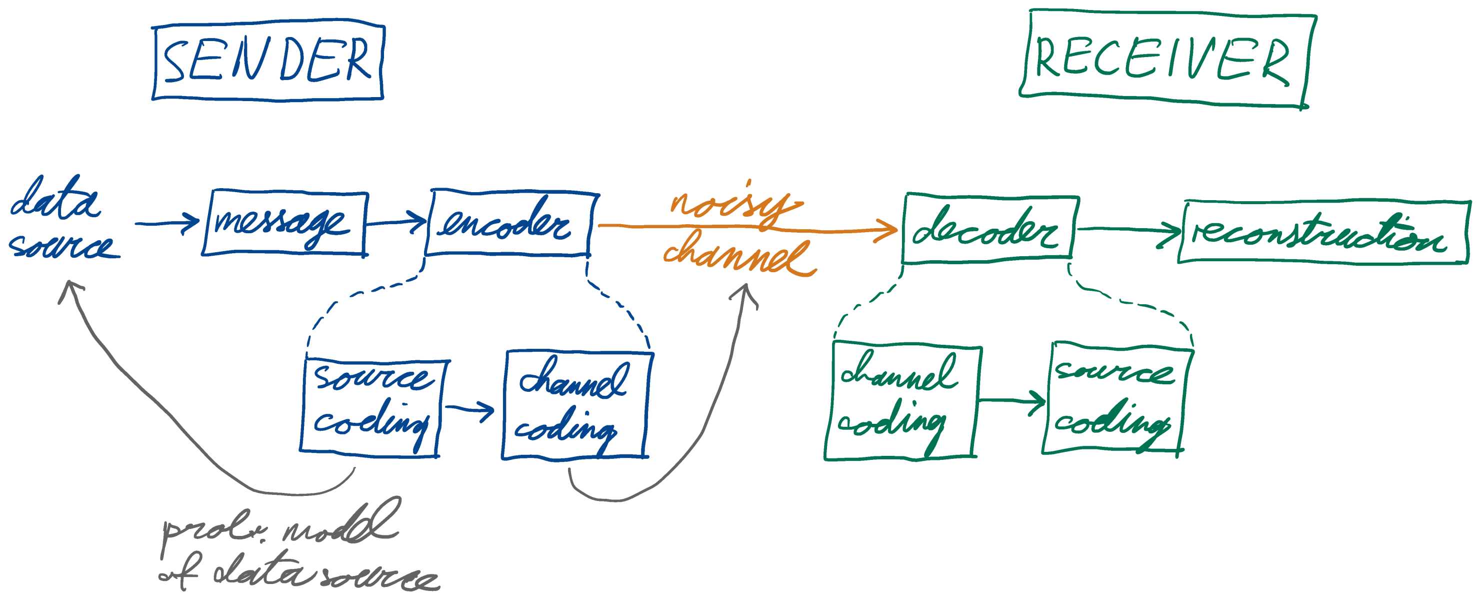
Lecture notes from the course "Data Compression With Deep Probabilistic Models" by Prof. Robert Bamler, University of Tuebingen, 22 May 2021; more course materials at https://robamler.github.io/teaching/compress21/

Probabilistic Models of Data Sources

Reminder: the big picture



→ "Don't transmit what you can predict."

Qualitatively: better prob. models ⟹ better compression performance

Quantifying the Modeling Error: The Kullback-Leibler Divergence

- consider general lossless compression setup: (i.e., no longer to symbol codes)
- data source generates messages $\underline{x}$ with prob. $p_{data}(\underline{x})$

  (p_{data}: prob. dist. over entire messages (not just single symbols))
- **Def:** bit rate $R(\underline{x})$:= total no. of bits in compressed rep. of $\underline{x}$ for a given lossless compression method

(lowest theoret. possible)

- optimal expected bit rate

$\mathbb{E}_{p_{data}}[-\log p_{data}]$

< 1 bit (typically irrelevant)

$$\mathbb{E}_{\underline{x} \sim p_{data}} \left[R_{opt}(\underline{x}) \right] = H(p_{data}) + \cancel{\varepsilon}$$

consider entire set possible messages as alphabet → $\underline{x}$ is a single symbol
→ $L_{opt} = H[p_{data}] + \varepsilon$

- reminder: to reach optimal expected bit rate, a compression method has to satisfy $R_{opt}(\underline{x}) = -\log p_{data}(\underline{x})$ $(+\varepsilon)$ $\forall \underline{x}$

Problem: in practice, we don't know p_{data}

→ distinguish

$\boxed{p_{data}}$ vs. $\boxed{p_{model}}$

true dist. of the data
↳ we don't know this
↳ but we may have samples from p_{data}
⇒ we can evaluate empirical averages, to estimate true expectation values

(for now:) assume we can explicitly evaluate $p_{model}(\underline{x})$ $\forall \underline{x}$

⇒ optimal (w.r.t. p_{model}) lossless compression code: $R(\underline{x}) = -\log p_{model}(\underline{x})$ $\forall \underline{x}$

⇒ expected bit rate

$$\mathbb{E}_{\underline{x} \sim p_{data}} \left[R(\underline{x}) \right] = \mathbb{E}_{\underline{x} \sim p_{data}} \left[-\log p_{model}(\underline{x}) \right]$$

$$= H(p_{data}, p_{model})$$

"cross entropy"

→ we have to minimize $H(p_{data}, p_{model})$ over parameters of p_{model}.

$\Rightarrow$ Overhead due to $p_{model} \neq p_{data}$:

Kullback-Leibler Divergence

$$D_{KL}(p_{data} \,\|\, p_{model}) = \underbrace{H(p_{data}, p_{model})}_{\text{actual bit rate}} - \underbrace{H(p_{data})}_{\text{theoret. lower bound}}$$

$$= \mathbb{E}_{\underline{x} \sim p_{data}}\left[\log \frac{p_{data}(\underline{x})}{p_{model}(\underline{x})}\right]$$

- Problem 3.1(c): prove that $D_{KL}(p\|q) \geq 0$
  ("Gibbs Theorem")

So far: $p_{model}(\underline{x}) = p(k) \prod_{i=1}^{k} \underbrace{p(x_i)}_{\text{prob. of a single symbol}}$

$\uparrow$ $\mathfrak{X}^*$

$\rightarrow$ assumed that symbols are "i.i.d.":

<u>independent</u> & <u>identically</u> distributed

difficult part: we haven't been able to model <u>correlations</u> between symbols

easy to drop this restriction:

$$p_{model}(\underline{x}) = p(k) \prod_{i=1}^{k} p_{\underline{i}}(x_i)$$

$\rightarrow$ prefix codes still work

Interlude: Probability Theory & Random Variables

<u>Goal</u>: <u>efficiently</u> model <u>correlations</u> between parts of the message

- sample space Ω: (abstract) space of "all states of the world"
- event $E \subset \Omega$: "event E occurs" = "the world is in a state $\omega \in E$"
- probability measure: $P: \Sigma \to [0, 1]$ ← σ-algebra on Ω (set of subsets of Ω)
  - ↳ $P(\Omega) = 1$
  - ↳ $P(\emptyset) = 0$
  - ↳ $P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$ if $\{E_i\}$ are pairwise disjoint
  - $P\left(\bigcup_{i=1}^{k} E_i\right) = \sum_{i=1}^{k} P(E_i)$ ——— " ———

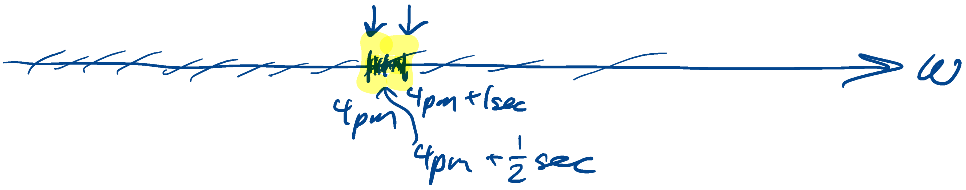
<u>Remark</u>: for <u>continuous</u> states, $P(\{\omega\})$ for a single $\omega \in \Omega$ typically doesn't make much sense.

<u>e.g.</u>: $\Omega \in \mathbb{R}$, $\omega \in \Omega$ is the arrival time of a bus

<u>Q</u>: with what probability that the bus arrives <u>exactly</u> at 4pm today

$P(\{\underbrace{\text{"4pm"}}_{\text{expressed as a real number according to some standard}}\}) = 0$

$P([\text{"4pm"}, \text{"4pm + 1sec"}])$ can be >0



• Random variable: $X: \Omega \to \mathbb{R}$

<u>Example:</u> "Simplified Game of Monopoly"

↳ $\Omega = \{ (a, b) : a, b \in \{1, 2, 3\}\}$

value of red die — value of blue die

↳ $\Sigma = \cancel{\mathcal{P}(\Omega)} = 2^\Omega := \{\text{all subsets of } \Omega\}$

↳ $P(E) = \frac{|E|}{|\Omega|} = \frac{|E|}{9}$

↳ Random variables:

- value of red die: $X_r((a,b)) = a$
- value of blue die: $X_b((a,b)) = b$
- sum of red + blue die: $X_s((a,b)) = a + b \in \{2,3,4,5,6\}$

<u>Properties of a single random variable</u>

↳ expectation value (of discrete r.v.)

$$\mathbb{E}_P[X] = \sum_{\omega \in \Omega} P(\{\omega\}) X(\omega)$$

$$= \sum_{x \in X(\Omega)} P(\underbrace{X^{-1}(x)}_{=\{\omega \in \Omega : X(\omega) = x\}}) \, x$$

$$\mathbb{E}_P[X_r] = \mathbb{E}_P[X_b] = 2$$

$$\mathbb{E}_P[X_s] = 4$$

↳ expectation value (of a continuous r.v.)

$$\mathbb{E}_P[X] = \int X(\omega) \, dP(\omega) = \int_{-\infty}^{\infty} x \, p(x) \, dx$$

integration measure

in this course

probability density fct.

$p(x) \geq 0 \; \forall x$

$\int_{-\infty}^{\infty} p(x) \, dx = 1$

$p(x)$ can be > 1

↳ prob. dist. of a r.v.: • $P(X = x) = P(\{\omega \in \Omega : X(\omega) = x\})$

• $P(X): \mathbb{R} \mapsto [0, 1]$, $x \mapsto P(X = x)$

"the fct. $P(X = \cdot)$"

Properties of two r.v.s

↳ joint probability distribution X & Y:

$$P(X=x,\ Y=y) = P(\{\omega\in\Omega : X(\omega)=x \wedge Y(\omega)=y\})$$

$$P(X,Y): \mathbb{R}\times\mathbb{R} \mapsto [0,1],\quad (x,y)\mapsto P(X=x,\ Y=y)$$

↳ Def: 2 r.v.s X & Y are (statistically) independent iff:

$$P(X,Y) = P(X)\,P(Y)$$

"marginal distribution"

(i.e.: $P(X=x,\ Y=y) = P(X=x)\,P(Y=y)\quad \forall x,y$)

- X_r, X_b are independent
- X_r, X_s are not independent:

e.g.: $P(X_r=1,\ X_s=3) = P(\{(1,2)\}) = \frac{1}{9}$

but $P(X_r=1)\,P(X_s=3) = \frac{1}{3}\times \underbrace{P(\{(1,2),\ (2,1)\})}_{2/9} = \frac{2}{27} \neq \frac{1}{9}$

Conditional Probability Distribution

↳ for events: "conditional prob. of event E_2 given event E_1"

$$P(E_2\,|\,E_1) = \frac{P(E_1\cap E_2)}{P(E_1)}$$

$$\left(\Rightarrow P(\underbrace{\neg E_2}_{\neg E_2 = \Omega\setminus E_2}\,|\,E_1) = \frac{P(E_2\setminus E_1)}{P(E_1)} \Rightarrow P(E_2|E_1)+P(\neg E_2|E_1)=1\right)$$

↳ for r.v.: $P(Y\,|\,X) = \dfrac{P(X,Y)}{P(Y)}$

$$P(Y=y\,|\,X=x) = \frac{P(X=x,\ Y=y)}{P(Y=y)}$$

"What is the prob. of Y being y if I already know that $X = x$?"

→ if X, Y are indep: $P(X,Y) = P(X)P(Y)$

$$\Rightarrow P(Y|X) = \frac{P(X,Y)}{P(X)} \underset{\uparrow}{=} \frac{P(X)P(Y)}{P(X)} = P(Y)$$

for indep. r.v.s

Important: writing $P(Y|X)$ does not imply causality, i.e. it does not mean that X is the cause of Y.

→ even if X is the cause of Y, we can still calculate:

$$P(X|Y) = \frac{P(X,Y)}{P(Y)} = \frac{P(X)P(Y|X)}{\sum_{x'} P(X=x')P(Y|X=x')}$$

"Bayesian inference"

Chain rule of probabilities: (follows directly from def. of cond. prob.)

$$P(X,Y) = P(X)P(Y|X) = P(Y)P(X|Y)$$

$$P(X,Y,Z) = P(X)P(Y|X)P(Z|X,Y) = \ldots$$

Back to source coding

Problem 3.2: You'll implement a compression method for natural language; it will exploit correlations between symbols (chars).

message: $\underline{X} = (X_1, X_2, X_3, \ldots, X_k)$

$$P(\underline{X}) \underset{\uparrow}{=} P(X_1)P(X_2|X_1)P(X_3|X_1,X_2)\cdots P(X_k|X_1,X_2 \ldots X_{k-1})$$

chain rule (always correct)

Issue: this general (exact) factorization of the joint distribution is not computationally feasible: $P(X_k \mid X_1, X_2, \dots, X_{k-1})$ is an extremely complicated fct.

→ need ways to:

capture relevant correlations while still maintaining
· compact model representation
· computational efficiency

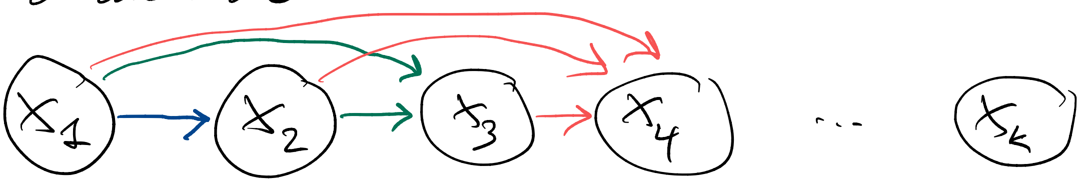
→ general strategy: enforce conditional independence:

for r.v.s X, Y, Z: Y & Z are cond. indep given X iff:

$$P(Z \mid X, Y) = P(Z \mid X)$$

(exercise: $\Leftrightarrow P(Y, Z \mid X) = P(Y \mid X)\, P(Z \mid X)$)

· general chain rule:

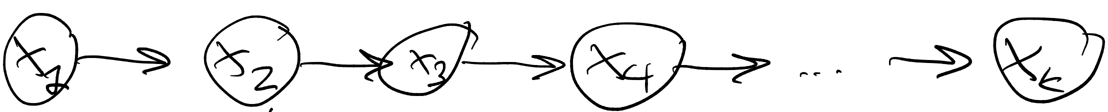


$$P(\underline{X}) = P(X_1)\, P(X_2 \mid X_1)\, P(X_3 \mid X_1, X_2)\, P(X_4 \mid X_1, X_2, X_3) \dots$$

3 possible simplifications

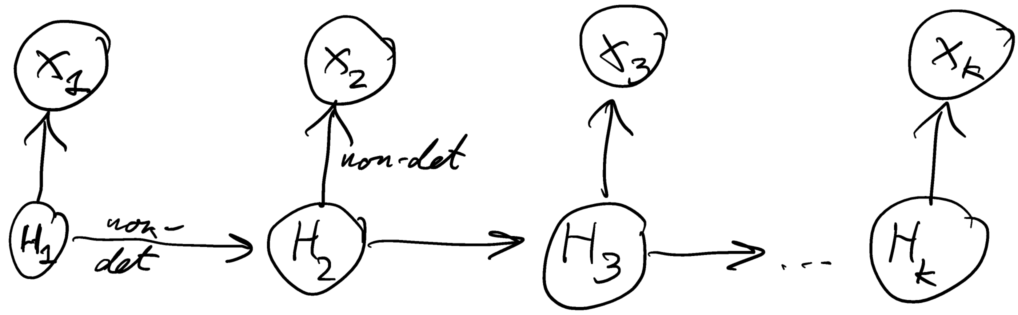
ⓐ if X_i are generated in sequence by some memoryless process → Markov Process

→ assumes cond. indep. of X_i & all X_j with $j < i-1$ given X_{i-1}



$$P(\underline{X}) = \prod_{i=1}^{k} P(X_i \mid X_{i-1})$$

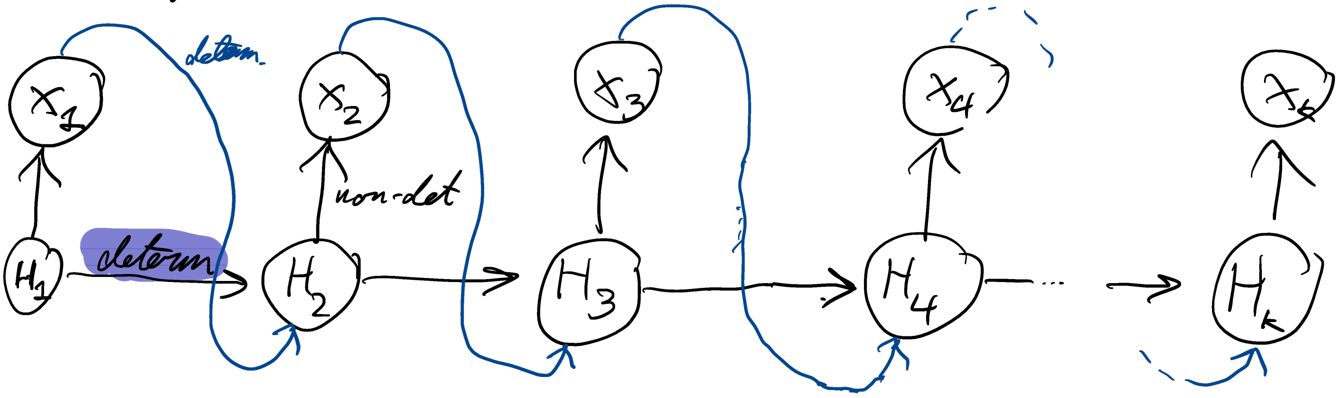
(b) Hidden Markov Model



(exercise: this can capture long-range correlations
→ in a model like this X_1 & X_3 do not have to be cond. indep given X_2)

→ difficult for compression (e.g. using bits-back coding)

(c) Autoregressive Model

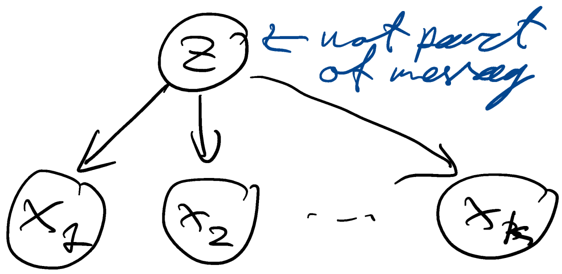


H_{i+1} is a determ fct. of H_i, X_i

☺ can capture long-range correlations
(i.e. X_2, X_4 are not cond. indep given X_3)

☹ hard to parallelize

→ next video: latent variable models



☺ can capture correlations between X_i's

☺ can be parallelized

☹ how to use this for compression → bits-back coding