Lecture 4, Part 1:

A Primer on Probability Theory

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/`.

Recap: Why We Need Good Probabilistic Models

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- Bound on practical compression performance: cross entropy

$$\boxed{\text{expected bit rate} \geq H\big(p_{\text{data}}(\mathbf{x}), p_{\text{model}}(\mathbf{x})\big) := -\sum_{\mathbf{x}} p_{\text{data}}(\mathbf{x}) \log p_{\text{model}}(\mathbf{x})}$$

- Overhead due to $p_{\text{model}} \neq p_{\text{data}}$: Kullback-Leibler divergence (aka relative entropy)

$$\boxed{D_{\text{KL}}\big(p_{\text{data}}(\mathbf{x}) \,\big\|\, p_{\text{model}}(\mathbf{x})\big) := H\big(p_{\text{data}}(\mathbf{x}), p_{\text{model}}(\mathbf{x})\big) - H\big[p_{\text{data}}(\mathbf{x})\big]}$$

  - For low overhead, we need p_{model} to approximate p_{data}
  - But so far: only simplistic p_{model}s that ignore *correlations* between symbols
- **This part:** mathematical language for probabilistic models
- **Next part:** information-theoretical quantification of correlations
- **Then:** machine learning models that describe correlations

Ingredients of a Probabilistic Model

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- *sample space* Ω (abstract space of "all states of the world")
  - subsets $E \subseteq \Omega$: *"events"* ("event E occurs" $\iff$ "the world is in some state $\omega \in E$")
- *probability measure:* a function $P : \Sigma \to [0, 1]$ where

essentially means that all statements below are well defined

  - Σ is a so-called $\boxed{\sigma\text{-algebra}}$ on Ω. (a set of all "expressible" events $E \subseteq \Omega$)
  - $P(\emptyset) = 0$ and $P(\Omega) = 1$.
  - countable additivity: $P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$ if all E_i are pairwise disjoint.

since we can set $E_i = \emptyset \;\forall i > k$

  - therefore, for finite sums: $P\left(\bigcup_{i=1}^{k} E_i\right) = \sum_{i=1}^{k} P(E_i)$ if all E_i are pairwise disjoint.
  - therefore: $P(E) + P(\Omega \setminus E) = P(\Omega) = 1 \quad \forall E \in \Sigma$.
  - therefore: $P(E_1) \leq P(E_2)$ if $E_1 \subseteq E_2$ (and $E_1, E_2 \in \Sigma$)

since $E_2 = E_1 \cup (E_2 \setminus E_1)$ if $E_2 \supseteq E_1$

Examples of Probability Measures

1. Simplified Game of Monopoly: (throw two fair three-sided dice)
   - sample space: $\Omega = \{1,2,3\}^2 = \big\{(1,1),(1,2),(1,3),(2,1),(2,2),(2,3),(3,1),(3,2),(3,3)\big\}$
   - sigma algebra: $\Sigma = 2^\Omega := \big\{\text{all subsets of } \Omega \text{ (including } \emptyset \text{ and } \Omega)\big\}$
   - probability measure P: for all $E \subseteq \Sigma$, let $P(E) := |E|/|\Omega| = |E|/9$

     reflects our assumption that all $\omega \in \Omega$ have equal probability since the dice are fair

Examples of Probability Measures (cont'd)

1. Simplified Game of Monopoly
2. Wait times for the next three buses from "Sternwarte": (e.g., measured in minutes)
   - sample space (in a simple model): $\Omega = \{(x_1, x_2, x_3) \in \mathbb{R}^3 \quad \text{where} \quad 0 \leq x_1 \leq x_2 \leq x_3\}$
   - sigma algebra: all "measurable subsets" of Ω
     (essentially, all subsets of Ω except for extremely pathological exceptions)
   - probability measure P: complicated function, but we know it satisfies certain relations, e.g.,
     $$P(\text{"next bus departs in at most 5 minutes"}) = P(\text{"next bus departs in at most 2 minutes"}) + P(\text{"next bus departs in between 2 and 5 minutes"}).$$
   - **Question:** what is the probability that the next bus departs in *exactly* 3 minutes?
     i.e., what is $P\big((\{3 \text{ min}\} \times \mathbb{R}^2) \cap \Omega\big)\cancel{?} = 0$ ← The probability should be continuous ⇒ if $P((\{3\text{min}\}\times\mathbb{R}^2)\cap\Omega) =: p > 0$ then $\exists \varepsilon > 0$ s.t. $P((\{x\}\times\mathbb{R}^2)\cap\Omega) \geq \frac{p}{2}$ $\forall x \in [3\text{min}-\varepsilon, 3\text{min}+\varepsilon] =: \mathcal{I}$
     ⇒ we can pick some integer $n > 2/p$ and n distinct points $x_i \in \mathcal{I}$
     $\Rightarrow P\big((\bigcup_{i=1}^n \{x_i\}\times\mathbb{R}^2)\cap\Omega\big) = \sum_{i=1}^n P((\{x_i\}\times\mathbb{R}^2)\cap\Omega) \geq n\frac{p}{2} > \frac{2}{p}\frac{p}{2} = 1 = P(\Omega)$
     ⇒ violates axioms.
   - **Question:** what is the probability that the next bus departs in between 2 and 5 minutes?
     $$P\big((\underbrace{[2 \text{ min}, 5 \text{ min}]}_{=:\mathcal{I}} \times \mathbb{R}^2) \cap \Omega\big) = P\Big(\bigcup_{x_1 \in \mathcal{I}} \big((\{x_1\} \times \mathbb{R}^2) \cap \Omega\big)\Big) \cancel{\overset{?}{=}} \sum_{x_1 \in \mathcal{I}} P\big((\{x_1\} \times \mathbb{R}^2) \cap \Omega\big) = 0$$
     argument not applicable since $\mathcal{I}$ is not countable

Random Variables

- Often, we we're not interested in a *full* description of the state $\omega \in \Omega$, but only in certain properties of it.
- **Definition:** "random variable": function $X : \Omega \to \mathbb{R}$ (not necessarily injective)

  can also be a different value space, e.g., $\mathbb{R}^d$ for some integer d, or $\mathfrak{X}^k$ for some alphabet $\mathfrak{X}$ & message length k

Examples:

1. Simplified Game of Monopoly; $\Omega = \big\{(a,b) \quad \text{where} \quad a, b \in \{1,2,3\}\big\}$
   - total value: $X_{\text{sum}}\big((a,b)\big) = a + b \in \{2,3,4,5,6\}$
   - value of the red die: $X_{\text{red}}\big((a,b)\big) = a$
   - value of the blue die: $X_{\text{blue}}\big((a,b)\big) = b$
2. In our bus schedule model from before; $\Omega = \{(x_1,x_2,x_3) \in \mathbb{R}^3 \quad \text{where} \quad 0 \leq x_1 \leq x_2 \leq x_3\}$
   - Time between the next bus and the one after it: $X_{\text{gap}}\big((x_1,x_2,x_3)\big) = x_2 - x_1$

Properties of Individual Random Variables

- "Probability that a random variable X has some given value x":
  $P(X = x) := P\big(X^{-1}(x)\big) = P\big(\{\omega \in \Omega : X(\omega) = x\}\big)$
  - Example 1 (Simplified Game of Monopoly): $P(X_{\text{sum}} = 3) =$
  - Example 2 (bus schedule): $P(X_{\text{gap}} = 20 \text{ minutes}) =$
- When we write just $P(X)$, then we mean the *function* that maps $x \mapsto P(X = x)$. in the discrete case
  (more precisely: $P(X)$ denotes a *probability measure* on the space of X) i.e., the measure that assigns probability $P(X\in E)$ to an event $E \subseteq X(\Omega)$ (defined on next slide)
- **Expectation value** of a random variable X under a model P
  - discrete case: $\mathbb{E}_P[X] := \sum_{\omega\in\Omega} P(\{\omega\})\, X(\omega) = \sum_{x\in X(\Omega)} P(X{=}x)\, x$

    more generally: $\mathbb{E}_P[f(X)] = \begin{cases} \sum_{x\in X(\Omega)} P(X=x) f(x) \\ \int_\Omega f(\omega)\, dP(\omega) \end{cases}$

    examples: $\mathbb{E}_P[X_{\text{red}}] = 2$; $\mathbb{E}_P[X_{\text{blue}}] = 2$; $\mathbb{E}_P[X_{\text{sum}}] = 4 = \mathbb{E}_P[X_{red} + X_{blue}] = \mathbb{E}_P[X_{red}] + \mathbb{E}_P[X_{blue}]$

    "$\mathbb{E}$" is a linear operation
  - continuous case: $\mathbb{E}_P[X] := \int_\Omega X(\omega)\, dP(\omega)$ (see next slide)

Properties of Individual Random Variables (cont'd)

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- *Cumulative Density Function (CDF):* $P(X \le x) := P\big(\{\omega \in \Omega : X(\omega) \le x\}\big)$
  - Example 1 (Simplified Game of Monopoly): $P(X_{\text{sum}} \le 3) = P(X_{sum}=2) + P(X_{sum}=3) = \frac{1}{9} + \frac{2}{9} = \frac{3}{9} = \frac{1}{3}$
  - Example 2 (bus schedule): $P(X_{\text{gap}} \le 20 \text{ minutes}) \in [0, 1]$ (nonzero in general)
- Analogous definitions for: $P(X < x)$, $P(X \ge x)$, $P(X > x)$, $P(X \in \text{some set})$, ...
- *Probability Density Function (PDF)* of a real-valued random variable X: (in 1 dimension)
  $p(x) := \frac{d}{dx} P(X \le x)$ (if derivative exists)
  $\to$ expectation value: $\mathbb{E}_P[X] = \int X(\omega)\, dP(\omega) = \int_{-\infty}^{\infty} x\, p(x)\, dx$
  (if a density $p(x)$ exists)

More general definition of a PDF (also for higher dimensions):
p is a PDF of P if $\mathbb{E}_P[f(X)] = \int p(x) f(x)\, dx$ for all ("measurable") functions f

Multiple Random Variables

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- **Definition:** *joint probability distribution* of two random variables X and Y:
  $P(X{=}x, Y{=}y) := P\big(\{\omega \in \Omega : X(\omega) = x \ \wedge\ Y(\omega) = y\}\big)$
- **Notation:** "$P(X, Y)$": function that maps $(x, y) \mapsto P(X{=}x, Y{=}y)$
  (more precisely: $P(X, Y)$ denotes a *probability measure* on the product space of X and Y)

  again, more precisely: $P(X,Y)$ denotes the probability measure on the product space $X(\Omega)\times Y(\Omega)$ that assigns probability $P((X,Y)\in E)$ to an event $E \subseteq X(\Omega)\times Y(\Omega)$
- If we know $P(X, Y)$, then we can calculate $P(X) = \sum_y P(X, Y{=}y)$ (for discrete Y)

  i.e. $Y(\Omega)$ is finite or countably infinite

  $$\begin{aligned} \forall x \in X(\Omega):\ P(X=x) &= P(\{\omega\in\Omega: X(\omega)=x\}) \\ &= P\Big(\bigcup_y \{\omega\in\Omega: X(\omega)=x \wedge Y(\omega)=y\}\Big) \\ &= \sum_y P(\{\omega\in\Omega: X(\omega)=x \wedge Y(\omega)=y\}) \\ &= \sum_y P(X=x, Y=y) \end{aligned}$$

  $\Rightarrow$ in short, we write: $P(X) = \sum_y P(X, Y=y)$
  - This process is called "marginalization".
  - for continuous random variables: $p(X) = \int p(X, y)\, dy$

Statistical Independence

- **Definition:** X and Y are *(statistically) independent* iff: $P(X, Y) = P(X)\,P(Y)$
  (i.e., if $P(X \in \mathbb{X}, Y \in \mathbb{Y}) = P(X \in \mathbb{X})\,P(Y \in \mathbb{Y})\ \ \forall \mathbb{X}, \mathbb{Y}$)
- **Examples** (Simplified Game of Monopoly):
  - X_{red} and X_{blue} are statistically independent.
  - X_{red} and X_{sum} are *not* statistically independent. (proof: Problem 4.1)
- **Definition:** *conditional* independence of X and Y given Z: see later

Conditional Probability Distributions: Examples

"X & Y are *not* statistically independent" $\iff$ "knowing X reveals something about Y"

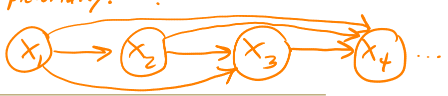
Examples: (Simplified Game of Monopoly; $P(E) = \frac{|E|}{9}$)

		1	2	3	4	5	6
	$x =$	1	2	3	4	5	6
What are the (marginal) probability distributions $P(X_{\text{red}})$ and $P(X_{\text{sum}})$ of the red die and the sum, respectively?	$P(X_{\text{red}} = x) =$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0	0
	$P(X_{\text{sum}} = x) =$	0	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{1}{3}$	$\frac{2}{9}$	$\frac{1}{9}$
Assume that you only accept throws where the red die comes up with value 1, and you keep rethrowing both dice until this condition is satisfied. What is the probability distribution of X_{sum} in your first accepted throw? We call this the *conditional* probability distribution $P(X_{\text{sum}} \mid X_{\text{red}} = 1)$.	$P(X_{\text{sum}} = x \mid X_{\text{red}} = 1) =$	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0
Now you only accept throws where the sum of both dies is at least 5. What is the conditional probability distribution of X_{red}?	$P(X_{\text{red}} = x \mid X_{\text{sum}} \geq 5) =$	0	$1/3$	$2/3$	0	0	0
Finally, assume you only accept throws where $X_{\text{blue}} = 1$. What is the conditional probability distribution of X_{red}?	$P(X_{\text{red}} = x \mid X_{\text{blue}} = 1) =$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0	0

Conditional Probability Distributions: Definition

- **Definition:** "conditional probability of event E_2 *given* event E_1": $P(E_2 \mid E_1) := \frac{P(E_1 \cap E_2)}{P(E_1)}$
  - Thus, $P(E_2 \mid E_1)$ is a (properly normalized) probability distribution w.r.t. the first parameter,
    i.e., $P(E_2 \mid E_1) + P(\Omega \setminus E_2 \mid E_1) = \dfrac{P(E_2 \cap E_1) + P((\Omega \setminus E_2) \cap E_1)}{P(E_1)} = \dfrac{P(E_1)}{P(E_1)} = 1.$
- **Definition:** "conditional probability distribution of a random variable Y *given* another random variable X": $P(Y \mid X) := \frac{P(X,Y)}{P(X)}$ i.e., $P(Y = y \mid X = x) := \frac{P(X=x, Y=y)}{P(X=x)}\ \ \forall x, y$
  - Thus, if X and Y are statistically independent *(but only then!)*:
    $P(Y \mid X) = \frac{P(X,Y)}{P(X)} = \frac{P(X)\,P(Y)}{P(X)} = P(Y)$ ("knowing X reveals no new information about Y")
  - In the general case: "chain rule" of probability theory: (follows directly from above definition)
    $P(X_1, X_2, X_3, \ldots) = \underbrace{P(X_1)\ P(X_2 \mid X_1)}_{= P(X_1, X_2)}\ P(X_3 \mid X_1, X_2) \ldots$

pictorially:



Conditional Independence

- **Reminder:** X and Z are (statistically) independent $:\Longleftrightarrow\ P(X,Z) = P(X)\,P(Z)$
- **Analogous definition:**

  $\boxed{X \text{ and } Z \text{ are } \textit{conditionally} \text{ independent } \textit{given } Y \;:\Longleftrightarrow\; P(X,Z\,|\,Y) = P(X\,|\,Y)\,P(Z\,|\,Y)}$

  - **equivalently:** chain rule simplifies:
    $P(X,Y,Z) = P(X)\,P(Y\,|\,X)\,P(Z\,|\,X,Y) = P(Y)\,P(X\,|\,Y)\,P(Z\,|\,Y)$

    $$X \longrightarrow Y \longrightarrow Z$$

    edge does not exist if X, Z are cond. indep. given Y

- **Problem Set 5:** comparison to normal (i.e., unconditional) independence
- **Problem Set 10:** propagation of information along $X \longrightarrow Y \longrightarrow Z$ ("data processing inequality")

Warning: Conditionality $\neq$ Causation

- We'll often specify a joint probability distribution as, e.g., $P(X,Y) = P(X)\,P(Y\,|\,X)$.
- But just because we write "$P(Y\,|\,X)$", this does *not* necessarily mean that X is the cause of Y.
- **Example:** (Simplified Game of Monopoly):
  - X_{red} and X_{blue} can be considered to cause X_{sum}.
  - But, in the examples three slides ago, we were still able to calculate, e.g., $P(X_{\text{red}}\,|\,X_{\text{sum}})$.
    (i.e., the probability of the *cause* X_{red} given its *effect* X_{sum})

  $$P(X_{red}\,|\,X_{sum}) = \frac{P(X_{red}, X_{sum})}{P(X_{sum})} = \frac{P(X_{red}, X_{sum})}{\sum_{x'} P(X_{red}=x', X_{sum})} = \frac{P(X_{red})\,P(X_{sum}\,|\,X_{red})}{\sum_{x'} P(X_{red}=x')\,P(X_{sum}\,|\,X_{red}=x')}$$

  $\rightarrow$ This is called "posterior inference". (more in Lectures 7 and 8) (or "Bayesian inference")
- Causality goes beyond the scope of a probabilistic model;
  understanding causal structures generally requires *interventions* in the generative process.

Outlook

- **Problem 4.1:** probability measures & statistical independence
- **Next part:**
  - information-theoretical quantification of *correlations*
  - machine-learning models that can capture correlations

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

Lecture 4, Part 2:

Mutual Information and Taxonomy of Probabilistic (Machine-Learning) Models

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/`.

Recap: Random Variables, Conditional Probabilities

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- **Random variables**: (uppercase letters $X, Y, Z, \ldots$)
  - Think "placeholders" for values: $P(X_i)$ is a probability measure for symbol X_i.
  - $P(X{=}x)$: probability ($\in [0,1]$) that the random variable X assumes value x.
  - *Expectation value:* $\mathbb{E}_P[f(X)] = \sum_x P(X{=}x)\, f(x)$ (discrete case)
- **Multiple random variables:**
  - *joint distribution:* $P(X, Y)$ $P(X) = \sum_y P(X, Y{=}y)$
  - *marginal distributions:* $P(X)$, $P(Y)$ $\leftarrow P(Y) = \sum_x P(X{=}x, Y)$
  - *conditional distribution:* $P(Y \mid X) = \frac{P(X,Y)}{P(X)}$ ("How is Y distributed if I know the value of X?")
- **Statistical (in-)dependencies between random variables:**
  i.e., knowing X does not tell me anything about Y
  - *(unconditional) (statistical) independence:* if $P(X, Y) = P(X)\,P(Y)$ $(\Longleftrightarrow P(Y \mid X) = P(Y))$
  - *conditional independence:* if $P(X, Z \mid Y) = P(X \mid Y)\,P(Y \mid Y)$ $(\Longleftrightarrow P(Z \mid X, Y) = P(Z \mid Y))$
  i.e., if I already know Y, then the additional knowledge of X does not tell me anything about Z
- **Goal now:** *quantify* statistical dependencies

Quantification of Statistical Dependencies

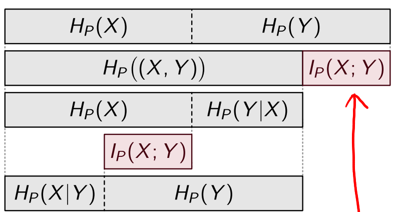
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

- **Use information theory:**
  - information content of the statement "$X = x$": $-\log_2 P(X{=}x)$
  - entropy of a random variable X under a model P: $H_P(X) := \mathbb{E}_P[-\log_2 P(X{=}x)]$
  - analogously: joint and conditional information content and entropy (see Problems 4.2 and 4.3).
  $\rightarrow$ definitions are as you'd expect, but properties are somewhat subtle
- **Entropy is *subadditive*:** $\forall$ random variables X & Y:

  optimal expected bit rate

  $$\boxed{H_P((X,Y)) \leq H_P(X) + H_P(Y)}$$ (proof: Problem 4.4)

  optimal expected bit rate if we ignore correlations between X&Y (→ Problem 5.2(a))
  - equality holds iff X and Y are statistically independent (proof: Problem 2.3 (b))
- **Thus:** *wrongfully* assuming independence (to simplify a model) leads to an overhead in bit rate:



(figure adapted from MacKay book)

Def. *"mutual information"*: $\boxed{I_P(X;Y) := H_P(X) + H_P(Y) - H_P((X,Y)) \geq 0}$ (Problem 4.4)

Modeling Statistical Dependencies

- Assume that the message is a sequence of symbols: $\mathbf{X} = (X_1, X_2, \ldots, X_k)$
- Subadditivity of entropies: $H(\mathbf{X}) \leq \sum_{i=1}^{k} H(X_i)$

optimal expected bit rate if we use a perfect model

optimal expected bit rate if we model the symbols as being statistically independent (proof: problem 5.2 (a))

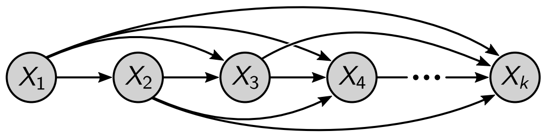
- **Thus:** instead of modeling each symbol X_i independently, we should model the message $\mathbf{X}$ *as a whole* (without completely sacrificing computational efficiency).
  - autoregressive models (e.g., Problem 3.3)
  - latent variable models (~~planned for Problem Set 6;~~ also: basis for variational autoencoders)

    Lectures 7+8 ↖ Lecture 9

Probabilistic Models at Scale

- *All* probability distributions P over messages $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ satisfy the *chain rule:*

  $P(\mathbf{X}) = P(X_1)\,P(X_2 \,|\, X_1)\,P(X_3 \,|\, X_1, X_2)\,P(X_4 \,|\, X_1, X_2, X_3) \cdots P(X_k \,|\, X_1, X_2, \ldots, X_{k-1})$



- **Example:** assume each symbol is from alphabet $\mathfrak{X} = \{1, 2, 3\}$.
  - How many model parameters do we need to specify an arbitrary distribution $P(X_1)$? $\rightarrow 2$

    (one per symbol $x \in \mathfrak{X}$ to specify $P(X=x)$, but $P(X=3) = 1 - P(X=1) - P(X=2)$ can be inferred from normalization)
  - How many parameters for an arbitrary conditional distribution $P(X_2 \,|\, X_1)$? $\rightarrow 2 \times 3 = 6$

    2 parameters as above per distribution $P(X_2 \,|\, X_1 = x_1)$ $\forall x_1 \in \mathfrak{X}$
  - How many parameters for an arbitrary conditional distribution $P(X_k \,|\, X_1, X_2, \ldots, X_k)$? $\rightarrow O(|\mathfrak{X}|^k)$

    EXPONENTIAL!

Expressive Yet Efficient Probabilistic Models

- **Goal:** Find approximation to arbitrary models $P(\mathbf{X})$ that
  - captures relevant correlations
  - but is still *computationally efficient*:

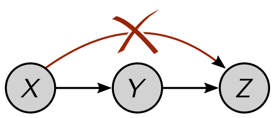
    $\rightarrow$ *reasonably compact* representation of the model in memory

    $\rightarrow$ *reasonably efficient evaluation* of probabilities $P(\mathbf{X} = \mathbf{x})$

    $\rightarrow$ suitable for entropy coding (later)
- **General Strategy:** enforce *conditional independence*:

  X & Z *are conditionally independent given* Y $:\Longleftrightarrow P(X, Z \,|\, Y) = P(X \,|\, Y)\,P(Z \,|\, Y)$

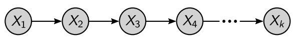
  $\Longleftrightarrow P(X, Y, Z) = P(X)\,P(Y \,|\, X)\,P(Z \,|\, Y)$ (proof: Problem 5.1 (a))



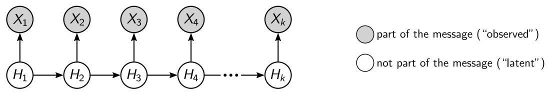
Four Kinds of Scalable Probabilistic Models

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

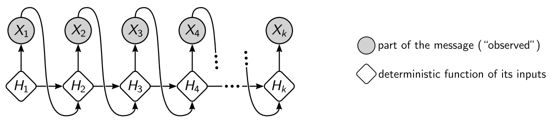
(1) Markov Process



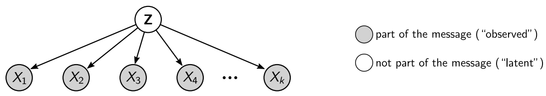
(2) Hidden Markov Model



(3) Autoregressive Model



(4) Latent Variable Model

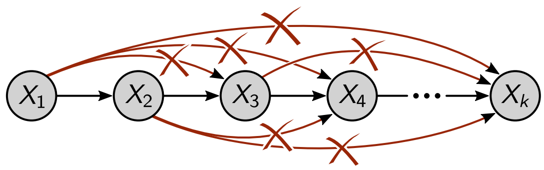


(1) Markov Process

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Modeling assumption: symbols X_i are generated by a *memoryless* process.

- Each symbol X_i depends on its immediate precessor X_{i-1} but not on any earlier symbols:



$$\boxed{P(\mathbf{X}) = P(X_1)\, P(X_2 \,|\, X_1)\, P(X_3 \,|\, X_2)\, P(X_4 \,|\, X_3) \cdots P(X_k \,|\, X_{k-1})}$$

- i.e., for all $j < i$, the symbols X_{i+1} and X_j are *conditionally independent* given X_i.

☺ only $O(k\,|\mathfrak{X}|^2)$ (or even $O(|\mathfrak{X}|^2)$) model parameters; if we use the same conditional distribution $P(X_{i+1}|X_i)$ $\forall i$

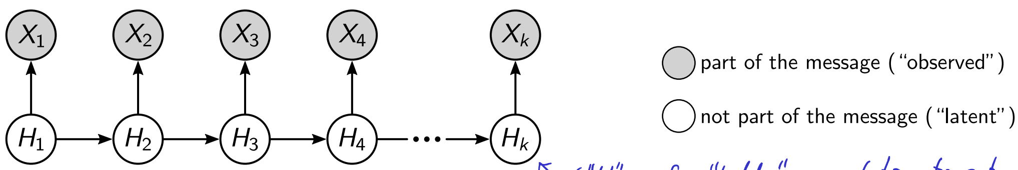
☹ simplistic assumption; e.g., in English text, the string "the" is very frequent.

$\Rightarrow P_{\text{data}}(X_i = \text{'e'} \,|\, X_{i-2} = \text{'t'}, X_{i-1} = \text{'h'}) > P_{\text{data}}(X_i = \text{'e'} \,|\, X_{i-1} = \text{'h'})$ (i.e., *not* cond. indep.)

(2) Hidden Markov Model

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Modeling assumption: there is some memoryless *hidden* process, which is observed *indirectly*.



("H" is for "hidden"; no relation to entropy)

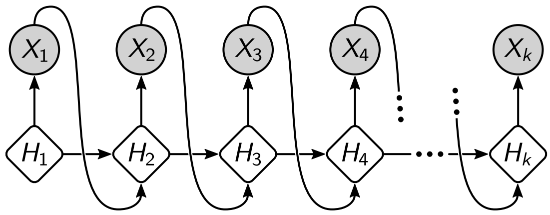
$$\boxed{P(\mathbf{X}) = \int P(\mathbf{X}, \mathbf{H})\, d\mathbf{H} \quad \text{with} \quad P(\mathbf{X}, \mathbf{H}) = P(H_1)\, P(X_1 \,|\, H_1) \prod_{i=2}^{k} P(H_i \,|\, H_{i-1})\, P(X_i \,|\, H_i)}$$

☺ can model long-range correlations, i.e., X_i, X_{i-2} *not* cond. indep. given X_{i-1} (exercise);

☹ bit-rate overhead: in order to model $P(X_i \,|\, H_i)$, decoder has to first decode H_i, even though it's not part of the message (solution: "bits-back coding", see Lecture 7).

(3) Autoregressive Model → see Problem 3.2

Modeling assumption: memoryless hidden process with (typically) *deterministic* transitions;
but: transitions are also conditioned on the previous *symbol*. ← otherwise model would be fully factorized



part of the message ("observed")

deterministic function of its inputs

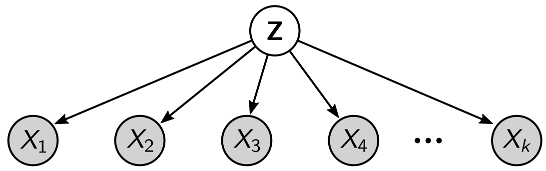
$$P(\mathbf{X}) = \prod_{i=1}^{k} P(X_i \,|\, H_i) \quad \text{where} \quad H_1 = \text{fixed}; \quad H_i = f(H_{i-1}, X_{i-1})$$

☺ no compression overhead for reconstructing H_i (see Problem 3.3);

☹ encoding & decoding are not parallelizable (⇒ slow on modern hardware). ← (also an issue for Markov chains & hidden Markov models)

(4) Latent Variable Model

Modeling assumption: there is some unobserved higher level of abstraction **Z**.



part of the message ("observed")

not part of the message ("latent")

$$P(\mathbf{X}) = \int P(\mathbf{X}, \mathbf{Z})\, d\mathbf{Z} \quad \text{where} \quad P(\mathbf{X}, \mathbf{Z}) = P(\mathbf{Z}) \prod_{i=1}^{k} P(X_i \,|\, \mathbf{Z})$$

☺ can model long-range correlations (see Problem 5.2 (c));

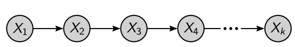
☺ parallelizable;

☹ bit-rate overhead for encoding **Z** (solution: "bits-back coding", see Lecture 7).

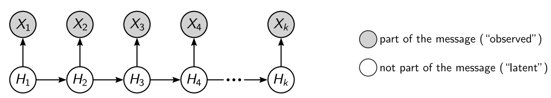
Summary: 4 Kinds of Scalable Probabilistic Models

- Each architecture makes different assumption about conditional independence of symbols.

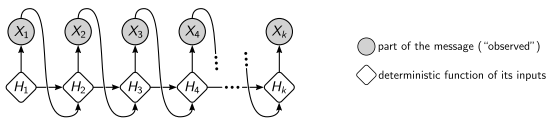
(1) Markov Process



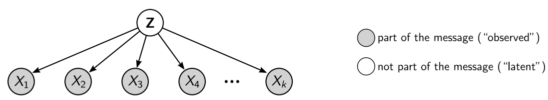
(2) Hidden Markov Model



(3) Autoregressive Model

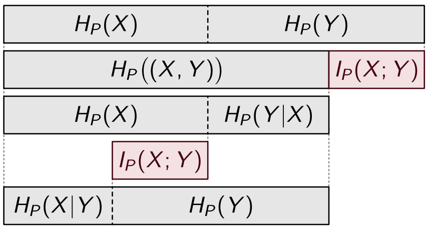


(4) Latent Variable Model



Outlook

- **Problem Set 4:**



- **Now and next 3 lectures:** lossless compression with deep probabilistic models
  - Different model architectures require different entropy coding algorithms.
- **Afterwards:** Lossy compression