



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



- ▶ Bound on practical compression performance: cross entropy

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

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

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

- ▶ For low overhead, we need p_{model} to approximate p_{data}
- ▶ But so far: only simplistic p_{models} 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



- ▶ *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 $P : \Sigma \rightarrow [0, 1]$ where
 - ▶ Σ is a so-called σ -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.
 - ▶ 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 $E_1, E_2 \in \Sigma$)

Examples of Probability Measures



1. Simplified Game of Monopoly: (throw two fair three-sided dice)

- ▶ sample space: $\Omega = \{1, 2, 3\}^2 = \{(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (2, 3), (3, 1), (3, 2), (3, 3)\}$
- ▶ sigma algebra: $\Sigma = 2^\Omega := \{\text{all subsets of } \Omega \text{ (including } \emptyset \text{ and } \Omega\}\}$
- ▶ 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": (o.g., measured in minutes)

- ▶ sample space (in a simple model): $\Omega = \{(x_1, x_2, x_3) \in \mathbb{R}^3 \text{ where } 0 \leq x_1 \leq x_2 \leq x_3\}$
- ▶ 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 $P(\{\{3 \text{ min}\} \times \mathbb{R}^2\} \cap \Omega)$?
*The probability should be continuous \Rightarrow if $P(\{3 \text{ min}\} \times \mathbb{R}^2) \cap \Omega) =: p > 0$
then $\exists \varepsilon > 0$ s.t. $P(\{x \in \mathbb{R}^2 \mid 3 - \varepsilon \leq x \leq 3 + \varepsilon\} \cap \Omega) \geq \frac{p}{2}$ $\forall x \in [3 - \varepsilon, 3 + \varepsilon]$ $=: \mathcal{I}$*
- ▶ **Question:** what is the probability that the next bus departs in between 2 and 5 minutes?
*We can pick some integer $n \geq 2/p$ and n distinct points $x_i \in \mathcal{I}$
 $\Rightarrow P(\{x_i \in \mathcal{I} \mid x_i \in \mathbb{R}^2\} \cap \Omega)$
 $= \sum_{i=1}^n P(\{x_i \in \mathcal{I} \mid x_i \in \mathbb{R}^2\} \cap \Omega)$
 $\geq n \cdot \frac{p}{2} > \frac{2}{2} \cdot \frac{p}{2} = 1 = P(\Omega)$
since \mathcal{I} is not countable
*argument not applicable**

Random Variables



- ▶ Often, 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 $X : \Omega \rightarrow \mathbb{R}$ (not necessarily injective)

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

Examples:

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

Properties of Individual Random Variables



- ▶ “Probability that a random variable X has some given value x ”: $P(X = x) := P(X^{-1}(x)) = P(\{\omega \in \Omega : X(\omega) = x\})$

▶ 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 $x \mapsto P(X = x)$.

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

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

► continuous case: $\mathbb{E}_P[X] := \int_{\Omega} X(\omega) dP(\omega)$ (see next slide)

Properties of Individual Random Variables (cont'd)



- ▶ **Cumulative Density Function (CDF)**: $P(X \leq x) := P(\{\omega \in \Omega : X(\omega) \leq x\})$

► Example 1 (Simplified Game of Monopoly): $P(X_{\text{sum}} \leq 3) = P(X_{\text{sum}}=2) + P(X_{\text{sum}}=3) = \frac{1}{9} + \frac{2}{9} = \frac{3}{9} = \frac{1}{3}$

► Example 2 (bus schedule): $P(X_{\text{gap}} \leq 20 \text{ minutes}) \in [0, 1]$ (nonzero in general)

- ▶ Analogous definitions for: $P(X < x)$, $P(X \geq x)$, $P(X > x)$, $P(X \in \text{some set})$, ...

- ▶ **Probability Density Function (PDF)** of a real-valued random variable X : (in 1 dimension)

$p(x) := \frac{d}{dx} P(X \leq x)$ (if derivative exists)

→ 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



- ▶ **Definition:** joint probability distribution of two random variables X and Y :

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

- ▶ **Notation:** “ $P(X, Y)$ ”: function that maps $(x, y) \mapsto P(X=x, Y=y)$ ↪ $P(X, Y)$ denotes the probability measure on the product space $X(\Omega) \times Y(\Omega)$ that assigns probability $P((X, Y) \in E)$ (more precisely: $P(X, Y)$ denotes a probability measure on the product space of X and Y) 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)

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

▶ 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 $P(X \in \mathbb{X}, Y \in \mathbb{Y}) = P(X \in \mathbb{X}) P(Y \in \mathbb{Y}) \quad \forall \mathbb{X}, \mathbb{Y}$)
- ▶ **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; $P(E) = \frac{|E|}{9}$)

What are the (marginal) probability distributions $P(X_{\text{red}})$ and $P(X_{\text{sum}})$ of the red die and the sum, respectively?

$x =$	1	2	3	4	5	6
$P(X_{\text{red}}=x) =$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	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 $P(X_{\text{sum}} | X_{\text{red}}=1)$.

$P(X_{\text{sum}}=x X_{\text{red}}=1) =$	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} ?

$P(X_{\text{red}}=x X_{\text{sum}} \geq 5) =$	0	$\frac{1}{3}$	$\frac{1}{3}$	0	0	0
---	---	---------------	---------------	---	---	---

Finally, assume you only accept throws where $X_{\text{blue}} = 1$. What is the conditional probability distribution of X_{red} ?

$P(X_{\text{red}}=x X_{\text{blue}}=1) =$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0	0
---	---------------	---------------	---------------	---	---	---

Conditional Probability Distributions: Definition



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

$$P(X_1, X_2, X_3, \dots) = \underbrace{P(X_1) P(X_2 | X_1) P(X_3 | X_1, X_2) \dots}_{= P(X_1, X_2)} \quad \text{pictorially: } \dots \rightarrow X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \rightarrow \dots$$

Conditional Independence

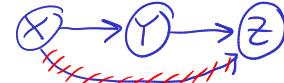
- ▶ **Reminder:** X and Z are (statistically) independent : $\iff P(X, Z) = P(X) P(Z)$

- ▶ **Analogous definition:**

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

- ▶ **equivalently:** chain rule simplifies:

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



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 $X \rightarrow Y \rightarrow Z$ ("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})

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

→ 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



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



- ▶ **Random variables:** (uppercase letters X, Y, Z, \dots)
 - ▶ 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: $P(X, Y) \quad \text{blue arrow} \quad P(X) = \sum_Y P(X, Y=y)$
 - ▶ marginal distributions: $P(X), P(Y) \quad \text{blue arrow} \quad P(Y) = \sum_X P(X=x, Y)$
 - ▶ conditional distribution: $P(Y | X) = \frac{P(X, Y)}{P(X)}$ (“How is Y distributed if I know the value of $X?$ ”)
 - i.e., knowing X does not tell me anything about Y*
- ▶ **Statistical (in-)dependencies between random variables:**
 - ▶ (unconditional) (statistical) independence: if $P(X, Y) = P(X) P(Y)$ ($\iff P(Y | X) = P(Y)$)
 - ▶ conditional independence: if $P(X, Z | Y) = P(X | Y) P(Z | Y)$ ($\iff P(Z | X, Y) = P(Z | Y)$)
 - 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



- ▶ **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).
 - definitions are as you'd expect, but proportions are somewhat subtle*
- ▶ **Entropy is subadditive:** \forall random variables X & Y :

$H_P((X, Y))$	$H_P(Y)$
$H_P((X, Y))$	$I_P(X; Y)$
$H_P(X)$	$H_P(Y X)$
$I_P(X; Y)$	
$H_P(X Y)$	$H_P(Y)$

 - ▶ 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:*

Def. “*mutual information*”: $I_P(X; Y) := H_P(X) + H_P(Y) - H_P((X, Y)) \geq 0$ (Problem 4.4)

$H_P(X)$	$H_P(Y)$
$H_P((X, Y))$	$I_P(X; Y)$
$H_P(X)$	$H_P(Y X)$
$I_P(X; Y)$	
$H_P(X Y)$	$H_P(Y)$

(figure adapted from MacKay book)

Modeling Statistical Dependencies



- ▶ Assume that the message is a sequence of symbols: $\mathbf{X} = (X_1, X_2, \dots, X_k)$
- ▶ Subadditivity of entropies: $\underbrace{H(\mathbf{X})}_{\text{optimal expected bit rate if we model the symbols as being statistically independent (proof: problem 5.2 (a))}} \leq \sum_{i=1}^k H(X_i)$
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

Robert Bamler - Lecture 4, Part 2 of the course "Data Compression With and Without Deep Probabilistic Models" - Summer Term of 2023 - more course materials at <https://robamler.github.io/teaching/compress23/>

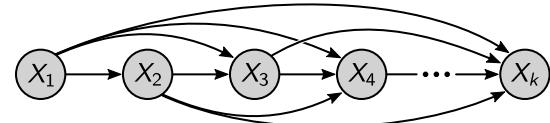
| 3

Probabilistic Models at Scale



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

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



- ▶ **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 $P(X_2 | X_1=x)$ $\forall x \in \mathfrak{X}$
- ▶ How many parameters for an arbitrary conditional distribution $P(X_k | X_1, X_2, \dots, X_{k-1})$? $\rightarrow O(1|\mathfrak{X}|^k)$
EXONENTIAL

Robert Bamler - Lecture 4, Part 2 of the course "Data Compression With and Without Deep Probabilistic Models" - Summer Term of 2023 - more course materials at <https://robamler.github.io/teaching/compress23/>

| 4

Expressive Yet Efficient Probabilistic Models



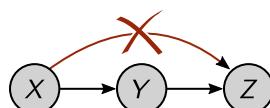
- ▶ **Goal:** Find approximation to arbitrary models $P(\mathbf{X})$ that

- ▶ captures relevant correlations
- ▶ but is still *computationally efficient*:
 - *reasonably compact* representation of the model in memory
 - *reasonably efficient evaluation* of probabilities $P(\mathbf{X} = \mathbf{x})$
 - suitable for entropy coding (later)

- ▶ **General Strategy:** enforce *conditional independence*:

$$X \text{ & } Z \text{ are } \textit{conditionally independent given } Y : \iff P(X, Z | Y) = P(X | Y) P(Z | Y)$$

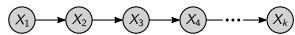
$$\iff P(X, Y, Z) = P(X) P(Y | X) P(Z | Y) \quad (\text{proof: Problem 5.1 (a)})$$



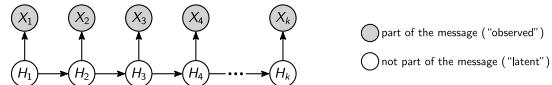
Four Kinds of Scalable Probabilistic Models



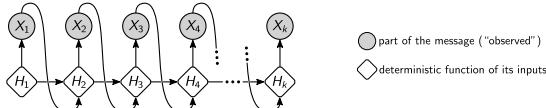
(1) Markov Process



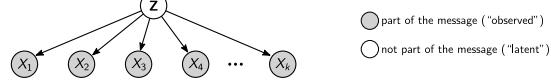
(2) Hidden Markov Model



(3) Autoregressive Model



(4) Latent Variable Model

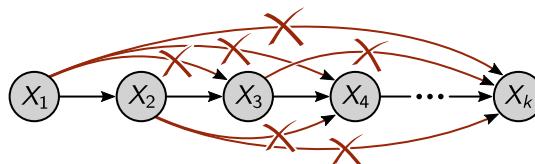


(1) Markov Process



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

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



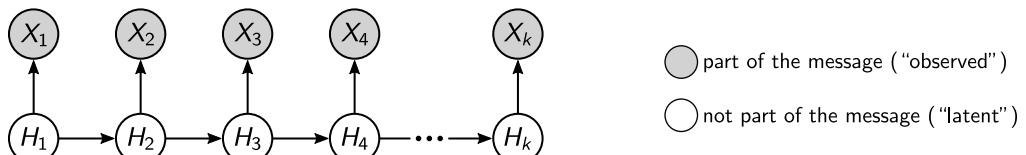
$$P(\mathbf{X}) = P(X_1) P(X_2 | X_1) P(X_3 | X_2) P(X_4 | X_3) \cdots P(X_k | X_{k-1})$$

- i.e., for all $j < i$, the symbols X_{i+1} and X_j are *conditionally independent* given X_i .
- ☺ only $O(k |\mathcal{X}|^2)$ (or even $O(|\mathcal{X}|^2)$) model parameters;
- ☹ simplistic assumption; e.g., in English text, the string "the" is very frequent.
 $\Rightarrow P_{\text{data}}(X_i = 'e' | X_{i-2} = 't', X_{i-1} = 'h') > P_{\text{data}}(X_i = 'e' | X_{i-1} = 'h')$ (i.e., *not cond. indep.*)

(2) Hidden Markov Model



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



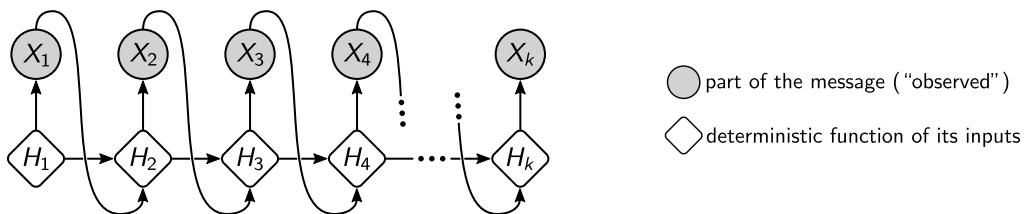
$$P(\mathbf{X}) = \int P(\mathbf{X}, \mathbf{H}) d\mathbf{H} \quad \text{with} \quad P(\mathbf{X}, \mathbf{H}) = P(H_1) P(X_1 | H_1) \prod_{i=2}^k P(H_i | H_{i-1}) P(X_i | H_i)$$

- ☺ 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 \rightarrow see Problem 3.2



Modeling assumption: memoryless hidden process with (typically) *deterministic* transitions;
 but: transitions are also conditioned on the previous symbol.



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

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

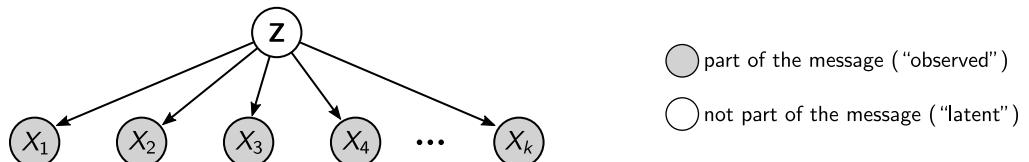
😢 encoding & decoding are not parallelizable (\Rightarrow slow on modern hardware).

(also an issue for
Markov chains & hidden
Markov models)

(4) Latent Variable Model



Modeling assumption: there is some unobserved higher level of abstraction \mathbf{Z} .



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

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

😊 parallelizable;

😢 bit-rate overhead for encoding \mathbf{Z} (solution: "bits-back coding", see Lecture 7).

Summary: 4 Kinds of Scalable Probabilistic Models

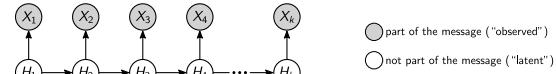


- ▶ Each architecture makes different assumption about conditional independence of symbols.

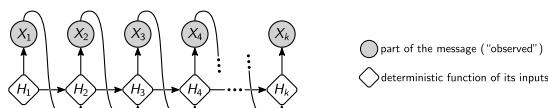
(1) Markov Process



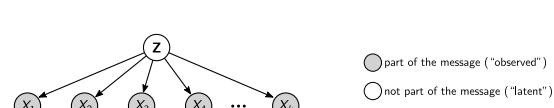
(2) Hidden Markov Model



(3) Autoregressive Model



(4) Latent Variable Model





► **Problem Set 4:**

$H_P(X)$	$H_P(Y)$
$H_P((X, Y))$	$I_P(X; Y)$
$H_P(X)$	$H_P(Y X)$
$I_P(X; Y)$	
$H_P(X Y)$	$H_P(Y)$

► **Now and next 3 lectures:** lossless compression with deep probabilistic models

- Different model architectures require different entropy coding algorithms.

► **Afterwards:** Lossy compression