Probability Theory, Mutual Information, and Taxonomy of Probabilistic models

Robert Bamler · 12 May 2022

This lecture is a part 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/

Recap From Last Week (1 of 3): Two probability distributions

- p_{data}: true probability distribution of the data generative process
  - typically unknown, i.e., we can't *evaluate* the true probability $p_{\text{data}}(\mathbf{x})$ of a given message $\mathbf{x}$;
  - but we may have access to a data set $\mathcal{D}$ of empirical samples from p_{data}.

    $\Rightarrow$ then we can *estimate* expectations under p_{data} as follows:

    $$\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}\big[f(\mathbf{x})\big] = \lim_{|\mathcal{D}|\to\infty} \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x}\in\mathcal{D}} f(\mathbf{x}) \qquad \text{(assuming i.i.d. samples and expectation exists)}$$
- p_{model}: probabilistic model of the data source
  - approximates p_{data};
  - let's assume, for now, that we can evaluate $p_{\text{data}}(\mathbf{x}) \in [0, 1]$ for any given message $\mathbf{x}$.

Recap From Last Week (2 of 3): Entropy vs. Cross Entropy

Consider the expected bitrate $\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[R_C(\mathbf{x})]$ of a lossless compression code C:

- **fundamental lower bound:** true entropy of the data source:

  $$\boxed{\text{Entropy: } H[p_{\text{data}}] = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[-\log_2 p_{\text{data}}(\mathbf{x})]}$$

  Intrinsic property of the data source (i.e., independent of our model).

  We can't reach this bound because we can't optimize C for p_{data}.

  We can't even calculate $H[p_{\text{data}}]$ because we can't evaluate $p_{\text{data}}(\mathbf{x})$.
- **practically achievable expected bit rate:** cross entropy between p_{data} & p_{model}:

  $$\boxed{\text{Cross entropy: } H(p_{\text{data}}, p_{\text{model}}) = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[-\log_2 p_{\text{model}}(\mathbf{x})]}$$

  Assumes that the code C is optimal for p_{model}, which is more realistic.

  We can estimate $H(p_{\text{data}}, p_{\text{model}})$ (assuming that we can evaluate $p_{\text{model}}(\mathbf{x})$).

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Recap From Last Week (3 of 3): Kullback-Leibler (KL) Divergence

We need accurate probabilistic models to achieve good compression performance.

- **Modeling error:** How many additional bits do we need to transmit (in expectation) due to an inaccurate model?

Kullback-Leibler Divergence:

$$D_{\mathrm{KL}}(p_{\mathrm{data}} \,||\, p_{\mathrm{model}}) := \underbrace{H(p_{\mathrm{data}}, p_{\mathrm{model}})}_{\text{practically achievable expected bit rate with } p_{\text{model}}} - \underbrace{H[p_{\mathrm{data}}]}_{\text{expected bit rate if we had a perfect model}}$$

- Problem 3.1: explicit proof that $D_{\mathrm{KL}}(p, q) \geq 0$ for any p and q ("Gibb's theorem")
- Problem 3.2: fit p_{model} to a data set by minimizing $D_{\mathrm{KL}}(p_{\mathrm{data}}, p_{\mathrm{model}})$ numerically
  - To reach low $D_{\mathrm{KL}}(p_{\mathrm{data}}, p_{\mathrm{model}})$, we need an expressive model architecture.

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Interlude: Probability Theory & Random Variables

What makes up a probabilistic model:

- *sample space* Ω (abstract space of "all states of the world")
  - *event* $E \subset \Omega$*:* "event E occurs" := "the world is in some state $\omega \in E$".
- *probability measure:* a function $P : \Sigma \to [0, 1]$ where
  - Σ is a so-called σ-algebra on Ω. (a set of all "expressible" events $E \subset \Omega$)
  - $P(\emptyset) = 0$ and $P(\Omega) = 1$. i.e., the following statements are well-defined
  - 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: $P\left(\bigcup_{i=1}^{k}\right) = \sum_{i=1}^{k} P(E_i)$ if all E_i are pairwise disjoint. (proof: set $E_i = \emptyset \; \forall i > k$)
  - therefore: $P(E) + P(\Omega \setminus E) = P(\Omega) = 1 \quad \forall E \in \Sigma$.
  - therefore: $P(E_1) \leq P(E_2)$ if $E_1 \subseteq E_2$ (because $P(E_2) = P(E_1 \cup \overbrace{(E_2 \setminus E_1)}^{\in\Sigma \text{ if } E_1, E_2 \in \Sigma}) = P(E_1) + \overbrace{P(E_2 \setminus E_1)}^{\geq 0}$)

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

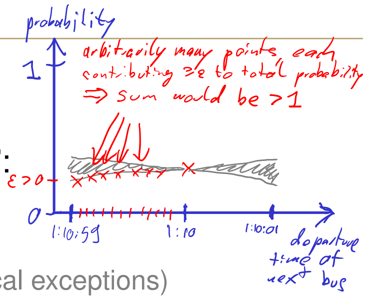
Examples of Probability Measures

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 $E \subset \Sigma$, let $P(E) := |E|/|\Omega| = |E|/9$

value of red die

value of blue die

Examples of Probability Measures (cont'd)



2. Departure times of the next three buses from "Sternwarte":
   - sample space (in a simple model): $\Omega = \mathbb{R}^3$
   - sigma algebra: all "measurable subsets" of $\mathbb{R}^3$
     (think of this as all subsets of $\mathbb{R}^3$ except for extremely pathological exceptions)
   - probability measure P: complicated function, but we know it satisfies certain relations, e.g.,

$$P(\text{“next bus departs before 1.15 pm”}) = P(\text{“next bus departs before 1.10 pm”}) + P(\text{“next bus departs between 1.10 and 1.15 pm”}).$$

   - **Question:** what is the probability that the next bus departs *exactly* at 1.10 pm?
     I.e., what is $P\big(\{1.10\text{ pm}\} \times \mathbb{R}^2\big)$?
   - **Question:** what is the probability that the next bus departs between 1.10 pm and 1.15 pm?

$$P\big(\underbrace{[1.10\text{ pm}, 1.15\text{ pm}]}_{=:\mathcal{I}} \times \mathbb{R}^2\big) = P\Big(\bigcup_{x_1\in\mathcal{I}} \underbrace{\{x_1\}}_{\leftarrow\text{ not a countable set}} \times \mathbb{R}^2\Big) \cancel{\overset{?}{=}} \sum_{x_1\in\mathcal{I}} P\big(\{x_1\} \times \mathbb{R}^2\big) = 0$$

Random Variables

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

Examples:

1. Simplified Game of Monopoly; $\Omega = \big\{(a, b) \;\text{ where }\; a, b \in \{1, 2, 3\}\big\}$
   - 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; $\Omega = \mathbb{R}^3$
   - 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":
  $P(X = x) := P\big(X^{-1}(x)\big) = P\big(\{\omega \in \Omega : X(\omega) = x\}\big)$ $P(E) = |E|/|\Omega| = |E|/9$
  - Example 1 (Simplified Game of Monopoly): $P(X_{\text{sum}} = 3) = P(\{(1,2), (2,1)\}) = \frac{2}{9}$
  - Example 2 (bus schedule): $P(X_{\text{gap}} = 20\text{ minutes}) = 0$ (by same argument as on slide 6)
- When we write just $P(X)$, then we mean the *function* that maps $x \to P(X = x)$.
- **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: $\mathbb{E}_P[X_{\text{red}}] = 2$; $\mathbb{E}_P[X_{\text{blue}}] = 2$; $\mathbb{E}_P[X_{\text{sum}}] = \mathbb{E}_P[X_{\text{red}} + X_{\text{blue}}] = \mathbb{E}_P[X_{\text{red}}] + \mathbb{E}_P[X_{\text{blue}}]$
    $= 2 + 2 = 4$
  - 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):* $P(X \le x) := P\big(\{\omega \in \Omega : X(\omega) \le x\}\big)$
  - Example 1 (Simplified Game of Monopoly): $P(X_{\text{sum}} \le 3) = P(\{(1,2),(2,1),(1,1)\}) = \frac{3}{9} = \frac{1}{3}$
  - Example 2 (bus schedule): $P(X_{\text{gap}} \le 20 \text{ minutes}) \in [0,1]$ (can be nonzero)
- 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:
  $p(x) := \frac{d}{dx} P(X \le x)$ (if derivative exists)
  $\rightarrow$ expectation value: $\mathbb{E}_P[X] = \quad X(\omega)\, dP(\omega) = {}_{-\infty}^{\infty} x\, p(x)\, dx$
  (if a density $p(x)$ exists)

Multiple Random Variables: Joint & Marginal Probability Distributions

- Def. *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 $(x, y) \to P(X = x, Y = x)$)
- If we know $P(X, Y)$, then we can calculate $P(X) = \quad {}_{y} P(X, Y = y)$ (for discrete Y)

i.e. $\{y : P(Y=y) \neq 0\}$ is finite or countably infinite

Proof: $\forall x: \sum_y P(X=x, Y=y) = \sum_y P(\{\omega \in \Omega : X(\omega) = x \wedge Y(\omega) = y\})$
$= P(\bigcup_y \{ \quad " \quad \})$
$= P(\{\omega \in \Omega : X(\omega) = x\}) = P(X = x)$

  - This process is called "marginalization".
  - for continuous random variables: $P(X) = \quad P(X, Y = y)\, dy$

Multiple Random Variables: Statistical Independence

- Def.: X and Y are *(statistically) independent* if and only if: $P(X, Y) = P(X)\, P(Y)$
  (i.e., if $P(X = x, Y = y) = P(X = x)\, P(Y = y) \;\; \forall x, y$)
- **Examples** (Simplified Game of Monopoly):
  - X_{red} and X_{blue} are statistically independent.
  - X_{red} and X_{sum} are *not* statistically independent. (proof: Problem 3.1)
- Def.: *conditional* independence of X and Y given Z: see later

Conditional Probability Distributions

“X & Y are *not* statistically independent” $\Leftrightarrow$ “knowing X reveals something about Y”

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

	$x =$	1	2	3	4	5	6
What are the (marginal) probability distributions $P(X_{red})$ and $P(X_{sum})$ of the red die and the sum, respectively?	$P(X_{red} = x) =$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0	0
	$P(X_{sum} = x) =$	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 $P(X_{sum} \mid X_{red} = 1)$.	$P(X_{sum} = x \mid X_{red} = 1) =$	0	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0
Now you only accept throws where the sum of both dies is 3. What is the conditional probability distribution of X_{red}?	$P(X_{red} = x \mid X_{sum} = 3) =$	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0	0
Finally, assume you only accept throws where $X_{blue} = 1$. What is the conditional probability distribution of X_{red}?	$P(X_{red} = x \mid X_{blue} = 1) =$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	0	0	0

→ note: lower entropy than $P(X_{sum})$

→ lower entropy than $P(X_{red})$

Problem Set: prove that this is not always true but in some sense "typical".

Conditional Probability Distributions (cont’d)

- Def. “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 with respect to the first parameter, i.e., $P(E_2 \mid E_1) + P(\Omega \setminus E_2 \mid E_1) = \frac{P(E_2 \cap E_1) + P((\Omega \setminus E_2) \cap E_1)}{P(E_1)} = \frac{P(E_1)}{P(E_1)} = 1$.
- Def. “conditional probability distribution of random variable Y *given* random variable X”: $P(Y \mid X) := \frac{P(X,Y)}{P(X)}$ i.e., $P(Y = y \mid X = x) := \frac{P(X=x, Y=y)}{P(X=x)}$ $\forall x, y$

  → exercise: redo last slide by strictly following this definition
  - 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 def.)

$$P(X_1, X_2, X_3, \ldots) = \underbrace{\underbrace{P(X_1)\, P(X_2 \mid X_1)}_{= P(X_1, X_2)}\, P(X_3 \mid X_1, X_2)}_{\to P(X_1, X_2, X_3)}\, P(X_4 \mid X_1, X_2, X_3) \cdots$$

Warning: Conditionality $\neq$ Causation

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

$$P(X_{red} \mid X_{sum}) = \frac{P(X_{red}, X_{sum})}{P(X_{sum})} = \frac{P(X_{red}, X_{sum})}{\sum_{a=1}^{3} P(X_{red} = a, X_{sum})} = \frac{P(X_{red})\, P(\underset{\text{effect}}{X_{sum}} \mid \underset{\text{cause}}{X_{red}})}{\sum_{a=1}^{3} P(X_{red} = a)\, P(X_{sum} \mid X_{red} = a)}$$

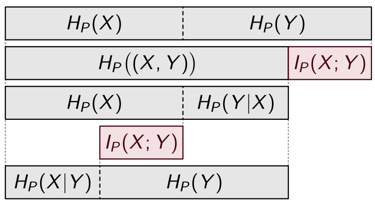
→ This is called “posterior inference”. (more in Lectures 6 and 7)

Next Step:

tying it back to information theory

Information Content and Entropy of Random Variables

- Definitions as you'd expect:
  - information content of the statement "$X = x$": $-\log_2 P(X=x)$
  - entropy of a random variable X under a model P: $H_P(X) := \mathbb{E}_P\big[\overbrace{-\log_2 P(X)}^{\text{random variable of which we take the expectation}}\big]$
  - joint and conditional information content and entropy: see Problems 4.2 and 4.3.
- Subadditivity of entropies: $\forall$ random vars X and Y:
  $\boxed{H_P\big((X,Y)\big) \leq H_P(X) + H_P(Y)}$ (proof: Problem 4.4)
  - equality holds if X and Y are statistically independent (proof: Problem 2.3 (b))
  - Thus, *wrongfully* assuming independence (to simplify the model) leads to a compression overhead of $I_P(X;Y)$ bits:



(figure adapted from MacKay book)

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

Deep Probabilistic Models: Overview and Taxonomy

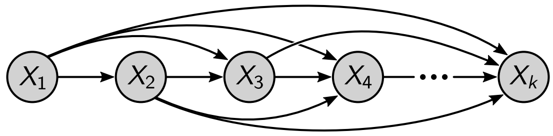
- Assume that the message is a sequence of symbols: $\mathbf{X} = (X_1, X_2, \ldots, X_k)$
- Subadditivity of entropies: $\underbrace{H(\mathbf{X})}_{\text{optimal expected bit rate}} \leq \underbrace{\sum_{i=1}^{k} H(X_i)}_{\text{optimal expected bit rate if we model the symbols as statistically independent}}$

  (Proof: Problem Set 5)
- 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 3.2)
  - latent variable models (planned for Problem Set 6; also: basis for variational autoencoders)

Probabilistic Models at Scale

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

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



- Assume each symbol is from alphabet $\mathfrak{X} = \{1, 2, 3\}$.
  - How many model parameters do we need to specify an arbitrary distribution $P(X_1)$? $|\mathfrak{X}| - 1$
  - How many parameters for an arbitrary conditional distribution $P(X_2 \mid X_1)$? $\rightarrow |\mathfrak{X}|^2 - 1$
  - How many parameters for an arbitrary conditional distribution $P(X_k \mid X_1, X_2, \ldots, X_k)$?

$\rightarrow |\mathfrak{X}|^k - 1$ $\rightarrow$ exponential growth $\rightarrow$ not scalable !

Expressive Yet Efficient Probabilistic Models

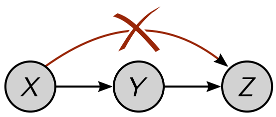
- **Goal:** Find approximation to arbitrary models $P(\mathbf{X})$ that
  - captures relevant correlations
  - but is still *computationally efficient*:
    $\rightarrow$ *reasonably compact* representation of the model in memory
    $\rightarrow$ *reasonably efficient evaluation* of probabilities $P(\mathbf{X} = \mathbf{x})$

weaker statement than simple statistical independence: $P(X, Z) = P(X)\,P(Z)$

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

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

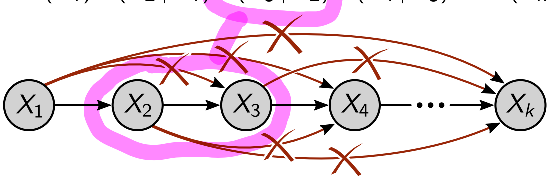
$\Longleftrightarrow P(X, Y, Z) = P(X)\,P(Y \mid X)\,P(Z \mid Y)$ (proof: Problem Set 5)



Four Approximation Schemes

(a) **Markov Process:** assume symbols X_i are generated by a *memoryless* process

- Each symbol X_{i+1} is conditioned on the immediately preceding symbol X_i but not on any earlier symbols: $P(\mathbf{X}) = P(X_1)\,P(X_2 \mid X_1)\,P(X_3 \mid X_2)\,P(X_4 \mid X_3) \cdots P(X_k \mid X_{k-1})$



next character:
"h" $\rightarrow$?
"th" $\rightarrow$ maybe "e"

- 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

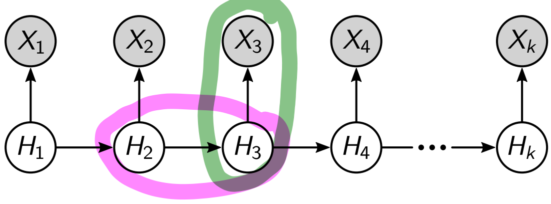
simplistic assumption; e.g., in English text, the string "the" is very frequent.

$\Rightarrow P(X_{i+1} = \text{'e'} \mid X_i = \text{'h'}, X_{i-1} = \text{'t'}) > P(X_{i+1} = \text{'e'} \mid X_i = \text{'h'})$ (i.e., *not* cond. indep.)

Four Approximation Schemes (cont'd)

(b) **Hidden Markov Model:**

- Morkov Process of "hidden" states H_i that are only indirectly observed



part of the message ("observed")

not part of the message ("latent")

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

can model long-range correlations (exercise)

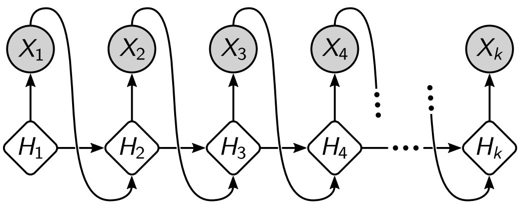
overhead: in order to model $P(X_i\,|\,H_i)$, decoder has to decode H_i, even though it's not part of the message (solution: "bits-back coding", Lecture 6)

Four Approximation Schemes (cont'd)

deterministic function
(e.g., neural network)

(c) **Autoregressive Model:**

- similar to a hidden Markov model, but the hidden process is *deterministic*: $H_{i+1} = f(H_i, X_i)$



part of the message ("observed")

deterministic function of its inputs

$$P(\mathbf{X}) = P(X_1\,|\,H_1) \mathop{}\limits_{i=2}^{k} P(X_i\,|\,H_i) \quad \text{where} \quad H_1 = \text{fixed};\ H_{i+1} = f(H_i, X_i)$$

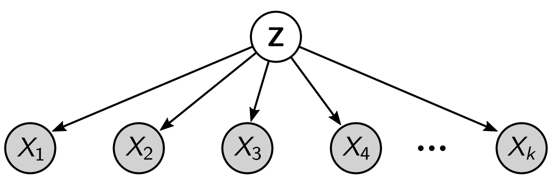
no compression overhead for reconstructing **H** (see Problem 3.2)

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

Four Approximation Schemes (cont'd)

(d) **Latent Variable Models:** "explain" message **X** by some unobserved **Z**

- Intuition: **Z** captures message at a higher level of abstraction (e.g., the "topic" of a piece of text, or basic shapes in an image)



part of the message ("observed")

not part of the message ("latent")

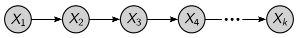
$$P(\mathbf{X}) = \quad P(\mathbf{X},\mathbf{Z})\,d\mathbf{Z} \quad \text{with} \quad P(\mathbf{X},\mathbf{Z}) = P(\mathbf{Z}) \mathop{}\limits_{i=1}^{k} P(X_i\,|\,\mathbf{Z})$$

can model correlations (see Problem Set 6) and is parallelizable

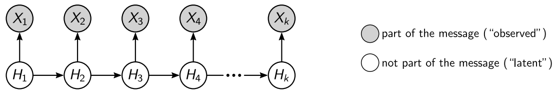
compression overhead for encoding **Z** (solution: "bits-back coding", Lecture 6)

Recap: Four Approximation Schemes

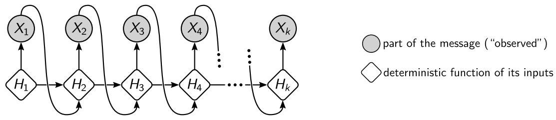
Markov Process



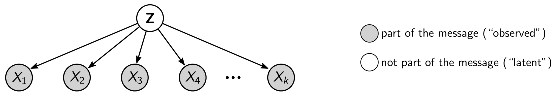
Hidden Markov Model



Autoregressive Model

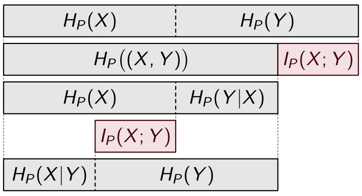


Latent Variable Model



Outlook

- **Problem Set:**



- **Next ∼4 weeks:** lossless compression with deep probabilistic models
  → Different model architectures require different compression algorithms.
  - Problem Set 3 (discussed tomorrow): compressing English text with a learnt autoregressive model (using recurrent neural networks)
  - Lecture 5 (next week): stream codes with first-in-first out vs. last-in-first-out
  - Lecture 6: (net-)optimal lossless compression with latent variable models
  - Lectures 7 and 8: deep-learning based latent variable models
- **Afterwards:** Lossy compression → will also build on these information theoretical concepts, especially on mutual information