

# STA414: Statistical Methods for Machine Learning II

## Lecture Notes

Yiliu Cao

May 3, 2024

### Contents

<b>1</b>	<b>Probabilistic models</b>	<b>3</b>
1.1	An overview of probabilistic models	3
1.2	Statistical decision theory	4
<b>2</b>	<b>Graphical Models</b>	<b>6</b>
2.1	Introduction to graphical models	6
2.2	Directed Acyclic Graphical Models	7
2.3	Undirected Graphical Models	10
<b>3</b>	<b>Exact Inference</b>	<b>13</b>
3.1	Introduction to exact inference	13
3.2	Sum-product algorithm	15
<b>4</b>	<b>Message passing</b>	<b>16</b>
4.1	TrueSkill latent variable models	16
4.2	Message Passing	16
4.3	Belief Propagation on Trees	18
<b>5</b>	<b>Monte Carlo Methods</b>	<b>19</b>
5.1	Introduction to Monte Carlo Methods	19
5.2	Three Sampling Methods	20
<b>6</b>	<b>Markov Chain Monte Carlo</b>	<b>22</b>
6.1	Introduction to Markov Chains	22
6.2	Stationary distribution of a (homogeneous) Markov Chain	24
<b>7</b>	<b>Variational Inference</b>	<b>25</b>
7.1	Introduction to Variational Inference	25
7.2	Kullback-Leibler divergence	27
<b>8</b>	<b>Appendix</b>	<b>28</b>
8.1	Example 1	28
8.2	Example 2	29
8.3	Derivations 1	30

8.4	Example 3 . . . . .	30
8.5	Derivations 2 . . . . .	32
8.6	Derivations 3 . . . . .	32
8.7	Example 4 . . . . .	33

# 1 Probabilistic models

## 1.1 An overview of probabilistic models

We have the random vector  $X = (X_1, X_2, \dots, X_n)$ , and we want to compute the relationship between each random variable. The joint distribution is  $p(x) = p(x_1, \dots, x_d)$ . Denote the input data  $\mathbf{x}$  (high-dimensional), and output  $y$  (discrete or continuous). In general, we have two models:

**Regression:**

$$p(y|x) = \frac{p(x, y)}{p(x)} = \frac{p(x, y)}{\int p(x, y) dy}$$

**Classification/Clustering:**

$$p(c|x) = \frac{p(x, c)}{\sum_c p(x, c)}$$

## Observed vs Unobserved random variables

**Supervised classification(learning):**

- We **KNOW** what to predict
- **Supervised Dataset:**  $\{x^{(i)}, c^{(i)}\}_{i=1}^N \sim p(x, c)$
- The class labels are observed.

**Unsupervised classification(learning):**

- We do **NOT KNOW** what to predict
- **Unsupervised Dataset:**  $\{x^{(i)}\}_{i=1}^N \sim p(x) = \sum_c p(x, c)$
- We only observe the inputs  $\mathbf{x}$

In order to estimate the unknown distribution  $p(x)$ , we have few assumptions:

1. **IID Data:** we assume the samples  $x^{(i)}$  are independent and identically distributed.
2. **Parametrized distribution:**  $p(x|\theta)$  comes from a parametrized family  $\mathcal{P} = \{p(x|\theta) : \theta \in \Theta\}$

## Maximum Likelihood Estimation(MLE)

MLE is the method to estimate the parameters of an assume probability distribution, given some observed data. Technically, we can use MLE to estimate any parameters we want. More specifically:

- Let  $x^{(i)} \sim p_* = p(x|\theta_*)$  for  $i = 1, \dots, N$  be i.i.d. random variables.
- The joint of  $\mathcal{D} = \{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$  is  $p(\mathcal{D}|\theta_*) = \prod_i p(x^{(i)}|\theta_*)$ .

- Assume we observe data  $\mathcal{D}$  and  $\theta_*$  is unknown. The likelihood function is:

$$\mathcal{L}(\theta; \mathcal{D}) = p(\mathcal{D}|\theta) = \prod_{i=1}^N p(x^{(i)}|\theta)$$

- The log-likelihood function:

$$\ell(\theta; \mathcal{D}) = \log \mathcal{L}(\theta; \mathcal{D}) = \sum_{i=1}^N \log p(x^{(i)}|\theta)$$

Here is an example of MLE

## Sufficient Statistics and Exponential Families

A **sufficient statistics** is a function of the data that conveys exactly the same information about the parameter as the entire data.

In addition, we can writing any exponential family member in the form:

$$p(x|\eta) = h(x) \exp\{\eta^\top T(x) - A(\eta)\}$$

where

$T(x)$  : sufficient statistics

$\eta$  : natural parameter

$A(\eta)$  : log-partition function

$h(x)$  : carrying measure

Moreover, let  $X \sim p(x|\eta)$ , then we have  $E[T(X)] = A'(\eta)$

One example of exponential family

## 1.2 Statistical decision theory

Suppose we have an input vector  $x$  and the corresponding target, we want to predict the label given a new input. Notice that here we assume the output is the label/class which is discrete. However, the output can also be continuous (regression).

Intuitively, for a given new input  $x$ , we have:

$$p(\mathcal{C}_k|x) = \frac{p(x|\mathcal{C}_k)p(\mathcal{C}_k)}{p(x)}$$

We then pick the  $\mathcal{C}_k$  with the highest probability.

**Decision Rule:** Divide the input space to  $\mathcal{R}_1$  &  $\mathcal{R}_2$  such that all points in  $\mathcal{R}_k$  are assigned to class  $\mathcal{C}_k$ . We want to make mistakes as less as possible; equivalently, we want to minimize the **misclassification rate**.

For  $k \in \{1, 2\}$ :

$$p(\text{mistake}) = p(x \in \mathcal{R}_2, \mathcal{C}_1) + p(x \in \mathcal{R}_1, \mathcal{C}_2) = \int_{\mathcal{R}_1} p(x, \mathcal{C}_2) dx + \int_{\mathcal{R}_2} p(x, \mathcal{C}_1) dx \quad (1.1)$$

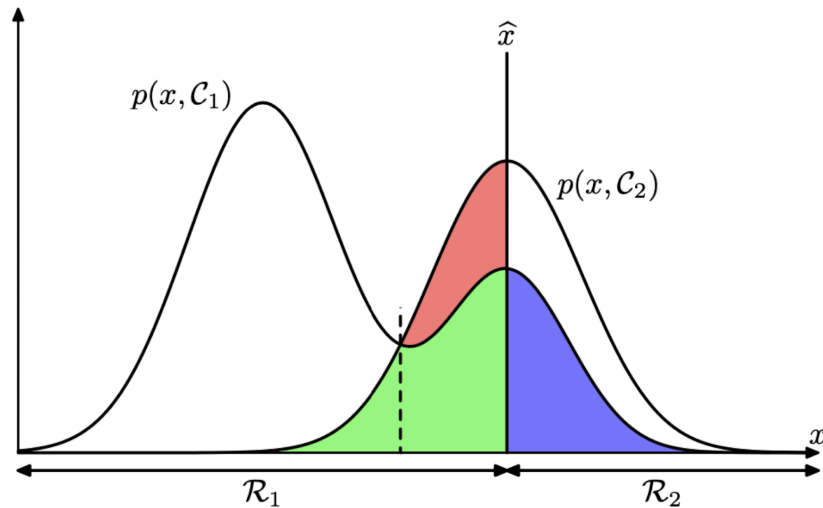


Figure 1.1: Misclassification Rate

1. **RedGreen regions:** inputs that belong to  $\mathcal{C}_2$  but assigns to  $\mathcal{R}_1$  as they are under  $p(x, \mathcal{C}_2)$ .
2. **Blue regions:** inputs that belong to  $\mathcal{C}_1$  but assigns to  $\mathcal{R}_2$  as they are under  $p(x, \mathcal{C}_1)$
3. Therefore, for any data  $\mathbf{x}$ , if  $p(x, \mathcal{C}_1) > p(x, \mathcal{C}_2)$ , then we assign this point to  $\mathcal{C}_1$ , vice-versa. Therefore,  $\mathcal{R} = \{x : p(x, \mathcal{C}_1) > p(x, \mathcal{C}_2)\}$

We want to minimize the misclassification error rate  $\Rightarrow$  minimize the **loss**

## Loss Function

**Loss function** measures the loss incurred by taking of any available decisions.

### For discrete case

we denote  $L_{ij}$  as the  $(k, j)$  element of the loss matrix. We want to minimize the expected loss

Therefore:

$$\begin{aligned}\mathbb{E}[L] &= \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(x, \mathcal{C}_k) dx \\ &= \sum_j \int_{\mathcal{R}_j} \sum_k L_{kj} p(x, \mathcal{C}_k) dx\end{aligned}$$

Define  $g_j(x) = \sum_k L_{kj} p(x, \mathcal{C}_k)$ . Notice that  $g_j(x) \geq 0$  and

$$\mathbb{E}[L] = \sum_j \int_{\mathcal{R}_j} g_j(x) dx$$

Thus, minimizing  $\mathbb{E}[L]$  is equivalent to choosing

$$\mathcal{R}_j = \{x : g_j(x) < g_i(x) \text{ for all } i \neq j\} \quad (1.2)$$

$$\Rightarrow \mathcal{R}_j = \left\{ x : \sum_k L_{kj} p(\mathcal{C}_k | x) < \sum_k L_{ki} p(\mathcal{C}_k | x) \text{ for all } i \neq j \right\} \quad (1.3)$$

## For regression

- Consider the input/target  $(x, t)$ , where  $t$  is continuous and the joint density is  $p(x, t)$
- The regression function is  $y(x)$
- The loss function is  $L(y(x), t) = (y(x) - t)^2$

Therefore the expected loss will be:

$$\begin{aligned} \mathbb{E}[L] &= \iint L(y(x), t) p(x, t) dx dt \\ &= \iint (y(x) - \mathbb{E}[t | x])^2 p(x, t) dx dt + \iint (\mathbb{E}[t | x] - t)^2 p(x, t) dx dt \end{aligned}$$

Full derivations here

The second term is the conditional variance of  $t|x$  and does not depend on  $y(x)$  and hence the expected loss is minimized when  $y(x) = \mathbb{E}[t|x]$ . Therefore, we can see that the loss function will change the decision rule significantly; however, we can always reject the option or not making a decision.

## 2 Graphical Models

### 2.1 Introduction to graphical models

Remember our goal is to specify the joint distribution  $N$  random variables  $p(x_1, \dots, x_N) = p(x)$ . If we assume each  $x_i$  is binary such that  $x_i \in \{0, 1\}$ , then we need  $2^N - 1$  parameters to specify  $p(x)$ . For example,  $p(x_1 = 0, x_2 = 0, \dots, x_N = 0)$  or  $p(x_1 = 1, x_2 = 0, \dots, x_N = 0)$ .

Equivalently, we can specify the joint distribution  $p(x)$  as:

$$\begin{aligned} p(x_1, x_2, \dots, x_N) &= \prod_{j=1}^N p(x_j | x_1, x_2, \dots, x_{j-1}) \\ &= p(x_1 | x_0) p(x_2 | x_1, x_0) \dots \end{aligned}$$

Thus total number of parameters is  $1 + 2 + 4 + \dots + 2^{N-1} = 2^N - 1$

We can see that it requires a huge number of parameters to specify the joint distribution.

We want to draw relationships between variables.

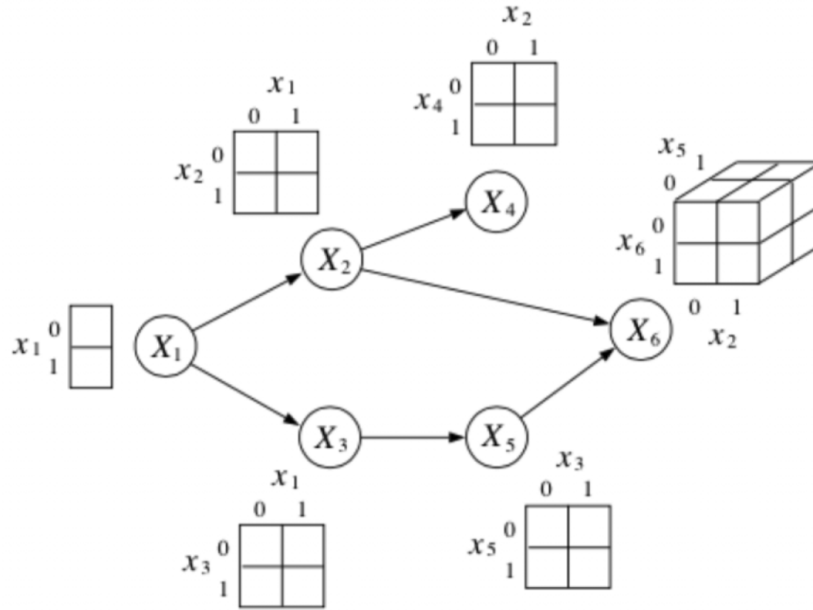


Figure 2.1: An example of conditional probability tables(CPT)

## Condition independence

For three random variables  $x_A, x_B, x_C$ , if  $x_A, x_B$  are conditionally independent given  $x_C$ , then we write  $x_A \perp x_B \mid x_C$ . The following conditions are equivalent:

- $x_A \perp x_B \mid x_C$
- $p(x_A, x_B \mid x_C) = p(x_A \mid x_C)p(x_B \mid x_C)$
- $p(x_A \mid x_B, x_C) = p(x_A \mid x_C)$
- $p(x_B \mid x_A, x_C) = p(x_B \mid x_C)$

## 2.2 Directed Acyclic Graphical Models

A directed cyclic graphical model encode a particular form of factorization of the joint distribution. The form of factorization is various.

Figure 2 shows an example of conditional probability. From the graph, we only need  $2^1 * 4 + 2^0 + 2^2 = 13 < 2^6 - 1$  parameters.

**D-separation:** If  $C$  d-separates  $A$  and  $B$ , then  $x_A \perp x_B \mid x_C \forall a \in A, b \in B$

## Bayes ball algorithm

Bayes ball determines the conditional independence/dependence in a DAG (I personally found this part most ambiguous). There are three fundamental Bayes ball algorithms which are causal chain, common cause and explaining away. For each one, we will under it intuitively by drawing a story.

### 1. Causal chain

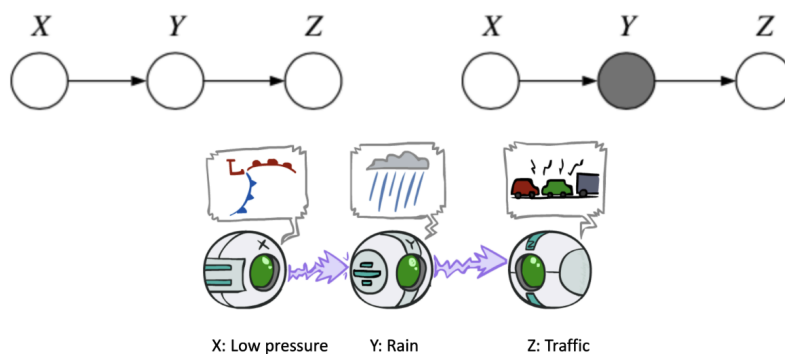


Figure 2.2: An illustration of causal chain

$$\begin{aligned}
 p(z \mid x, y) &= \frac{p(x, y, z)}{p(x, y)} \\
 &= \frac{p(x)p(y \mid x)p(z \mid y)}{p(x)p(y \mid x)} \\
 &= p(z \mid y) \\
 \Rightarrow X \text{ and } Z \text{ d-separated given } Y
 \end{aligned}$$

## 2. Common cause

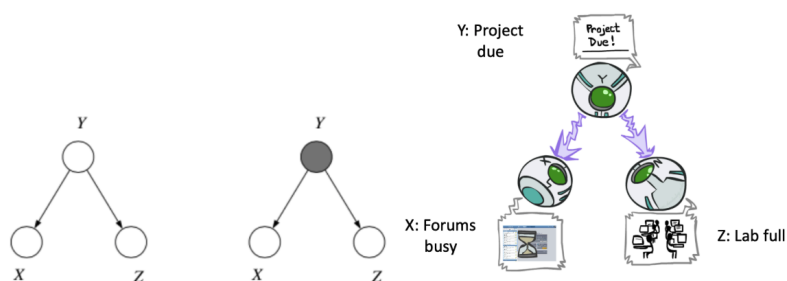


Figure 2.3: An illustration of common cause

$$\begin{aligned}
 p(x, z \mid y) &= \frac{p(x, y, z)}{p(y)} \\
 &= \frac{p(y)p(x \mid y)p(z \mid y)}{p(y)} \\
 &= p(x \mid y)p(z \mid y) \\
 \Rightarrow X \text{ and } Z \text{ d-separated given } Y
 \end{aligned}$$

## 3. Explaining away



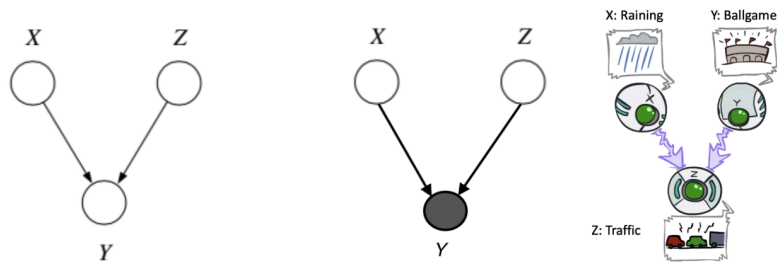


Figure 2.4: An illustration of explaining away

$$\begin{aligned}
 p(z \mid x, y) &= \frac{p(x)p(z)p(y \mid x, z)}{p(x)p(y \mid x)} \\
 &= \frac{p(z)p(y \mid x, z)}{p(y \mid x)} \neq p(z \mid y) \\
 \Rightarrow X \text{ and } Z \text{ are NOT d-separated given } Y
 \end{aligned}$$

In general, the Bayes ball works as follows:

1. Shade all nodes  $x_C$  (these are observed)
2. Place "balls" at each node in  $x_A$  (or  $x_B$ )
3. Let the "balls" "bounce" around according to some rules. If any of the balls reach any of the nodes in  $x_B$  from  $x_A$  then  $x_A \not\perp x_B \mid x_C$ . Otherwise  $x_A \perp x_B \mid x_C$

### Example

*Question:* Is  $x_2 \perp x_3 \mid \{x_1, x_6\}$

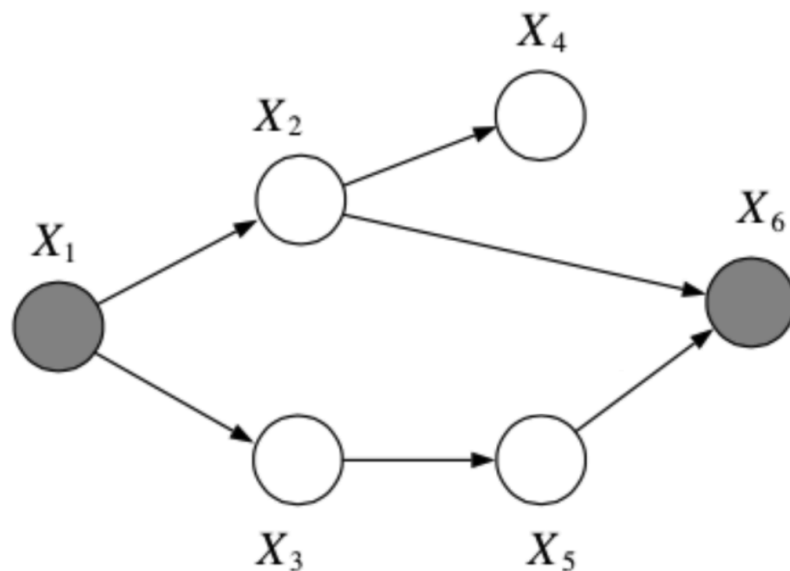


Figure 2.5: An illustration of explaining away

*Answer:* No. By Bayes ball algorithm,  $x_2$  can travel to  $x_5$ , and hence can travel to  $x_3$ .

## Moralization

Like I said, I personally do not like Bayes algorithm. Instead, another one called "moralization" is more straightforward and easier to use. We can follow the procedure:

1. **Draw the ancestral graph**

We only keep the ancestor of the mentioned nodes. That said, in the previous example, we only keep the ancestors of  $\{x_2, x_3, x_1, x_6\}$ . Hence we have the entire graph except the node  $x_4$ . Note the ancestors includes **their parents, parents' parents etc.**

2. **"Moralize" the ancestral graph by "marrying" the parents**

If two nodes have the same children, such as  $x_2$  and  $x_5$ , then we draw a line between these two nodes.

3. **"Disorient" the graph**

Ignore the directions by replacing the arrows to edges.

4. **Delete the givens and their edges**

In the previous example, the givens are the  $x_1$  and  $x_4$ .

5. **Find the answer**

After we finished the step 1 to 4, we then justify whether the two nodes are connected or disconnected. If **connected**, then the two nodes are conditionally **dependent**. Otherwise **disconnected**, the two nodes are conditionally **independent**. In the previous example, we can easily find that  $x_2$  and  $x_3$  are d-separated by  $x_1$  and  $x_6$ .

*Question:* What about the marginal independence, such as  $x_2 \perp x_3$ ?

*Answer:* We use the same way as above without step 4.

## 2.3 Undirected Graphical Models

The undirected graphical models are also called the **Markov random fields (MRFs)**. Compare to graphical models, we have no more directed edges; instead, the dependencies are now described as undirected graphs. Moreover, **Markov blanket** is the set of nodes that makes  $X_i$  conditionally independent of all other nodes. **Clique** is a subset of nodes that every two nodes are connected by an edge. **Maximal clique** a clique that can not be extended by including one more adjacent vertex.

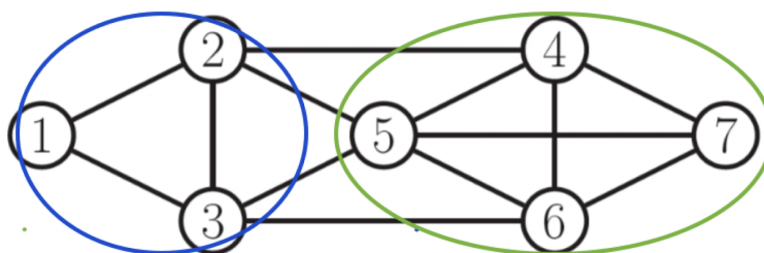


Figure 2.6: An example of Markov random fields:  $\{1, 2, 3\}$  is a clique and  $\{4, 5, 6, 7\}$  is a maximal clique

## Distribution induced by MRFs

- Let  $X = (X_1, \dots, X_m)$  be the set of all random variables in our graph  $G$ .
- Let  $\mathcal{C}$  be the set of all maximal cliques of  $G$ .
- The distribution  $p$  of  $X$  factorizes with respect to  $G$  if

$$p(x) \propto \prod_{C \in \mathcal{C}} \psi_C(x_C)$$

for some nonnegative potential functions  $\psi_C$ , where  $x_C = (x_i)_{i \in C}$ .

The density can be factorized to cliques is also called the **Hammersley-Clifford Theorem**.

**Global markov properties:**  $X_A \perp X_B \mid X_S$  if the sets  $A$  and  $B$  are separated by  $S$  in  $G$  (every path from  $A$  to  $B$  has to pass  $S$ ). Therefore, the joint distribution described on above is:

$$p(x) \propto \psi_{1,2,3}(x_1, x_2, x_3) \psi_{2,3,5}(x_2, x_3, x_5) \psi_{2,4,5}(x_2, x_4, x_5) \\ \times \psi_{3,5,6}(x_3, x_5, x_6) \psi_{4,5,6,7}(x_4, x_5, x_6, x_7)$$

Not all DAGMs can be represented as MRFs.

## Relations between exponential families and MRFs

- Consider a parametric family of factorized distributions

$$p(x \mid \theta) = \frac{1}{Z(\theta)} \prod_{C \in \mathcal{C}} \psi_C(x_C \mid \theta_C), \quad \theta = (\theta_C)_{C \in \mathcal{C}}.$$

- We can write this in an exponential form:

$$p(x \mid \theta) = \exp\left\{\sum_{C \in \mathcal{C}} \log \psi_C(x_C \mid \theta_C) - \underbrace{\log Z(\theta)}_{=A(\theta)}\right\}$$

- Suppose the potentials have a log-linear form

$$\log \psi_C(x_C \mid \theta_C) = \theta_C^\top \phi_C(x_C)$$

we then get the exponential family

$$p(x \mid \theta) = \exp\left\{\sum_{C \in \mathcal{C}} \theta_C^\top \phi_C(x_C) - \underbrace{\log Z(\theta)}_{=A(\theta)}\right\}$$

*Question:* When the potentials have a log-linear form?

*Solution:* Suppose we have the random vector  $x_1$  and  $x_2$ , and they are all binary. Then

there are four possible values of  $(x_1, x_2)$ :  $(0, 0)$ ,  $(1, 0)$ ,  $(0, 1)$ ,  $(1, 1)$ . We take

$$\theta_{1,2} := \begin{bmatrix} \log \psi_{1,2}(0, 0) \\ \log \psi_{1,2}(0, 1) \\ \log \psi_{1,2}(1, 0) \\ \log \psi_{1,2}(1, 1) \end{bmatrix} \in \mathbb{R}^4$$

and let  $\psi_{1,2}(x_1, x_2)$  be the function that satisfies

$$\phi_{1,2}(0, 0) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \phi_{1,2}(0, 1) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \phi_{1,2}(1, 0) = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \quad \phi_{1,2}(1, 1) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

$$\Rightarrow \phi_{1,2}(x_1, x_2) = \begin{bmatrix} (1-x_1)(1-x_2) \\ (1-x_1)x_2 \\ x_1(1-x_2) \\ x_1x_2 \end{bmatrix}$$

We then have a linear form:

$$\log \psi_C(x_{12} \mid \theta_{12}) = \theta_{12}^\top \phi_{12}(x_{12})$$

## Ising model

The Ising model is an example of MRFs, which is used to model magnets. It has a form of potential function:

$$\psi_{st}(x_s, x_t) = e^{J_{st}x_sx_t}$$

Equivalently:

$$\begin{aligned} \psi_{st}(-1, -1) &= \psi_{st}(1, 1) = e^{J_{st}} \\ \psi_{st}(1, -1) &= \psi_{st}(-1, 1) = e^{-J_{st}} \end{aligned}$$

$$\psi_{st}(x_s, x_t) = 0 \text{ if the two nodes are not connected}$$

In terms of the distribution in Ising model, we also include the node potential  $\psi_s(x_s) = e^{b_sx_s}$ :

$$p(x) \propto \prod_{s \sim t} \psi_{st}(x_s, x_t) \prod_s \psi_s(x_s) = \exp \left\{ \sum_{s \sim t} J_{st}x_sx_t + \sum_s b_sx_s \right\}$$

### Recall: Multivariate normal distribution

$X = (X_1, \dots, X_m) : \mu \in \mathbb{R}^m$  and  $\Sigma$  symmetric positive definite  $m \times m$  matrix. Write  $X \sim N_m(\mu, \Sigma)$  if the density of the vector  $X$  is

$$f(\mathbf{x}; \mu, \Sigma) = \frac{1}{(2\pi)^{m/2}} (\det \Sigma)^{-1/2} \exp \left( -\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right).$$

Denote  $K = \Sigma^{-1}$  then

$$f(\mathbf{x}; \mu, \Sigma) \propto \prod_s e^{-\frac{1}{2} K_{ss}(x_s - \mu_s)^2} \prod_{s < t} e^{-K_{st}(x_s - \mu_s)(x_t - \mu_t)}$$

Intuitively, we can also visualize the conditional independencies between each variables (similar to Bayes ball). Just like the Ising model, here we can use concentration matrix  $K$  to represent the relationship between variables.

Conditional independence:

- $X_i \perp X_j$  if and only if  $\Sigma_{ij} = 0$ .
- $X_i \perp X_j \mid X_C$  if and only if  $\Sigma_{ij} - \Sigma_{i,C} \Sigma_{C,C}^{-1} \Sigma_{C,j} = 0$
- Let  $R = V \setminus \{i, j\}$ . The following are equivalent:
  - $X_i \perp X_j \mid X_R$
  - $\Sigma_{ij} - \Sigma_{i,R} \Sigma_{R,R}^{-1} \Sigma_{R,j} = 0$
  - $(\Sigma^{-1})_{ij} = 0$

Hence we have the **Gaussian Graphical models**

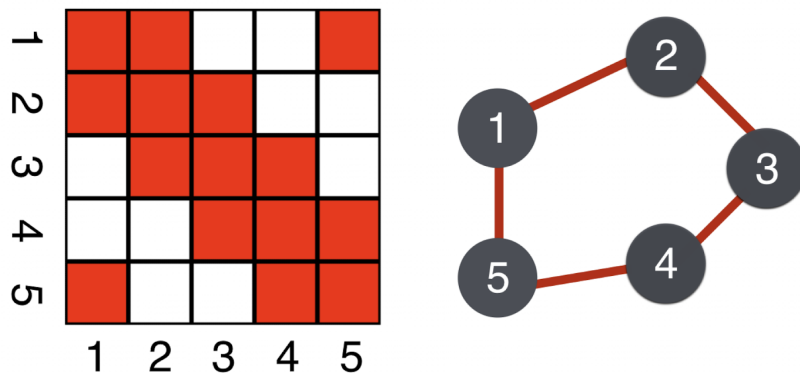


Figure 2.7: An example of Gaussian Graphical model

$K_{ij} = 0$  if and only if  $X_i \perp X_j \mid X_{rest}$ . For example,  $X_1 \perp X_4 \mid X_{rest}$

## 3 Exact Inference

### 3.1 Introduction to exact inference

We want to explore the inference from the probabilistic graphical models. In notations, we have

- $x_E$  represent the observed evidence
- $x_F$  represent the unobserved variable we want to infer
- $x_R = x \setminus \{x_E, x_F\}$  represent the remaining variables

For the conditional probability  $p(x_F | x_E)$ , by Bayes theorem:

$$p(x_F | x_E) = \frac{p(x_F, x_E)}{p(x_E)} = \frac{p(x_F, x_E)}{\sum_{x_F} p(x_F, x_E)}$$

Moreover, we can also marginalize the extraneous variables to have the marginal joint distribution:

$$p(x_F, x_E) = \sum_{x_R} p(x_F, x_E, x_R)$$

The below equation give us an intuition of exact inference, which is to marginalize other variables. However, when the number of variables increases, the order we marginalize will affect the computational cost. Hence we have to choose an appropriate **elimination order**. More specifically, we want to have the exact inference for one variable in DAGMs or MRFs.

**Example 3.1.1.** Suppose we have the simple chain for four variables A, B, C, D

$$A \rightarrow B \rightarrow C \rightarrow D$$

where:

$$x_F = \{D\}, x_E = \{A, B, C\}, x_R = \{\}$$

We want to compute the exact inference of  $D$ ,  $p(D)$  In the simple chain settings, we can express the joint distribution as:

$$\begin{aligned} p(D) &= \sum_{A,B,C} p(A, B, C, D) \\ &= \sum_C \sum_B \sum_A p(A)p(B | A)p(C | B)p(D | C) \end{aligned}$$

If we choose an elimination order:

$$\begin{aligned} p(D) &= \sum_{A,B,C} p(A, B, C, D) \\ &= \sum_C p(D | C) \left( \sum_B p(C | B) \left( \sum_A p(A)p(B | A) \right) \right) \\ &= \sum_C p(D | C) \sum_B p(C | B) \sum_A p(A)p(B | A) \\ &= \sum_C p(D | C) \sum_B p(C | B)p(B) \\ &= \sum_C p(D | C)p(C) \\ &= \sum_C p(D, C) \end{aligned}$$

The above example give us an intuition that we can firstly marginalize the node with no children ( $C$  first). Equivalently, we start the nodes that comes early in the induced ordering of the DAG.

### 3.2 Sum-product algorithm

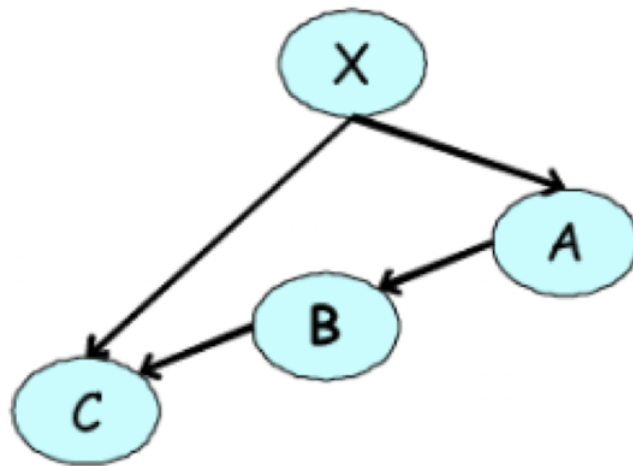


Figure 3.1: An example of Sum-product inference

**Example 3.2.1.**

To express the joint distribution, we introduce the nonnegative factors  $\psi$ :

$$\begin{aligned}
 p(A, B, C) &= \sum_X p(A, B, C, X) \\
 &= \sum_X p(X) p(A | X) p(B | A) p(C | B, X) \\
 &= \sum_X \psi_1(X) \psi_2(A, X) \psi_3(A, B) \psi_4(X, B, C) \\
 &= \psi_3(A, B) \sum_X \psi_1(X) \psi_2(A, X) \psi_4(X, B, C) \\
 &= \psi_3(A, B) \tau(A, B, C)
 \end{aligned}$$

We eventually end with a new factor  $\tau$  and it is an example of how sum-product algorithm works.

In general, we compute the conditional distribution  $p(x_F | x_E)$  using the **sum-product algorithm**. Denote  $\mathcal{F}$  as the set of potentials or factors.

$$p(x_F | x_E) \propto \tau(x_F, x_E) = \sum_{x_R} \prod_{C \in \mathcal{F}} \psi_C(x_C)$$

For DAGMs,  $\mathcal{F}$  is  $\{i\} \cup \text{parents}(i)$ ,  $\forall$  nodes  $i$

For MRFs,  $\mathcal{F}$  is all maximal cliques.

We will see more about exact inference on message passing

## 4 Message passing

### 4.1 TrueSkill latent variable models

**Latent variables** are variables that can only be inferred indirectly through a mathematical model from other observable variables that can be directly observed or measured.

**TrueSkill** model is a player ranking system for competitive games. Our goal is to infer the skills of players in a competitive game, based on the results of the games. Hence, each player's skill is a **latent variable**, and we assume they are fixed. Moreover, we also assume each player's skill is independent such as independent Gaussian prior.

For each game, the probability that player  $i$  beats player  $j$  is

$$p(i \text{ beats } j) = \sigma(z_i - z_j), \text{ where}$$

$$\sigma(y) = \frac{1}{1 + \exp(-y)}$$

Hence the entire joint likelihood of players and games is:

$$\begin{aligned} & p(z_1, z_2, \dots, z_N, \text{ game 1, game 2, .. game } T) \\ &= \underbrace{\left[ \prod_{i=1}^N p(z_i) \right]}_{\text{prior}} \underbrace{\left[ \prod_{\text{games}} p(i \text{ beats } j \mid z_i, z_j) \right]}_{\text{likelihood}} \end{aligned}$$

However, it is hard to compute the posterior

$$\begin{aligned} & p(z_1, z_2 \mid \text{ game 1, game 2, } \dots \text{ game } T) \\ &= \int \dots \int p(z_1, z_2, z_3 \dots z_N \mid x) dz_3 \dots dz_N \end{aligned}$$

**Message passing** can help us to compute the posterior.

### 4.2 Message Passing

From the last section, we can see that determining an optimal elimination order is hard, and the resulting marginalization can be still costly. Therefore, we introduce the tress as any elimination ordering from leaves to the root is optimal.

#### Inference in trees

A graph is  $G = (\mathcal{V}, \mathcal{E})$ , where  $\mathcal{V}$  is vertices and  $\mathcal{E}$  is edges. A leaf in a tree is the nodes where it has only one neighbor. We can regard a tress as MRFs, with joint distribution:

$$p(x_1, x_2, \dots, x_n) = \underbrace{\frac{1}{Z}}_{\text{normalization}} \underbrace{\prod_{i \in \mathcal{V}} \psi(x_i)}_{\text{nodes}} \underbrace{\prod_{(i,j) \in \mathcal{E}} \psi_{ij}(x_i, x_j)}_{\text{edges}}$$



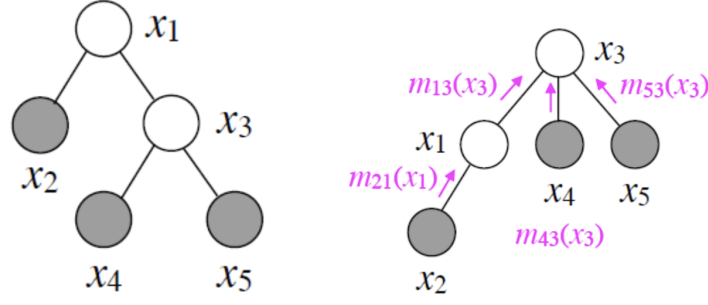


Figure 4.1: An example of tree and message passing

**Example 4.2.1.** Above shows a tree where  $x_E = \{\bar{x}_2, \bar{x}_4, \bar{x}_5\}$ ,  $x_R = x_1$ . We want to compute  $p(x_3 | x_E)$ :

$$\begin{aligned} p(x_3 | x_E) &= \frac{p(x_3, x_E)}{\sum_{x'_3} p(x'_3, x_E)} \\ &= \frac{1}{Z_E} p(x_3, x_E) \end{aligned}$$

Express  $p(x_3, x_E)$ , we have

$$\begin{aligned} &= \sum_{x_1} \psi_1(x_1) \psi_2(\bar{x}_2) \psi_3(x_3) \psi_4(\bar{x}_4) \psi_5(\bar{x}_5) \psi_{12}(x_1, \bar{x}_2) \psi_{13}(x_1, x_3) \psi_{34}(x_3, \bar{x}_4) \psi_{35}(x_3, \bar{x}_5) \\ &= \frac{1}{Z_E} \underbrace{\psi_4(\bar{x}_4) \psi_{34}(x_3, \bar{x}_4)}_{m_{43}(x_3)} \underbrace{\psi_5(\bar{x}_5) \psi_{35}(x_3, \bar{x}_5)}_{m_{53}(x_3)} \psi_3(x_3) \sum_{x_1} \psi_1(x_1) \psi_{13}(x_1, x_3) \underbrace{\psi_2(\bar{x}_2) \psi_{12}(x_1, \bar{x}_2)}_{m_{21}(x_1)} \\ &= \frac{1}{Z_E} m_{43}(x_3) m_{53}(x_3) \psi_3(x_3) \underbrace{\sum_{x_1} \psi_1(x_1) \psi_{13}(x_1, x_3) m_{21}(x_1)}_{m_{13}(x_3)} \\ &= \frac{1}{Z_E} \psi_3(x_3) m_{43}(x_3) m_{53}(x_3) m_{13}(x_3) \\ &= \frac{\psi_3(x_3) m_{43}(x_3) m_{53}(x_3) m_{13}(x_3)}{\sum_{x'_3} \psi_3(x'_3) m_{43}(x'_3) m_{53}(x'_3) m_{13}(x'_3)} \end{aligned}$$

## Sum-product algorithm

The above example shows how we can pass the message and eventually compute the conditional distribution. For each passed message such as  $m_{43}(x_3)$ , it uses the sum-product algorithm we introduced in the chapter of exact inference. To summarize, we have:

- If  $x_j$  **unobserved**, the message sent from variable  $j$  to  $i \in N(j)$  is

$$m_{j \rightarrow i}(x_i) = \sum_{x_j} \psi_j(x_j) \psi_{ij}(x_i, x_j) \prod_{k \in N(j) \setminus i} m_{k \rightarrow j}(x_j)$$

- If  $x_j$  is **observed**, the message is

$$m_{j \rightarrow i}(x_i) = \psi_j(\bar{x}_j) \psi_{ij}(x_i, \bar{x}_j) \prod_{k \in N(j) \setminus i} m_{k \rightarrow j}(\bar{x}_j)$$

- Once the message passing stage is complete, we can compute our beliefs as

$$b(x_i) = p(x_i | x_E) \propto \psi_i(x_i) \prod_{j \in N(i)} m_{j \rightarrow i}(x_i).$$

### 4.3 Belief Propagation on Trees

The above algorithm shows how the message is passed from the leaves to the root. In a formal setting, we have four steps:

1. Choose the root  $r$  arbitrarily.
2. Pass the message from leaves to  $r$ .
3. **Pass the message from the root to the leafs.** Note that here we only pass the message from the root to the **unobserved** nodes.
4. Compute the beliefs/marginals for all unobserved nodes.

**Example 4.3.1.** We can still infer the previous example (image), and shows the complete belief propagation process. From Figure 4.1, our goal is to compute the beliefs

$$p(x_3 | \bar{x}_2, \bar{x}_4, \bar{x}_5)$$

and

$$p(x_1 | \bar{x}_2, \bar{x}_4, \bar{x}_5)$$

**Step 1:** Choose  $x_3$  to be our root.

**Step 2:** Pass the message from leaves to  $x_3$ .

$$\begin{aligned} m_{5 \rightarrow 3}(x_3) &= \psi_5(\bar{x}_5) \psi_{35}(x_3, \bar{x}_5) \\ m_{2 \rightarrow 1}(x_1) &= \psi_2(\bar{x}_2) \psi_{12}(x_1, \bar{x}_2) \\ m_{4 \rightarrow 3}(x_3) &= \psi_4(\bar{x}_4) \psi_{34}(x_3, \bar{x}_4) \\ m_{1 \rightarrow 3}(x_3) &= \sum_{x_1} \psi_1(x_1) \psi_{13}(x_1, x_3) m_{2 \rightarrow 1}(x_1) \end{aligned}$$

**Step 3:** Pass the message from the root to other unobserved nodes:

$$m_{3 \rightarrow 1}(x_1) = \sum_{x_3} \psi_3(x_3) \psi_{13}(x_1, x_3) m_{4 \rightarrow 3}(x_3) m_{5 \rightarrow 3}(x_3)$$

**Step 4:** Calculate the beliefs for  $x_1$  and  $x_3$ .

$$\begin{aligned} b(x_1) &\propto \psi_1(x_1) m_{2 \rightarrow 1}(x_1) m_{3 \rightarrow 1}(x_1) \\ b(x_3) &\propto \psi_3(x_3) m_{1 \rightarrow 3}(x_3) m_{4 \rightarrow 3}(x_3) m_{5 \rightarrow 3}(x_3) \end{aligned}$$

### Loopy Belief Propagation

The inference we just introduced has limits that the graph (MRF) must be a tree. However, in most cases such as TrueSkill models, we do not have such prerequisites. What we do instead is to keep passing the message until convergence. This is called the Loopy Belief Propagation. It works as follows:

- Initialize all messages uniformly:

$$m_{i \rightarrow j}(x_j) = (1/k, \dots, 1/k)$$

where  $k$  is the number of states  $x_j$  can take.

- Keep running BP updates until **the messages** "converges":

$$m_{j \rightarrow i}(x_i) = \sum_{x_j} \psi_j(x_j) \psi_{ij}(x_i, x_j) \prod_{k \in N(j) \setminus i} m_{k \rightarrow j}(x_j)$$

and (sometimes) normalized for stability.

- It will generally not converge, but often works fine.
- Compute beliefs  $b(x_i) \propto \psi_i(x_i) \prod_{j \in N(i)} m_{j \rightarrow i}(x_i)$ .

## Max-product belief propagation

The main difference between the sum-product and max-product belief propagation is that we are now **maximizing** over  $x_j$  instead of summarizing them.

$$m_{j \rightarrow i}(x_i) = \max_{x_j} \psi_j(x_j) \psi_{ij}(x_i, x_j) \prod_{k \in N(j) \setminus i} m_{k \rightarrow j}(x_j)$$

Now the beliefs are called max-marginals:

$$\hat{b}(x_i) = \max_{x_i} p(x_i, x_{\setminus i}) \propto \psi_i(x_i) \prod_{j \in N(i)} m_{j \rightarrow i}(x_i).$$

MAP inference: take  $\hat{x}_i := \arg \max_{x_i} \hat{b}(x_i)$  for all  $i \notin E$ .

Here is a numerical example of message passing.

## 5 Monte Carlo Methods

### 5.1 Introduction to Monte Carlo Methods

In formal definitions, **Monte Carlo methods** is a **collection of methods** involving the use of simulated random numbers to estimate some functions of a probability distribution. Simply put, we use Monte Carlo methods to draw samples from an unknown distribution  $p(x)$ . After obtaining the samples, we can then estimate  $p(x)$  via different methods such as moments, KDE, etc.

A sample from a distribution  $p(x)$  is a single realization  $x$  whose probability distribution is  $p(x)$ . However, we assume that the density form  $p(x)$  can be evaluated within a multivariate constant  $Z$ . That said:

$$p(x) = \frac{\tilde{p}(x)}{Z} \propto \tilde{p}(x)$$

Here,  $p(x)$  is very hard to evaluate due to the multivariate constant  $Z$ . Hence our primary goal is  $\tilde{p}(x)$ .

## Main objectives of sampling

Using the Monte Carlo methods, we aim to solve one or both of the following problems:

1. Generate the samples  $\{x^{(i)}\}_{i=1}^R$  from the unknown distribution  $p(x)$ .
2. To estimate the expectation of functions,  $\psi(x)$ , under the distribution  $p(x)$

$$\Phi = \mathbb{E}_{x \sim p(x)}[\phi(x)] = \int \phi(x)p(x)dx$$

**\*Note:** The notation  $\mathbb{E}_{x \sim p(x)}[\phi(x)]$  represent the expected value of the test function  $\psi(x)$  and  $x$  are the samples generated from  $p(x)$ .

## Simple Monte Carlo

Given  $\{x^{(r)}\}_{r=1}^R \sim p(x)$  we can estimate the expectation  $\mathbb{E}_{x \sim p(x)}[\phi(x)]$  using the estimator  $\hat{\phi}$ :

$$\Phi := \mathbb{E}_{x \sim p(x)}[\phi(x)] \approx \frac{1}{R} \sum_{r=1}^R \phi(x^{(r)}) := \hat{\Phi}$$

$\hat{\Phi}$  follows the same from the Law of Large Numbers (LLN).

Moreover,  $\hat{\Phi}$  is **unbiased** and its variance is  $\frac{1}{R}\text{var}[\phi(x)]$ , proof can be found here.

## 5.2 Three Sampling Methods

### Ancestral Sampling

Ancestral sampling follows the same algorithm we saw from the conditional dependencies of DAGMs. We start to sample the nodes with no parents and then sample any other conditional distribution at next steps.

**Example 5.2.1.** Suppose we have a DAGM whose joint distribution can be factorized as follows:

$$\begin{aligned} p(x_{1,\dots,5}) &= \prod_i^5 p(x_i \mid \text{parents}(x_i)) \\ &= p(x_1) p(x_2 \mid x_1) p(x_3 \mid x_1) p(x_4 \mid x_2, x_3) p(x_5 \mid x_3) \end{aligned}$$

The sampling process works as follows:

1. Start by sampling from  $p(x_1)$ .
2. Then sample from  $p(x_2 \mid x_1)$  and  $p(x_3 \mid x_1)$ .
3. Then sample from  $p(x_4 \mid x_2, x_3)$ .
4. Finally, sample from  $p(x_5 \mid x_3)$ .

For example, to sample  $p(x_2|x_1)$ , we take  $x_1$  as the time of a day and  $x_2$  as the temperature. Different time on a day will have different distribution of temperature. For example, we can assume  $x_2 \sim N(T(x_1), 3^2)$ .

However, the main shortcuts of ancestral sampling is that it can not compute the normlizing constant  $Z$ , where

$$p(x) = \frac{\tilde{p}(x)}{Z}$$

$$Z = \int \tilde{p}(x) dx$$

To compute  $Z$ , one **bad idea** is to divide  $x$  in to many intervals. However, it will be very hard if we have high-dimensional data.

To accommodate the high-dimensioanal errors, we have two other sampling methods which are **Importance Sampling** and **Rejection Sampling**.

## Importance Sampling

Instead of sampling  $p(x)$ , we can choose  $q(x)$  which is easier to sample from (like Gaussian). Our goal is still to estimate the expectation of  $\phi(x)$ . That said, we have:

$$p(x) = \frac{\tilde{p}(x)}{Z_p}$$

$$q(x) = \frac{\tilde{q}(x)}{Z_q}$$

We then generate  $R$  samples from  $q(x)$ :  $\{x^{(r)}\}_{r=1}^R$ . If the samples are also generated from  $p(x)$ , then we can use simple Monte Carlo; however, these are generated from  $q(x)$ .

To correct this, we introduce the weights, which works as:

1. we define weights as  $\tilde{w}_r = \frac{\tilde{p}(x^{(r)})}{\tilde{q}(x^{(r)})} = \frac{Z_p}{Z_q} \frac{p(x^{(r)})}{q(x^{(r)})}$  and notice that

$$\frac{1}{R} \sum_{r=1}^R \tilde{w}_r \approx \mathbb{E}_{x \sim q(x)} \left[ \frac{\tilde{p}(x)}{\tilde{q}(x)} \right] = \frac{Z_p}{Z_q} \int \frac{p(x)}{q(x)} q(x) dx = \frac{Z_p}{Z_q}$$

2. Finally, we rewrite our estimator under  $q$

$$\phi = \int \phi(x) p(x) dx = \int \phi(x) \cdot \frac{p(x)}{q(x)} \cdot q(x) dx \approx \frac{1}{R} \sum_{r=1}^R \phi(x^{(r)}) \frac{p(x^{(r)})}{q(x^{(r)})} = (*)$$

3. However, the estimator relies on  $p$ . It can only rely on  $\tilde{p}$  and  $\tilde{q}$ .

$$(*) = \frac{Z_q}{Z_p} \frac{1}{R} \sum_{r=1}^R \phi(x^{(r)}) \cdot \frac{\tilde{p}(x^{(r)})}{\tilde{q}(x^{(r)})} = \frac{Z_q}{Z_p} \frac{1}{R} \sum_{r=1}^R \phi(x^{(r)}) \cdot \tilde{w}_r$$

$$\approx \frac{\frac{1}{R} \sum_{r=1}^R \phi(x^{(r)}) \cdot \tilde{w}_r}{\frac{1}{R} \sum_{r=1}^R \tilde{w}_r} = \sum_{r=1}^R \phi(x^{(r)}) \cdot w_r = \hat{\Phi}_{iw}$$

where  $w_r = \frac{\tilde{w}_r}{\sum_{r=1}^R \tilde{w}_r}$  and  $\hat{\Phi}_{iw}$  is our importance weighted estimator.

## Rejection Sampling

The common thing between importance sampling and rejection sampling is that we use introduce an easier distribution  $q(x)$  to estimate  $p(x)$ . However, in rejection sampling, we assume:

$$c\tilde{q}(x) > \tilde{p}(x), \forall x$$

The algorithm works as follows:

1. Generate two random numbers  $x$  and  $u$ .
  - 1.1  $x$  is generated from  $q(x)$ .
  - 1.2  $u$  is generated uniformly from the interval  $[0, c\tilde{q}(x)]$ .
2. Accept or reject the sample  $x$  by comparing the value of  $u$  with  $\tilde{p}(x)$ 
  - 2.1 **Reject**  $x$  if  $u > \tilde{p}(x)$
  - 2.2 Otherwise  $x$  is accepted;  $x$  is added to our set of samples  $\{x^{(r)}\}$ .

The reason why Rejection Sampling work is that  $\mathbb{P}_{x \sim q}(x \in A \mid u \leq \tilde{p}(x)) = \mathbb{P}_{x \sim p}(x \in A)$  (*Full derivation here*).

In other words, rejection sampling also has shortcuts. In high-dimensional case, the  $c$  will be very large and hence the acceptance rate will be very small. Moreover, such  $c$  will also be hard to define.

## 6 Markov Chain Monte Carlo

### 6.1 Introduction to Markov Chains

In most cases, we assume the samples are independent. However, it is a poor assumption as this may rarely happen in reality. Therefore, we assume the data are dependent and hence we have **sequential data**. More specifically, we can use **markov chain** to model the sequential data.

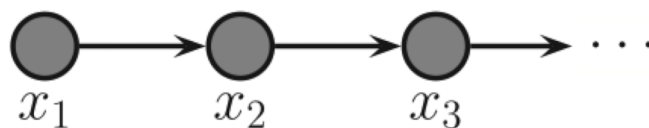


Figure 6.1: First order Markov Chain

Figure 6.1 shows an example of First order Markov chain.

$$\begin{aligned}
 p(x_t \mid x_{1:t-1}) &= p(x_t \mid x_{t-1}) \\
 &= \prod_{t=1}^T p(x_t \mid x_{t-1})
 \end{aligned}$$

From the graph, we can see that the first order means each node is conditioned on the previous node. We can therefore generalize to higher order Markov Chains.

- For second-order Markov Chains:

$$p(x_t | x_{1:t-1}) = p(x_t | x_{t-1}, x_{t-2})$$

- For m-order Markov Chains:

$$p(x_{1:T}) = p(x_t | x_{t-1:t-m})$$

We have two further definitions:

**Stationary(homogeneous) Markov Chain:** The distribution generating the data does not change over time:

$$p(x_{t+1} = y | x_t = x) = p(x_{t+2} = y | x_{t+1} = x)$$

**Non-stationary Markov Chain:**  $p(x_{t+1} = y | x_t = x)$  depends on time  $t$ .

## Transition Matrix

If we assume  $x_t$  is discrete:  $x_t \in \{1, 2, \dots, K\}$ . Then the conditional distribution  $p(x_t | x_{t-1})$  can be written as a  $K \times K$  matrix, which is called the **transition (stochastic) matrix**  $A$ . Each element on  $A$  represents the probability of transition from  $t - 1$  to  $t$ , i.e.

$$A_{ij} = p(x_t = j | x_{t-1} = i), \quad A \in K \times K$$

We also have:

$$\begin{aligned} p(x_t = j) &= \sum_i p(x_t = j | x_{t-1} = i) p(x_{t-1} = i) \\ &= \sum_i A_{ij} p(x_{t-1} = i) \end{aligned}$$

The sum for each row in  $A$  equals to **1** as it represent all the possibilities of changing from  $t - 1$  to  $t$  given  $i$  at  $t - 1$ .

$$\sum_i A_{ij} = 1$$

*An example of transition matrix can be found [here](#).*

The  $n$ -step transition matrix  $A(n)$  is the probability of transferring from two states in  $n$  steps, where  $n$  can be considered as "n" times. In equation:

$$A_{ij}(n) = p(x_{t+n} = j | x_t = i)$$

Since the one-step transition matrix considers  $t - 1$  and  $t$  (or equivalently  $t$  and  $t + 1$ )

$$A(1) = A$$

**Chapman-Kolmogorov equations:**

$$\begin{aligned} A_{ij}(m+n) &= \sum_{k=1}^K A_{ik}(m) A_{kj}(n) \\ \Leftrightarrow A(m+n) &= A(m)A(n) \end{aligned}$$

It says we can find an intermediate state  $k$  between  $i$  and  $j$ , which takes  $m$  steps from  $i$  to  $k$ , and  $n$  steps from  $k$  to  $j$ . It is similar to lattice paths on combinatorics.  
 $\Rightarrow A(n) = A(1 + n - 1) = A \cdot A(n - 1) = A \cdot A \cdot A(n - 2) = \dots = A^n$ .

## Markov Language Models

Language models can be represented as Markov Chains. That said, each word is a state, and the probability of transition from one word to another can be also represented as transition matrix.

$$\begin{aligned}
 p(x_{1:T} | \theta) &= \pi(x_1) A(x_1, x_2) \cdots A(x_{T-1}, x_T) \\
 &= \underbrace{\prod_{j=1}^K \pi_j^{1[x_1=j]}}_{\text{prob. starting with } x_1} \prod_{t=2}^T \prod_{j=1}^K \prod_{k=1}^K A_{jk}^{1[x_t=k, x_{t-1}=j]}
 \end{aligned}$$

### 6.2 Stationary distribution of a (homogeneous) Markov Chain

The stationary<sup>1</sup> distribution is the **long-term** distribution over states. In notation, we have  $\pi_t(j) = p(x_t = j)$ , to be the probability of being in state  $j$  on time  $t$ .

- We have the initial distribution which is given by  $\pi_0 \in \mathbb{R}^K$
- We then represent  $\pi_1$  as:

$$\begin{aligned}
 \pi_1(j) &= \sum_{i=1}^K p(x_1 = j | x_0 = i) \pi_0(i) = \sum_{i=1}^K A_{ij} \pi_0(i) = \sum_{i=1}^K (A^\top)_{ji} \pi_0(i) \\
 &\Rightarrow \pi_1 = A^\top \pi_0
 \end{aligned}$$

- We can therefore generalize to  $\pi_t$ .

$$\pi_t = A^\top \pi_{t-1} = A^\top A^\top \pi_{t-2} = \cdots = (A^\top)^t \pi_0$$

- If we do this for many steps; eventually, the distribution of  $\pi_t$  may converge:

$$\pi = A^\top \pi$$

which is the stationary distribution of the Markov chain.

#### Recall: Eigenvalues and Eigenvectors

Let  $T$  be a transformation,  $\mathbf{v}$  be a vector and  $\lambda$  be a constant. If we have:

$$T\mathbf{v} = \lambda\mathbf{v}$$

Then  $\mathbf{v}$  is the eigenvector of  $T$  and  $\lambda$  is the eigenvalue.

Therefore, it follows that the stationary distribution  $\pi$  is an eigenvector of  $A^\top$  with eigenvalue 1. In other words,  $A$  and  $A^\top$  have the same eigenvalues and  $A\mathbf{1} = \mathbf{1}$  as the row sum of  $A$  is 1. Thus, due to the equality of characteristics polynomials, 1 is also the eigenvalue of  $A^\top$ . More importantly, **stationary distribution is NOT unique**.

<sup>1</sup>the "stationary" in "stationary distribution" is not as same as in "stationary Markov chain." Therefore, we usually called it as homogeneous Markov chains.



## Detailed balance equation

Markov Chain is called:

- **irreducible** if we can get from any state to any other state.
- **regular** if  $A^n$  has positive entries for some  $n$ .
- **time reversible** if there exists a distribution  $\pi$  such that

$$\underbrace{\pi_i A_{ij} = \pi_j A_{ji}}_{\text{detailed balance equation}} \quad \text{for all } i, j.$$

In other words, detailed balance means the probability of transition from one state to another and reverse back is **equally possible**.

**Theorem 6.2.1.** If a markov chain with transition matrix  $A$  satisfies the detailed balance with respect to distribution  $\pi$ , then  $\pi$  is a stationary distribution.

## 7 Variational Inference

### 7.1 Introduction to Variational Inference

**variational inference** is an inference method which can help us to approximate a tractable inference. "Variational" refers to the technique in which complex probability distributions are approximated using simpler, parameterized distributions by optimizing over the parameters. We already see a similar application in rejection sampling, where  $q(x)$  is the simpler distribution and help us to estimate the desired distribution,  $p(x)$ .

### The posterior distribution of latent variable models

Recall the joint distribution  $p(x, z) = p(x)p(z|x)$ .  $x$  are the observed data points,  $z$  are the unobserved (latent) data points. Moreover:

$$\underbrace{p(z|x)}_{\text{posterior}} = \frac{\overbrace{p(x|z)}^{\text{likelihood}} \overbrace{p(z)}^{\text{prior}}}{\underbrace{p(x)}_{\text{normalizing factor}}}$$

**Example 7.1.1.** The TrueSkill model we introduced earlier is indeed a latent variable model.

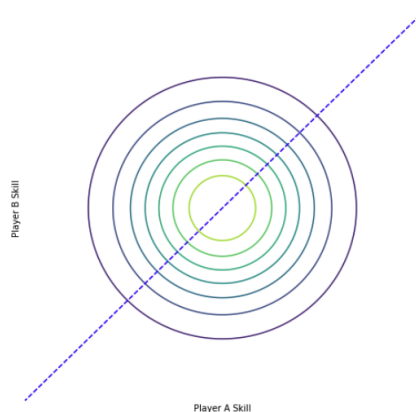


Figure 7.1: Contour plot between two players

The above graph shows the contour graph between the two players. Since we assume both players follow a Gaussian distribution, we are uncertain about players' skills and hence we observe perfect "circles." Each contour represents different values of the pdf. It goes higher when we move to the center. The blue dashed line represents equal skills between the two players. Here is a 3D view of the bivariate Gaussian distribution.

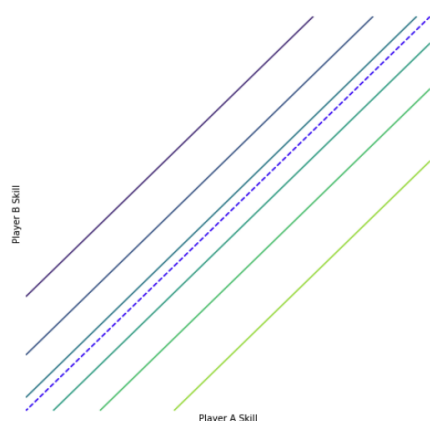


Figure 7.2: 2D projection of likelihood

The above graph shows the 2D projection of the likelihood  $p(i \text{ beats } j | z_i, z_j) = \sigma(z_i - z_j)$ .

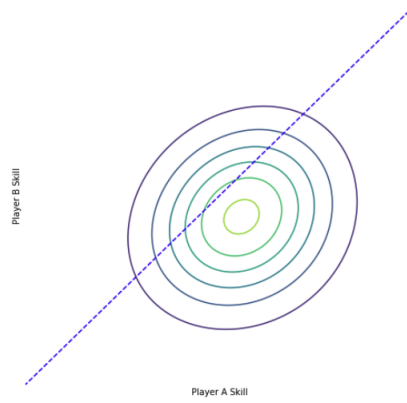


Figure 7.3: The posterior

The contours are not perfect "circles" anymore, which are not Gaussian anymore. However, the center moves to the left and we conclude that the guess player A is better than B.

## 7.2 Kullback-Leibler divergence

From the TrueSkill model, the normalizing factor  $p(x) = \int p(x, z) dz$  is intractable when  $z$  is high-dimensional. To estimate  $p(x)$ , like previously explained, our procedure is to approximate it using a simpler distribution  $q(x)$ . We measure the closeness/similarity between  $p$  and  $q$  using **KL Divergence**.

$$\text{KL}(q(z) \| p(z | x)) = \int q(z) \log \frac{q(z)}{p(z | x)} dz = \mathbb{E}_{z \sim q} \log \frac{q(z)}{p(z | x)}$$

## 8 Appendix

### 8.1 Example 1

The Bernoulli Naïve Bayes model parameterized by  $\theta$  and  $\pi$  defines the following joint probability of  $x$  and  $c$ ,

$$p(x, c | \theta, \pi) = p(c | \pi) p(x | c, \theta) = p(c | \pi) \prod_{j=1}^D p(x_j | c, \theta),$$

where  $x_j | c, \theta \sim \text{Bernoulli}(\theta_{jc})$ , i.e.  $p(x_j | c, \theta) = \theta_{jc}^{x_j} (1 - \theta_{jc})^{1-x_j}$ , and  $c | \pi$  follows a simple categorical distribution, i.e.  $p(c | \pi) = \pi_c$ .

**Solution:**

For  $\mathbf{x}^1$ , its likelihood function is:

$$L(\theta, \pi; x^1, c) = \prod_{c=1}^{10} [p(x^1, c | \theta, \pi)]^{1\{c^1=c\}} \quad (8.1)$$

$$= \prod_{c=1}^{10} \left[ p(c | \pi) \prod_{j=1}^{784} p(x_j^1 | c, \theta) \right]^{1\{c^1=c\}} \quad (8.2)$$

$$= \prod_{c=1}^{10} \left[ \pi_c \prod_{j=1}^{784} \theta_{jc}^{x_j^1} (1 - \theta_{jc})^{1-x_j^1} \right]^{1\{c^1=c\}} \quad (8.3)$$

Therefore, the joint likelihood function for  $\mathbf{x}^1, \dots, \mathbf{x}^n$  is:

$$L(\theta, \pi) = \prod_{i=1}^n \prod_{c=1}^{10} \left[ \pi_c \prod_{j=1}^{784} \theta_{jc}^{x_j^i} (1 - \theta_{jc})^{1-x_j^i} \right]^{1\{c^i=c\}} \quad (8.4)$$

$$\Rightarrow l(\theta, \pi) = \log \left( \prod_{i=1}^n \prod_{c=1}^{10} \left[ \pi_c \prod_{j=1}^{784} \theta_{jc}^{x_j^i} (1 - \theta_{jc})^{1-x_j^i} \right]^{1\{c^i=c\}} \right) \quad (8.5)$$

$$= \sum_{i=1}^n \sum_{c=1}^{10} 1\{c^i = c\} \left\{ \log(\pi_c) + \sum_{j=1}^{784} [x_j^i \log(\theta_{jc}) + (1 - x_j^i) \log(1 - \theta_{jc})] \right\} \quad (8.6)$$

$$= \sum_{i=1}^n \sum_{c=1}^9 1\{c^i = c\} \left\{ \log(\pi_c) + \sum_{j=1}^{784} [x_j^i \log(\theta_{jc}) + (1 - x_j^i) \log(1 - \theta_{jc})] \right\} \quad (8.7)$$

$$+ \sum_{i=1}^n 1\{c^i = 10\} \left\{ \log(1 - \sum_{c=1}^9 \pi_c) + \sum_{j=1}^{784} [x_j^i \log(\theta_{j,10}) + (1 - x_j^i) \log(1 - \theta_{j,10})] \right\} \quad (8.8)$$

If we pick any  $c \in [C]$  and  $j \in [D]$ :

$$\frac{\partial l(\theta, \pi)}{\partial \theta_{jc}} = \sum_{i=1}^n 1\{c^i = c\} \left( \frac{x_j^i}{\theta_{jc}} - \frac{1 - x_j^i}{1 - \theta_{jc}} \right) \quad (8.9)$$

Letting it equal to zero, we have:

$$\hat{\theta}_{jc} = \frac{\sum_{i=1}^n 1\{c^i = c\} x_j^i}{\sum_{i=1}^n 1\{c^i = c\}} \quad (8.10)$$

For  $\pi_c$ :

$$\frac{\partial l(\theta, \pi)}{\partial \pi_c} = \sum_{i=1}^n 1\{c^i = c\} \frac{1}{\pi_c} - \sum_{i=1}^n 1\{c^i = 10\} \frac{1}{1 - \sum_{c=1}^9 \pi_c} \quad (8.11)$$

Letting  $n_c = \sum_{i=1}^n 1\{c^i = c\}$ , when it equals to zero, we have:

$$n_c(1 - \sum_{c=1}^9 \hat{\pi}_c) = \hat{\pi}_c n_{10}, \quad \text{where } 1 \leq c \leq 9 \quad (8.12)$$

Summation on both sides over  $1 \leq c \leq 9$ , we have:

$$\sum_{c=1}^9 n_c(1 - \sum_{c=1}^9 \hat{\pi}_c) = \sum_{c=1}^9 \hat{\pi}_c n_{10} \quad (8.13)$$

$$\Rightarrow (n - n_{10})(1 - \sum_{c=1}^9 \hat{\pi}_c) = n_{10} \sum_{c=1}^9 \hat{\pi}_c \quad (8.14)$$

$$\sum_{c=1}^9 \hat{\pi}_c = \frac{n - n_{10}}{n} \quad (8.15)$$

Substituting back, we will have:

$$\hat{\pi}_c = \frac{n_c}{n}, \quad 1 \leq c \leq 9 \quad (8.16)$$

## 8.2 Example 2

We can write this distribution as an exponential family

$$p(x | \theta) = \theta^x (1 - \theta)^{1-x} \quad (8.17)$$

$$= \exp \{x \log(\theta) + (1 - x) \log(1 - \theta)\} \quad (8.18)$$

$$= \exp \left\{ x \log \left( \frac{\theta}{1 - \theta} \right) + \log(1 - \theta) \right\} \quad (8.19)$$

Here,

$$T(x) = x$$

$$\eta = \log \left( \frac{\theta}{1 - \theta} \right)$$

$$A(\eta) = \log(1 + e^\eta)$$

$$h(x) = 1$$

Notice that  $A'(\eta) = \frac{e^\eta}{1+e^\eta} = \theta$  is the mean of  $T(X) = X$  and  $A''(\eta) = \frac{e^\eta}{(1+e^\eta)^2} = \theta(1 - \theta)$  is the variance of  $X$ .

### 8.3 Derivations 1

We add and subtract  $\mathbb{E}[t \mid x]$  and write

$$\begin{aligned}
 \mathbb{E}[L] &= \iint (y(x) - t)^2 p(x, t) dx dt \\
 &= \iint (y(x) - \mathbb{E}[t \mid x] + \mathbb{E}[t \mid x] - t)^2 p(x, t) dx dt \\
 &= \iint (y(x) - \mathbb{E}[t \mid x])^2 p(x, t) dx dt + \iint (\mathbb{E}[t \mid x] - t)^2 p(x, t) dx dt \\
 &\quad + 2 \iint (y(x) - \mathbb{E}[t \mid x])(\mathbb{E}[t \mid x] - t) p(x, t) dx dt
 \end{aligned}$$

The last term is zero since

$$\begin{aligned}
 &\iint (y(x) - \mathbb{E}[t \mid x])(\mathbb{E}[t \mid x] - t) p(x, t) dx dt \\
 &= \iint (y(x) - \mathbb{E}[t \mid x])(\mathbb{E}[t \mid x] - t) p(t \mid x) p(x) dx dt \\
 &= \int (y(x) - \mathbb{E}[t \mid x]) \underbrace{\left\{ \int (\mathbb{E}[t \mid x] - t) p(t \mid x) dt \right\}}_{=0} p(x) dx = 0
 \end{aligned}$$

### 8.4 Example 3

This shape of tree is from Figure 4.1. To have concrete numbers, suppose all variables are binary  $\{0, 1\}$  and take  $\psi_i(x_i) \equiv 1$  with

$$\psi_{12} = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}, \quad \psi_{13} = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \quad \psi_{34} = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}, \quad \psi_{35} = \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix}$$

We have

$$p(x_1, x_2, x_3, x_4, x_5) = \frac{1}{Z} \prod_{i=1}^5 \psi_i(x_i) \psi_{12}(x_1, x_2) \psi_{13}(x_1, x_3) \psi_{34}(x_3, x_4) \psi_{35}(x_3, x_5).$$

Since  $x_E = \{x_2, x_4, x_5\}$ , we can just fix the values of three variables:  $\bar{x}_2 = 1, \bar{x}_4 = 1, \bar{x}_5 = 0$ . Then

$$p(x_1, 1, x_3, 1, 0) = \frac{1}{Z} \psi_{12}(x_1, 1) \psi_{13}(x_1, x_3) \psi_{34}(x_3, 1) \psi_{35}(x_3, 0).$$

This gives

$$\begin{aligned}
 p(0, 1, 0, 1, 0) &= \frac{1}{Z} 2 \cdot 2 \cdot 1 \cdot 1 = \frac{4}{Z} \\
 p(0, 1, 1, 1, 0) &= \frac{1}{Z} 2 \cdot 1 \cdot 2 \cdot 1 = \frac{4}{Z} \\
 p(1, 1, 0, 1, 0) &= \frac{1}{Z} 1 \cdot 1 \cdot 1 \cdot 1 = \frac{1}{Z} \\
 p(1, 1, 1, 1, 0) &= \frac{1}{Z} 1 \cdot 2 \cdot 2 \cdot 1 = \frac{4}{Z}
 \end{aligned}$$

We can also compute the joint distribution:

$$\begin{aligned}
 p(x_1, x_3 \mid x_2 = 1, x_4 = 1, x_5 = 0) &= \frac{p(x_1, 1, x_3, 1, 0)}{\sum_{x'_1, x'_3=0}^1 p(x'_1, 1, x'_3, 1, 0)} \\
 &= \frac{\frac{1}{Z} \psi_{12}(x_1, 1) \psi_{13}(x_1, x_3) \psi_{34}(x_3, 1) \psi_{35}(x_3, 0)}{\frac{1}{Z}(4 + 4 + 1 + 4)} \\
 &= \frac{\psi_{12}(x_1, 1) \psi_{13}(x_1, x_3) \psi_{34}(x_3, 1) \psi_{35}(x_3, 0)}{13} \\
 &= \frac{1}{13} \begin{bmatrix} 4 & 4 \\ 1 & 4 \end{bmatrix}
 \end{aligned}$$

For the message passing:

$$\begin{aligned}
 m_{2 \rightarrow 1}(x_1) &= \psi_2(1) \psi_{12}(x_1, 1) = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \\
 m_{4 \rightarrow 3}(x_3) &= \psi_4(1) \psi_{34}(x_3, 1) = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \\
 m_{5 \rightarrow 3}(x_3) &= \psi_5(0) \psi_{35}(x_3, 0) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}
 \end{aligned}$$

Since  $x_3$  is not observed we have

$$\begin{aligned}
 m_{3 \rightarrow 1}(x_1) &= \sum_{x_3} \psi_3(x_3) \psi_{13}(x_1, x_3) m_{4 \rightarrow 3}(x_3) m_{5 \rightarrow 3}(x_3) \\
 &= \psi_3(0) \psi_{13}(x_1, 0) m_{4 \rightarrow 3}(0) m_{5 \rightarrow 3}(0) + \psi_3(1) \psi_{13}(x_1, 1) m_{4 \rightarrow 3}(1) m_{5 \rightarrow 3}(1) \\
 &= \begin{bmatrix} 4 \\ 5 \end{bmatrix}
 \end{aligned}$$

From this we get

$$b(x_1) = p(x_1 \mid \bar{x}_2 = 1, \bar{x}_4 = 1, \bar{x}_5 = 0) \propto \psi_1(x_1) m_{2 \rightarrow 1}(x_1) m_{3 \rightarrow 1}(x_1) = \begin{bmatrix} 8 \\ 5 \end{bmatrix}$$

and so  $p(x_1 = 1 \mid \bar{x}_2 = 1, \bar{x}_4 = 1, \bar{x}_5 = 0) = \frac{5}{13}$ .

To compute  $b(x_3) = p(x_3 \mid \bar{x}_2 = 1, \bar{x}_4 = 1, \bar{x}_5 = 0)$  we need to compute the message  $m_{1 \rightarrow 3}$

$$m_{1 \rightarrow 3}(x_3) = \sum_{x_1} \psi_1(x_1) \psi_{13}(x_1, x_3) m_{2 \rightarrow 1}(x_1) = \begin{bmatrix} 5 \\ 4 \end{bmatrix}$$

This gives

$$b(x_3) \propto \psi_3(x_3) m_{1 \rightarrow 3}(x_3) m_{4 \rightarrow 3}(x_3) m_{5 \rightarrow 3}(x_3) = \begin{bmatrix} 5 \\ 8 \end{bmatrix}$$

giving that  $p(x_3 = 1 \mid \bar{x}_2 = 1, \bar{x}_4 = 1, \bar{x}_5 = 0) = \frac{8}{13}$ .

## 8.5 Derivations 2

### Unbiaseness

If the vectors  $\{x^{(r)}\}_{r=1}^R$  are generated independently from  $p(x)$ , then the expectation of  $\hat{\phi}$  is  $\Phi$ . Indeed,

$$\begin{aligned}\mathbb{E}[\hat{\Phi}] &= \mathbb{E}\left[\frac{1}{R} \sum_{r=1}^R \phi(x^{(r)})\right] = \frac{1}{R} \sum_{r=1}^R \mathbb{E}[\phi(x^{(r)})] \\ &= \frac{1}{R} \sum_{r=1}^R \mathbb{E}_{x \sim p(x)}[\phi(x)] = \frac{R}{R} \mathbb{E}_{x \sim p(x)}[\phi(x)] \\ &= \Phi\end{aligned}$$

### Variance

As the number of samples of  $R$  increases, the variance of  $\hat{\phi}$  will decrease with rate  $\frac{1}{R}$

$$\begin{aligned}\text{var}[\hat{\phi}] &= \text{var}\left[\frac{1}{R} \sum_{r=1}^R \phi(x^{(r)})\right] = \frac{1}{R^2} \text{var}\left[\sum_{r=1}^R \phi(x^{(r)})\right] \\ &= \frac{1}{R^2} \sum_{r=1}^R \text{var}[\phi(x^{(r)})] = \frac{R}{R^2} \text{var}[\phi(x)] = \frac{1}{R} \text{var}[\phi(x)]\end{aligned}$$

## 8.6 Derivations 3

**WTS:**  $\mathbb{P}_{x \sim q}(x \in A \mid u \leq \tilde{p}(x)) = \mathbb{P}_{x \sim p}(x \in A)$

Recall:

- Note:  $\mathbb{P}(u \leq \tilde{p}(x) \mid x) = \frac{\tilde{p}(x)}{c\tilde{q}(x)}$  (remember we assume  $\tilde{p}(x) < x\tilde{q}(x)$ ).
- $\forall A \subseteq \mathcal{X} : \mathbb{P}_{x \sim p}(x \in A) = \int_A p(x)dx = \int \mathbf{1}_{\{x \in A\}} p(x)dx = \mathbb{E}_{x \sim p}[\mathbf{1}_{\{x \in A\}}]$
- Law of total expectation  $\mathbb{E}[\mathbb{E}[Z \mid \mathcal{H}]] = \mathbb{E}Z$

We then have:

$$\begin{aligned}\mathbb{P}_{x \sim q}(x \in A \mid u \leq \tilde{p}(x)) &= \mathbb{P}_{x \sim q}(x \in A, u \leq \tilde{p}(x)) / \mathbb{E}_{x \sim q}[\mathbb{P}(u \leq \tilde{p}(x) \mid x)] \\ &= \mathbb{E}_{x \sim q}[\mathbf{1}_{\{x \in A\}} \mathbb{P}(u \leq \tilde{p}(x) \mid x)] / \mathbb{E}_{x \sim q}\left[\frac{\tilde{p}(x)}{c\tilde{q}(x)}\right] \\ &= \mathbb{E}_{x \sim q}\left[\mathbf{1}_{\{x \in A\}} \frac{\tilde{p}(x)}{c\tilde{q}(x)}\right] / \frac{Z_p}{cZ_q} = \mathbb{P}_{x \sim p}(x \in A) \frac{Z_p}{cZ_q} / \frac{Z_p}{cZ_q} \\ &= \mathbb{P}_{x \sim p}(x \in A)\end{aligned}$$



## 8.7 Example 4

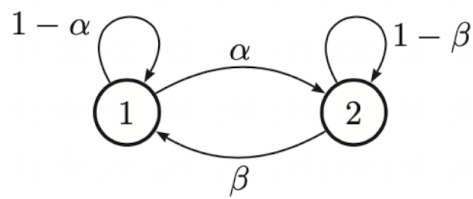


Figure 8.1: An example of transition matrix

$$A = \begin{bmatrix} 1 - \alpha & \alpha \\ \beta & 1 - \beta \end{bmatrix}$$