Deep Generative Models

Structured density

Hamid Beigy

Sharif University of Technology

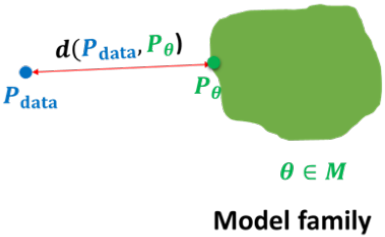
February 16, 2025

Table of contents

Introduction

1. **A Generative model** (GM) is a probability distribution $p(\mathbf{x})$.
   - A statistical GM is a trainable probabilistic model, $p_\theta(\mathbf{x})$.
   - A deep GM is a statistical generative model parametrized by a neural network.
2. A generative model needs
   - Data ($\mathbf{x}$): Complex, unstructured samples such as images, speech, molecules, text, etc.
   - Prior knowledge: parametric form (e.g., Gaussian, mixture, softmax), loss function (e.g., maximum likelihood, divergence), optimization algorithm, etc.

$\mathbf{x}_i \sim P_{\text{data}}$
$i = 1, 2, \ldots, n$



Credit: Aditya Grover

Density estimation

1. Density estimation is the problem of reconstructing the probability density function using a set of given data points.
2. Let $\mathbf{x}_1, \ldots, \mathbf{x}_m \sim p(\mathbf{x})$ be the training set.
3. The goal is to recover the underlying probability density function generating this dataset.
4. Let $\mathbf{x}_1, \ldots, \mathbf{x}_m$ be identically independently distributed random variables. Hence,

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_m) = \prod_{k=1}^{m} p(\mathbf{x}_k)$$

5. Density can be estimated using two approaches:
   - Parametric approach
   - Non-parametric approach

Parametric density estimation approach

Parametric density estimation approach

1. Let us to approximate the density function $p(\mathbf{x})$ using density function $p_\theta(\mathbf{x})$.
2. θ is parameters of $p_\theta(\mathbf{x})$.
3. There are many approaches for estimating θ such as
   - maximum likelihood method (ML)
   - maximum a posteriori probability (MAP)
   - method of moments
   - Bayesian estimation method

Example

- Let x_i be a one-dimensional real valued random variable.
- Let $p_\theta(\mathbf{x}) = \mathcal{N}(\mu, \sigma^2)$ be the target pdf, where $\theta = \{\mu, \sigma^2\}$ is its parameters.
- The goal is to estimate parameters $\theta = \{\mu, \sigma^2\}$.

1. Let $p_\theta(x) = \mathcal{N}(\mu, \sigma^2)$. Then $\theta = \{\mu, \sigma^2\}$.
2. The likelihood equals

$$L(\theta) = p_\theta(x_1, \ldots, x_n) = \prod_{k=1}^{m} p_\theta(x_k)$$

$$\mathrm{LL}(\theta) = \ln L(\theta) = \sum_{k=1}^{m} p_\theta(x_k)$$

3. By differentiating $\mathrm{LL}(\theta)$ with respect to θ and setting to zero, we obtain

$$\widehat{\mu}_m = \frac{1}{m} \sum_{k=1}^{m} x_k$$

$$\widehat{\sigma}_m^2 = \frac{1}{m} \sum_{k=1}^{m} (x_k - \widehat{\mu}_m)^2$$

4. Then the resulting density function is

$$p_\theta(x) = \frac{1}{\sqrt{2\pi\widehat{\sigma}_m^2}} e^{-\frac{1}{2}\left(\frac{x-\widehat{\mu}_m}{\widehat{\sigma}_m}\right)^2}$$

Definition (Bias of an estimator)

Let $\widehat{\theta}$ be a point estimator for θ. The bias of point estimator $\widehat{\theta}$ is defined by

$$\text{Bias}\,(\widehat{\theta}) = \mathbb{E}\left[\widehat{\theta}\right] - \theta.$$

Definition (Unbiased estimator)

Let $\widehat{\theta}$ be a point estimator for θ. We say that the point estimator $\widehat{\theta}$ is an unbiased estimator of θ if for all values of θ, we have

$$\text{Bias}\,(\widehat{\theta}) = 0.$$

Example (Unbiased estimator)

Let $\widehat{\mu}_m = \frac{1}{m}\sum_{k=1}^{m} \mathbf{x}_k$, then $\widehat{\mu}_m$ is an unbiased estimator.

$$\begin{aligned}\text{Bias}\,(\widehat{\mu}_m) &= \mathbb{E}[\widehat{\mu}_m] - \mu = \mathbb{E}\left[\frac{1}{m}\sum_{k=1}^{m}\mathbf{x}_k\right] - \mu \\ &= \frac{1}{m}\sum_{k=1}^{m}\mathbb{E}[\mathbf{x}_k] - \mu = \mu - \mu = 0.\end{aligned}$$

Example (Biased estimator)

Let $\widehat{\sigma}_m^2 = \frac{1}{m}\sum_{k=1}^{m}(\mathbf{x}_k - \widehat{\mu}_m)^2$, then $\widehat{\sigma}_m^2$ is a biased estimator.

$$
\begin{aligned}
\text{Bias}\,(\widehat{\sigma}_m^2) &= \mathbb{E}\left[\frac{1}{m}\sum_{k=1}^{m}(\mathbf{x}_k - \widehat{\mu}_m)^2\right] - \sigma^2 \\
&= \frac{1}{m}\sum_{k=1}^{m}\mathbb{E}\left[\left(\mathbf{x}_k - \frac{1}{m}\sum_{j=1}^{m}\mathbf{x}_j\right)^2\right] - \sigma^2 \\
&= \frac{1}{m}\sum_{k=1}^{m}\mathbb{E}\left[\mathbf{x}_k^2 - \frac{2}{m}\mathbf{x}_k\sum_{j=1}^{m}\mathbf{x}_j + \frac{1}{m^2}\sum_{k=1}^{m}\mathbf{x}_k\sum_{j=1}^{m}\mathbf{x}_j\right] - \sigma^2 \\
&= \frac{1}{m}\sum_{k=1}^{m}\left[\frac{m-2}{m}\mathbb{E}[\mathbf{x}_k^2] - \frac{2}{m}\sum_{j\neq k}\mathbb{E}[\mathbf{x}_k\mathbf{x}_j] + \frac{1}{m^2}\sum_{j=1}^{m}\sum_{k\neq j}\mathbb{E}[\mathbf{x}_k\mathbf{x}_j] + \frac{1}{m^2}\sum_{j=1}^{m}\mathbb{E}[\mathbf{x}_j^2]\right] - \sigma^2 \\
&= \frac{1}{m}\sum_{k=1}^{m}\left[\frac{m-2}{m}(\mu^2+\sigma^2) - \frac{2(m-1)}{m}\mu^2 + \frac{m(m-1)}{m^2}\mu^2 + \frac{1}{m}(\mu^2+\sigma^2)\right] - \sigma^2 \\
&= \frac{1}{m}\sum_{k=1}^{m}\left[\left(\frac{m-1}{m}\right)\sigma^2\right] - \sigma^2 = \left(\frac{m-1}{m}\right)\sigma^2 - \sigma^2 \neq 0.
\end{aligned}
$$

Definition (Mean squared error of an estimator)

The mean squared error (MSE) of a point estimator $\widehat{\theta}$, shown by MSE $(\widehat{\theta})$, is defined as

$$\text{MSE}\ (\widehat{\theta}) = \mathbb{E}\left[\left(\widehat{\theta} - \theta\right)^2\right].$$

Example

Let $\mathbf{x}_1, \ldots, \mathbf{x}_m$ be a random sample from a distribution with mean $\mathbb{E}[\mathbf{x}_i] = \theta$ and variance $\text{var}[\mathbf{x}_i] = \sigma^2$. Consider two estimators for θ

$$\widehat{\theta}_1 = \mathbf{x}_1 \qquad\qquad \widehat{\theta}_2 = \frac{1}{m}\sum_{k=1}^{m} \mathbf{x}_k.$$

These two estimators are both unbiased. Hence, we study their MSE:

$$\text{MSE}\ (\widehat{\theta}_1) = \mathbb{E}\left[\left(\widehat{\theta}_1 - \theta\right)^2\right] = \mathbb{E}\left[(\mathbf{x}_1 - \mathbb{E}[\mathbf{x}_1])^2\right] = \text{var}[\mathbf{x}_1] = \sigma^2.$$

$$\text{MSE}\ (\widehat{\theta}_2) = \mathbb{E}\left[\left(\widehat{\theta}_2 - \theta\right)^2\right] = \mathbb{E}\left[\left(\frac{1}{m}\sum_{k=1}^{m} \mathbf{x}_k - \theta\right)^2\right] = \frac{\sigma^2}{m}.$$

Thus, MSE $(\widehat{\theta}_1) >$ MSE $(\widehat{\theta}_2)$. Hence, $\widehat{\theta}_2$ is better.

Theorem

Let $\widehat{\theta}$ is a point estimator for θ. Then $\text{MSE}\ (\widehat{\theta}) = \text{var}[\widehat{\theta}] + \text{Bias}\ (\widehat{\theta})^2$

Proof.

We can write

$$\begin{aligned}
\text{MSE}\ (\widehat{\theta}) &= \mathbb{E}\left[\left(\widehat{\theta} - \theta\right)^2\right] \\
&= \mathbb{E}\left[\left(\widehat{\theta} - \mathbb{E}\left[\widehat{\theta}\right] + \mathbb{E}\left[\widehat{\theta}\right] - \theta\right)^2\right] \\
&= \underbrace{\mathbb{E}\left[\left(\widehat{\theta} - \mathbb{E}\left[\widehat{\theta}\right]\right)^2\right]}_{=\text{var}[\widehat{\theta}]} + 2\underbrace{\left(\widehat{\theta} - \mathbb{E}\left[\widehat{\theta}\right]\right)}_{=0} \cdot \left(\mathbb{E}\left[\widehat{\theta}\right] - \theta\right) + \left(\underbrace{\mathbb{E}\left[\widehat{\theta}\right] - \theta}_{\text{Bias}\ (\widehat{\theta})}\right)^2 \\
&= \text{var}[\widehat{\theta}] + \text{Bias}\ (\widehat{\theta})^2.
\end{aligned}$$

□

This decomposition is also known as the **bias-variance trade-off**.

Definition (Consistency of an estimator)

Let $\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_n, \ldots$ be a sequence of point estimators of θ. We say $\widehat{\theta}_n$ is a **consistent estimator** of θ, if

$$\lim_{n\to\infty} p(|\widehat{\theta}_n - \theta| \geq \epsilon) = 0, \text{ for all } \epsilon > 0.$$

Example (Consistency of sample average)

Let $\mathbf{x}_1, \ldots, \mathbf{x}_m$ be a random sample from a distribution with mean $\mathbb{E}[\mathbf{x}_i] = \theta$ and variance $\text{var}[\mathbf{x}_i] = \sigma^2$. Consider the following estimator for θ

$$\widehat{\theta}_m = \frac{1}{m}\sum_{k=1}^{m} \mathbf{x}_k.$$

We have found that $\text{MSE}\,(\widehat{\theta}_m) = \frac{\sigma^2}{m}$. Thus,

$$\lim_{m\to\infty} \text{MSE}\,(\widehat{\theta}_m) \to 0.$$

Hence, this estimator is consistent.

Theorem (Consistency of an estimator)

Let $\widehat{\theta}_1, \widehat{\theta}_2, \ldots$ be a sequence of point estimators of θ. If $\lim_{n\to\infty} \text{MSE}\,(\widehat{\theta}_n) = 0$, then $\widehat{\theta}_n$ is **a consistent estimator** *of θ.*

Proof.

We can write

$$\begin{aligned} \mathrm{p}(|\widehat{\theta}_n - \theta| \geq \epsilon) &= \mathrm{p}(|\widehat{\theta}_n - \theta|^2 \geq \epsilon^2) \\ &\leq \frac{\mathbb{E}\left[\left(\widehat{\theta}_n - \theta\right)^2\right]}{\epsilon^2} \qquad \text{Using Markov's inequality} \\ &= \frac{\text{MSE}\,(\widehat{\theta}_n)}{\epsilon^2}, \end{aligned}$$

which goes to 0 as $n \to \infty$ by the assumption. □

Note: Let $\mathbf{x}$ be a nonnegative random variable and $a > 0$, then $p(\mathbf{x} \geq a) \leq \frac{\mathbb{E}[\mathbf{x}]}{a}$.

Definition (Convergence in Probability)

A sequence of random variables $\mathbf{z}_1, \mathbf{z}_2, \ldots$ **converges in probability** to a random variable $\mathbf{z}$, shown by $\mathbf{z}_n \xrightarrow{p} \mathbf{z}$, if

$$\lim_{n\to\infty} p(|\mathbf{z}_n - \mathbf{z}| \geq \epsilon) = 0, \qquad \text{for all } \epsilon > 0.$$

This implies that the distribution is concentrating at the targeting point.

Lemma

Let $\widehat{\theta}$ be an estimator of θ. If Bias $(\widehat{\theta}) \to 0$ *and* $\text{var}[\widehat{\theta}] \to 0$, *then $\widehat{\theta} \xrightarrow{p} \theta$, i.e. $\widehat{\theta}$ is a* **consistent** *estimator of θ.*

Definition (Convergence in Distribution)

1. Let $F_1, F_2, \ldots$ be the corresponding CDFs of $\mathbf{z}_1, \mathbf{z}_2, \ldots$.
2. For a random variable $\mathbf{z}$ with CDF F, we say $\mathbf{z}_n$ **converges in distribution** to a random variable $\mathbf{z}$, shown by $\mathbf{z}_n \xrightarrow{d} \mathbf{z}$, if

$$\lim_{n \to \infty} F_n(\mathbf{x}) = F(\mathbf{x}),$$

3. This implies that F_n converge to the CDF of a fixed random variable.

Stochastic order notations

Definition

For a sequence of numbers a_n (indexed by n), we write

1. $a_n = o(1)$ if $\lim_{n\to\infty} a_n \to 0$. For another sequence b_n, we write $a_n = o(b_n)$ if $\frac{a_n}{b_n} = o(1)$.
2. $a_n = O(1)$ if for all large n, there exists a constant C such that $|a_n| < C$. For another sequence b_n, we write $a_n = O(b_n)$ if $\frac{a_n}{b_n} = O(1)$.

Example

1. Let $a_n = \frac{2}{n}$. Then $a_n = o(1)$ and $a_n = O(\frac{1}{n})$.
2. Let $b_n = n + 5 + \log n$. Then $b_n = O(n)$ and $b_n = o(n^2)$ and $b_n = o(n^3)$.
3. Let $c_n = 1000n + 10^{-10}n^2$. Then $c_n = O(n^2)$ and $b_n = o(n^2 \cdot \log n)$.

The O and o notations give us a way to compare convergence/divergence rate of a sequence of (non-random) numbers.

The O_p and o_p are similar notations to O and o but are designed for random numbers.

Definition

For a sequence of random variables $\mathbf{x}_n$, we write

1. $\mathbf{x}_n = o_p(1)$ if for any $\epsilon > 0$,

$$\lim_{n\to\infty} p(|\mathbf{x}_n| > \epsilon) \to 0$$

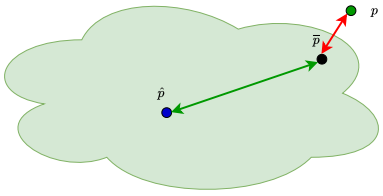
   Namely, $p(|\mathbf{x}_n| > \epsilon) = o_p(1)$ for any $\epsilon > 0$.
   Let a_n be a nonrandom sequence, we write $\mathbf{x}_n = o_p(a_n)$ if $\frac{\mathbf{x}_n}{a_n} = o_p(1)$.
2. $\mathbf{x}_n = O_p(1)$ if for any $\epsilon > 0$, there exists a constant C such that

$$p(|\mathbf{x}_n| > C) < \epsilon.$$

   We write $\mathbf{x}_n = O_p(a_n)$ if $\frac{\mathbf{x}_n}{a_n} = O_p(1)$.

Is the parametric approach a good one? We analyze the quality of estimation in the parametric approach for Gaussian distribution.

1. We quantify $p_{\theta_n}(\mathbf{x}) - p(\mathbf{x})$.



2. Since the sample mean $\widehat{\mu} \xrightarrow{p} \overline{\mu} = \mathbb{E}[\mathbf{x}]$ and the sample variance $\widehat{\sigma}^2 \xrightarrow{p} \overline{\sigma}^2 = \text{var}[\mathbf{x}]$, we define another density function

$$\overline{p}_{\theta}(\mathbf{x}) = \frac{1}{\sqrt{2\pi\overline{\sigma}^2}} e^{-\frac{1}{2}\left(\frac{\mathbf{x}-\overline{\mu}_n}{\overline{\sigma}_n}\right)^2}$$

3. The estimated density function is

$$p_{\theta_n}(\mathbf{x}) = \frac{1}{\sqrt{2\pi\widehat{\sigma}^2}} e^{-\frac{1}{2}\left(\frac{\mathbf{x}-\widehat{\mu}_n}{\widehat{\sigma}_n}\right)^2}$$

1. Using $\overline{p}_\theta(\mathbf{x})$, we have

$$p_{\theta_n}(\mathbf{x}) - p(\mathbf{x}) = p_{\theta_n}(\mathbf{x}) - \overline{p}_\theta(\mathbf{x}) + \overline{p}_\theta(\mathbf{x}) - p(\mathbf{x})$$

2. The first difference $p_{\theta_n}(x) - \overline{p}_\theta(x)$ is something that converges to 0 because the sample mean and variance converges to their population counterparts. Namely, we have $p_{\theta_n}(\mathbf{x}) \xrightarrow{p} \overline{p}_\theta(\mathbf{x})$.

3. However, the second difference $\overline{p}_\theta(\mathbf{x}) - p(\mathbf{x})$ never goes to 0 unless the true pdf is Gaussian.

$$p_{\theta_n}(x) = \frac{1}{\sqrt{2\pi\widehat{\sigma}^2}} e^{-\frac{1}{2}\left(\frac{x-\widehat{\mu}_n}{\widehat{\sigma}_n}\right)^2}$$

4. It can be shown that the convergence rate of $p_{\theta_n}(x) - \overline{p}_\theta(x)$ equals to

$$p_{\theta_n}(x) - \overline{p}_\theta(x) = O_p\left(\frac{1}{\sqrt{n}}\right).$$

5. **This will help us understand when a parametric approach may be better than a nonparametric one**.

Gaussian mixture model

1. Let the parametric model be

$$p_{\theta_n}(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \sigma_k^2)$$

$$\sum_{k=1}^{K} \pi_k = 1$$

2. We compute parameters $\theta = \{\mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2, \pi_1, \ldots, \pi_K\}$ based on training data.
3. We use EM algorithm to estimate the parameters.
4. The convergence rate of $p_{\theta_n}(x) - \overline{p}_{\theta}(x)$ equals to

$$p_{\theta_n}(x) - \overline{p}_{\theta}(x) = O_p\left(\frac{1}{\sqrt{n}}\right).$$

Product of experts

1. Let $p(\mathbf{x})$ be a high-dimensional probability distribution.
2. We can approximate $p(\mathbf{x})$ using the product of several one-dimensional distributions.
3. This model is called the **product of experts** (PoE).
4. Let n expert models $p_{\theta_1}(\mathbf{x}), \ldots, p_{\theta_m}(\mathbf{x})$, each parameterized by $\theta_1, \ldots, \theta_m$, respectively.
5. The probability distribution of the PoE can be expressed as:

$$p_\theta(\mathbf{x}) = \frac{\prod_k p_{\theta_k}(\mathbf{x})}{\sum_{\mathbf{z}} \prod_k p_{\theta_k}(\mathbf{z})}. \tag{1}$$

   where $\theta = \{\theta_1, \ldots, \theta_m\}$.
6. We will study the training algorithm for finding $\theta = \{\theta_1, \ldots, \theta_m\}$ later.

Nonparametric density estimation approach

1. For simplicity, we assume that $x_i \in [0,1]$. So $p(x) > 0$ in interval $[0,1]$.
2. We also assume that $p(x) > 0$ is smooth and $|\, p(x)' | \leq L$ for all x.
3. In histogram we partition interval $[0,1]$ into M bins (B_k) of equal widths as

$$B_k = \left[\frac{k-1}{M}, \frac{k}{M}\right]$$

4. Then, we count the number of samples in a bin as density estimate.
5. Hence, for any point $x \in B_l$, the density estimator from the histogram will be

$$\widehat{p}_n(x) = \frac{|B_l|}{n} \times \frac{1}{\text{len}\,(B_l)} = \frac{M}{n} \sum_{i=1}^{n} \mathbb{I}[x_i \in B_l]$$

6. The histogram density estimator has the following bounds (Drive the following bounds.)

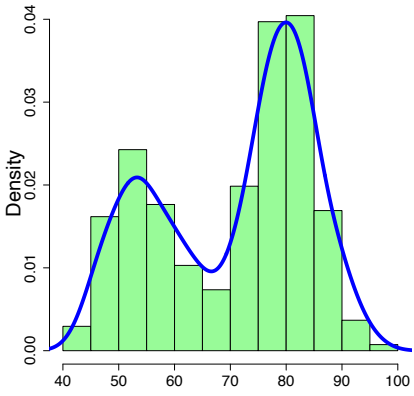
$$\text{Bias}\,(\widehat{p}_n(x)) \leq \frac{L}{M}$$

$$\text{var}[\,\widehat{p}_n(x)] = M\frac{p(x^*)}{n} + \frac{(\,p(x^*))^2}{n}$$

$$\text{MSE}\,(\widehat{p}_n(x)) \leq \frac{L}{M} + M\frac{p(x^*)}{n} + \frac{(\,p(x^*))^2}{n}$$

1. To balance the bias and variance, we choose M that minimizes the MSE, which leads to

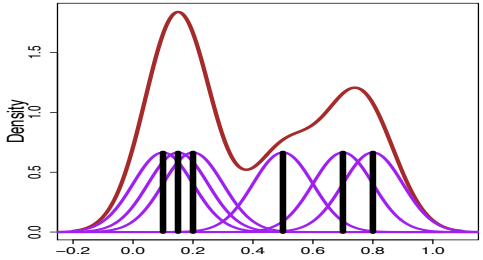
$$M_{opt} = \left(\frac{n \times L^2}{p(x^*)} \right)$$



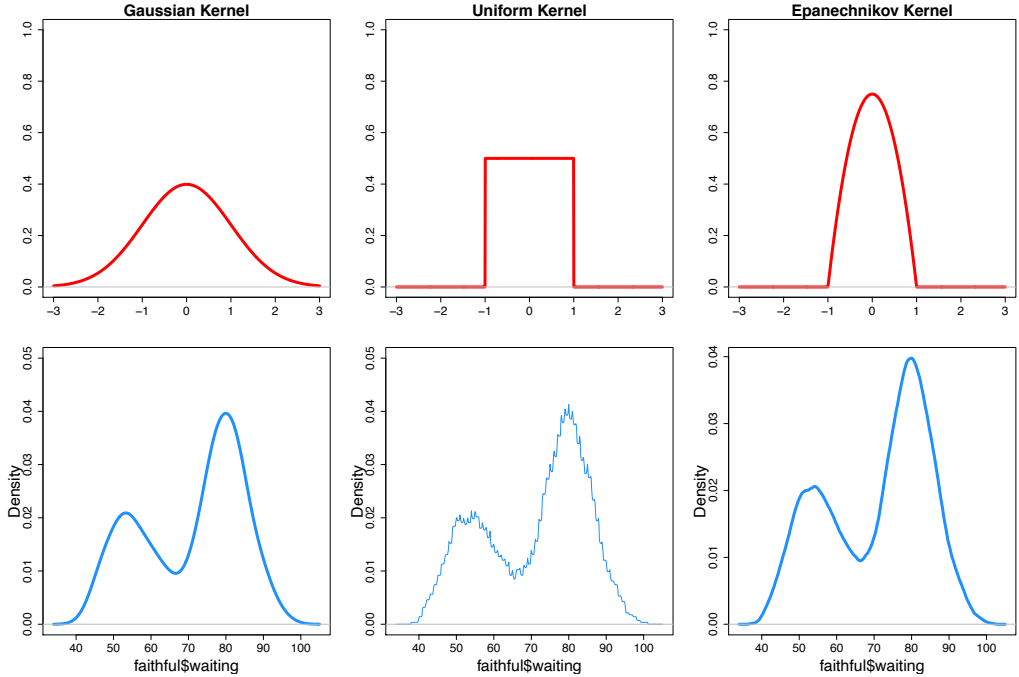
1. The KDE is a function of

$$\widehat{p}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)$$

2. $K(x)$ is kernel function and is a smooth, symmetric function such as Gaussian.
   - $K(x)$ is symmetric.
   - $\int K(x)dx = 1$
   - $\lim_{|x|\to\infty} K(x) = 0$
3. $h > 0$ is called the smoothing bandwidth that controls the amount of smoothing.



Kernel density estimator



1. The bias of KDE is

$$\text{Bias}\ (\widehat{p}_n(x_0)) = \frac{1}{2}h^2 \frac{d^2\, p(x_0)}{dx^2}\mu_K + o(h^2) \qquad \mu_K = \int y^2 K(y)dy$$

2. This means that when we allow $h \to 0$, the bias is shrinking at a rate $O(h^2)$.

3. The upper bound of variance of KDE is

$$\text{var}[\,\widehat{p}_n(x_0)] = \frac{1}{nh}\, p(x_0)\sigma_K^2 + o\left(\frac{1}{nh}\right) \qquad \sigma_K^2 = \int K^2(y)dy$$

4. Putting both bias and variance together, we obtain MSE of KDE:

$$\text{MSE}\ (\widehat{p}_n(x_0)) = O(h^4) + O\left(\frac{1}{nh}\right)$$

5. The optimal bandwidth equals to

$$h_{opt} = C_1 n^{-\frac{1}{5}}$$

6. This choice of smoothing bandwidth leads to a MSE at rate

$$\text{MSE}\ (\widehat{p}_n(x_0)) = O\left(n^{-\frac{1}{5}}\right)$$

Structured density

The number of parameters of density estimators

1. Let $\mathbf{x} = \{x_1, \ldots, x_d\}$ be an $d-$dimensional random variable where $x_i \in \{0, 1\}$.
2. How many parameters do we need to estimate the density function?

Sample	x_d	x_{d-1}	$\ldots$	x_2	x_1
1	0	0	$\ldots$	0	0
2	0	0	$\ldots$	0	1
3	0	0	$\ldots$	1	0
4	0	0	$\ldots$	1	1
			$\vdots$		
2^d	1	1	$\ldots$	1	1

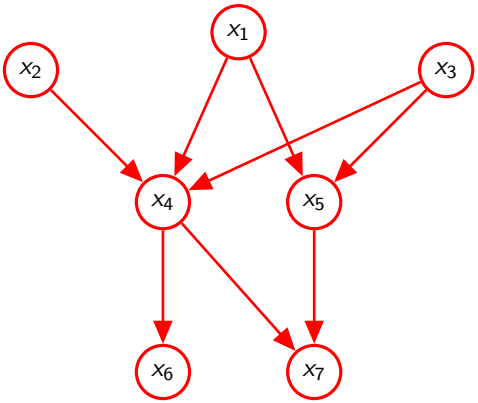
3. How can we decrease the number of parameters?

How to find the structure of density functions?

1. One way is to use probabilistic graphical models.
2. A (probabilistic) graphical model defines a family of probability distributions over a set of random variables, by means of a graph.
3. These models offer several useful properties:
   - They provide a simple way to visualize the structure of a probabilistic model and can be used to design and motivate new models.
   - Insights into the properties of the model, including conditional independence properties, can be obtained by inspection of the graph.
   - Complex computations, required to perform inference and learning in sophisticated models, can be expressed in terms of graphical manipulations, in which underlying mathematical expressions are carried along implicitly.

Graph

1. A graph $G = (V, E)$ comprises nodes (vertices) V connected by links (edges or arcs) E.
   - Each node represents a random variable (or group of random variables).
   - Each link express probabilistic relationships between these variables.
   - The graph captures joint distribution over random variables and can be decomposed into a product of factors each depending only on a subset of the variables.

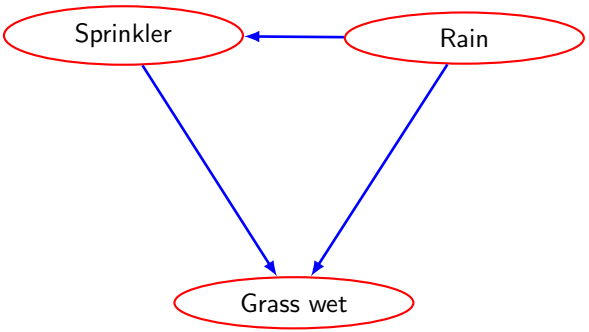


Probabilistic graphical model

1. Some types of probabilistic graphical models:
   - Bayesian networks,
   - Markov random fields,
   - Factor graphs
2. Important problems in probabilistic graphical models:
   - Structure learning,
     - Constraint-based approach
     - Score-based approach
     - Hybrid-approach
   - Parameter learning
   - Probabilistic inference : Compute marginal probabilities $p(x \mid)$

Structured density function

Rain	Sprinkler T	Sprinkler F
F	0.4	0.6
T	0.01	0.99



Sprinkler T	Sprinkler F
0.2	0.8

Sprinkler	rain	Grass wet T	Grass wet F
F	F	0.4	0.6
F	T	0.01	0.99
T	F	0.01	0.99
T	T	0.01	0.99

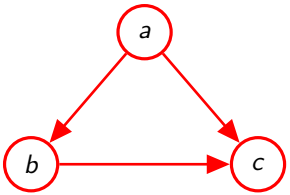
Structured density

Bayesian networks

1. Let $p(a, b, c)$ be joint distribution over three variables a, b, and c.
2. By application of the product rule of probability, we can write the joint distribution as

$$p(a, b, c) = p(c \mid a, b)\, p(a, b)$$
$$p(a, b, c) = p(c \mid a, b)\, p(b \mid a)\, p(a)$$

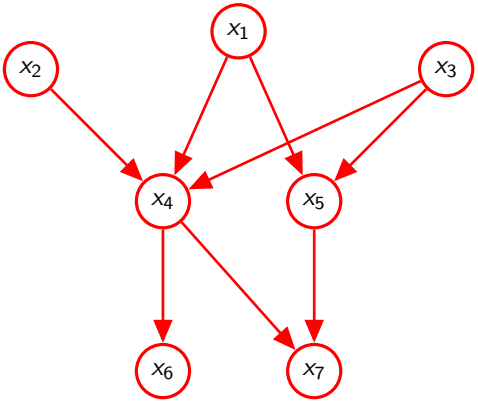
3. This decomposition holds for any choice of the joint distribution.



4. An interesting point: $p(a, b, c)$ is symmetrical with respect to a, b, and c, whereas $p(c \mid a, b)\, p(b \mid a)\, p(a)$ is not.
5. Generalization to K variables:

$$p(x_1, \ldots, x_K) = p(x_K \mid x_1, \ldots, x_{K-1}) \ldots p(x_2 \mid x_1)$$

1. Consider the following Bayesian networks



2. The joint distribution of all $x_1, \ldots, x_7$ variables is

$$p(x_1, \ldots, x_7) = p(x_1)\, p(x_2)\, p(x_3)\, p(x_4 \mid x_1, x_2, x_3)\, p(x_5 \mid x_1, x_3)\, p(x_6 \mid x_4)\, p(x_7 \mid x_4, x_5).$$

3. For a graph with K nodes, the joint distribution is

$$p(x_1, \ldots, x_K) = \prod_{k=1}^{K} p(x_k \mid \mathrm{pa}_k).$$

Conditional independence

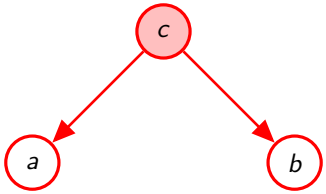
1. An important concept for probability distributions over multiple variables is **conditional independence**.
2. For three variables a, b, c, and suppose $p(a \mid b, c)$ does not depend on the value of b.

$$p(a \mid b, c) = p(a \mid c)$$

3. a is conditionally independent of b given c.

$$\begin{aligned} p(a, b \mid c) &= p(a \mid b, c)\, p(b \mid c) \\ &= p(a \mid c)\, p(b \mid c). \end{aligned}$$

4. A shorthand notation for conditional independence $a \perp\!\!\!\perp b \mid c$

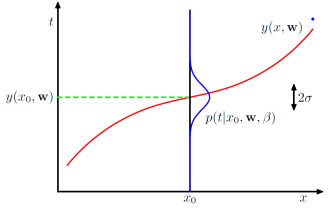


$$\begin{aligned} p(a, b, c) &= p(a \mid c)\, p(b \mid c)\, p(c) \\ p(a, b \mid c) &= \frac{p(a, b, c)}{p(c)} \\ &= p(a \mid c)\, p(b \mid c). \end{aligned}$$

Obtaining the conditional independence property $a \perp\!\!\!\perp b \mid c$.

1. Consider the regression model in which
   - $\mathbf{x} = (x_1, \ldots, x_m)$ is set of m iid observations
   - $\mathbf{t} = (t_1, \ldots, t_m)$ is the corresponding target values
   - t_k is actual value plus a Gaussian noise value with precision β.
2. Let $y(x, \mathbf{w})$ be the predicted function and the goal is to make predictions of target variable t for new input x.

$$p(t \mid x, \mathbf{w}, \beta) = \mathcal{N}\big(t \mid y(x, \mathbf{w}), \beta^{-1}\big)$$



3. Using training data $\{\mathbf{x}, \mathbf{t}\}$, we can determine $\mathbf{w}$ and β by MLE.

$$p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta) = \prod_{k=1}^{K} \mathcal{N}\big(t_k \mid y(x_k, \mathbf{w}), \beta^{-1}\big)$$

1. Let introduce a prior distribution over parameters $\mathbf{w}$ as

$$p(\mathbf{w} \mid \alpha) = \mathcal{N}\left(\mathbf{w} \mid 0, \alpha^{-1}\mathbf{I}\right)$$

   where α is the precision of the distribution.

2. The posterior distribution for $\mathbf{w}$ can be estimated using MAP as

$$p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \alpha, \beta)\, p(\mathbf{w} \mid \alpha).$$

3. In Bayesian regression model, for a new point x, we need to predict value t as

$$p(t \mid x, \mathbf{x}, \mathbf{t}) = \int p(t \mid x, \mathbf{w})\, p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}) dw.$$

   where we assume that parameters α and β are fixed and known in advance.

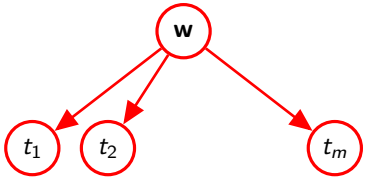
4. The random variables are parameters $\mathbf{w}$ and observed data $\mathbf{t} = (t_1, \ldots, t_m)$.

5. In addition, this model contains input data $\mathbf{x} = (x_1, \ldots, x_m)$ and parameters α and β.

1. By focusing only on random variables, the joint distribution is

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{k=1}^{m} p(t_k \mid \mathbf{w}).$$

2. The conditional distributions $p(t_k \mid \mathbf{w})$ (for $k = 1, \ldots, m$) is



3. The random variables in this model are t
   - the vector of coefficients $\mathbf{w}$
   - the observed data $\mathbf{t} = (t_1, \ldots, t_m)$.
4. Other parameters are not random variables
   - the input data $\mathbf{x} = (x_1, \ldots, x_m)$
   - the noise precision β and the hyper-parameter α.

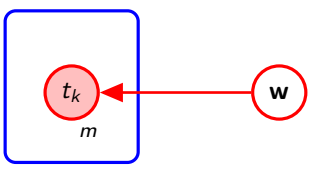
1. The joint distribution $p(\mathbf{t}, \mathbf{w})$ is

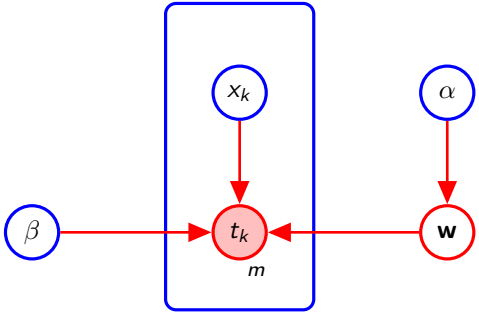
$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{k=1}^{m} p(t_k \mid \mathbf{w}).$$

2. Sometimes it is helpful to make the parameters of a model, as well as its random variables, explicit.

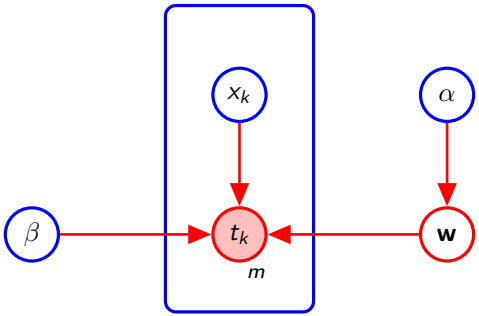
$$p(\mathbf{t}, \mathbf{w} \mid \mathbf{x}, \alpha, \beta) = p(\mathbf{w} \mid \alpha) \prod_{k=1}^{m} p(t_k \mid \mathbf{w}, x_k, \beta).$$

3. We can represent it in graphical notations.





1. Having observed values $\{t_k\}$ we can evaluate the posterior distribution of **w**



$$p(\mathbf{w} \mid \mathbf{t}) \propto p(\mathbf{w}) \prod_{k=1}^{m} p(t_k \mid \mathbf{w})$$

2. Let new input $\widehat{x}$ is given and we wish to find the corresponding probability distribution for $\widehat{t}$ conditioned on the observed data.

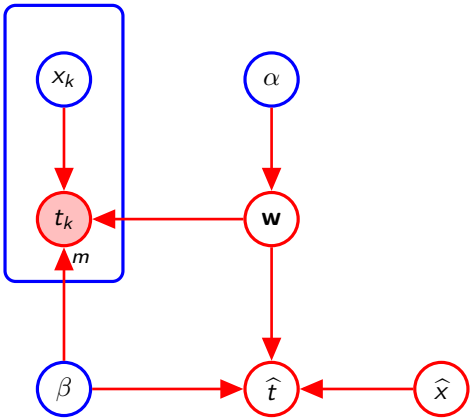
3. The joint distribution of **all random variables** conditioned on **deterministic parameters** is

$$p(\widehat{t}, \mathbf{t}, \mathbf{w} \mid \widehat{x}, \mathbf{x}, \alpha, \beta) = \left[\prod_{k=1}^{m} p(t_k \mid x_k, \mathbf{w}, \beta) \right] p(\mathbf{w} \mid \alpha)\, p(t_k \mid \widehat{x}, \mathbf{w}, \beta)$$

1. The joint distribution of all random variables conditioned on deterministic parameters is

$$p(\widehat{t}, \mathbf{t}, \mathbf{w} \mid \widehat{x}, \mathbf{x}, \alpha, \beta) = \left[\prod_{k=1}^{m} p(t_k \mid x_k, \mathbf{w}, \beta)\right] p(\mathbf{w} \mid \alpha)\, p(t_k \mid \widehat{x}, \mathbf{w}, \beta)$$

2. The corresponding graphical model is

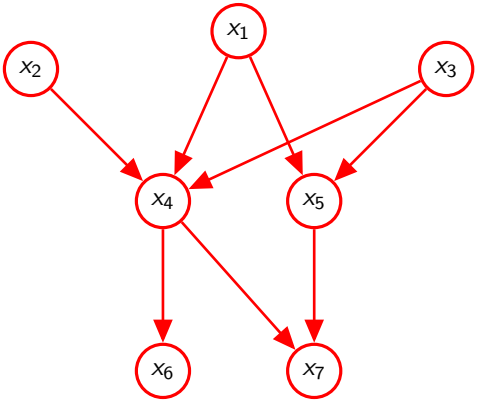


$$p(\widehat{t} \mid \widehat{x}, \mathbf{x}, \alpha, \beta) = \int p(\widehat{t}, \mathbf{t}, \mathbf{w} \mid \widehat{x}, \mathbf{x}, \alpha, \beta) dw$$

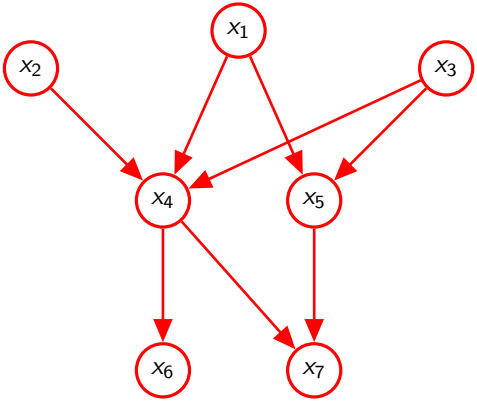
3. We are implicitly setting the random variables in $\mathbf{t}$ to the specific values observed in the data set.

Generative models

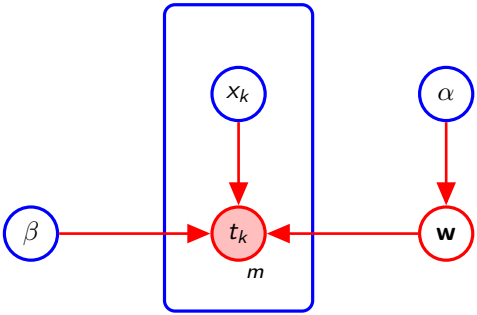
1. In some situations we wish to draw samples from a given probability distribution.
2. Let $p(x_1, \ldots, x_d)$ be the joint distribution over d variables.
3. The goal is to draw a sample $(x_1, \ldots . x_d)$ from the joint distribution.
4. To do this (suppose that the variables have been ordered such that there are no links from any node to any lower numbered node),
   - 4.1 Start with the lowest-numbered node and draw a sample from $p(x_1)$, and call $\widehat{x}_1$.
   - 4.2 For a node x_k, draw a sample from the conditional distribution $p(x_k \mid \mathrm{pa}_k)$
   - 4.3 Continue until the last variable is being sampled.



1. To obtain a sample from some marginal distribution corresponding to a subset of the variables:
   1.1 we simply take the sampled values for the required nodes and
   1.2 ignore the sampled values for the remaining nodes.

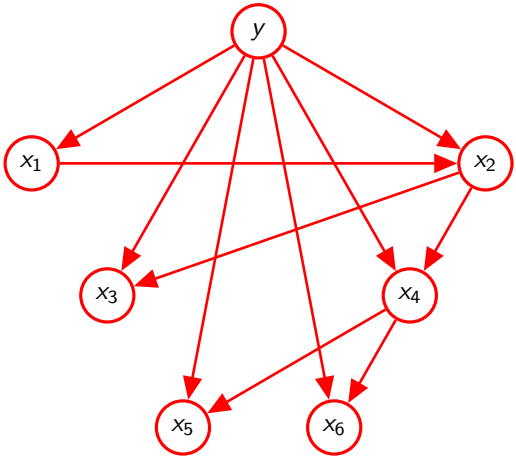


1. Consider the following graphical model: **Is it generative?**



2. This model is not generative because **there is no probability distribution associated with the input variable** x.
3. So it is not possible to generate synthetic data points from this model.
4. Can we make the above model generative?
5. We could make it generative by introducing a suitable prior distribution $p(x)$, at the expense of a more complex model.

1. Consider the following graphical model.



2. How do you compute $p(y \mid x_5)$?

3. The joint distribution $p(y, x_1, x_2, x_3, x_4, x_5, x_6)$ equals to

$$p(y, x_1, x_2, x_3, x_4, x_5, x_6) = p(y)\, p(x_1 \mid y)\, p(x_2 \mid x_1, y)\, p(x_3 \mid x_2, y)$$
$$p(x_4 \mid x_2, y)\, p(x_5 \mid x_4, y)\, p(x_6 \mid x_4, y)$$

$$
\begin{aligned}
p(y \mid x_5) &\propto \sum_{x_1}\sum_{x_2}\sum_{x_3}\sum_{x_4}\sum_{x_6} p(y)\, p(x_1 \mid y)\, p(x_2 \mid x_1, y)\, p(x_3 \mid x_2, y)\, p(x_4 \mid x_2, y)\, p(x_5 \mid x_4, y)\, p(x_6 \mid x_4, y) \\
&= \sum_{x_1}\sum_{x_2}\sum_{x_4} p(y)\, p(x_1 \mid y)\, p(x_2 \mid x_1, y)\, p(x_4 \mid x_2, y)\, p(x_5 \mid x_4, y) \underbrace{\sum_{x_3} p(x_3 \mid x_2, y) \underbrace{\sum_{x_6} p(x_6 \mid x_4, y)}_{=1}}_{=1} \\
&= p(y) \sum_{x_1} p(x_1 \mid y) \sum_{x_2} p(x_2 \mid x_1, y) \underbrace{\sum_{x_4} p(x_4 \mid x_2, y)\, p(x_5 \mid x_4, y)}_{m_4(x_2)} \\
&= p(y) \sum_{x_1} p(x_1 \mid y) \underbrace{\sum_{x_2} p(x_2 \mid x_1, y) m_4(x_2)}_{m_2(x_1)} \\
&= p(y) \underbrace{\sum_{x_1} p(x_1 \mid y) m_2(x_1)}_{m_1} = p(y) m_1.
\end{aligned}
$$

The order of summations is important.

Consider ordering x_4, x_1, x_2, y, x_3.

$$p(x_3 \mid x_5) \propto \sum_y p(y) \sum_{x_2} p(x_3 \mid x_2, y) \sum_{x_1} p(x_2 \mid x_1, y)\, p(x_1 \mid y) \underbrace{\sum_{x_4} p(x_4 \mid x_2, y)\, p(x_5 \mid x_4, y)}_{m_4(x_2, y)}$$

$$= \sum_y p(y) \sum_{x_2} p(x_3 \mid x_2, y) \underbrace{\sum_{x_1} p(x_2 \mid x_1, y)\, p(x_1 \mid y) m_4(x_2, y)}_{m_1(x_2, y)}$$

$$= \sum_y p(y) \underbrace{\sum_{x_2} p(x_3 \mid x_2, y) m_1(x_2, y)}_{m_2(y)}$$

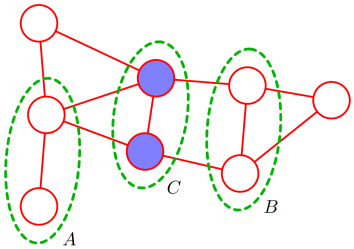
$$= \underbrace{\sum_y p(y) m_2(y)}_{m_y}.$$

Structured density

Markov Random Fields

Markov random field

1. A Markov random field, also known as a Markov network or an undirected graphical model, has
   - a set of nodes each of which corresponds to a variable or group of variables and
   - a set of links each of which connects a pair of nodes.
2. The links are undirected, that is they do not carry arrows.



3. In above undirected graph every path from any node in set A to any node in set B passes through at least one node in set C. Hence,

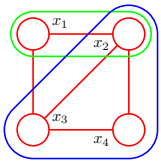
$$A \perp\!\!\!\perp B \mid C$$

Factorization properties

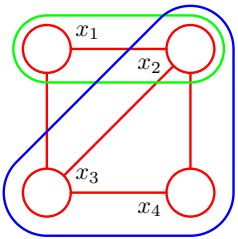
1. We need a factorization rule for undirected graphs that correspond to the conditional independence test.
2. Consider two nodes x_i and x_j that are not connected by a link, then these variables must be conditionally independent given all other nodes in the graph.
3. This conditional independence property can be expressed as

$$p(x_i, x_j \mid \mathbf{x}_{\setminus\{i,j\}}) = p(x_i \mid \mathbf{x}_{\setminus\{i,j\}})\, p(x_j \mid \mathbf{x}_{\setminus\{i,j\}})$$

4. The factorization of the joint distribution must be such that x_i and x_j do not appear in the same factor in order for the conditional independence property to hold for all possible distributions belonging to the graph.
5. This leads us to consider a graphical concept called a **clique**.
6. A **maximal clique** is a clique such that it is not possible to include any other nodes from the graph in the set without it ceasing to be a clique.



1. Consider the following graph



Two-nodes cliques

- $\{x_1, x_2\}$
- $\{x_2, x_3\}$
- $\{x_3, x_4\}$
- $\{x_4, x_2\}$
- $\{x_1, x_3\}$

Two maximal cliques

- $\{x_1, x_2, x_3\}$
- $\{x_2, x_3, x_4\}$

Factorization properties

1. We can define the factors in the decomposition of the joint distribution to be functions of the variables in the cliques.
2. We can consider functions of the maximal cliques, because other cliques must be subsets of maximal cliques.
3. If $\{x_1, x_2, x_3\}$ is a maximal clique and we define an arbitrary function over this clique, then including another factor defined over a subset of these variables would be redundant.
4. Let us denote a clique by C and the set of variables in that clique by $\mathbf{x}_C$.
5. The joint distribution is written as a product of **potential functions** $\psi(\mathbf{x}_C) > 0$ over the maximal cliques of the graph.

$$p(\mathbf{x}) = \frac{1}{Z} \prod_C \psi(\mathbf{x}_C)$$

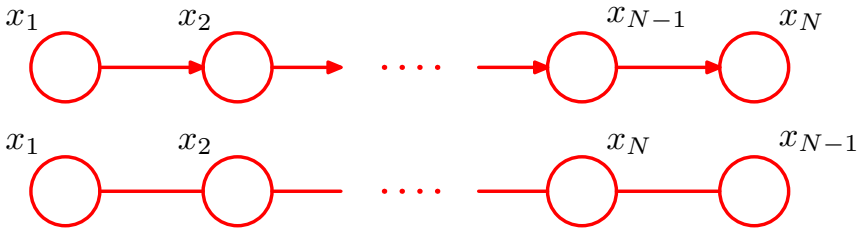
6. The quantity Z, called the **partition function**, is a normalization constant given by (for discrete variables)

$$Z = \sum_{\mathbf{x}} \prod_C \psi(\mathbf{x}_C)$$

to ensure the distribution $p(\mathbf{x})$ is correctly normalized.

Relation between directed and undirected graphs

1. Consider the following graphs



2. For the directed graph, we have

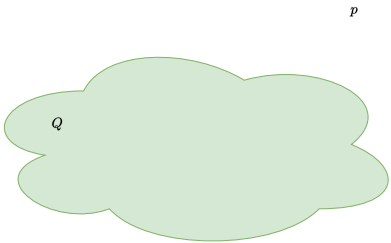
$$p(\mathbf{x}) = p(x_1)\, p(x_2 \mid x_1)\, p(x_3 \mid x_2) \ldots p(x_N \mid x_{N-1})$$

3. For the undirected graph, we have

$$p(\mathbf{x}) = \frac{1}{Z}\psi_{1,2}(x_1, x_2)\psi_{2,3}(x_2, x_3) \ldots \psi_{N-1,N}(x_{N-1}, x_N)$$

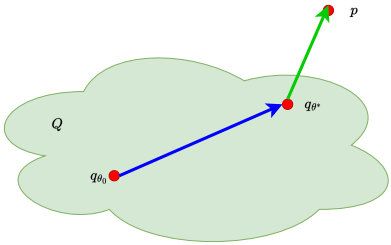
Variational inference

1. Let D be the data set.
2. Let $p(\mathbf{x}) \triangleq p(\mathbf{x} \mid D)$ be the true but intractable distribution.
3. Let $q_\theta(\mathbf{x})$ be some approximation chosen from some tractable family Q such as multi-variate Gaussian.
4. We assume $q_\theta(\mathbf{x})$ has some free parameters which we want to optimize so as to make $q_\theta(\mathbf{x})$ "similar to" $p(\mathbf{x})$.



5. An obvious cost function is to try minimize the difference between $q_\theta(\mathbf{x})$ and $p(\mathbf{x})$.

1. An obvious cost function is to try minimize the KL divergence between $q_\theta(\mathbf{x})$ and $p(\mathbf{x})$.



$$\begin{aligned} \mathrm{D}_{KL}(\, p(\mathbf{x}) \,||\, q_\theta(\mathbf{x})) &= \sum_{\mathbf{x}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q_\theta(\mathbf{x})} \\ &= \mathbb{E}_{p(\mathbf{x})}\left[\log \frac{p(\mathbf{x})}{q_\theta(\mathbf{x})}\right] \end{aligned}$$

2. This is hard to compute, since $\mathbb{E}_{p(\mathbf{x})}$ is assumed to be intractable.

1. A natural alternative is the reverse KL divergence.

$$\begin{aligned} \mathrm{D}_{KL}(\, q_\theta(\mathbf{x}) \,||\, p(\mathbf{x})) &= \sum_{\mathbf{x}} q(\mathbf{x}) \log \frac{q_\theta(\mathbf{x})}{p(\mathbf{x})} \\ &= \mathbb{E}_{q_\theta(\mathbf{x})} \left[\log \frac{q_\theta(\mathbf{x})}{p(\mathbf{x})} \right] \end{aligned}$$

2. The main advantage of the objective function is that computing $\mathbb{E}_{q_\theta(\mathbf{x})}$ is tractable.
3. Equation $\mathbb{E}_{q_\theta(\mathbf{x})} \left[\log \frac{q_\theta(\mathbf{x})}{p(\mathbf{x})} \right]$ **is not tractable** because evaluating $p(\mathbf{x})$ point-wise is hard since it requires $Z = \int_{\mathbf{x}} p(\mathbf{x})$.
4. Using un-normalized distribution $\widetilde{p}(\mathbf{x}) \triangleq p(\mathbf{x} \mid D) = p(\mathbf{x}) Z$, it is tractable to compute.
5. Then, we define the objective function as

$$J(\, q_\theta(\mathbf{x})) = \mathrm{D}_{KL}(\, q_\theta(\mathbf{x}) \,||\, \widetilde{p}(\mathbf{x}))$$

1. Then, we define the objective function as

$$J(\, q_\theta(\mathbf{x})) = \mathrm{D}_{KL}(\, q_\theta(\mathbf{x}) \,||\, \widetilde{p}(\mathbf{x}))$$

2. The above KL was abused because $\widetilde{p}(\mathbf{x})$ is not a valid distribution.

$$\begin{aligned} J(\, q_\theta(\mathbf{x})) &= \sum_{\mathbf{x}} q(\mathbf{x}) \log \frac{q(\mathbf{x})}{\widetilde{p}(\mathbf{x})} \\ &= \sum_{\mathbf{x}} q(\mathbf{x}) \log \frac{q(\mathbf{x})}{Z\, p(\mathbf{x})} \\ &= \sum_{\mathbf{x}} q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} - logZ \\ &= \mathrm{D}_{KL}(\, q_\theta(\mathbf{x}) \,||\, p(\mathbf{x})) - \log Z \end{aligned}$$

3. Z is a constant, by minimizing $J(\, q_\theta(\mathbf{x}))$, we will force $q_\theta(\mathbf{x})$ to become close to $p(\mathbf{x})$.

1. Since KL divergence is always non-negative, $J(\, q(\mathbf{x}))$ is an upper bound on $\log Z$.

$$\begin{aligned} J(\, q_\theta(\mathbf{x})) &= \mathrm{D}_{KL}(\, q_\theta(\mathbf{x}) \,||\, p(\mathbf{x})) - logZ \\ &\geq -logZ \end{aligned}$$

2. The value of $\log Z$ is called evidence lower bound (ELBO).
3. Alternatively, we can try to maximize the following quantity, called energy functional.

$$\begin{aligned} L(\, q_\theta(\mathbf{x})) &= -J(\, q_\theta(\mathbf{x})) \\ &= -\mathrm{D}_{KL}(\, q_\theta(\mathbf{x}) \,||\, p(\mathbf{x})) + logZ \\ &\leq \log Z. \end{aligned}$$

Alternative interpretation of $J(\, q_\theta(\mathbf{x}))$

1. The objective function $J(\, q_\theta(\mathbf{x}))$ can be written as

$$\begin{aligned} J(\, q_\theta(\mathbf{x})) &= \mathbb{E}_{q_\theta(\mathbf{x})}[\log\, q_\theta(\mathbf{x})] + \mathbb{E}_{q_\theta(\mathbf{x})}[\log\, \widetilde{p}(\mathbf{x})] \\ &= H(\, q_\theta(\mathbf{x})) + \mathbb{E}_{q_\theta(\mathbf{x})}[E(\mathbf{x})] \end{aligned}$$

   where $E(\mathbf{x}) = -\log\, \widetilde{p}(\mathbf{x})$ is **energy**.

2. Thus, $J(\, q_\theta(\mathbf{x}))$ is expected energy minus Entropy of the system.

3. In statistical physics, $J(\, q_\theta(\mathbf{x}))$ is called the variational free energy or the Helmholtz free energy.

1. Let $p(\mathbf{x})$ and $q(\mathbf{x})$ be two k-dimensional Gaussian distribution.

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}|\Sigma_p|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_p)^T\Sigma_p^{-1}(\mathbf{x}-\mu_p)\right)$$
$$q(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}|\Sigma_q|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_q)^T\Sigma_q^{-1}(\mathbf{x}-\mu_q)\right)$$

2. Then, KL divergence can be written as

$$\begin{aligned}
D_{KL}(\,p(\mathbf{x})\,||\,q(\mathbf{x})) &= \mathbb{E}_{p(\mathbf{x})}[\log\, p(\mathbf{x}) - \log\, q(\mathbf{x})] \\
&= \mathbb{E}_{p(\mathbf{x})}\left[\frac{1}{2}\log\frac{|\Sigma_q|}{|\Sigma_p|} - \frac{1}{2}(\mathbf{x}-\mu_p)^T\Sigma_p^{-1}(\mathbf{x}-\mu_p) + \frac{1}{2}(\mathbf{x}-\mu_q)^T\Sigma_q^{-1}(\mathbf{x}-\mu_q)\right] \\
&= \frac{1}{2}\mathbb{E}_{p(\mathbf{x})}\left[\log\frac{|\Sigma_q|}{|\Sigma_p|}\right] - \frac{1}{2}\mathbb{E}_{p(\mathbf{x})}\left[(\mathbf{x}-\mu_p)^T\Sigma_p^{-1}(\mathbf{x}-\mu_p)\right] \\
&+ \frac{1}{2}\mathbb{E}_{p(\mathbf{x})}\left[(\mathbf{x}-\mu_q)^T\Sigma_q^{-1}(\mathbf{x}-\mu_q)\right] \\
&= \frac{1}{2}\log\frac{|\Sigma_q|}{|\Sigma_p|} - \frac{1}{2}\mathbb{E}_{p(\mathbf{x})}\left[(\mathbf{x}-\mu_p)^T\Sigma_p^{-1}(\mathbf{x}-\mu_p)\right] \\
&+ \frac{1}{2}\mathbb{E}_{p(\mathbf{x})}\left[(\mathbf{x}-\mu_q)^T\Sigma_q^{-1}(\mathbf{x}-\mu_q)\right]
\end{aligned}$$

3. $(\mathbf{x}-\mu_p)^T\Sigma_p^{-1}(\mathbf{x}-\mu_p)$ is scaler: $\mathrm{tr}\big((\mathbf{x}-\mu_p)^T\Sigma_p^{-1}(\mathbf{x}-\mu_p)\big) = \mathrm{tr}\big((\mathbf{x}-\mu_p)(\mathbf{x}-\mu_p)^T\Sigma_p^{-1}\big)$.

1. The expectation and trace can be interchanged to get,

$$= \frac{1}{2}\,\mathrm{tr}\big(\mathbb{E}_{p(\mathbf{x})}\big[(\mathbf{x}-\mu_p)(\mathbf{x}-\mu_p)^T\Sigma_p^{-1}\big]\big)$$
$$= \frac{1}{2}\,\mathrm{tr}\big(\mathbb{E}_{p(\mathbf{x})}\big[(\mathbf{x}-\mu_p)(\mathbf{x}-\mu_p)^T\big]\Sigma_p^{-1}\big)$$

2. We know $\Sigma_p = \mathbb{E}_{p(\mathbf{x})}\big[(\mathbf{x}-\mu_p)(\mathbf{x}-\mu_p)^T\big]$. Simplifying it to

$$\frac{1}{2}\,\mathrm{tr}\big(\mathbb{E}_{p(\mathbf{x})}\big[(\mathbf{x}-\mu_p)(\mathbf{x}-\mu_p)^T\big]\Sigma_p^{-1}\big) = \frac{1}{2}\,\mathrm{tr}\big(\Sigma_p\Sigma_p^{-1}\big)$$
$$= \frac{1}{2}\,\mathrm{tr}(I_k) = \frac{k}{2}$$

3. By using matrix cookbook, the third term is also equals to

$$\mathbb{E}_{p(\mathbf{x})}\big[(\mathbf{x}-\mu_q)^T\Sigma_q^{-1}(\mathbf{x}-\mu_q)\big] = (\mu_p-\mu_q)^T\Sigma_q^{-1}(\mu_p-\mu_q) + \mathrm{tr}\big(\Sigma_q^{-1}\Sigma_p\big)$$

4. Combining all this we get,

$$\mathrm{D}_{KL}(\,p(\mathbf{x})\;||\;q(\mathbf{x})) = \frac{1}{2}\left\{\log\frac{|\Sigma_q|}{|\Sigma_p|} - k + (\mu_p-\mu_q)^T\Sigma_q^{-1}(\mu_p-\mu_q) + \mathrm{tr}\big(\Sigma_q^{-1}\Sigma_p\big)\right\}$$

5. **What happens if we have not distributions explicitly?**

Mean field variational inference

1. In mean field variational inference, we assume that the variational family factorizes,

$$q(x_1, \ldots, x_d) = \prod_{j=1}^{d} p(x_j)$$

2. The goal is to solve this optimization problem:

$$\min_{q_1, \ldots, q_d} \mathrm{D}_{KL}(q \,||\, p)$$

3. We optimize over the parameters of each marginal distribution q_i.
4. The standard way of performing this optimization problem is via coordinate descent over the q_j.
5. Interestingly, the optimization problem for one coordinate has a simple closed form solution.

References

Reading

1. Chapter 21 of Machine Learning: A Probabilistic Perspective (Murphy 2012).
2. Chapter 10 of Probabilistic Machine Learning: Advanced Topics (Murphy 2023).
3. Chapter 8 of Pattern Recognition and Machine Learning (C. M. Bishop 2006).
4. Chapter 11 of Deep Learning: Foundations and Concepts (C. M. Bishop and H. Bishop 2024).
5. Chapter 7 of All of Statistics: A Concise Course in Statistical Inference (Wasserman 2010).

References

- Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag.
- Bishop, Christopher M. and Hugh Bishop (2024). *Deep Learning: Foundations and Concepts*. Second edition. Springer.
- Murphy, Kevin P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.
- – (2023). *Probabilistic Machine Learning: Advanced Topics*. The MIT Press.
- Wasserman, Larry (2010). *All of Statistics: A Concise Course in Statistical Inference*. Springer.

Questions?