Probabilistic Graphical Models in Computer Vision (IN2329)

Csaba Domokos

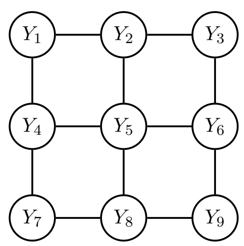
Summer Semester 2015/2016

2. Expectation-maximization algorithm

Agenda for today's lecture *

In the **previous lecture** we learnt about

- Probability space
- Conditional probability
- Independence, conditional independence



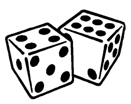
Today we are going to learn about

1. Random variables $(Y_1, \ldots, Y_9)$
2. Probability distributions
   - Joint distribution $(p(y_1, \ldots, y_9))$
   - Marginal distribution $(p(y_1))$
   - Conditional distribution $(p(y \mid x))$
   - Expectation
3. Expectation-maximization algorithm

Random variables

Example: throwing two "fair" dice *

We have the *sample space* $\Omega = \{(i,j) : 1 \leqslant i,j \leqslant 6\}$ and the *(uniform) probability measure* $P(\{(i,j)\}) = \frac{1}{36}$, where $(\Omega, \mathcal{P}(\Omega), P)$ forms a *probability space*.



In many cases it would be more natural to consider *attributes* of the outcomes. A **random variable** is a way of reporting an *attribute* of the *outcome*.

Le us consider the *sum of the numbers showing on the dice*, defined by define the **mapping** $X : \Omega \to \Omega'$, $X(i,j) = i + j$, where $\Omega' = \{2, 3, \ldots, 12\}$.

It can be seen that this mapping leads a *probability space* $(\Omega', \mathcal{P}(\Omega'), P')$, such that $P' : \mathcal{P}(\Omega') \to [0,1]$ is defined as

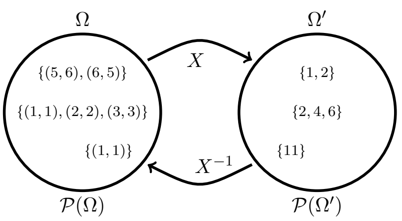
$$P'(A') = P(\{(i,j) : X(i,j) \in A'\}) \ .$$

Example: $P'(\{11\}) = P(\{(5,6),(6,5)\}) = \frac{2}{36}$.

Preimage mapping

Let $X : \Omega \to \Omega'$ be an arbitrary *mapping*. The **preimage mapping** $X^{-1} : \mathcal{P}(\Omega') \to \mathcal{P}(\Omega)$ is defined as

$$X^{-1}(A') = \{\omega \in \Omega : X(\omega) \in A'\} \ .$$



Random variable

Let $(\Omega, \mathcal{A})$ and $(\Omega', \mathcal{A}')$ *measurable spaces*. A *mapping* $X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ is called **measurable mapping**, if

$$X^{-1}(A') = \{\omega \in \Omega : X(\omega) \in A'\} \in \mathcal{A} \ .$$

A *measurable mapping* $X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{A}')$ is called **random variable**.

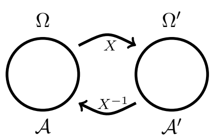
Let $X : (\Omega, \mathcal{A}) \to (\Omega' \subseteq \mathbb{R}, \mathcal{A}')$ be a *random variable* and P a *measure* over $\mathcal{A}$. Then

$$P'(A') := P_X(A') \triangleq P(X^{-1}(A'))$$

defines a measure over $\mathcal{A}'$.

P_X is called the **image measure** of P by X.

Specially, if P is a *probability measure* then P_X is a *probability measure* over $\mathcal{A}'$. (See Exercise.)



Example: throwing two "fair" dice *

We are given two *sample spaces* $\Omega = \{(i,j) : 1 \leqslant i,j \leqslant 6\}$ and $\Omega' = \{2, 3, \ldots, 12\}$. We assume the *(uniform) probability measure* P over $(\Omega, \mathcal{P}(\Omega))$. Define a mapping $X : (\Omega, \mathcal{P}(\Omega)) \to (\Omega', \mathcal{P}(\Omega'))$, where $X(i,j) = i + j$.

Question: Is X a random variable?

$$X^{-1}(A') = \{\omega \in \Omega : X(\omega) \in A'\} \in \mathcal{P}(\Omega)$$

is satisfied, since for any $\omega' \in \Omega'$ one can find an $\omega \in \Omega$ such that $X(\omega) = \omega'$. Therefore X is *measurable*, thus it is a *random variable*. Moreover, P is a *probability measure*, hence the *image measure*

$$P_X(A') \triangleq P(X^{-1}(A'))$$

is a *probability measure* on $(\Omega', \mathcal{P}(\Omega'))$.

Example: $P_X(\{2,4,5\}) = P(X^{-1}(\{2,4,5\})) = P(\{(1,1),(1,3),(2,2),(3,1),(1,4),(2,3),(3,2),(4,1)\}) = \frac{8}{36} = \frac{2}{9}$.

Probability distributions

Probability distribution

Note that a *random variable* is a measurable mapping from a probability space to a measure space. It is *neither a variable nor random*.

Let $X : (\Omega, \mathcal{A}, P) \to (\Omega' \subseteq \mathbb{R}, \mathcal{A}')$ be a *random variable*. Then the *image measure* P_X of P by X is called **probability distribution**.

We use the notation $P(x)$ for $P(X = x)$, where

$$P(x) := P(X = x) \triangleq P(\{\omega \in \Omega : X(\omega) = x\}) .$$

Similarly, $P(X < x) \triangleq P(\{\omega \in \Omega : X(\omega) < x\})$.

Let $X : (\Omega, \mathcal{A}, P) \to (\Omega' \subseteq \mathbb{R}, \mathcal{A}')$ be a *random variable*. Then $F_X : \mathbb{R} \to \mathbb{R}$

$$F_X(x) \triangleq P(X < x) , \quad x \in \mathbb{R}$$

is called **cumulative distribution function** (cdf.) of X.
Each probability measure is *uniquely defined* by its distribution function.

Density function

Let $F_X : \mathbb{R} \to \mathbb{R}$ be the *cumulative distribution function* of a *random variable* X.
A *measurable function* $f_X(x)$ is called a **density function** of X, if

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t , \quad x \in \mathbb{R} .$$

A **measurable function** we mean to be a function with *improper Riemann-integral*.

A *random variable* $X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ is said to be **discrete random variable** if Ω' is *countable*.

Continuous random variable

A random variable $X : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{A}')$ is called **continuous random variable**, if it has a density function $f_X(x)$. Then the followings are held:

1. $f_X(x)$ is non-negative,
2. $\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = 1$,
3. $P(a \leqslant X < b) \triangleq F_X(a \leqslant X < b) = \int_a^b f_X(x)\mathrm{d}x$.

Proof.

1. F_X is non-negative and monotonously increasing, thus $f_X(x) \geqslant 0$.
2. $$\int_{-\infty}^{\infty} f_X(x)\mathrm{d}x = F_X(\infty) - F_X(-\infty) = 1 - 0 = 1 .$$
3. $$F_X(a \leqslant X < b) = F_X(b) - F_X(a) = \int_{-\infty}^{b} f_X(x)\mathrm{d}x - \int_{-\infty}^{a} f_X(x)\mathrm{d}x = \int_a^b f_X(x)\mathrm{d}x.$$

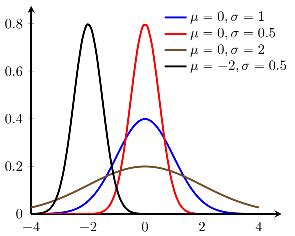
□

The Normal (Gaussian) distribution *

A *continuous* random variable $X : \mathbb{R} \to \mathbb{R}$ with density function

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

is said the have **Normal distribution** (or **Gaussian distribution** with parameters $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+$.



Standard normal distribution: $\mu = 0$ and $\sigma = 1$.

Joint distribution

Suppose a probability space $(\Omega, \mathcal{A}, P)$. Let $X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ and $Y : (\Omega, \mathcal{A}) \to (\Omega'', \mathcal{A}'')$ be *discrete* random variables, where $x_1, x_2, \ldots$ denote the values of X and $y_1, y_2, \ldots$ denote the values of Y.

We introduce the notation

$$p_{ij} \triangleq P(X = x_i, Y = y_j) \quad i, j = 1, 2, \ldots$$

for the probability of the *events*

$$\{X = x_i, Y = y_j\} := \{\omega \in \Omega : X(\omega) = x_i \text{ and } Y(\omega) = y_j\} .$$

These probabilities p_{ij} form a *distribution*, called the **joint distribution** of X and Y.

Therefore,

$$\sum_i \sum_j p_{ij} = 1 .$$

Marginal distributions

Suppose a probability space $(\Omega, \mathcal{A}, P)$. Let $X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ and $Y : (\Omega, \mathcal{A}) \to (\Omega'', \mathcal{A}'')$ be *discrete* random variables, where $x_1, x_2, \ldots$ denote the values of X and $y_1, y_2, \ldots$ denote the values of Y.

The *distributions* defined by the probabilities

$$p_i \triangleq P(X = x_i) \quad \text{and} \quad q_j \triangleq P(Y = y_j)$$

are called the **marginal distributions** of X and of Y, respectively.

Let us consider the *marginal distribution* of X. Then

$$p_i = P(X = x_i) = \sum_j P(X = x_i, Y = y_j) = \sum_j p_{ij} .$$

Similarly, the *marginal distribution* of Y is given by

$$q_j = P(Y = y_j) = \sum_i P(X = x_i, Y = y_j) = \sum_i p_{ij} .$$

Example: marginal distribution *

Consider two producing machines creating identical product in a factory. Assume we are given the following table with probabilities

	Machine I	Machine II	
The product is good	0.56	0.41	0.97
The product is waste	0.01	0.02	0.03
	0.57	0.43	1

The marginal distributions of discrete random variables corresponding to the values of {good, waste} and {I, II} are shown in the last column and last row, respectively.

The following also holds

$$\sum_i p_i = \sum_i P(X = x_i) = \sum_i \sum_j P(X = x_i, Y = y_i) = \sum_i \sum_j p_{ij} = 1 .$$

Joint density

Suppose a probability space $(\Omega, \mathcal{A}, P)$. Let $X : (\Omega, \mathcal{A}) \to (\Omega' \subseteq \mathbb{R}, \mathcal{A}')$ and $Y : (\Omega, \mathcal{A}) \to (\Omega'' \subseteq \mathbb{R}, \mathcal{A}'')$ be random variables. The **joint cumulative distribution function** of X and Y, denoted by $F_{XY} : \mathbb{R}^2 \to \mathbb{R}$, is defined as

$$F_{XY}(x, y) \triangleq P(X < x, Y < y) , \quad x, y \in \mathbb{R} .$$

If both X and Y are *continuous random variables*, then the **joint density function** $f_{XY} : \mathbb{R}^2 \to \mathbb{R}$ is defined by

$$F_{XY}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(u, v) \mathrm{d}u \mathrm{d}v .$$

The *joint density function* $f_{XY}(x, y)$ also satisfies the following property:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(u, v) \mathrm{d}u \mathrm{d}v = 1 .$$

Marginal densities

Suppose a probability space $(\Omega, \mathcal{A}, P)$. Let $X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ and $Y : (\Omega, \mathcal{A}) \to (\Omega'', \mathcal{A}'')$ be random variables with *joint cumulative distribution function* $F_{XY} : \mathbb{R}^2 \to \mathbb{R}$. The **marginal cumulative distribution functions** of X and Y are given by

$$F_X(x) := F_{XY}(x, \infty) = \lim_{y \to \infty} F_{XY}(x, y) , \quad \text{and}$$
$$F_Y(y) := F_{XY}(\infty, y) = \lim_{x \to \infty} F_{XY}(x, y) .$$

If both X and Y are *continuous random variables* with the *joint density function* $f_{XY}(x, y)$, then the **marginal density functions** $f_X, f_Y : \mathbb{R} \to \mathbb{R}$ are defined as

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) \mathrm{d}y \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y) \mathrm{d}x .$$

Conditional distribution

Suppose a probability space $(\Omega, \mathcal{A}, P)$. Let X and Y be *discrete random variables*, where $x_1, x_2, \ldots$ denote the values of X and $y_1, y_2, \ldots$ denote the values of Y.

The **conditional distribution** of X given Y is defined by

$$P(X = x_i \mid Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)} = \frac{p_{ij}}{\sum_k p_{kj}} = \frac{p_{ij}}{q_j} .$$

Therefore, $\sum_i P(X = x_i \mid Y = y_j) = \sum_i \frac{p_{ij}}{\sum_k p_{kj}} = 1$ is also held.

The **conditional cumulative distribution function** is defined as

$$F_{X|Y}(x \mid y) \triangleq \lim_{h \to 0} P(X < x \mid y \leqslant Y < y + h)$$
$$= \lim_{h \to 0} \frac{P(X < x, y \leqslant Y < y + h)}{P(y \leqslant Y < y + h)} .$$

Conditional density

Suppose a probability space $(\Omega, \mathcal{A}, P)$. Let X and Y be random variables with *joint density function* $f_{XY}(x, y)$. If the *marginal density function* $f_Y(y) \neq 0$, then the **conditional density function** of X given Y is defined as

$$f_{X|Y}(x \mid y) = \frac{f_{XY}(x, y)}{f_Y(y)} .$$

Expectation

Expectation

The *expectation* of a random variable is intuitively the long-run average value of repetitions of the experiment it represents.

Let X be a *discrete random variable* taking values $x_1, x_2, \ldots$ with probabilities $p_1, p_2, \ldots$, respectively. The **expectation** (or **expected value**) of X is defined as

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} x_i p_i ,$$

assuming that this series is *absolutely convergent* (that is $\sum_{i=1}^{\infty} |x_i| p_i$ is convergent).

Example: throwing two "fair" dice and the value of X is is *the sum the numbers showing on the dice*.

$$\mathbb{E}[X] = 2\frac{1}{36} + 3\frac{2}{36} + 4\frac{3}{36} + 5\frac{4}{36} + 6\frac{5}{36}$$
$$+ 7\frac{6}{36} + 8\frac{5}{36} + 9\frac{4}{36} + 10\frac{3}{36} + 11\frac{2}{36} + 12\frac{1}{36} = 7 .$$

Expectation

Let X be a *(continuous) random variable* with *density function* $f_X(x)$. The **expectation** of X is defined as

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f_X(x) \mathrm{d}x ,$$

assuming that this integral is *absolutely convergent* (that is the value of the integral $\int_{-\infty}^{\infty} |x \cdot f_X(x)| \mathrm{d}x = \int_{-\infty}^{\infty} |x| \cdot f_X(x) \mathrm{d}x$ is finite).

Suppose a random variable X with density function $f_X(x)$. The **expected value of a function** $g(x) : \mathbb{R} \to \mathbb{R}$ is defined as

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f_X(x) \mathrm{d}x ,$$

assuming that this integral is absolutely convergent.

Conditional expectation

A **random vector** $\mathbf{X} = (X_1, \ldots, X_n)$ is a vector whose components are random variables. If all X_i are discrete, then $\mathbf{X}$ is called a **discrete random vector**.

Let (X, Y) be a *discrete random vector*. The **conditional expectation** of X given the event $\{Y = y\}$ is defined as

$$\mathbb{E}[X \mid Y = y] = \sum_{i=1}^{\infty} x_i P(X = x_i \mid Y = y) ,$$

assuming that this series is absolutely convergent.

Let (X, Y) be a *(continuous) random vector* with *conditional density function* $f_{X|Y}(x \mid y)$. The **conditional expectation** of X given the event $\{Y = y\}$ is defined as

$$\mathbb{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x \cdot f_{X|Y}(x \mid Y = y) \mathrm{d}x ,$$

assuming that this integral is absolutely convergent.

Conditional expectation

Suppose a *(continuous) random vector* (X, Y) with *conditional density function* $f_{X|Y}(x \mid y)$. The **conditional expectation of a function** $g(x) : \mathbb{R} \to \mathbb{R}$ given the event $\{Y = y\}$ is defined as

$$\mathbb{E}[g(X) \mid Y = y] = \int_{-\infty}^{\infty} g(x) \cdot f_{X|Y}(x \mid Y = y)\mathrm{d}x \ ,$$

assuming that this integral is absolutely convergent.

Summary *

- A **random variable** $X : (\Omega, \mathcal{A}, P) \to (\Omega' \subseteq \mathbb{R}, \mathcal{A}', P_X)$ is a measurable mapping from a probability space to a measure space.
- The image measure P_X of P by X is called **probability distribution**.
- The function $F_X : \mathbb{R} \to \mathbb{R}$, $F_X(x) = P(x < X)$ is called **cumulative distribution function** of X.
- A measurable function $f_X(x)$ is called **density function** of X, if

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t \ .$$

- Probability distributions and densities
  - Joint distribution: $p_{XY}(x, y)$
  - Marginal distribution: $p_X(x)$
  - Conditional distribution: $p_{X|Y}(x \mid y)$
- The **expected value** is intuitively the long-run average value of repetitions of the experiment.

The Expectation-maximization algorithm

Latent variables

Suppose we are given a set of *i.i.d.* (i.e. independent and identically distributed) data samples $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ represented by a matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$. The samples are drawn from a model (e.g., mixture of Gaussians) given by its parameters $\boldsymbol{\theta}$.

There are mainly two applications of the EM algorithm:

1. The data has **missing values** due to limitations of the observation.
2. The **likelihood function can be simplified** by assuming missing values.

Latent variables gathering the missing values are represented by a matrix $\mathbf{Z}$.

We generally want to maximize the **posterior probability**

$$\boldsymbol{\theta}^* \in \operatorname*{argmax}_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \mathbf{X}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{\mathbf{Z}} p(\boldsymbol{\theta}, \mathbf{Z} \mid \mathbf{X}) \ .$$

Alternatively, one can maximize the log-likelihood

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{X}) = \ln p(\mathbf{X} \mid \boldsymbol{\theta}) = \ln \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) \ .$$

Jensen's inequality *

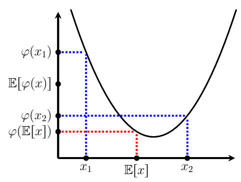
Reminder: A function $f : \mathbb{R}^n \to \mathbb{R}$ is **convex**, if $\forall a, b \in \mathbb{R}^n$, $\forall t \in [0, 1]$

$$f(ta + (1 - t)b) \leqslant tf(a) + (1 - t)f(b)$$

holds. A function f is said to be **concave** if $-f$ is convex.

Assume a random vector $\mathbf{X}$ and a convex function φ, then

$$\varphi\left(\mathbb{E}[\mathbf{X}]\right) \leqslant \mathbb{E}\left[\varphi(\mathbf{X})\right] \ .$$



Proof of Jensen's inequality *

For a discrete random variable X taking values $x_1, x_2, \dots$ with probabilities $p_1, p_2, \dots$, one can obtain

$$\varphi(\mathbb{E}[X]) = \varphi\left(\sum_{i=1}^{\infty} x_i p_i\right) \triangleq L\left(\sum_{i=1}^{\infty} x_i p_i\right) = a\left(\sum_{i=1}^{\infty} x_i p_i\right) + b \ ,$$

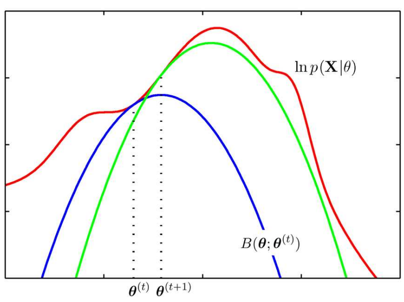
where $L : \mathbb{R} \leftarrow \mathbb{R}$, $L(x) = ax + b$ is an *affine function* corresponding to the **tangent line** of φ at $\mathbb{E}[X]$.

$$= \sum_{i=1}^{\infty} p_i(ax_i + b) - \sum_{i=1}^{\infty} p_i b + b = \sum_{i=1}^{\infty} p_i(ax_i + b) = \sum_{i=1}^{\infty} p_i L(x_i)$$

$$\leqslant \sum_{i=1}^{\infty} p_i \varphi(x_i) = \mathbb{E}[\varphi(X)] \ .$$

The overview of the EM algorithm

The idea: start with a guess $\boldsymbol{\theta}^{(t)}$ for the parameters, calculate an easily computed lower bound $B(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)})$ that touches the function $\ln p(\mathbf{X} \mid \boldsymbol{\theta})$, and maximize that bound instead. This procedure generally converges to a **local maximizer** $\hat{\boldsymbol{\theta}}$.



Lower bound maximization *

First we derive the lower bound $B(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)})$.

$$\ln p(\mathbf{X} \mid \boldsymbol{\theta}) = \ln \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) = \ln \sum_{\mathbf{Z}} q^{(t)}(\mathbf{Z}) \underbrace{\frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q^{(t)}(\mathbf{Z})}}_{g(\mathbf{Z})}$$

where $q^{(t)}(\mathbf{Z})$ is an arbitrary probability distribution of the latent variables $\mathbf{Z}$.

$$= \ln \mathbb{E}\underbrace{\left[\frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q^{(t)}(\mathbf{Z})}\right]}_{g(\mathbf{Z})} \geqslant \mathbb{E}\left[\ln \frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q^{(t)}\mathbf{Z}}\right]$$

$$= \sum_{\mathbf{Z}} q^{(t)}(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q^{(t)}(\mathbf{Z})} \triangleq B(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) \ .$$

Lagrange multiplier *

Suppose two functions $f, g : \mathbb{R}^D \to \mathbb{R}$ having continuous first partial derivatives. We consider the following optimization problem

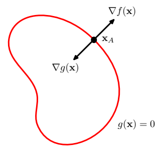
$$\max f(\mathbf{x}) \quad \text{subject to } g(\mathbf{x}) = 0 \,.$$

It is convenient to study the **Lagrangian function**, defined as

$$L(\mathbf{x}, \lambda) \triangleq f(\mathbf{x}) + \lambda g(\mathbf{x}) \,,$$

where $\lambda \neq 0$ is called a **Lagrange multiplier**.

Geometric interpretation of a Lagrange multiplier *



The constraint $g(\mathbf{x}) = 0$ forms a $D - 1$ dimensional surface in $\mathbb{R}^D$. Suppose $\mathbf{x}$ and a nearby point $\mathbf{x} + \boldsymbol{\varepsilon}$ lying on the surface $g(\mathbf{x}) = 0$. Based on the Taylor expansion of g around $\mathbf{x}$ we get

$$g(\mathbf{x} + \boldsymbol{\varepsilon}) \approx g(\mathbf{x}) + \boldsymbol{\varepsilon}^T \nabla g(\mathbf{x}) \quad \Rightarrow \quad \boldsymbol{\varepsilon}^T \nabla g(\mathbf{x}) \approx 0 \,.$$

In the limit $\|\boldsymbol{\varepsilon}\| \to 0$, we have $\boldsymbol{\varepsilon}^T \nabla g(\mathbf{x}) = 0$, which means that $\nabla g(\mathbf{x})$ **is normal to the constraint surface**, since $\boldsymbol{\varepsilon}$ is parallel to the surface.

At an optimal $\mathbf{x}_A$ lying on the constraint surface, $\nabla f(\mathbf{x}_A)$ **must be orthogonal to the surface**, otherwise we could increase the value of f by moving along the constraint surface. Therefore, there exist a **Lagrange multiplier** λ such that

$$\nabla f + \lambda \nabla g = 0$$

which can be equivalently written as $\nabla_x L = 0$. Note that $\frac{\partial}{\partial \lambda} L = 0$ leads to the constraint $g(\mathbf{x}) = 0$.

Finding an optimal bound *

We want to find the *best* lower bound, defined as the bound $B(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)})$ that touches the objective function $\ln p(\mathbf{X} \mid \boldsymbol{\theta})$ at $\boldsymbol{\theta}^{(t)}$.

The optimal bound at the current guess $\boldsymbol{\theta}^{(t)}$ can be found by maximizing

$$B(\boldsymbol{\theta}^{(t)}; \boldsymbol{\theta}^{(t)}) = \sum_{\mathbf{Z}} q^{(t)}(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}^{(t)})}{q^{(t)}(\mathbf{Z})}$$

with respect to the distribution $q^{(t)}(\mathbf{Z})$.

Introducing a *Lagrange multiplier* λ to enforce $\sum_{\mathbf{Z}} q^{(t)}(\mathbf{Z}) = 1$, the objective becomes

$$h(q^{(t)}) = \sum_{\mathbf{Z}} q^{(t)}(\mathbf{Z}) \ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}^{(t)}) - \sum_{\mathbf{Z}} q^{(t)}(\mathbf{Z}) \ln q^{(t)}(\mathbf{Z}) + \lambda \left(\sum_{\mathbf{Z}} q^{(t)}(\mathbf{Z}) - 1 \right) .$$

Finding an optimal bound *

$$h(q^{(t)}) = \sum_{\mathbf{Z}} q^{(t)}(\mathbf{Z}) \ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}^{(t)}) - \sum_{\mathbf{Z}} q^{(t)}(\mathbf{Z}) \ln q^{(t)}(\mathbf{Z}) + \lambda \left(\sum_{\mathbf{Z}} q^{(t)}(\mathbf{Z}) - 1 \right) .$$

Setting the derivative of h w.r.t. $q^{(t)}(\mathbf{Z})$ to 0, we obtain

$$\frac{\partial}{\partial q^{(t)}(\mathbf{Z})} h = \ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}^{(t)}) - \ln q^{(t)}(\mathbf{Z}) - 1 - \lambda = 0 \,.$$

$$p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}^{(t)}) \exp(-1 - \lambda) = q^{(t)}(\mathbf{Z}) \tag{1}$$

$$\exp(-1 - \lambda) \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}^{(t)}) = \sum_{\mathbf{Z}} q^{(t)}(\mathbf{Z}) = 1$$

$$\exp(-1 - \lambda) = \frac{1}{\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}^{(t)})} = \frac{1}{p(\mathbf{X} \mid \boldsymbol{\theta}^{(t)})} \,.$$

Therefore, substituting back into Eq. (1), we get

$$q^{(t)}(\mathbf{Z}) = \frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}^{(t)})}{p(\mathbf{X} \mid \boldsymbol{\theta}^{(t)})} = p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{(t)}) \,. \tag{2}$$

Finding an optimal bound *

The resulting optimal bound at $\boldsymbol{\theta}^{(t)}$ indeed touches the objective function:

$$B(\boldsymbol{\theta}^{(t)}; \boldsymbol{\theta}^{(t)}) = \sum_{\mathbf{Z}} q^{(t)}(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}^{(t)})}{q^{(t)}(\mathbf{Z})}$$

By substituting Eq. (2), we get

$$= \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{(t)}) \ln \underbrace{\frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}^{(t)})}{p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{(t)})}}_{p(\mathbf{X} \mid \boldsymbol{\theta}^{(t)})}$$

$$= \ln p(\mathbf{X} \mid \boldsymbol{\theta}^{(t)}) \underbrace{\sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{(t)})}_{=1}$$

$$= \ln p(\mathbf{X} \mid \boldsymbol{\theta}^{(t)}) \,.$$

Maximizing the bound *

We want to maximize $B(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)})$ with respect to $\boldsymbol{\theta}$.

$$B(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) = \sum_{\mathbf{Z}} q^{(t)}(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q^{(t)}(\mathbf{Z})}$$

$$= \sum_{\mathbf{Z}} q^{(t)}(\mathbf{Z}) \ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) - \sum_{\mathbf{Z}} q^{(t)}(\mathbf{Z}) \ln q^{(t)}(\mathbf{Z}) \,.$$

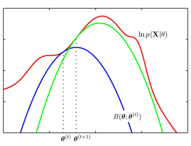
We need to consider the first term only

$$\sum_{\mathbf{Z}} q^{(t)}(\mathbf{Z}) \ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{(t)}) \ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})$$

$$= \mathbb{E}[\ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) \mid \mathbf{X}, \boldsymbol{\theta}^{(t)}] \triangleq Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) \,.$$

$$\boldsymbol{\theta}^{(t+1)} \in \operatorname*{argmax}_{\boldsymbol{\theta}} B(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) = \operatorname*{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) \,.$$

The EM algorithm



1: Choose an initial setting for the parameters $\boldsymbol{\theta}^{(0)}$
2: $t \to 0$
3: **repeat**
4: $\quad t \to t + 1$
5: $\quad$ **E step**. Evaluate $q^{(t-1)}(\mathbf{Z}) \triangleq p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{(t-1)})$
6: $\quad$ **M step**. Evaluate $\boldsymbol{\theta}^{(t)}$ given by

$$\boldsymbol{\theta}^{(t)} = \operatorname*{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) \,,$$

where $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) \triangleq \mathbb{E}[\ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) \mid \mathbf{X}, \boldsymbol{\theta}^{(t-1)}]$

$$= \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{(t-1)}) \ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})$$

7: **until** convergence of either the parameters $\boldsymbol{\theta}$ or the log likelihood $\mathcal{L}(\boldsymbol{\theta}; \mathbf{X})$

Summary *

- We have finished the overview of Probability theory.
- The **Expectation-maximization algorithm** is an iterative method for parameter estimation of *maximum likelihood*, where the model also depends on *latent variables*.

In the **next lecture** we will learn about

- The EM algorithm for Mixtures of Gaussians
- Introduction to Graphical models:
  - *Directed* graphical models: Bayesian network
  - *Undirected* graphical models: Markov random field

Literature *

Probability theory

1. Marek Capiński and Ekkerhard Kopp. *Measure, Integral and Probability*. Springer, 1998
2. Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009

The Expectation-maximization algorithm

3. A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977
4. Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006
5. Frank Dellaert. The expectation maximization algorithm. Technical Report GIT-GVU-02-20, Georgia Institute of Technology, Atlanta, GA, USA, 2002
6. Shane M. Haas. The expectation-maximization and alternating minimization algorithms. Unpublished, 2002
7. Yihua Chen and Maya R. Gupta. EM demystified: An expectation-maximization tutorial. Technical Report UWEETR-2010-0002, University of Washington, Seattle, WA, USA, 2009