Probabilistic Graphical Models in Computer Vision (IN2329)

Csaba Domokos

Summer Semester 2017

Recap $*$

Image

Source: Berkeley Segmentation Dataset

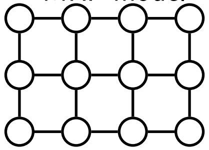
$I \colon \mathcal{V} \subset \mathbb{Z}^2 \to [0, 255]^{3 \times |\mathcal{V}|}$

Labeling

$L \colon \mathcal{V} \to \mathcal{L}^{|\mathcal{V}|}$

We may consider $P(L)$, by defining random variables $L_i = Y_i : [0, 255]^3 \to \mathcal{L}$ for all $i \in \mathcal{V}$ and modeling the joint distribution $p(\mathbf{y})$.

MRF model



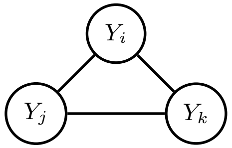
Factorization

$$p(\mathbf{y}) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \psi_c(\mathbf{y_c})$$

We want to find the best labeling: $\mathbf{y}^* \in \operatorname{argmax}_{\mathbf{y}} p(\mathbf{y})$.

Recap: Factor graphs *

Let us consider the following MRF model:



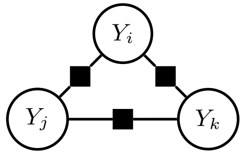
The factorization is given as

$$
\begin{aligned}
p(\mathbf{y}) =& \psi_{ijk}(y_i, y_j, y_k) \\
=& \psi'_i(y_i) \cdot \psi'_j(y_j) \cdot \psi'_k(y_k) \cdot \psi'_{ij}(y_i, y_j) \cdot \psi'_{ik}(y_i, y_k) \\
& \cdot \psi'_{jk}(y_j, y_k) \cdot \psi'_{ijk}(y_i, y_j, y_k) \,.
\end{aligned}
$$

Assume a factorization having with pairwise terms only:

$$
\begin{aligned}
p(\mathbf{y}) =& \psi_i(y_i) \cdot \psi_j(y_j) \cdot \psi_k(y_k) \cdot \psi_{ij}(y_i, y_j) \cdot \psi_{ik}(y_i, y_k) \cdot \psi_{jk}(y_j, y_k) \cdot \psi_{ijk}(y_i, y_j, y_k) \\
=& 1 \cdot 1 \cdot 1 \cdot 1 \cdot \psi_{ij}(y_i, y_j) \cdot \psi_{ik}(y_i, y_k) \cdot \psi_{jk}(y_j, y_k) \cdot 1 \\
=& \psi_A(y_i, y_j) \cdot \psi_B(y_i, y_k) \cdot \psi_C(y_j, y_k) \,.
\end{aligned}
$$

This is explicitly shown by the factor graph



Agenda for today's lecture *

Today we are going to learn about

1. Graphical models
   - Conditional random fields (CRF)
   - Inference for graphical models
2. Formulation of **binary image segmentation**
3. Probability theory
   - Continuous random variables, probability density functions (PDF)
   - Expectation
4. `Expectation-maximization algorithm`

Conditional random field

We have discussed the joint distribution

$$p(\mathbf{y}) = \frac{1}{Z} \prod_{F \in \mathcal{F}} \psi_F(\mathbf{y}_{N(F)}) \ ,$$

but we often have access to measurements $\mathbf{X} = \mathbf{x}$, hence the **conditional distribution** $p(\mathbf{Y} = \mathbf{y} \mid \mathbf{X} = \mathbf{x})$ could be directly modeled, too.

This can be expressed compactly using **conditional random fields** (CRF) with the factorization

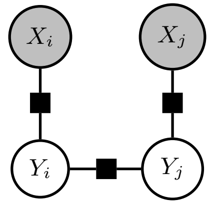
$$\begin{aligned} p(\mathbf{y} \mid \mathbf{x}) = & \frac{p(\mathbf{y}, \mathbf{x})}{p(\mathbf{x})} = \frac{p(\mathbf{y}, \mathbf{x})}{\sum_{\mathbf{y}' \in \mathcal{Y}} p(\mathbf{y}', \mathbf{x})} = \frac{\frac{1}{Z} \prod_{F \in \mathcal{F}} \psi_F(\mathbf{y}_{N(F)}; \mathbf{x}_{N(F)})}{\sum_{\mathbf{y}' \in \mathcal{Y}} \frac{1}{Z} \prod_{F \in \mathcal{F}} \psi_F(\mathbf{y}'_{N(F)}; \mathbf{x}_{N(F)})} \\ = & \frac{1}{Z(\mathbf{x})} \prod_{F \in \mathcal{F}} \psi_F(\mathbf{y}_{N(F)}; \mathbf{x}_{N(F)}) \ . \end{aligned}$$

Conditional random field

$$p(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{F \in \mathcal{F}} \psi_F(\mathbf{y}_F; \mathbf{x}_F)$$

with the **partition function** depending on $\mathbf{x}$

$$Z(\mathbf{x}) = \sum_{\mathbf{y} \in \mathcal{Y}} \prod_{F \in \mathcal{F}} \psi_F(\mathbf{y}_F; \mathbf{x}_F) \ .$$



Shaded variables: The observations $\mathbf{X} = \mathbf{x}$.

Note that the potentials become also functions of (part of) $\mathbf{x}$, i.e. $\psi_F(\mathbf{y}_F; \mathbf{x}_F)$ instead of just $\psi_F(\mathbf{y}_F)$. Nevertheless, $\mathbf{X}$ is **not** part of the probability model, i.e. it is not treated as random vector.

Potentials and energy functions

We typically would like to infer marginal probabilities $p(\mathbf{Y}_F = \mathbf{y}_F \mid \mathbf{x})$ for some factors $F \in \mathcal{F}$.

Assuming $\psi_F : \mathcal{Y}_F \to \mathbb{R}^+$, where $\mathcal{Y}_F = \times_{i \in N(F)} \mathcal{Y}_i$ is the product domain of the variables adjacent to F, instead of *potentials*, we can also work with **energies**.

We define an **energy function** $E_F : \mathcal{Y}_F \to \mathbb{R}$ for each factor $F \in \mathcal{F}$:

$$E_F(\mathbf{y}_F; \mathbf{x}_F) = -\log(\psi_F(\mathbf{y}_F; \mathbf{x}_F)) \quad \Leftrightarrow \quad \psi_F(\mathbf{y}_F; \mathbf{x}_F) = \exp(-E_F(\mathbf{y}_F; \mathbf{x}_F)) \ .$$

$$\begin{aligned} p(\mathbf{y} \mid \mathbf{x}) &= \frac{1}{Z(\mathbf{x})} \prod_{F \in \mathcal{F}} \psi_F(\mathbf{y}_F; \mathbf{x}_F) = \frac{1}{Z(\mathbf{x})} \exp(-\sum_{F \in \mathcal{F}} E_F(\mathbf{y}_F; \mathbf{x}_F)) \\ &= \frac{1}{Z(\mathbf{x})} \exp(-E(\mathbf{y}; \mathbf{x})) \ . \end{aligned}$$

Hence, $p(\mathbf{y} \mid \mathbf{x})$ is completely determined by $E(\mathbf{y}; \mathbf{x})$.

Inference

The goal is to make predictions $\mathbf{y} \in \mathcal{Y}$, *as good as possible*, about unobserved properties for a given data instance $\mathbf{x} \in \mathcal{X}$.

Suppose we are given a *graphical model* (e.g., a factor graph). The **inference** means the procedure to estimate the *probability distribution*, encoded by the *graphical model*, for a *given data* (or observation).

Probabilistic inference: Given a graphical model and the observation x, find the value of the *log partition function* and the *marginal distributions* for each factor,

$$\log Z(\mathbf{x}) = \log \sum_{\mathbf{y} \in \mathcal{Y}} \exp(-E(\mathbf{y}; \mathbf{x})) \ ,$$

$$\mu_F(y_F) = p(\mathbf{Y}_F = \mathbf{y}_F \mid \mathbf{x}) \quad \forall F \in \mathcal{F}, \ \forall \mathbf{y}_F \in \mathcal{Y}_F \ .$$

This typically includes variable marginals, i.e. $\mu_i = p(y_i \mid \mathbf{x})$, to make a single prediction y_i for all variables $i \in \mathcal{V}$.

MAP inference

Maximum A Posteriori (MAP) inference: Given a graphical model and the observation $\mathbf{x}$, find the *state* $\mathbf{y}^* \in \mathcal{Y}$ *of maximum probability*

$$\mathbf{y}^* \in \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}} p(\mathbf{Y} = \mathbf{y} \mid \mathbf{x}) \ .$$

Both inference problems are known to be NP-hard for general graphs and factors, but they can be tractable if the underlying graphical model is suitably restricted.

Energy minimization

Assuming a finite $\mathcal{X}$, the goal is to solve $\mathbf{y}^* \in \operatorname{argmax}_{\mathbf{y}\in\mathcal{Y}} p(\mathbf{y} \mid \mathbf{x})$.

$$\begin{aligned}
\underset{\mathbf{y}\in\mathcal{Y}}{\operatorname{argmax}}\ p(\mathbf{y} \mid \mathbf{x}) &= \underset{\mathbf{y}\in\mathcal{Y}}{\operatorname{argmax}}\ \frac{1}{Z(\mathbf{x})} \exp(-E(\mathbf{y};\mathbf{x})) \\
&= \underset{\mathbf{y}\in\mathcal{Y}}{\operatorname{argmax}}\ \exp(-E(\mathbf{y};\mathbf{x})) \\
&= \underset{\mathbf{y}\in\mathcal{Y}}{\operatorname{argmax}}\ -E(\mathbf{y};\mathbf{x}) \\
&= \underset{\mathbf{y}\in\mathcal{Y}}{\operatorname{argmin}}\ E(\mathbf{y};\mathbf{x})\ .
\end{aligned}$$

Energy minimization can be interpreted as solving for the most likely state of factor graph, i.e. MAP inference.

In practice, one typically models the energy function directly.

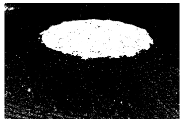
Summary *

- A **Conditional random field** is an *undirected graphical model*, which expresses compactly $p(\mathbf{y} \mid \mathbf{x})$ for some observation $\mathbf{X} = \mathbf{x}$.
- The **inference** means the procedure to estimate the *probability distribution*, encoded by the *graphical model*, for a *given data*.
- Given a *graphical model* and the *observation* $\mathbf{x}$, **MAP inference** means to find the *state* $\mathbf{y}^* \in \mathcal{Y}$ *of maximum probability*

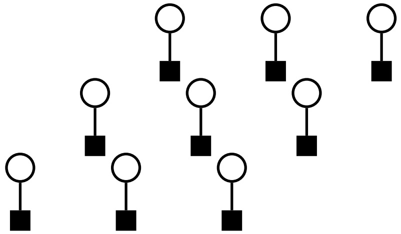
$$\mathbf{y}^* \in \underset{\mathbf{y}\in\mathcal{Y}}{\operatorname{argmax}}\, p(\mathbf{Y} = \mathbf{y} \mid \mathbf{x})\ .$$

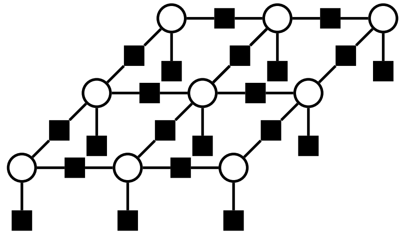
Binary image segmentation


Input image


Unary terms only


Unary and pairwise terms





Conditional independences are specified by a factor graph $G = (\mathcal{V}, \mathcal{F}, \mathcal{E}')$, where all pixels have influence only on the neighboring ones (i.e. $\mathcal{E}$ consists of 4-neighboring connections).

Binary image segmentation

The conditional distribution factorizes (up to pairwise factors) as

$$p(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{i \in \mathcal{V}} \psi_i(y_i; x_i) \prod_{i \in \mathcal{V},\, j \in N(i)} \psi_{ij}(y_i, y_j; x_i, x_j)$$

with

$$Z(\mathbf{x}) = \sum_{\mathbf{y} \in \{0,1\}^{\mathcal{V}}} \prod_{i \in \mathcal{V}} \psi_i(y_i; x_i) \prod_{i \in \mathcal{V},\, j \in N(i)} \psi_{ij}(y_i, y_j; x_i, x_j) ,$$

where $N(i) = \{j \in \mathcal{V} : (i, j \in \mathcal{E})\}$.

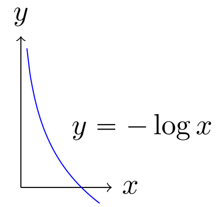
The corresponding energy function $E : \{0,1\}^{\mathcal{V}} \times \mathcal{X} \to \mathbb{R}$:

$$E(\mathbf{y}; \mathbf{x}) = \sum_{i \in \mathcal{V}} E_i(y_i; x_i) + \sum_{i \in \mathcal{V},\, j \in N(i)} E_{ij}(y_i, y_j; x_i, x_j) .$$

Unary energy terms

In order to define energy functions for unary factors, one can consider a set of functions $\phi_i : \mathcal{Y}_i \times \mathcal{X}_i \to [0;1]$:

$$E_i(y_i; x_i) = -\log \phi_i(y_i; x_i) \quad \text{for all } i \in V \ .$$



Assuming that we are provided with the *foreground* and *background distributions*, based on image intensities, $p_f(x)$ and $p_b(x)$, respectively. Then a common way to define the *unary terms* $E_i(y_i; x_i)$ is as follows:

$$E_i(y_i; x_i) = \begin{cases} -\log p_b(x_i) & \text{if } y_i = 0 \\ -\log p_f(x_i) & \text{otherwise .} \end{cases}$$

Pairwise energy terms

For pairwise factor energies we use the **Potts model** here, that is

$$E_{ij}(y_i, y_j; x_i, x_j) := E_{ij}(y_i, y_j) = w_{ij}[\![y_i \neq y_j]\!] = \begin{cases} 0, & \text{if } y_i = y_j \\ w_{ij}, & \text{otherwise.} \end{cases}$$

The parameters $w_{ij} \in \mathbb{R}$ can also be set to the same value, that is $w_{ij} = w$ for all $(i,j) \in \mathcal{E}$.

The resulting **energy function** given as

$$\begin{aligned} E(\mathbf{y}; \mathbf{x}) &= \sum_{i \in \mathcal{V}} E_i(y_i; x_i) + \sum_{i \in \mathcal{V},\, j \in N(i)} E_{ij}(y_i, y_j; x_i, x_j) \\ &= \sum_{i \in \mathcal{V}} -\log \phi_i(y_i; x_i) + \sum_{i \in \mathcal{V},\, j \in N(i)} w_{ij}[\![y_i \neq y_j]\!] \ . \end{aligned}$$

Continuous random variable *

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.

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*.

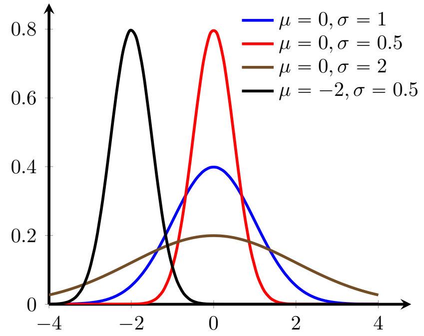
A random variable $X : (\Omega, \mathcal{A}, P) \to (\mathbb{R}, \mathcal{A}')$ is called **continuous random variable**, if it has a density function $f_X(x)$.

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 $\sigma \in \mathbb{R}_+$.



We also use the notation

$$\mathcal{N}(x \mid \mu, \sigma) \triangleq \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) .$$

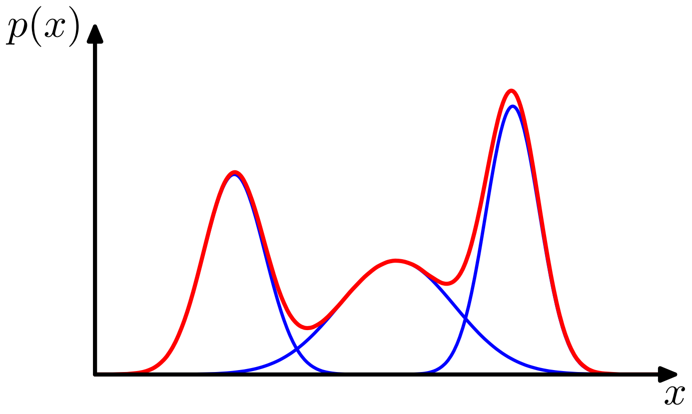
Mixture of Gaussians

While the Gaussian distribution has some important analytical properties, it suffers from limitations when it comes to modelling real data sets.

However the **linear combination of Gaussians** can give rise to very complex densities.
Let us consider a superposition of K Gaussian densities

$$p(x) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(x \mid \mu_k, \sigma_k) \; ,$$

which is called a **mixture of Gaussians**.



Mixture of three Gaussians

The parameters π_k are called **mixing coefficients**.

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

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

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 *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 density *

Suppose a probability space $(\Omega, \mathcal{A}, P)$. Let X and Y be *continuous 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

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*.

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

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).

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.

Expected value of a function

Suppose a *(discrete) random variable* X taking values $x_1, x_2, \ldots$ with probabilities $p_1, p_2, \ldots$, respectively. The **expected value of a function** $g(x) : \mathbb{R} \to \mathbb{R}$ is defined as

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

assuming that this series is absolutely convergent.

Suppose a *(discrete) random vector* (X, Y) with joint probabilities p_{ij}. The **conditional expectation of a function** $g(x) : \mathbb{R} \to \mathbb{R}$ given the event $\{Y = y_j\}$ is defined as

$$\mathbb{E}[g(X) \mid Y = y] = \sum_{i=1}^{\infty} g(x_i) \cdot P(X = x_i \mid Y = y_j) = \sum_{i=1}^{\infty} g(x_i) \cdot \frac{p_{ij}}{q_j} \ ,$$

assuming that this series is absolutely convergent.

Latent variables

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

Basically, 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 $*$

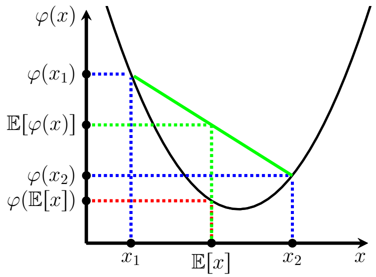
Reminder: A function $f : \mathbb{R}^n \to \mathbb{R}$ is **convex**, if $\forall x_1, x_2 \in \mathbb{R}^n$, $\forall t \in [0, 1]$

$$f(tx_1 + (1-t)x_2) \leqslant tf(x_1) + (1-t)f(x_2)$$

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

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



Proof of Jensen's inequality *

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

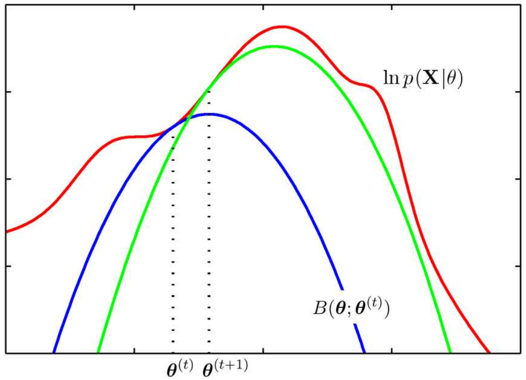
$$\varphi(\mathbb{E}[X]) = \varphi\left(\sum_{i=1}^{\infty} x_i p_i\right) \triangleq l\left(\sum_{i=1}^{\infty} x_i p_i\right) = a\left(\sum_{i=1}^{\infty} x_i p_i\right) + b\,,$$

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** $\hat{\boldsymbol{\theta}}$.



Source: C. Bishop: PRML, 2006.

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

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

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

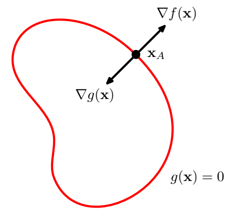
$$L(\mathbf{x}, \lambda) \triangleq f(\mathbf{x}) + \lambda g(\mathbf{x}) \ ,$$

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

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



Source: C. Bishop: PRML, 2006.

In the limit $\|\boldsymbol{\varepsilon}\| = \sqrt{\boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}} \to 0$, 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

$$\begin{aligned}
&= \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{(t)}) \ln \underbrace{\frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}^{(t)})}{p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{(t)})}}_{p(\mathbf{X} \mid \boldsymbol{\theta}^{(t)})} \\
&= \ln p(\mathbf{X} \mid \boldsymbol{\theta}^{(t)}) \underbrace{\sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{(t)})}_{=1} \\
&= \ln p(\mathbf{X} \mid \boldsymbol{\theta}^{(t)}) \; .
\end{aligned}$$

Maximizing the bound *

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

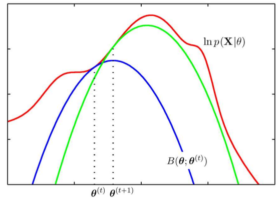
$$
\begin{aligned}
B(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}) &= \sum_{\mathbf{Z}} q^{(t)}(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q^{(t)}(\mathbf{Z})} \\
&= \sum_{\mathbf{Z}} q^{(t)}(\mathbf{Z}) \ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) - \sum_{\mathbf{Z}}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

We need to consider the first term only

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

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

The EM algorithm



Source: C. Bishop: PRML, 2006.

Choose an initial setting for the parameters $\boldsymbol{\theta}^{(0)}$
$t \rightarrow 0$
repeat
    $t \rightarrow t+1$
    E step. Evaluate $q^{(t-1)}(\mathbf{Z}) \triangleq p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{(t-1)})$
    M step. Evaluate $\boldsymbol{\theta}^{(t)}$ given by

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

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

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

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

Summary *

- The **Expectation-maximization algorithm** is an iterative method for parameter estimation of *maximum likelihood*, where the model also depends on *latent variables*.
- We are still focusing on the solution of the problem **binary image segmentation**. To this end we want to *minimize* the *energy function* $E : \{0,1\}^{\mathcal{V}} \times \mathcal{X} \to \mathbb{R}$:

$$E(\mathbf{y}; \mathbf{x}) = \sum_{i \in \mathcal{V}} -\log \phi_i(y_i; x_i) + \sum_{i \in \mathcal{V},\, j \in N(i)} w_{ij} [\![y_i \neq y_j]\!] \,,$$

  where $\phi_i(y_i; x_i)$ can be obtained by applying the `EM algorithm`.

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

- The `EM algorithm` for *Mixtures of Gaussians*
- *Energy minimization* for **binary image segmentation** via *graph cut*

Literature *

Conditional random field

1. Sebastian Nowozin and Christoph H. Lampert. Structured prediction and learning in computer vision. *Foundations and Trends in Computer Graphics and Vision*, 6(3–4), 2010
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