Pairwise energy terms ................................................................. 19

PDF  
  Continuous random variables .................................................. 20
  Continuous random variable * .................................................. 20
  The Normal (Gaussian) distribution * ....................................... 21
  Mixture of Gaussians ............................................................. 22
  Joint density * ....................................................................... 23
  Marginal densities * ............................................................... 24
  Conditional density * .............................................................. 25

Expectation  
  Expectation ............................................................................. 26
  Expectation ............................................................................. 27
  Conditional expectation .......................................................... 28
  Expected value of a function ................................................... 29

EM algorithm  
  The Expectation-maximization algorithm .................................. 30
  Latent variables ....................................................................... 31
  Jensen’s inequality * ............................................................... 32
  Proof of Jensen’s inequality * .................................................. 33
  The overview of the EM algorithm ......................................... 34
  Lower bound maximization * ................................................... 35
  Lagrange multiplier * ............................................................. 36
  Geometric interpretation of a Lagrange multiplier * ................. 37
  Finding an optimal bound * ..................................................... 38
  Finding an optimal bound * ..................................................... 39
  Finding an optimal bound * ..................................................... 40
  Maximizing the bound * ........................................................ 41
  The EM algorithm ................................................................. 42
  Summary * ............................................................................. 43
  Literature * ............................................................................ 44
Recap *

Image Labeling

Source: Berkeley Segmentation Dataset

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

$L : \mathcal{V} \rightarrow \mathcal{L}^{|\mathcal{V}|}$

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

MRF model

Factorization

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

We want to find the best labeling: $y^* \in \text{argmax}_y p(y)$. 
Recap: Factor graphs

Let us consider the following MRF model:

The factorization is given as

\[ p(y) = \psi_{ijk}(y_i, y_j, y_k) \]
\[ = \psi'_i(y_i) \cdot \psi'_j(y_j) \cdot \psi'_k(y_k) \cdot \psi'_{ij}(y_i, y_j) \cdot \psi'_{ik}(y_i, y_k) \cdot \psi'_{jk}(y_j, y_k) \cdot \psi'_{ijk}(y_i, y_j, y_k) \]

Assume a factorization having with pairwise terms only:

\[ p(y) = \psi_i(y_i) \cdot \psi_j(y_j) \cdot \psi_k(y_k) \cdot \psi_{ij}(y_i, y_j) \cdot \psi_{ik}(y_i, y_k) \cdot \psi_{jk}(y_j, y_k) \cdot \psi_{ijk}(y_i, y_j, y_k) \]
\[ = 1 \cdot 1 \cdot 1 \cdot \psi_{ij}(y_i, y_j) \cdot \psi_{ik}(y_i, y_k) \cdot \psi_{jk}(y_j, y_k) \cdot \psi_{ijk}(y_i, y_j, y_k) \]

This is explicitly shown by the factor graph
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

\[ p(y) = \frac{1}{Z} \prod_{F \in \mathcal{F}} \psi_F(y_{N(F)}) , \]

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

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

\[
p(y \mid x) = \frac{p(y, x)}{p(x)} = \frac{p(y, x)}{\sum_{y' \in Y} p(y', x)} = \frac{1}{Z(x)} \prod_{F \in \mathcal{F}} \frac{1}{Z} \prod_{F \in \mathcal{F}} \psi_F(y_{N(F)}; x_{N(F)})
\]

\[
= \frac{1}{Z(x)} \prod_{F \in \mathcal{F}} \psi_F(y_{N(F)}; x_{N(F)}) .
\]
Conditional random field

\[ p(y \mid x) = \frac{1}{Z(x)} \prod_{F \in \mathcal{F}} \psi_F(y_F; x_F) \]

with the partition function depending on \( x \)

\[ Z(x) = \sum_{y \in \mathcal{Y}} \prod_{F \in \mathcal{F}} \psi_F(y_F; x_F) . \]

Note that the potentials become also functions of (part of) \( x \), i.e. \( \psi_F(y_F; x_F) \) instead of just \( \psi_F(y_F) \). Nevertheless, \( X \) 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 \( p(Y_F = y_F \mid x) \) 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 \( F \in \mathcal{F} \):

\[ E_F(y_F; x_F) = -\log(\psi_F(y_F; x_F)) \quad \Leftrightarrow \quad \psi_F(y_F; x_F) = \exp(-E_F(y_F; x_F)) . \]

\[ p(y \mid x) = \frac{1}{Z(x)} \prod_{F \in \mathcal{F}} \psi_F(y_F; x_F) = \frac{1}{Z(x)} \exp\left(-\sum_{F \in \mathcal{F}} E_F(y_F; x_F)\right) \]

\[ = \frac{1}{Z(x)} \exp(-E(y; x)) . \]
**Inference**

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

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,

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

\[
\mu_F(y_F) = p(Y_F = y_F \mid x) \quad \forall F \in \mathcal{F}, \forall y_F \in \mathcal{Y}_F.
\]

This typically includes variable marginals, i.e. \( \mu_i = p(y_i \mid x) \), 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 \( x \), find the state \( y^* \in \mathcal{Y} \) of maximum probability

\[
y^* \in \arg\max_{y \in \mathcal{Y}} p(Y = y \mid x).
\]

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 $y^* \in \arg\max_{y \in \mathcal{Y}} p(y \mid x)$.

$$\arg\max_{y \in \mathcal{Y}} p(y \mid x) = \arg\max_{y \in \mathcal{Y}} \frac{1}{Z(x)} \exp(-E(y; x))$$

$$= \arg\max_{y \in \mathcal{Y}} \exp(-E(y; x))$$

$$= \arg\max_{y \in \mathcal{Y}} -E(y; x)$$

$$= \arg\min_{y \in \mathcal{Y}} E(y; x) .$$

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 $p(y \mid x)$ for some observation $X = x$.
- 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 $x$, **MAP inference** means to find the state $y^* \in \mathcal{Y}$ of maximum probability

$$y^* \in \arg\max_{y \in \mathcal{Y}} p(Y = y \mid x) .$$
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).

The conditional distribution factorizes (up to pairwise factors) as

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

with

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

where $N(i) = \{ j \in \mathcal{V} : (i, j) \in \mathcal{E}) \}$. The corresponding energy function $E : \{0,1\}^\mathcal{V} \times \mathcal{X} \to \mathbb{R}$:

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

In order to define energy functions for unary factors, one can consider a set of functions $\phi_i : \mathcal{Y}_i \times \mathcal{X}_i \rightarrow [0; 1]$: $E_i(y_i; x_i) = -\log \phi_i(y_i; x_i)$ 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

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

$$= \sum_{i \in V} -\log \phi_i(y_i; x_i) + \sum_{i \in V, j \in N(i)} w_{ij}[y_i \neq y_j].$$
Continuous random variables

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

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

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 $f_X(x)$. 

---

IN2329 - Probabilistic Graphical Models in Computer Vision 3. Conditional random field & Expectation-maximization algorithm – 21 / 46
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 to 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 modeling 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.

The parameters $\pi_k$ 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

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

\[
  f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) \, dy \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y) \, dx.
\]

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

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

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 the sum the numbers showing on the dice.

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

Expectation

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

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

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

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

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

assuming that this series is absolutely convergent.

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

\[
E[X \mid Y = y] = \int_{-\infty}^{\infty} x \cdot f_{X \mid Y}(x \mid y) dx,
\]

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

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

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

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

assuming that this series is absolutely convergent.
Latent variables

Suppose we are given a set of \textit{i.i.d.} (i.e. independent and identically distributed) data samples \( \{x_1, \ldots, x_N\} \) represented by a matrix \( X \in \mathbb{R}^{D \times N} \). The samples are drawn from a \textit{model distribution} (e.g., mixture of Gaussians) given by its parameters \( \theta \).

Basically, there are mainly two applications of the EM algorithm:

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

Latent variables gathering the missing values are represented by a matrix \( Z \).

We generally want to maximize the \textbf{posterior probability}

\[
\theta^* \in \arg \max_{\theta} p(\theta \mid X) = \arg \max_{\theta} \sum_{Z} p(\theta, Z \mid X).
\]

Alternatively, one can maximize the \textbf{log-likelihood}

\[
\mathcal{L}(\theta; X) = \ln p(X \mid \theta) = \ln \sum_{Z} p(X, Z \mid \theta).
\]
Jensen’s inequality *

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

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

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

Assume a random vector $X$ and a convex function $\varphi$, then

$$\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)].$$
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 $l : \mathbb{R} \rightarrow \mathbb{R}$, $l(x) = ax + b$ is an affine function corresponding to the tangent line of $\varphi$ at $\mathbb{E}[X]$.

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

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

The overview of the EM algorithm

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

Lower bound maximization *

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

$$
\ln p(X | \theta) = \ln \sum_Z p(X, Z | \theta) = \ln \sum_Z q^{(t)}(Z) \frac{p(X, Z | \theta)}{q^{(t)}(Z)}
$$

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

$$
= \ln \mathbb{E}_{q^{(t)}(Z)} \left[ \frac{p(X, Z | \theta)}{q^{(t)}(Z)} \right] \geq \mathbb{E}_{q^{(t)}(Z)} \left[ \ln \frac{p(X, Z | \theta)}{q^{(t)}(Z)} \right]
$$

$$
= \sum_Z q^{(t)}(Z) \ln \frac{p(X, Z | \theta)}{q^{(t)}(Z)} \triangleq B(\theta; \theta^{(t)}).
$$

Lagrange multiplier *

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

$$
\max f(x)
$$

subject to $g(x) = 0$.

It is convenient to study the Lagrangian function, defined as

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

where $\lambda \neq 0$ is called a Lagrange multiplier.
Geometric interpretation of a Lagrange multiplier

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

\[
g(x + \epsilon) \approx g(x) + \epsilon^T \nabla g(x) \Rightarrow \epsilon^T \nabla g(x) \approx 0.
\]

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

At an optimal \( x_A \) lying on the constraint surface, \( \nabla f(x_A) \) 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 \( \lambda \) such that

\[
\nabla f + \lambda \nabla g = 0
\]

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

Finding an optimal bound

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

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

\[
B(\theta^{(t)}; \theta^{(t)}) = \sum_Z q^{(t)}(Z) \ln \frac{P(X, Z \mid \theta^{(t)})}{q^{(t)}(Z)}
\]

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

Introducing a Lagrange multiplier \( \lambda \) to enforce \( \sum_Z q^{(t)}(Z) = 1 \), the objective becomes

\[
h(q^{(t)}) = \sum_Z q^{(t)}(Z) \ln p(X, Z \mid \theta^{(t)}) - \sum_Z q^{(t)}(Z) \ln q^{(t)}(Z) + \lambda \left( \sum_Z q^{(t)}(Z) - 1 \right).
\]
Finding an optimal bound

\[ h(q^{(t)}) = \sum_{Z} q^{(t)}(Z) \ln p(x, z, \theta^{(t)}) - \sum_{Z} q^{(t)}(Z) \ln q^{(t)}(Z) + \lambda \left( \sum_{Z} q^{(t)}(Z) - 1 \right). \]

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

\[
\frac{\partial}{\partial q^{(t)}(Z)} h = \ln p(x, z, \theta^{(t)}) - \ln q^{(t)}(Z) - 1 - \lambda = 0.
\]

\[
p(x, z, \theta^{(t)}) \exp(-1 - \lambda) = q^{(t)}(Z)
\]

\[
\exp(-1 - \lambda) \sum_{Z} p(x, z, \theta^{(t)}) = \sum_{Z} q^{(t)}(Z) = 1
\]

\[
\exp(-1 - \lambda) = \frac{1}{\sum_{Z} p(x, z, \theta^{(t)})} = \frac{1}{p(x | \theta^{(t)})}.
\]

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

\[
q^{(t)}(Z) = \frac{p(x, z, \theta^{(t)})}{p(x | \theta^{(t)})} = p(z | x, \theta^{(t)}).
\]
Finding an optimal bound *

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

$$B(\theta^{(t)}; \theta^{(t)}) = \sum_Z q^{(t)}(Z) \ln \frac{p(X, Z | \theta^{(t)})}{q^{(t)}(Z)}$$

By substituting Eq. (2), we get

$$= \sum_Z p(Z | X, \theta^{(t)}) \ln \frac{p(X, Z | \theta^{(t)})}{p(Z | X, \theta^{(t)})} \frac{p(X | \theta^{(t)})}{p(X | \theta^{(t)})}$$

$$= \ln p(X | \theta^{(t)}) \sum_Z p(Z | X, \theta^{(t)}) \frac{p(X | \theta^{(t)})}{p(X | \theta^{(t)})}$$

$$= \ln p(X | \theta^{(t)}) .$$
Maximizing the bound *

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

$$B(\theta; \theta^{(t)}) = \sum_Z q^{(t)}(Z) \ln \frac{p(X, Z | \theta)}{q^{(t)}(Z)}$$

$$= \sum_Z q^{(t)}(Z) \ln p(X, Z | \theta) - \sum_Z q^{(t)}(Z) \ln q^{(t)}(Z).$$

We need to consider the first term only

$$\sum_Z q^{(t)}(Z) \ln p(X, Z | \theta) = \sum_Z p(Z | X, \theta^{(t)}) \ln p(X, Z | \theta)$$

$$= \mathbb{E} \left[ \ln p(X, Z | \theta) | X, \theta^{(t)} \right] \overset{\Delta}{=} Q(\theta, \theta^{(t)}).$$

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

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

$$
\theta^{(t)} = \arg\max_\theta Q(\theta, \theta^{(t-1)})
$$

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

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

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

$$E : \{0, 1\}^V \times X \rightarrow \mathbb{R}$$

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

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


The Expectation-maximization algorithm