Expectation Maximization

Seungjin Choi

Department of Computer Science and Engineering
Pohang University of Science and Technology
77 Cheongam-ro, Nam-gu, Pohang 37673, Korea
seungjin@postech.ac.kr

Outline

- Mathematical preliminaries
  - Jensen's inequality and Gibb's inequality
  - Entropy and mutual information
- Expectation maximization
- More on EM
  - Generalized EM and incremental EM
  - EM for exponential family

Convex Sets and Functions

Definition (Convex Sets)

Let C be a subset of $\mathbb{R}^m$. C is called a convex set if

$$\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y} \in C, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in C, \ \forall \alpha \in [0,1]$$

Definition (Convex Function)

Let C be a convex subset of $\mathbb{R}^m$. A function $f : C \mapsto \mathbb{R}$ is called a convex function if

$$f\left(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}\right) \leq \alpha f(\boldsymbol{x}) + (1-\alpha) f(\boldsymbol{y}) \quad \forall \boldsymbol{x}, \boldsymbol{y} \in C, \ \forall \alpha \in [0,1]$$

Jensen's Inequality

Theorem (Jensen's Inequality)

If $f(\boldsymbol{x})$ is a convex function and $\boldsymbol{x}$ is a random vector, then

$$\boxed{E\{f(\boldsymbol{x})\} \geq f\left(E\{\boldsymbol{x}\}\right).}$$

Note: Jensen's inequality can also be rewritten for a concave function, with the direction of the inequality reversed.

Proof of Jensen's Inequality

Need to show that $\sum_{i=1}^{N} p_i f(x_i) \geq f\left(\sum_{i=1}^{N} p_i x_i\right)$. The proof is based on the recursion, working from the right-hand side of this equation.

$$
\begin{aligned}
f\left(\sum_{i=1}^{N} p_i x_i\right) &= f\left(p_1 x_1 + \sum_{i=2}^{N} p_i x_i\right) \\
&\leq p_1 f(x_1) + \left[\sum_{i=2}^{N} p_i\right] f\left(\frac{\sum_{i=2}^{N} p_i x_i}{\sum_{i=2}^{N} p_i}\right) \qquad \left(\text{choose } \alpha = \frac{p_1}{\sum_{i=1}^{N} p_i}\right) \\
&\leq p_1 f(x_1) + \left[\sum_{i=2}^{N} p_i\right] \left\{\alpha f(x_2) + (1-\alpha) f\left(\frac{\sum_{i=3}^{N} p_i x_i}{\sum_{i=3}^{N} p_i}\right)\right\} \\
&\qquad \left(\text{choose } \alpha = \frac{p_2}{\sum_{i=2}^{N} p_i}\right) \\
&= p_1 f(x_1) + p_2 f(x_2) + \sum_{i=3}^{N} p_i f\left(\frac{\sum_{i=3}^{N} p_i x_i}{\sum_{i=3}^{N} p_i}\right),
\end{aligned}
$$

and so forth.

Information Theory

- Information theory answers two fundamental questions in communication theory
  - What is the ultimate data compression? $\longrightarrow$ entropy H.
  - What is the ultimate transmission rate of communication? $\longrightarrow$ channel capacity C.
- In the early 1940's, it was thought that increasing the transmission rate of information over a communication channel increased the probability of error $\longrightarrow$ *"This is not true."*
  Shannon surprised the communication theory community by proving that this was not true as long as the communication rate was below the channel capacity.
- Although information theory was developed for communications, it is also important to explain ecological theory of sensory processing. Information theory plays a key role in elucidating the goal of unsupervised learning.

Information and Entropy

- Information can be thought of as surprise, uncertainty, or unexpectedness. Mathematically it is defined by

$$\boxed{I = -\log P(i),}$$

  where $P(i)$ is the probability that the event labelled i occurs. The rare event gives large information and frequent event produces small information.

- Entropy is average information, i.e.,

$$\boxed{H = -\sum_{i=1}^{N} P(i) \log P(i).}$$

Example: Horse Race

Suppose we have a horse race with eight horses taking part. Assume that the probabilities of winning for the eight horse are

$$(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}).$$

Suppose that we wish to send a message to another person indicating which horse won the race.

How many bits are required to describe this for each of the horses?

3 bits for any of the horses?

No! The win probabilities are not uniform.
It makes sense to use shorter descriptions for the more probable horses and longer descriptions for the less probable ones so that we achieve a lower average description length. For example, we can use the following strings to represent the eight horses:

$$0, 10, 110, 1110, 111100, 111101, 111110, 111111.$$

The average description length in this case is 2 bits as opposed to 3 bits for the uniform code.
We calculate the entropy:

$$\begin{aligned} H(X) &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{16}\log\frac{1}{16} - 4\frac{1}{64}\log\frac{1}{64} \\ &= 2 \text{ bits.} \end{aligned}$$

The entropy of a random variable is a lower bound on the average number of bits required to represent the random variables and also on the average number of questions needed to identify the variable in a game of "twenty questions".

Entropy and Relative Entropy

- Entropy is the average information (a measure of uncertainty) that is defined by

$$\begin{aligned} H(x) &= -\sum_{x \in \mathcal{X}} p(x) \log p(x) \\ &= -E_p\{\log p(x)\}. \end{aligned}$$

- Relative entropy (Kullback-Leibler divergence) is a measure of distance between two distributions and is defined by

$$\begin{aligned} KL[p \| q] &= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\ &= E_p \left\{ \log \frac{p(x)}{q(x)} \right\}. \end{aligned}$$

Mutual Information

- Mutual information is the relative entropy between the joint distribution and the product of marginal distributions,

$$\begin{aligned} I(x,y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\left[\frac{p(x,y)}{p(x)p(y)}\right] \\ &= D\left[p(x,y)\|p(x)p(y)\right] \\ &= E_{p(x,y)}\left\{\log\left[\frac{p(x,y)}{p(x)p(y)}\right]\right\}. \end{aligned}$$

- Mutual information can be interpreted as the reduction in the uncertainty of x due to the knowledge of y, i.e.,

$$I(x,y) = H(x) - H(x|y),$$

where $H(x|y) = -E_{p(x,y)}\left\{\log p(x|y)\right\}$ is the conditional entropy

Gibb's Inequality

Theorem

$KL[p||q] \geq 0$ with equality iff $p = q$.

Proof: Consider the Kullback-Leibler divergence for discrete distributions:

$$\begin{aligned}
KL[p||q] &= \sum_i p_i \log \frac{p_i}{q_i} \\
&= -\sum_i p_i \log \frac{q_i}{p_i} \\
&\geq -\log \left[\sum_i p_i \frac{q_i}{p_i} \right] \quad \text{(by Jensen's inequality)} \\
&= -\log \left[\sum_i q_i \right] \\
&= 0.
\end{aligned}$$

More on Gibb's Inequality

In order to find the distribution p which minimizes $KL[p||q]$, we consider a Lagrangian

$$\mathcal{E} = KL[p||q] + \lambda\left(1 - \sum_i p_i\right) = \sum_i p_i \frac{p_i}{q_i} + \lambda\left(1 - \sum_i p_i\right).$$

Compute the partial derivative $\frac{\partial \mathcal{E}}{\partial p_k}$ and set to zero,

$$\frac{\partial \mathcal{E}}{\partial p_k} = \log p_k - \log q_k + 1 - \lambda = 0,$$

which leads to $p_k = q_k e^{\lambda - 1}$. It follows from $\sum_i p_i = 1$ that $\sum_i q_i e^{\lambda - 1} = 1$, which leads to $\lambda = 1$. Therefore $p_i = q_i$.
The Hessian, $\frac{\partial^2 \mathcal{E}}{\partial p_i^2} = \frac{1}{p_i}$, $\frac{\partial^2 \mathcal{E}}{\partial p_i \partial p_j} = 0$, is positive definite, which shows that $p_i = q_i$ is a genuine minimum.

The EM Algorithm

- Maximum likelihood parameter estimates
  - One definition of the "best" knob settings.
  - Often impossible to find directly.
- The EM algorithm:
  - Finds ML parameters when the original (hard) problem can be broken up into two (easy) pieces, i.e., finds the ML estimates of parameters for data in which some variables are unobserved.
  - A iterative method which consists of "Expectation step" (E-step) and "Maximization step" (M-step).
    1. Estimate some "missing" or "unobserved" data from observed data and current parameters (**E-step**)
    2. Using this "complete" data, find the maximum likelihood parameter estimates (**M-step**).
- For EM to work, two things have to be easy:
  1. Guessing (estimating) missing data form data we have and our current guess of parameters.
  2. Solving for the ML parameters directly given the complete data.

Auxiliary Function

Definition

$Q(\theta, \theta')$ is an auxiliary function for $\mathcal{L}(\theta)$ if the conditions

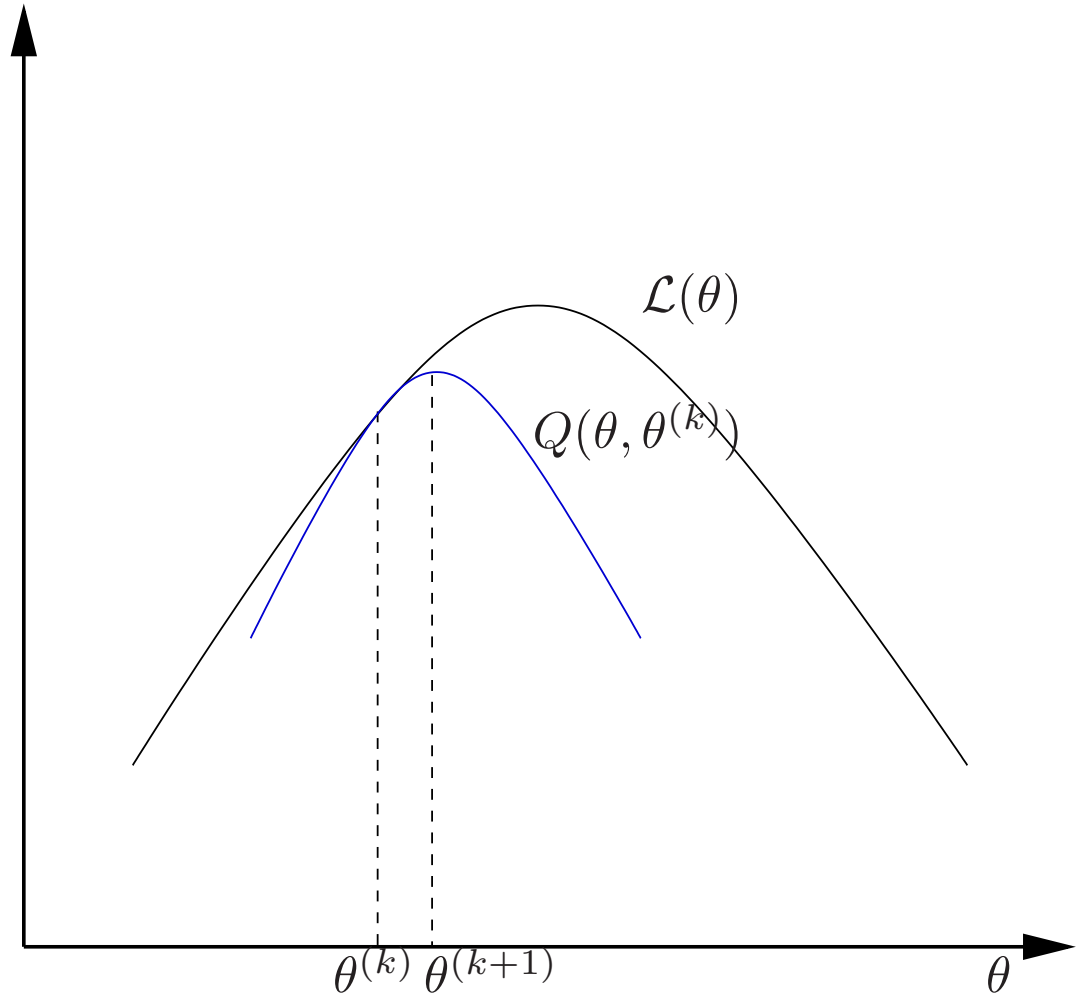
$$\begin{aligned} Q(\theta, \theta') &\leq \mathcal{L}(\theta), \\ Q(\theta, \theta) &= \mathcal{L}(\theta). \end{aligned}$$

Theorem

If Q is an auxiliary function, then $\mathcal{L}$ is nondecreasing under the update

$$Q\left(\theta, \theta^{(k+1)}\right) \leftarrow \arg\max_{\theta} Q\left(\theta, \theta^{(k)}\right).$$

A Graphical Illustration for an Auxiliary Function



A Lower Bound on the Likelihood

Consider a set of observed (visible) variables $\boldsymbol{x}$, a set of unobserved (hidden) variables $\boldsymbol{s}$, and model parameters θ.
Then the log-likelihood is given by

$$\begin{aligned}\mathcal{L}(\theta) &= \log p(\boldsymbol{x}|\theta) \\ &= \log \int p(\boldsymbol{x}, \boldsymbol{s}|\theta) d\boldsymbol{s}.\end{aligned}$$

Use Jensen's inequality for any distribution of hidden variables, $q(\boldsymbol{s})$, to obtain

$$\begin{aligned}\mathcal{L}(\theta) &= \log \int q(\boldsymbol{s}) \frac{p(\boldsymbol{x}, \boldsymbol{s}|\theta)}{q(\boldsymbol{s})} d\boldsymbol{s} \\ &\geq \int q(\boldsymbol{s}) \log \left[\frac{p(\boldsymbol{x}, \boldsymbol{s}|\theta)}{q(\boldsymbol{s})} \right] d\boldsymbol{s} \quad \text{(use Jensen's inequality)} \\ &= \mathcal{F}(q, \theta). \quad \text{(lower bound on the log-likelihood)}\end{aligned}$$

The Lower Bound $\mathcal{F}(q, \theta)$

Consider the lower bound $\mathcal{F}(q, \theta)$,

$$\begin{aligned} \mathcal{F}(q, \theta) &= \int q(\boldsymbol{s}) \log \left[\frac{p(\boldsymbol{x}, \boldsymbol{s}|\theta)}{q(\boldsymbol{s})} \right] d\boldsymbol{s} \\ &= \int q(\boldsymbol{s}) \log p(\boldsymbol{x}, \boldsymbol{s}|\theta) d\boldsymbol{s} + H(q), \end{aligned}$$

where $H(q) = -\int q(\boldsymbol{s}) \log q(\boldsymbol{s}) d\boldsymbol{s}$ (entropy of q) and $-\mathcal{F}(q, \theta)$ corresponds to variational free energy.

- In the EM algorithm, we alternately optimize $\mathcal{F}(q, \theta)$ w.r.t q and θ.
- It can be shown that this will never decrease $\mathcal{L}$.

EM as an Alternative Optimization

- E-step: Optimize $\mathcal{F}(q, \theta)$ w.r.t. the distribution over hidden variables given the parameters, i.e.,

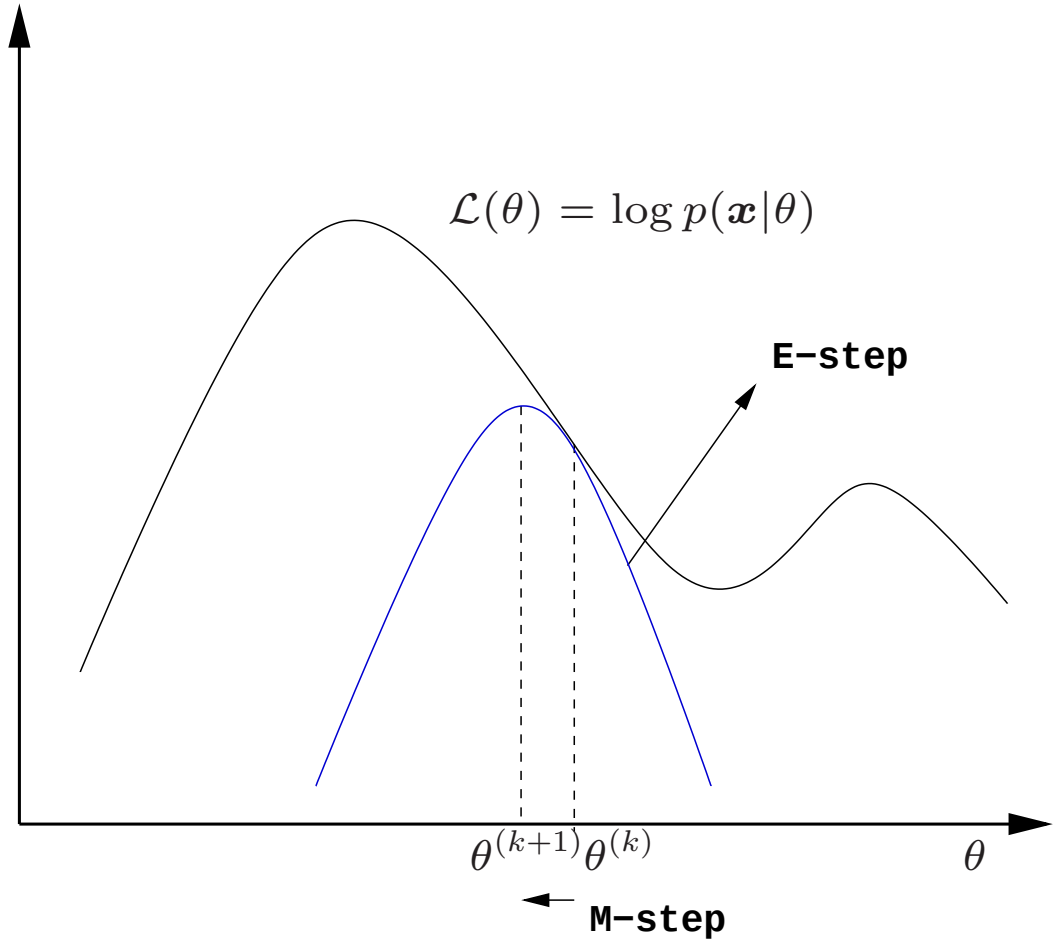
$$q^{(k+1)} = \arg\max_{q} \mathcal{F}\left(q, \theta^{(k)}\right).$$

- M-step: Maximize $\mathcal{F}(q, \theta)$ w.r.t. the parameters, given the hidden distribution, i.e.,

$$\begin{aligned} \theta^{(k+1)} &= \arg\max_{\theta} \mathcal{F}\left(q^{(k+1)}, \theta\right) \\ &= \arg\max_{\theta} \int q^{(k+1)}(\boldsymbol{s}) \log p(\boldsymbol{x}, \boldsymbol{s}|\theta) d\boldsymbol{s}, \end{aligned}$$

where $\log p(\boldsymbol{x}, \boldsymbol{s}|\theta)$ is the complete-data log-likelihood.

A Graphical Illustration of EM Behavior



The EM Algorithm never Decreases the Log-Likelihood

The difference between the log-likelihood and the lower bound is given by

$$\begin{aligned}
\mathcal{L}(\theta) - \mathcal{F}(q,\theta) &= \log p(\boldsymbol{x}|\theta) - \int q(\boldsymbol{s}) \log \left[\frac{p(\boldsymbol{x},\boldsymbol{s}|\theta)}{q(\boldsymbol{s})} \right] d\boldsymbol{s} \\
&= \log p(\boldsymbol{x}|\theta) - \int q(\boldsymbol{s}) \log \left[\frac{p(\boldsymbol{s}|\boldsymbol{x},\theta)p(\boldsymbol{x}|\theta)}{q(\boldsymbol{s})} \right] d\boldsymbol{s} \\
&= -\int q(\boldsymbol{s}) \log \left[\frac{p(\boldsymbol{s}|\boldsymbol{x},\theta)}{q(\boldsymbol{s})} \right] d\boldsymbol{s} \\
&= KL\left[q(\boldsymbol{s})||p(\boldsymbol{s}|\boldsymbol{x},\theta)\right].
\end{aligned}$$

This difference is zero only if $q(\boldsymbol{s}) = p(\boldsymbol{s}|\boldsymbol{x},\theta)$. (this is E-step)

$$\mathcal{L}\left(\theta^{(k)}\right) \underset{\text{E-step}}{=} \mathcal{F}\left(q^{(k+1)},\theta^{(k)}\right) \underset{\text{M-step}}{\leq} \mathcal{F}\left(q^{(k+1)},\theta^{(k+1)}\right) \underset{\text{Jensen}}{\leq} \mathcal{L}\left(\theta^{(k+1)}\right).$$

Generalized EM: Partial M-Steps

- The M-step of the algorithm may be only partially implemented, with the new estimate for the parameters improving the likelihood given the distribution found in the E-step, but not necessarily maximizing it.
- Such a partial M-step always results in the true likelihood improving as well.

Incremental EM: Partial E-Steps

- The unobserved variables are commonly independent, and influence the likelihood of parameters only through simple sufficient statistics. If these statistics can be updated incrementally when the distribution for one of the variables is re-calculated, it makes sense to immediately re-estimate the parameters before performing the E-step for the next unobserved variable, as this utilizes the new information immediately, speeding convergence.
- The proof holds even for the case of partial updates.
- Sparse or online versions of the EM algorithm would compute the posterior for a subset of the data points or as the data arrives, respectively. You can also update the posterior over a subset of the hidden variables, while holding others fixed.

Incremental EM

Assume that unobserved variables are independent, then we restrict the search for a maximum of $\mathcal{F}$ to distributions $q(\boldsymbol{s}) = \prod_t q_t(\boldsymbol{s}_t)$.
We can write $\mathcal{F}$ in the form $\mathcal{F}(q, \theta) = \sum_t \mathcal{F}_t(q_t, \theta)$ where

$$\mathcal{F}_t(q_t, \theta) = \langle \log p(\boldsymbol{s}_t, \boldsymbol{x}_t | \theta) \rangle_{q_t} + H(q_t).$$

Algorithm Outline: Incremental EM

E-step Choose a data point, $\boldsymbol{x}_t$.

$$\begin{aligned} \text{Set } q_l^{(k)} &= q_l^{(k-1)} \quad \text{for } l \neq t. \\ \text{Set } q_t^{(k)} &= \arg\max_{q_t} \mathcal{F}_t \left(q_t, \theta^{(k-1)} \right). \end{aligned}$$

M-step

$$\theta^{(k)} = \arg\max_{\theta} \mathcal{F} \left(q^{(k)}, \theta \right).$$

Sufficient Statistic

Definition

Let x be a random variable whose distribution depends on a parameter θ. A real-valued function $t(x)$ of x is said to be sufficient for θ if the condition distribution of x, is independent of θ. That is, t is sufficient for θ if

$$p(x|t,\theta) = p(x|t).$$

Exponential Family

Definition

A family of distributions is said to be a *k*-parameter exponential family if the probability density function for $\boldsymbol{x}$ has the form

$$\begin{aligned} p(\boldsymbol{x}|\theta) &= a(\theta)b(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^\top(\theta)\boldsymbol{t}(\boldsymbol{x})\right\}, \\ a^{-1}(\theta) &= \int b(\boldsymbol{x})\exp\left\{\boldsymbol{\eta}^\top(\theta)\boldsymbol{t}(\boldsymbol{x})\right\} d\boldsymbol{x}, \end{aligned}$$

where $\boldsymbol{\eta}(\theta) = [\eta_1(\theta), \ldots, \eta_k(\theta)]^\top$ (natural parameters) contains k functions of the parameter θ and the sufficient statistics $\boldsymbol{t}(\boldsymbol{x}) = [t_1(\boldsymbol{x}), \ldots, t_k(\boldsymbol{x})]^\top$ contains k functions of the data $\boldsymbol{x}$.

Alternative equivalent form is given by

$$\begin{aligned} p(\boldsymbol{x}|\theta) &= \exp\left\{\boldsymbol{\eta}^\top(\theta)\boldsymbol{t}(\boldsymbol{x}) + g(\theta) + h(\boldsymbol{x})\right\}, \\ g(\theta) &= -\log\int \exp\left\{\boldsymbol{\eta}^\top(\theta)\boldsymbol{t}(\boldsymbol{x}) + h(\boldsymbol{x})\right\} d\boldsymbol{x}. \end{aligned}$$

Exponential Family: Canonical Form

If $\boldsymbol{\eta}(\theta) = \theta$, then the exponential family is said to be in canonical form.

By defining a transformed parameter $\boldsymbol{\eta} = \boldsymbol{\eta}(\theta)$, it is always possible to convert a n exponential family to canonical form. The canonical form is non-unique, since $\boldsymbol{\eta}(\theta)$ can be multiplied by any nonzero constant, provided that $\boldsymbol{t}(\boldsymbol{x})$ is multiplied by that constant's reciprocal.

Let $\boldsymbol{x} \in \mathbb{R}^m$ be a random vector whose distribution belongs to the k-parameter exponential family, then the k-dimensional statistic $\boldsymbol{t}(\boldsymbol{x})$ is *sufficient* for the parameter θ.

Example of Exponential Family: Gaussian

Let us consider a univariate Gaussian distribution parameterized by its mean and variance:

$$\begin{aligned} p(x) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\} \exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x\right\}. \end{aligned}$$

This is a two-parameter exponential family with

$$\begin{aligned} a(\theta) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}, \quad b(x) = 1 \\ \boldsymbol{\eta} &= \left[-\frac{1}{2\sigma^2},\ \frac{\mu}{\sigma^2}\right]^T, \quad \boldsymbol{t} = \left[x^2,\ x\right]^T. \end{aligned}$$

Example of Exponential Family: Bernoulli

Consider Bernoulli distribution parameterized by $\mu \in [0, 1]$:

$$p(x) = \mu^x (1 - \mu)^{1-x}.$$

Matching it with the canonical form $p(x) = \exp\{\theta x + g(\theta) + h(x)\}$ yields

$$\begin{aligned} \theta &= \log \frac{\mu}{1-\mu}, \\ g(\theta) &= -\log\left(1 + e^{\theta}\right), \\ h(x) &= 0. \end{aligned}$$

Parameter values of the distribution are mapped to natural parameters via link function.

EM for Exponential Family

Given a complete data $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{s})$, we write the expected complete-data log-likelihood:

$$\begin{aligned}
\langle \mathcal{L}_c \rangle &= \left\langle \log p(\boldsymbol{x}, \boldsymbol{s}) | \boldsymbol{x}, \theta^{(k)} \right\rangle \\
&= \left\langle \boldsymbol{\eta}(\theta)^\top \boldsymbol{t}(\boldsymbol{z}) + g(\theta) + h(\boldsymbol{z}) | \boldsymbol{x}, \theta^{(k)} \right\rangle \\
&= \boldsymbol{\eta}(\theta)^\top \left\langle \boldsymbol{t}(\boldsymbol{z}) | \boldsymbol{x}, \theta^{(k)} \right\rangle + g(\theta) + \left\langle h(\boldsymbol{z}) | \boldsymbol{x}, \theta^{(k)} \right\rangle
\end{aligned}$$

EM algorithm for exponential families

- E-step: Requires only sufficient statistics

$$\boldsymbol{t}^{(k+1)} = \left\langle \boldsymbol{t}(\boldsymbol{z}) | \boldsymbol{x}, \theta^{(k)} \right\rangle .$$

- M-step:

$$\theta^{(k+1)} = \arg\max_{\theta} \left[\boldsymbol{\eta}(\theta)^\top \boldsymbol{t}^{(k+1)} + g(\theta) \right] .$$

References

- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977) "Maximum likelihood from incomplete data via the EM algorithm" (with discussion), *Journal of the Royal Statistical Society B*, vol. 39, pp. 1-38.
- Neal, R. M. and Hinton, G. E. (1999) "A view of the EM algorithm that justifies incremental, sparse, and other variants" *Learning in Graphical Models* (edited by M. Jordan), pp. 355-368.