Lecture 5: Parameter Estimation in Graphical Models

Tianpei Xie

Aug. 30th., 2022

Contents

1 Background knowledge

Recall the formulation of Bayesian network and Markov network

- Given directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $(s,t) \neq (t,s)$, the **directed graphical model** factorizes the joint distribution into a set of *factors* $\{p_s(x_s|x_{\pi(s)}) : s \in \mathcal{V}\}$ according to the ancestor relations defined in $\mathcal{G}$

$$p(x_1, \ldots, x_m) = \prod_{s \in \mathcal{V}} p_s(x_s|x_{\pi(s)}). \tag{1}$$

  This class of models are also referred as ***Bayesian networks*** [Koller and Friedman, 2009].

- Given undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $(s,t) = (t,s)$, the joint distribution of **Markov random fields** (**Markov network**) *factorize* as

$$p(x_1, \ldots, x_m) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C), \tag{2}$$

  where Z is a constant chosen to ensure that the distribution is normalized. The set $\mathcal{C}$ is often taken to be the *set of all* ***maximal cliques*** *of the graph*, i.e., the set of cliques that are *not* properly contained within any other clique. Note that any representation based on nonmaximal cliques can always be converted to one based on maximal cliques by redefining the compatibility function on a maximal clique to be the *product* over the compatibility functions on the *subsets* of that clique.

- The canonical representation of ***exponential famlity*** of distribution has the following form

$$\begin{aligned} p(x_1, \ldots, x_m) = p(\boldsymbol{x}; \boldsymbol{\eta}) &= \exp\left(\langle \boldsymbol{\eta}\,,\, \boldsymbol{\phi}(\boldsymbol{x}) \rangle - A(\boldsymbol{\eta})\right) h(\boldsymbol{x}) \nu(d\boldsymbol{x}) \\ &= \exp\left(\sum_{\alpha} \eta_\alpha \phi_\alpha(\boldsymbol{x}) - A(\boldsymbol{\eta})\right) \end{aligned} \tag{3}$$

  where ϕ is a feature map and $\boldsymbol{\phi}(\boldsymbol{x})$ defines a set of ***sufficient statistics*** (or ***potential functions***). The normalization factor is defined as

$$A(\boldsymbol{\eta}) := \log \int \exp\left(\langle \boldsymbol{\eta}\,,\, \boldsymbol{\phi}(\boldsymbol{x}) \rangle\right) h(\boldsymbol{x}) \nu(d\boldsymbol{x}) = \log Z(\boldsymbol{\eta})$$

  $A(\boldsymbol{\eta})$ is also referred as ***log-partition function*** or *cumulant function*. The parameters $\boldsymbol{\eta} = (\eta_\alpha)$ are called ***natural parameters*** or *canonical parameters*. The canonical parameter $\{\eta_\alpha\}$ forms a **natural (canonical) parameter space**

$$\Omega = \left\{ \boldsymbol{\eta} \in \mathbb{R}^d : A(\boldsymbol{\eta}) < \infty \right\} \tag{4}$$

- The exponential family is the unique solution of ***maximum entropy estimation*** problem:

$$\min_{q \in \Delta} \quad \mathbb{KL}\left(q \,\|\, p_0\right) \tag{5}$$

$$\text{s.t.} \quad \mathbb{E}_q\left[\phi_\alpha(X)\right] = \mu_\alpha \quad \forall\, \alpha \in \mathcal{I} \tag{6}$$

where $\mathbb{KL}\left(q \,\|\, p_0\right) = \int \log(\frac{q}{p_0}) q dx = \mathbb{E}_q\left[\log \frac{q}{p_0}\right]$ is the relative entropy or the Kullback-Leibler divergence of q w.r.t. p_0.

Here $\boldsymbol{\mu} = (\mu_\alpha)_{\alpha \in \mathcal{I}}$ is a set of ***mean parameters***. The space of mean parameters $\mathcal{M}$ is a *convex polytope* spanned by potential functions $\{\phi_\alpha\}$.

$$\mathcal{M} := \left\{\boldsymbol{\mu} \in \mathbb{R}^d : \exists q \text{ s.t. } \mathbb{E}_q\left[\phi_\alpha(X)\right] = \mu_\alpha \quad \forall\, \alpha \in \mathcal{I}\right\} = \text{conv}\left\{\phi_\alpha(x),\ x \in \mathcal{X},\ \alpha \in \mathcal{I}\right\} \tag{7}$$

- Note that $A(\boldsymbol{\eta})$ is a convex function and its gradient $\nabla A : \Omega \to \mathcal{M}^\circ$ is a bijection between the natural parameter space Ω and the **interior** of $\mathcal{M}$, $\mathcal{M}^\circ$; $\nabla A(\boldsymbol{\eta}) = \boldsymbol{\mu}$ based on the following equation

$$\frac{\partial A}{\partial \eta_\alpha} = \mathbb{E}_{\boldsymbol{\eta}}\left[\phi_\alpha(X)\right] := \int_{\mathcal{X}^m} \phi_\alpha(\boldsymbol{x}) q(\boldsymbol{x}; \boldsymbol{\eta}) d\boldsymbol{x} = \mu_\alpha \tag{8}$$

- Moreover $A(\boldsymbol{\eta})$ has a variational form

$$A(\boldsymbol{\eta}) = \sup_{\boldsymbol{\mu} \in \mathcal{M}} \left\{\langle \boldsymbol{\eta}\,,\, \boldsymbol{\mu}\rangle - A^*(\boldsymbol{\mu})\right\} \tag{9}$$

  where $A^*(\boldsymbol{\mu})$ is the conjugate dual function of A and it is defined as

$$A^*(\boldsymbol{\mu}) := \sup_{\boldsymbol{\eta} \in \Omega} \left\{\langle \boldsymbol{\mu}\,,\, \boldsymbol{\eta}\rangle - A(\boldsymbol{\eta})\right\} \tag{10}$$

  It is shown that $A^*(\boldsymbol{\mu}) = -H(q_{\boldsymbol{\eta}(\boldsymbol{\mu})})$ for $\boldsymbol{\mu} \in \mathcal{M}^\circ$ which is the negative entropy. $A^*(\boldsymbol{\mu})$ is also the optimal value for the **maximum likelihood estimation** problem on p. The exponential family can be reparameterized according to its mean parameters $\boldsymbol{\mu}$ via backward mapping $(\nabla A)^{-1} : \mathcal{M}^\circ \to \Omega$, called **mean parameterization**.

- The maximum likelihood estimation of exponential family is essentially the **dual problem** of the maximum entropy estimation (5).

$$\begin{aligned} &\max_{\boldsymbol{\eta}} \frac{1}{N} \sum_{n=1}^{N} \log q_{\boldsymbol{\eta}}(X_n) \\ \Rightarrow &\max_{\boldsymbol{\eta}} \langle \bar{\boldsymbol{\mu}}\,,\, \boldsymbol{\eta}\rangle - A(\boldsymbol{\eta}) \end{aligned} \tag{11}$$

  where $\bar{\boldsymbol{\mu}} = \hat{\mathbb{E}}[\boldsymbol{\phi}(X)] = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\phi}(X_n)$ fits the moment matching conditions. (11) is the right-hand side of the conjugate dual of log-partition function A^* in (10). Thus we have one statistical interpretation of this variational problem (10): A^* is the **optimal value of the rescaled log likelihood** (11).

  Also see that the gradient of log-likelihood function

$$\begin{aligned} \nabla_{\boldsymbol{\eta}} \frac{1}{N} \sum_{n=1}^{N} \log q_{\boldsymbol{\eta}}(X_n) &= \nabla_{\boldsymbol{\eta}}(\langle \hat{\boldsymbol{\mu}}\,,\, \boldsymbol{\eta}\rangle - A(\boldsymbol{\eta})) \\ &= \hat{\mathbb{E}}[\boldsymbol{\phi}(X)] - \mathbb{E}_{\boldsymbol{\eta}}\left[\boldsymbol{\phi}(X)\right] \\ &= \hat{\boldsymbol{\mu}} - \boldsymbol{\mu} = \text{sample mean} - \text{model mean} \end{aligned} \tag{12}$$

  This gives the moment matching condition $\mathbb{E}_{\boldsymbol{\eta}^*}\left[\boldsymbol{\phi}(X)\right] = \hat{\boldsymbol{\mu}}$ of MLE optimal solution $\boldsymbol{\eta}^*$. From the formula above, whenever $\hat{\boldsymbol{\mu}} \in \mathcal{M}^\circ$, there exists a **unique** maximum likelihood solution.

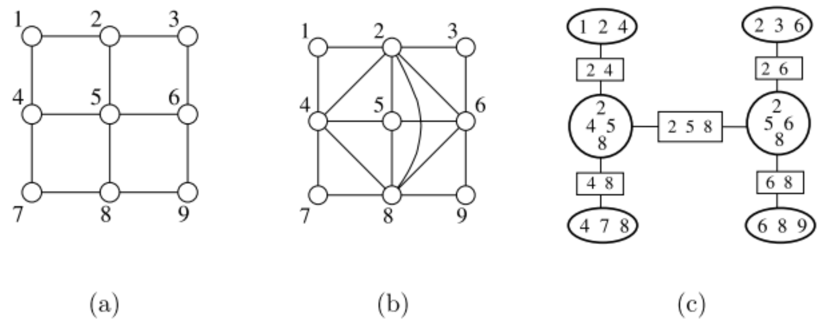

Fig. 2.11 Illustration of junction tree construction. (a) Original graph is a 3 × 3 grid. (b) Triangulated version of original graph. Note the two 4-cliques in the middle. (c) Corresponding junction tree for triangulated graph in (b), with maximal cliques depicted within ellipses, and separator sets within rectangles.

Figure 1: Triangulation and junction tree. [Wainwright et al., 2008]

- We can formulate the **KL divergence** between two distributions in exponential family Ω using its primal and dual form
  - **Primal-form**: given $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \Omega$
$$\begin{aligned}\mathbb{KL}\left(p_{\boldsymbol{\eta}_1} \,\|\, p_{\boldsymbol{\eta}_2}\right) \equiv \mathbb{KL}\left(\boldsymbol{\eta}_1 \,\|\, \boldsymbol{\eta}_2\right) &= A(\boldsymbol{\eta}_2) - A(\boldsymbol{\eta}_1) - \langle \boldsymbol{\mu}_1 \,,\, \boldsymbol{\eta}_2 - \boldsymbol{\eta}_1 \rangle \\ &\equiv A(\boldsymbol{\eta}_2) - A(\boldsymbol{\eta}_1) - \langle \nabla A(\boldsymbol{\eta}_1) \,,\, \boldsymbol{\eta}_2 - \boldsymbol{\eta}_1 \rangle\end{aligned} \tag{13}$$
  - **Primal-dual form**: given $\boldsymbol{\mu}_1 \in \mathcal{M}, \boldsymbol{\eta}_2 \in \Omega$
$$\mathbb{KL}\left(\boldsymbol{\mu}_1 \,\|\, \boldsymbol{\eta}_2\right) = A(\boldsymbol{\eta}_2) + A^*(\boldsymbol{\mu}_1) - \langle \boldsymbol{\mu}_1 \,,\, \boldsymbol{\eta}_2 \rangle \tag{14}$$
  - **Dual-form**: given $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathcal{M}$
$$\begin{aligned}\mathbb{KL}\left(\boldsymbol{\mu}_1 \,\|\, \boldsymbol{\mu}_2\right) &= A^*(\boldsymbol{\mu}_1) - A^*(\boldsymbol{\mu}_2) - \langle \boldsymbol{\eta}_2 \,,\, \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \rangle \\ &\equiv A^*(\boldsymbol{\mu}_1) - A^*(\boldsymbol{\mu}_2) - \langle \nabla A^*(\boldsymbol{\mu}_2) \,,\, \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \rangle\end{aligned} \tag{15}$$

2 Parameter estimation in fully observed models

The simplest case of parameter estimation corresponds to the case of fully observed data: a collection $\boldsymbol{X}^{1:n} := \{\boldsymbol{X}^1, \ldots, \boldsymbol{X}^n\}$ of n independent and identically distributed (i.i.d.) m-vectors, each sampled according to $p_{\boldsymbol{\eta}}$. Suppose that our goal to estimate the unknown parameter $\boldsymbol{\eta}$, which we view as a deterministic but nonrandom quantity for the moment. This problem is solved via maximum likelihood estimation. For exponential family, the MLE can be formulated as in (11), The optimal solution is **unique** and is specified by the ***moment matching conditions***.

2.1 Maximum likelihood for triangulated graphs

In this section, we focus on solving the MLE problem (11) on *triangulated graphs*. We say that a graph is ***triangulated*** if every cycle of length *four* or longer has a *chord*, meaning an edge joining a pair of nodes that are not adjacent on the cycle. **A key theorem** is that a graph $\mathcal{G}$ has a ***junction tree*** if and only if it is triangulated. See Figure 1 for an illustration.

For triangulated graphs, the MLE can be written as a ***closed-form function*** of the ***empirical marginals*** $\boldsymbol{\mu}$ [Wainwright et al., 2008]. For the sake of simplicity, let us consider the *simplest* triangulated graph: a tree $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ with discrete variables $\mathcal{X} = \{0, 1, \ldots, r-1\}$. Recall the pairwise Markov random field with indicator potentials

$$\phi_{s;j}(x_s) = \mathbb{1}\{x_s = j\} \tag{16}$$

Moreover, for each edge (s,t) and pair of values $(j,k) \in \mathcal{X} \times \mathcal{X}$, define the sufficient statistics

$$\phi_{st;jk}(x_s, x_t) = \mathbb{1}\{x_s = j \wedge x_t = k\} \tag{17}$$

The joint distribution is

$$p(x_1, \ldots, x_m; \boldsymbol{\eta}) = \exp\left(\sum_{s\in\mathcal{V}}\sum_{j\in\mathcal{X}} \eta_{s;j}\,\phi_{s;j}(x_s) + \sum_{(s,t)\in\mathcal{E}}\sum_{(j,k)\in\mathcal{X}\times\mathcal{X}} \eta_{st;jk}\,\phi_{st;jk}(x_s,x_t) - A(\boldsymbol{\eta})\right), \tag{18}$$

The mean parameter space $\mathcal{M}(\mathcal{G})$ is the ***marginal polytope*** over $\mathcal{G}$ since

$$\begin{aligned} \mathcal{M}(\mathcal{G}) &:= \left\{\boldsymbol{\mu} \in \mathbb{R}^d : \exists q \text{ s.t. } \mathbb{E}_q\left[\phi_\alpha(X)\right] = \mu_\alpha \quad \forall\, \alpha \in \mathcal{I}\right\} \\ \text{where } \mu_{st;jk} &= \mathbb{P}_{\boldsymbol{\eta}}(X_s = j \wedge X_t = k), \quad \forall s,t,j,k \\ \mu_{s;j} &= \mathbb{P}_{\boldsymbol{\eta}}(X_s = j), \quad \forall s,j \end{aligned}$$

defines a set of matching constraints on the marginal distribution of p within each factor.

Given an i.i.d. sample $\boldsymbol{X}^{1:n} := \{\boldsymbol{X}^1, \ldots, \boldsymbol{X}^n\}$, the ***empirical mean parameters*** correspond to the **singleton** and **pairwise marginal probabilities**:

$$\hat{\mu}_{s;j} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{X_s^i = j\right\} \text{ and } \hat{\mu}_{st;jk} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\left\{X_s^i = j \wedge X_t^i = k\right\} \tag{19}$$

For this particular exponential family, our assumption that $\hat{\boldsymbol{\mu}} \in \mathcal{M}^\circ$ means that the empirical marginals are all strictly *positive*. Now choose $\hat{\boldsymbol{\eta}}$ as

$$\hat{\eta}_{s,j} = \log \hat{\mu}_{s;j}, \quad \forall s \in \mathcal{V}, j \in \mathcal{X} \tag{20}$$

$$\hat{\eta}_{st;jk} = \log \frac{\hat{\mu}_{st;jk}}{\hat{\eta}_{s,j}\hat{\eta}_{t,k}}, \quad \forall (s,t) \in \mathcal{E}, (j,k) \in \mathcal{X} \times \mathcal{X}. \tag{21}$$

We now claim that $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}_{mle}$ by proving that $\hat{\boldsymbol{\eta}}$ satisfies the moment matching conditions. Substituting (20) and (21) into (18), the **joint distribution under** $\hat{\boldsymbol{\eta}}$ is

$$p(x_1, \ldots, x_m; \hat{\boldsymbol{\eta}}) = \prod_{s\in\mathcal{V}} \hat{\mu}_s(x_s) \prod_{(s,t)\in\mathcal{E}} \frac{\hat{\mu}_{s,t}(x_s,x_t)}{\hat{\mu}_s(x_s)\hat{\mu}_t(x_t)} \tag{22}$$

Note that the log-partition function is $A(\hat{\boldsymbol{\eta}}) = 0$. Moreover, the distribution $p(\boldsymbol{x}; \hat{\boldsymbol{\eta}})$ has its **marginal distributions** as the empirical quantities $\hat{\mu}_s$ and $\hat{\mu}_{s,t}(x_s, x_t)$. To show this, we use an inductive "leaf-stripping" argument since by marginalize over leaf node variable. For example, for leaf node u, it only connects to one edge (u, v), so averaging $p(\boldsymbol{x})$ over x_u produces

$$\begin{aligned}
\sum_{x_u} p(x_1, \ldots, x_m; \hat{\boldsymbol{\eta}}) &= \sum_{x_u} \hat{\mu}_u(x_u) \frac{\hat{\mu}_{u,v}(x_u, x_v)}{\hat{\mu}_u(x_u)\hat{\mu}_v(x_v)} \prod_{s \in \mathcal{V}-\{u\}} \hat{\mu}_s(x_s) \prod_{(s,t)\in\mathcal{E}-\{(u,v)\}} \frac{\hat{\mu}_{s,t}(x_s, x_t)}{\hat{\mu}_s(x_s)\hat{\mu}_t(x_t)} \\
&= \frac{\sum_{x_u} \hat{\mu}_{u,v}(x_u, x_v)}{\hat{\mu}_v(x_v)} \prod_{s \in \mathcal{V}-\{u\}} \hat{\mu}_s(x_s) \prod_{(s,t)\in\mathcal{E}-\{(u,v)\}} \frac{\hat{\mu}_{s,t}(x_s, x_t)}{\hat{\mu}_s(x_s)\hat{\mu}_t(x_t)} \\
&= \prod_{s \in \mathcal{V}-\{u\}} \hat{\mu}_s(x_s) \prod_{(s,t)\in\mathcal{E}-\{(u,v)\}} \frac{\hat{\mu}_{s,t}(x_s, x_t)}{\hat{\mu}_s(x_s)\hat{\mu}_t(x_t)} := p(\boldsymbol{x}_{-u}; \mathcal{T}_{-u})
\end{aligned}$$

And the result is of the same form on a tree $\mathcal{T}_{-u} := (\mathcal{V} - \{u\}, \mathcal{E} - \{(u, v)\})$. Thus we can use induction to marginalize all other variables from leaf to root except for x_s to obtain the result. $\blacksquare$

Thus, we show that **tree-based model** $p(\boldsymbol{x}; \hat{\boldsymbol{\eta}})$ under maximum likelihood estimator $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}_{mle}$ has an explicit closed-form expression (22).

3 Parameter estimation in partially observed models

A more challenging version of parameter estimation arises in the partially observed setting, in which the random vector $\boldsymbol{X} \sim p_{\boldsymbol{\eta}}$ is not observed directly, but *indirectly* via a "noisy" version $\boldsymbol{Y}$ of $\boldsymbol{X}$. The ***expectation-maximization (EM) algorithm*** provides a general approach to computing MLEs in this partially observed setting.

3.1 Exact EM algorithm in exponential families

Suppose that the set of random variables is partitioned into a vector $\boldsymbol{Y}$ of observed variables, and a vector $\boldsymbol{X}$ of unobserved variables, and the probability model is a *joint exponential family* distribution for $(\boldsymbol{X}, \boldsymbol{Y})$:

$$p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\eta}) = \exp\left(\langle \boldsymbol{\eta}\,,\, \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})\rangle - A(\boldsymbol{\eta})\right) \tag{23}$$

Given an observation $\boldsymbol{Y} = \boldsymbol{y}$, we can also form the conditional distribution

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\eta}) &= \frac{p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\eta})}{\int_{\boldsymbol{x}} p(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\eta})\nu(d\boldsymbol{x})} \\ &:= \exp\left(\langle \boldsymbol{\eta}\,,\, \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})\rangle - A_{\boldsymbol{y}}(\boldsymbol{\eta})\right) \end{aligned} \tag{24}$$

where the log-partition for fixed $\boldsymbol{y}$ is given as

$$A_{\boldsymbol{y}}(\boldsymbol{\eta}) = \log \int_{\boldsymbol{x} \in \mathcal{X}^m} \exp\left(\langle \boldsymbol{\eta}\,,\, \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})\rangle\right) \nu(d\boldsymbol{x}) \tag{25}$$

Note $\boldsymbol{\phi}(\cdot, \boldsymbol{y})$ for fixed $\boldsymbol{y}$ is function of latent variables $\boldsymbol{x}$.

The maximum likelihood estimation is for observed likelihood function on $\boldsymbol{Y}$, which is referred to as the ***incomplete log likelihood*** in the setting of EM. This incomplete log likelihood is given by the integral

$$\begin{aligned} \ell(\boldsymbol{\eta}; \boldsymbol{y}) &= \log \int_{\boldsymbol{x} \in \mathcal{X}^m} \exp\left(\langle \boldsymbol{\eta}\,,\, \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})\rangle - A(\boldsymbol{\eta})\right) \nu(d\boldsymbol{x}) \\ &= \underline{A_{\boldsymbol{y}}(\boldsymbol{\eta}) - A(\boldsymbol{\eta})} \end{aligned} \tag{26}$$

The **key** for EM is to obtain the *lower bound* of the incomplete log likelihood function (26). For fixed $\boldsymbol{y}$, the mean parameter space $\mathcal{M}_y$ is

$$\mathcal{M}_{\boldsymbol{y}} = \left\{ \boldsymbol{\mu} \in \mathbb{R}^d : \;\; \mathbb{E}_p\left[\boldsymbol{\phi}(X, \boldsymbol{y})\right] = \boldsymbol{\eta},\; \text{for some } p \right\} \tag{27}$$

where $p \in \Delta$ is any distribution on $\mathcal{X}^m$. From dual representation of A, we can obtain its variational form and its conjugate

$$A_{\boldsymbol{y}}(\boldsymbol{\eta}) = \sup_{\boldsymbol{\mu}_{\boldsymbol{y}} \in \mathcal{M}_{\boldsymbol{y}}} \left\{ \left\langle \boldsymbol{\eta}\,,\, \boldsymbol{\mu}_{\boldsymbol{y}} \right\rangle - A^*_{\boldsymbol{y}}(\boldsymbol{\mu}_{\boldsymbol{y}}) \right\} \tag{28}$$

$$A^*_{\boldsymbol{y}}(\boldsymbol{\mu}_{\boldsymbol{y}}) := \sup_{\boldsymbol{\eta} \in \Omega_{\boldsymbol{y}}} \left\{ \left\langle \boldsymbol{\mu}_{\boldsymbol{y}}\,,\, \boldsymbol{\eta} \right\rangle - A_{\boldsymbol{y}}(\boldsymbol{\eta}) \right\} \tag{29}$$

From weak duality, we can obtain the lower bound of incomplete log-likelihood

$$
\begin{aligned}
A_{\boldsymbol{y}}(\boldsymbol{\eta}) &\geq \langle \boldsymbol{\eta}\,,\, \boldsymbol{\mu}_{\boldsymbol{y}} \rangle - A^{*}_{\boldsymbol{y}}(\boldsymbol{\mu}_{\boldsymbol{y}}) \quad \forall \boldsymbol{\mu}_{\boldsymbol{y}} \in \mathcal{M}_{\boldsymbol{y}} \\
\Rightarrow \ell(\boldsymbol{\eta}; \boldsymbol{y}) = A_{\boldsymbol{y}}(\boldsymbol{\eta}) - A(\boldsymbol{\eta}) &\geq \underline{\langle \boldsymbol{\eta}\,,\, \boldsymbol{\mu}_{\boldsymbol{y}} \rangle - A^{*}_{\boldsymbol{y}}(\boldsymbol{\mu}_{\boldsymbol{y}}) - A(\boldsymbol{\eta}) := \mathcal{L}(\boldsymbol{\eta}, \boldsymbol{\mu}_{\boldsymbol{y}})}
\end{aligned} \tag{30}
$$

The *expectation-maximization (EM) algorithm* is a <u>**coordinate ascent algorithm**</u> that *maximize* the lower bound $\mathcal{L}(\boldsymbol{\eta}, \boldsymbol{\mu}_{\boldsymbol{y}})$:

$$
\textbf{E Step:} \quad \boldsymbol{\mu}_{\boldsymbol{y}}^{(t+1)} := \arg \max_{\boldsymbol{\mu}_{\boldsymbol{y}} \in \mathcal{M}_{\boldsymbol{y}}} \mathcal{L}(\boldsymbol{\eta}^{(t)}, \boldsymbol{\mu}_{\boldsymbol{y}}) \tag{31}
$$

$$
\textbf{M Step:} \quad \boldsymbol{\eta}^{(t+1)} := \arg \max_{\boldsymbol{\eta} \in \Omega_{\boldsymbol{y}}} \mathcal{L}(\boldsymbol{\eta}, \boldsymbol{\mu}_{\boldsymbol{y}}^{(t+1)}) \tag{32}
$$

To see why this is called EM algorithm, at E-step (31), the optimization becomes

$$
\max_{\boldsymbol{\mu}_{\boldsymbol{y}} \in \mathcal{M}_{\boldsymbol{y}}} \left\langle \boldsymbol{\eta}^{(t)}\,,\, \boldsymbol{\mu}_{\boldsymbol{y}} \right\rangle - A^{*}_{\boldsymbol{y}}(\boldsymbol{\mu}_{\boldsymbol{y}}) = A_{\boldsymbol{y}}(\boldsymbol{\eta}^{(t)})
$$

and the optimal solution is $\boldsymbol{\mu}_{\boldsymbol{y}}^{(t+1)} = \mathbb{E}_{\boldsymbol{\eta}^{(t)}}\left[\boldsymbol{\phi}(X, \boldsymbol{y})\right]$, which is exactly the expectation step of original EM. On the other hand, at M-step, the maximization is

$$
\max_{\boldsymbol{\eta} \in \Omega_{\boldsymbol{y}}} \left\langle \boldsymbol{\mu}_{\boldsymbol{y}}^{(t+1)}\,,\, \boldsymbol{\eta} \right\rangle - A(\boldsymbol{\eta}),
$$

which is a maximum log-likelihood estimation on the joint distribution using expected sufficient statistics $\boldsymbol{\mu}_{\boldsymbol{y}}^{(t+1)}$. Moreover, given that $\boldsymbol{\mu}_{\boldsymbol{y}}^{(t+1)} = \mathbb{E}_{\boldsymbol{\eta}^{(t)}}\left[\boldsymbol{\phi}(X, \boldsymbol{y})\right]$ is the optimal solution and the corresponding optimal value in E-step is exactly $A_{\boldsymbol{y}}(\boldsymbol{\eta}^{(t)})$, the equality is met and $\ell(\boldsymbol{\eta}^{(t)}; \boldsymbol{y}) = \mathcal{L}(\boldsymbol{\eta}^{(t)}, \boldsymbol{\mu}_{\boldsymbol{y}}^{(t+1)})$ at the end of E-step. Then the subsequent maximization of $\mathcal{L}$ with respect to $\boldsymbol{\eta}$ in the M-step is **guaranteed to *increase*** the log likelihood as well.

3.2 Variational EM

The main difficulty in EM is to compute t*he expected sufficient statistics* $\boldsymbol{\mu}_{\boldsymbol{y}}^{(t+1)} = \mathbb{E}_{\boldsymbol{\eta}^{(t)}}\left[\boldsymbol{\phi}(X, \boldsymbol{y})\right] \in \mathcal{M}_{\boldsymbol{y}}$ in the E-step, esp. when $\mathcal{M}_{\boldsymbol{y}}$ is complicated. An *alterative* solution is to reduce the search space $\mathcal{M}_{\boldsymbol{y}}$ to be within the space of tractable distribution $\mathcal{M}_{\mathcal{F}}(\mathcal{G}) \subseteq \mathcal{M}_{\boldsymbol{y}}$, via **mean field approximation**.

The <u>***variational EM***</u> via **mean field approximation** is

$$
\begin{aligned}
\textbf{Mean field E Step:} \quad & \boldsymbol{\mu}_{\boldsymbol{y}}^{(t+1)} := \arg \max_{\boldsymbol{\mu}_{\boldsymbol{y}} \in \mathcal{M}_{\mathcal{F}}(\mathcal{G})} \mathcal{L}(\boldsymbol{\eta}^{(t)}, \boldsymbol{\mu}_{\boldsymbol{y}}) \\
\textbf{M Step:} \quad & \boldsymbol{\eta}^{(t+1)} := \arg \max_{\boldsymbol{\eta} \in \Omega_{\boldsymbol{y}}} \mathcal{L}(\boldsymbol{\eta}, \boldsymbol{\mu}_{\boldsymbol{y}}^{(t+1)})
\end{aligned} \tag{33}
$$

which replace the E-step by replacing the exact mean parameter $\mathbb{E}_{\boldsymbol{\eta}^{(t)}}\left[\boldsymbol{\phi}(X, \boldsymbol{y})\right]$, under the current model $\boldsymbol{\eta}^{(t)}$, with the ***approximate*** set of mean parameters computed by a mean field algorithm.

The variational EM with mean field approximation is still a *coordinate ascent algorithm.* That is, it is guaranteed to <u>***maximize the lower bound***</u> $\mathcal{L}(\boldsymbol{\eta}, \boldsymbol{\mu}_{\boldsymbol{y}})$. However, because the E-step no longer closes the gap between incomplete likelihood function and the lower bound, it is **no longer** the case that the algorithm necessarily **goes uphill** in the latter quantity. Note that mean field approximation provides lower bound because $\mathcal{M}_{\mathcal{F}}(\mathcal{G}) \subseteq \mathcal{M}_{\boldsymbol{y}}$ is the inner bound of the orignal space. This is the reason why Mean field E-step can guarantee to improve the lower bound. Other approximation may not enjoy this property.

3.3 Variational Bayes

In the literature on the topic, the term "***variational Bayes***" has been reserved thus far for the application of the *mean-field variational method* to Bayesian inference. Let the data be partitioned into an observed component $\boldsymbol{Y}$ and an unobserved component $\boldsymbol{X}$, and assume that the complete data likelihood lies in some exponential family

$$p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\eta}) = \exp\left\{\langle \boldsymbol{\zeta}(\boldsymbol{\eta}) \, , \, \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y})\rangle - A(\boldsymbol{\zeta}(\boldsymbol{\eta}))\right\} \tag{34}$$

The function $\boldsymbol{\zeta} : \mathbb{R}^d \to \mathbb{R}^d$ provides some additional flexibility in the *parameterization* of the exponential family. We assume that the prior distribution over $\boldsymbol{\eta} \in H$ also lies in some exponential family, of the ***conjugate prior form***:

$$p(\boldsymbol{\eta}; \boldsymbol{\xi}, \lambda) = \exp\left\{\langle \boldsymbol{\xi} \, , \, \boldsymbol{\zeta}(\boldsymbol{\eta})\rangle - \lambda A(\boldsymbol{\zeta}(\boldsymbol{\eta})) - B(\boldsymbol{\xi}, \lambda)\right\} \tag{35}$$

Note that this exponential family is specified by the sufficient statistics $\{\boldsymbol{\zeta}(\boldsymbol{\eta}), -A(\boldsymbol{\zeta}(\boldsymbol{\eta}))\} \in \mathbb{R}^d \times \mathbb{R}$, with associated canonical parameters $(\boldsymbol{\xi}, \lambda) \in \mathbb{R}^d \times \mathbb{R}$. The log-partition function of prior B is defined as

$$B(\boldsymbol{\xi}, \lambda) = \log \int \exp\left\{\langle \boldsymbol{\xi} \, , \, \boldsymbol{\zeta}(\boldsymbol{\eta})\rangle - \lambda A(\boldsymbol{\zeta}(\boldsymbol{\eta}))\right\} d\boldsymbol{\eta}$$

The main task is to compute the ***marginalized log-likelihood function*** $\log p_{\boldsymbol{\xi}^*, \lambda^*}(\boldsymbol{y})$ where $(\boldsymbol{\xi}^*, \lambda^*)$ are hyperparameters on the prior.

$$\begin{aligned} \log p_{\boldsymbol{\xi}^*, \lambda^*}(\boldsymbol{y}) &:= \log \int \left[\int p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\eta}) d\boldsymbol{x}\right] p(\boldsymbol{\eta}; \boldsymbol{\xi}^*, \lambda^*) d\boldsymbol{\eta} \\ &= \log \int p(\boldsymbol{y}|\boldsymbol{\eta}) p(\boldsymbol{\eta}; \boldsymbol{\xi}^*, \lambda^*) d\boldsymbol{\eta} \\ &= \log \int p(\boldsymbol{y}|\boldsymbol{\eta}) p(\boldsymbol{\eta}; \boldsymbol{\xi}, \lambda) \frac{p(\boldsymbol{\eta}; \boldsymbol{\xi}^*, \lambda^*)}{p(\boldsymbol{\eta}; \boldsymbol{\xi}, \lambda)} d\boldsymbol{\eta} \end{aligned} \tag{36}$$

Recall Jenson's inequality: if X is a random variable and ϕ is a convex function, then

$$\phi(\mathbb{E}\left[X\right]) \leq \mathbb{E}\left[\phi(X)\right].$$

Since $-\log(x)$ is convex function, so

$$\begin{aligned} \log p_{\boldsymbol{\xi}^*, \lambda^*}(\boldsymbol{y}) &= \log\left(\mathbb{E}_{\boldsymbol{\xi}, \lambda}\left[p(\boldsymbol{y}|\boldsymbol{\eta}) \frac{p(\boldsymbol{\eta}; \boldsymbol{\xi}^*, \lambda^*)}{p(\boldsymbol{\eta}; \boldsymbol{\xi}, \lambda)}\right]\right) \\ &\geq \mathbb{E}_{\boldsymbol{\xi}, \lambda}\left[\log\left(p(\boldsymbol{y}|\boldsymbol{\eta}) \frac{p(\boldsymbol{\eta}; \boldsymbol{\xi}^*, \lambda^*)}{p(\boldsymbol{\eta}; \boldsymbol{\xi}, \lambda)}\right)\right] \\ &= \mathbb{E}_{\boldsymbol{\xi}, \lambda}\left[\log p(\boldsymbol{y}|\boldsymbol{\eta})\right] + \mathbb{E}_{\boldsymbol{\xi}, \lambda}\left[\frac{p(\boldsymbol{\eta}; \boldsymbol{\xi}^*, \lambda^*)}{p(\boldsymbol{\eta}; \boldsymbol{\xi}, \lambda)}\right] \end{aligned} \tag{37}$$

with equality for $(\boldsymbol{\xi}, \lambda) = (\boldsymbol{\xi}^*, \lambda^*)$. Recall from that from (26),

$$\log p(\boldsymbol{y}|\boldsymbol{\eta}) = A_{\boldsymbol{y}}(\boldsymbol{\eta}) - A(\boldsymbol{\eta})$$

The inequality (37) becomes

$$\log p_{\boldsymbol{\xi}^*,\lambda^*}(\boldsymbol{y}) \geq \mathbb{E}_{\boldsymbol{\xi},\lambda}\left[A_{\boldsymbol{y}}(\boldsymbol{\zeta}(\boldsymbol{\eta})) - A(\boldsymbol{\zeta}(\boldsymbol{\eta}))\right] + \mathbb{E}_{\boldsymbol{\xi},\lambda}\left[\frac{p(\boldsymbol{\eta};\boldsymbol{\xi}^*,\lambda^*)}{p(\boldsymbol{\eta};\boldsymbol{\xi},\lambda)}\right] \tag{38}$$

where $A_{\boldsymbol{y}}(\boldsymbol{\zeta}(\boldsymbol{\eta}))$ is the log-partition function of *condition distribution* $p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\eta})$ as (25).

The inequality (38) provides a *lower bound* of objective function. For every fixed $\boldsymbol{y}$, and each *realization* of $\boldsymbol{\eta} \in H$, we can obtain the mean parameter $\boldsymbol{\mu}(\boldsymbol{\eta}) = \mathbb{E}_{\boldsymbol{\eta}}\left[\boldsymbol{\phi}(X,\boldsymbol{y})|\boldsymbol{\eta}\right]$. Thus by the weak duality on **variational representation** (9) of $A_{\boldsymbol{y}}(\boldsymbol{\zeta}(\boldsymbol{\eta}))$ for any $\boldsymbol{\mu}(\boldsymbol{\eta}) \in \mathcal{M}_{\boldsymbol{y}}$ we can find *a further lower bound.*

$$\log p_{\boldsymbol{\xi}^*,\lambda^*}(\boldsymbol{y}) \geq \mathbb{E}_{\boldsymbol{\xi},\lambda}\left[\langle \boldsymbol{\mu}(\boldsymbol{\eta})\,,\, \boldsymbol{\zeta}(\boldsymbol{\eta})\rangle - A^*_{\boldsymbol{y}}(\boldsymbol{\mu}(\boldsymbol{\eta})) - A(\boldsymbol{\zeta}(\boldsymbol{\eta}))\right] + \mathbb{E}_{\boldsymbol{\xi},\lambda}\left[\frac{p(\boldsymbol{\eta};\boldsymbol{\xi}^*,\lambda^*)}{p(\boldsymbol{\eta};\boldsymbol{\xi},\lambda)}\right] \tag{39}$$

The **variational Bayes algorithm** is based on optimizing this lower bound (39) using only *product distributions* over the pair $(X,\boldsymbol{\eta})$, i.e. **mean field assumption**. Note that if we can generate $(X,\boldsymbol{\eta})$ from original joint distribution, the lower bound (39) is *tight* and it would equal to $\log p_{\boldsymbol{\xi}^*,\lambda^*}(\boldsymbol{y})$. Compared to (30), (39) add additional term on prior variations. Such optimization is often described as "***free-form***", in that beyond the assumption of a product distribution, the factors composing this product distribution are allowed to be arbitrary.

We now derive the variational Bayes algorithm as *coordinate ascent* over (39) under mean field product distributions. Denote $\overline{A} := \mathbb{E}_{\boldsymbol{\xi},\lambda}\left[A(\boldsymbol{\zeta}(\boldsymbol{\eta}))\right]$ and $\overline{\boldsymbol{\zeta}} := \mathbb{E}_{\boldsymbol{\xi},\lambda}\left[\boldsymbol{\zeta}(\boldsymbol{\eta})\right]$. Since, under mean field assumption, $\boldsymbol{\mu}$ is independent of $\boldsymbol{\eta}$, the optimization problem (39) can be simplified to

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi},\lambda}\left[\langle \boldsymbol{\mu}\,,\, \boldsymbol{\zeta}(\boldsymbol{\eta})\rangle - A^*_{\boldsymbol{y}}(\boldsymbol{\mu}) - A(\boldsymbol{\zeta}(\boldsymbol{\eta}))\right] + \mathbb{E}_{\boldsymbol{\xi},\lambda}\left[\frac{p(\boldsymbol{\eta};\boldsymbol{\xi}^*,\lambda^*)}{p(\boldsymbol{\eta};\boldsymbol{\xi},\lambda)}\right] \\
&= \left\langle \boldsymbol{\mu}\,,\, \overline{\boldsymbol{\zeta}}\right\rangle - A^*_{\boldsymbol{y}}(\boldsymbol{\mu}) - \overline{A} + \mathbb{E}_{\boldsymbol{\xi},\lambda}\left[\frac{p(\boldsymbol{\eta};\boldsymbol{\xi}^*,\lambda^*)}{p(\boldsymbol{\eta};\boldsymbol{\xi},\lambda)}\right]
\end{aligned} \tag{40}$$

Using the exponential form (35) of conjugate prior $p(\boldsymbol{\eta};\boldsymbol{\xi},\lambda)$, we have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi},\lambda}\left[\frac{p(\boldsymbol{\eta};\boldsymbol{\xi}^*,\lambda^*)}{p(\boldsymbol{\eta};\boldsymbol{\xi},\lambda)}\right] \\
&= \left\langle \overline{\boldsymbol{\zeta}}\,,\, \boldsymbol{\xi}^* - \boldsymbol{\xi}\right\rangle + \left\langle -\overline{A}\,,\, \lambda^* - \lambda\right\rangle - B(\boldsymbol{\xi}^*,\lambda^*) + B(\boldsymbol{\xi},\lambda)
\end{aligned} \tag{41}$$

Recall that B is log-partition function of exponential family prior, and that $-\overline{A} := \mathbb{E}_{\boldsymbol{\xi},\lambda}\left[-A(\boldsymbol{\zeta}(\boldsymbol{\eta}))\right]$ and $\overline{\boldsymbol{\zeta}} := \mathbb{E}_{\boldsymbol{\xi},\lambda}\left[\boldsymbol{\zeta}(\boldsymbol{\eta})\right]$ are the mean parameters of prior $p(\boldsymbol{\eta};\boldsymbol{\xi},\lambda)$. By the conjugate B^* can be written as

$$B^*(\overline{\boldsymbol{\zeta}}, -\overline{A}) = \left\langle \overline{\boldsymbol{\zeta}}\,,\, \boldsymbol{\xi}\right\rangle + \left\langle -\overline{A}\,,\, \lambda\right\rangle - B(\boldsymbol{\xi},\lambda) \tag{42}$$

Substituting (41) and (42) into (40), we have the **objective function** as

$$\begin{aligned}
&\left\langle \boldsymbol{\mu}\,,\, \overline{\boldsymbol{\zeta}}\right\rangle - A^*_{\boldsymbol{y}}(\boldsymbol{\mu}) - \overline{A} + \left\langle \overline{\boldsymbol{\zeta}}\,,\, \boldsymbol{\xi}^* - \boldsymbol{\xi}\right\rangle + \left\langle -\overline{A}\,,\, \lambda^* - \lambda\right\rangle - B(\boldsymbol{\xi}^*,\lambda^*) + B(\boldsymbol{\xi},\lambda) \\
&= \left\langle \boldsymbol{\mu} + \boldsymbol{\xi}^*\,,\, \overline{\boldsymbol{\zeta}}\right\rangle - A^*_{\boldsymbol{y}}(\boldsymbol{\mu}) + \left\langle \lambda^* + 1\,,\, -\overline{A}\right\rangle - B^*(\overline{\boldsymbol{\zeta}}, -\overline{A}) := \mathcal{L}(\boldsymbol{\mu}, \overline{\boldsymbol{\zeta}}, -\overline{A})
\end{aligned} \tag{43}$$

over $\boldsymbol{\mu} \in \mathcal{M}_{\boldsymbol{y}}$ and $(\overline{\boldsymbol{\zeta}}, -\overline{A}) \in \Omega_B$.

Finally, we have the **variational Bayes algorithm**:

$$\textbf{VB-E Step}: \quad \boldsymbol{\mu}^{(t+1)} := \arg\max_{\boldsymbol{\mu}\in\mathcal{M}_{\boldsymbol{y}}} \mathcal{L}(\boldsymbol{\mu}, \overline{\boldsymbol{\zeta}}^{(t)}, -\overline{A}^{(t)}) \tag{44}$$

$$\textbf{VB-M Step}: \quad (\overline{\boldsymbol{\zeta}}^{(t+1)}, -\overline{A}^{(t+1)}) := \arg\max_{(\overline{\boldsymbol{\zeta}}, -\overline{A})\in\Omega_B} \mathcal{L}(\boldsymbol{\mu}^{(t+1)}, \overline{\boldsymbol{\zeta}}, -\overline{A}) \tag{45}$$

We can further break down it. In **E-step**, the optimization is

$$\boldsymbol{\mu}^{(t+1)} := \arg\max_{\boldsymbol{\mu}\in\mathcal{M}_{\boldsymbol{y}}} \langle \boldsymbol{\mu}\,,\, \overline{\boldsymbol{\zeta}} \rangle - A^*_{\boldsymbol{y}}(\boldsymbol{\mu}) \tag{46}$$

Like EM, the optimal solution for this problem satisfies the moment matching condition

$$\boldsymbol{\mu}^{(t+1)} = \mathbb{E}_{\overline{\boldsymbol{\zeta}}^{(t)}} \left[\boldsymbol{\phi}(X, \boldsymbol{y}) | \overline{\boldsymbol{\zeta}}^{(t)} \right] \tag{47}$$

In the M-step, we have update the hyperparameters $(\boldsymbol{\xi}, \lambda)$ as

$$(\boldsymbol{\xi}^{(t+1)}, \lambda^{(t+1)}) = (\boldsymbol{\xi}^* + \boldsymbol{\mu}^{(t+1)}, \lambda^* + 1) \tag{48}$$

Then the new mean on parameters are

$$\overline{\boldsymbol{\zeta}}^{(t+1)} = \mathbb{E}_{\boldsymbol{\xi}^{(t+1)}, \lambda^{(t+1)}} \left[\boldsymbol{\zeta}(\boldsymbol{\eta}) \right] \tag{49}$$

From (47) and (49), we obtain the simplified form of **variational Bayes algorithm**:

$$\begin{aligned} \textbf{VB-E Step}: \quad \boldsymbol{\mu}^{(t+1)} &= \mathbb{E}_{\overline{\boldsymbol{\zeta}}^{(t)}} \left[\boldsymbol{\phi}(X, \boldsymbol{y}) | \overline{\boldsymbol{\zeta}}^{(t)} \right] \\ &= \int \boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{y}) \; p(\boldsymbol{x} | \boldsymbol{y}, \overline{\boldsymbol{\zeta}}^{(t)}) d\boldsymbol{x}, \end{aligned} \tag{50}$$

$$\begin{aligned} \textbf{VB-M Step}: \quad \overline{\boldsymbol{\zeta}}^{(t+1)} &= \mathbb{E}_{\boldsymbol{\xi}^{(t+1)}, \lambda^{(t+1)}} \left[\boldsymbol{\zeta}(\boldsymbol{\eta}) \right] \\ &= \int \boldsymbol{\zeta}(\boldsymbol{\eta}) \; p(\boldsymbol{\eta}; \boldsymbol{\xi}^{(t+1)}, \lambda^{(t+1)}) d\boldsymbol{\eta} \\ &\text{where } (\boldsymbol{\xi}^{(t+1)}, \lambda^{(t+1)}) = (\boldsymbol{\xi}^* + \boldsymbol{\mu}^{(t+1)}, \lambda^* + 1) \end{aligned} \tag{51}$$

References

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques.* MIT press, 2009.

Martin J Wainwright, Michael I Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.