

Introduction	
Variational principle for probabilistic learning	
Yet another justification	
More simplification of updates for mean-field family	
Examples	$y_1(\mathbf{x})$
Dirichlet Process Mixture	
On minimization of divergence measures	
Energy minimization justifications	
Variational learning with exponential family	
Mean parametrization and marginal polytopes	
Convex dualities	
The log-partition function and conjugate duality	
Belief Propagation vs. $y_P(\mathbf{x})$ Mean-field approximation	
Bibliographical notes	

# 1 — Variational Principle

Daniel Khashabi <sup>1</sup>  
 KHASHAB2@ILLINOIS.EDU

## 1.1 Introduction

The name “Variational” is a general name to call a very broad range of optimization-based formulation of problems, which are mostly inspired from “calculus of variations”. Thus one should NOT look at this method, as a black-box algorithm; instead it includes a very useful techniques for simplifying very broad range of problems into tractable optimization problems.

■ **Example 1.1 — Variational representation for solving linear systems. Problem:** Let’s say we are given a vector  $\mathbf{y} \in \mathbb{R}^n$  and a positive-definite matrix  $\mathbf{A} \in \mathbb{R}^{n \times n}$ , and we want to solve the given linear system  $\mathbf{A}\mathbf{x} = \mathbf{y}$ .

**Direct solution:** The direct solution could be found from the matrix inversion:

$$\mathbf{x}^* = \mathbf{A}^{-1}\mathbf{y}$$

**Equivalent variational solution:** Assume the following cost function:

$$J_{\mathbf{y}}(\mathbf{x}) = \frac{1}{2}\mathbf{x}^{\top}\mathbf{A}\mathbf{x} - \mathbf{y}^{\top}\mathbf{x}.$$

The notation  $J_{\mathbf{y}}(\mathbf{u})$  means that this cost function is a parametric form of the observation  $\mathbf{y}$  and the variable  $\mathbf{u}$ . We can show that this cost function is strictly convex and the minimum fix-point equals to  $\mathbf{x}^* = \arg \min_{\mathbf{x}} J_{\mathbf{y}}(\mathbf{u}) = \mathbf{A}^{-1}\mathbf{y}$ .

The above two solutions (direct and variational) are both equivalent. This

<sup>1</sup>This is part of my notes; to find the complete list of notes visit <http://web.engr.illinois.edu/~khashab2/learn.html>. This work is licensed under a Creative Commons Attribution-NonCommercial 3.0 License. This document is updated on July 7, 2014.

example shows how the prevalent problem of matrix inverse and linear systems can be posed as an optimization problem. ■

Not any variational problem has a unique answer. See the next example for an instance.

■ **Example 1.2 — Variational problems without unique answers.** The following question doesn't have answer.

$$\min_f \int_0^1 \left[ f + \left( \left( \frac{\partial f}{\partial x} \right) - 1 \right) \right] dx$$

The following question has infinite number of answers:

$$\min_f \int_0^1 \left[ \frac{1}{\sqrt{1 + f'^2}} - \frac{1}{\sqrt{2}} \right] dx$$

■ **Example 1.3 — The shortest distance between two points in a line!.** ■

## 1.2 Variational principle for probabilistic learning

Variational is a very useful approach to solving problems with intractable posterior distribution via some approximations. The idea is to replace the *intractable* posterior with a *tractable* approximation. Here I am going to symbolically introduce the variational principle for learning probabilistic models. First assume the simple graphical model in Figure 1.1, in which we assume  $\mathbf{X} \in \mathbb{R}^n$  and  $\mathbf{Z} \in \mathbb{R}^m$  which are two random variables. Our observations is  $\mathbf{X}$ , and using that, we want to learn and infer about the latent variable  $\mathbf{Z}$ . Let's also assume that we have a set of parameters  $\theta$ . Note that, this simple graphical model is just a symbolic representation of a our main big graphical, and in practice we might have a set of observations and latent variables. In probabilistic way we can show this inference as posterior on latent variables, conditioned on observations  $p(\mathbf{Z}|\mathbf{X}, \theta)$ . We assume that the main model is complicated that this posterior is intractable:

$$p(\mathbf{X}, \mathbf{Z}|\theta) = p(\mathbf{Z}|\mathbf{X}, \theta)p(\mathbf{X}|\theta). \quad (1.1)$$

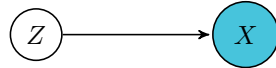



Figure 1.1: Symbolic connection between observation and latent variables. The blue variable is observation.

In variational learning we are going to approximate the real posterior distribution with a another distribution. Generally one can get an approximation to another distribution, by assuming some independence assumptions which simplify it into multiplication of several simpler distributions which are easier to

work with. For the sake of emphasis on *approximation*, it is common to denote this distribution with  $q(\cdot)$  instead of  $p(\cdot)$ :

$$q(\mathbf{Z}) = \prod_{i=1}^m q(Z_i) \quad (1.2)$$

There is not general rule for decomposition of variables into disjoint distributions; for any problem one should consider the interaction between variables and possibility of realistic decompositions in the distribution. Note that in practice the approximate distributions  $q(\mathbf{z})$  is a parametric form of  $\omega$  (variational parameters) which characterize this distribution; so it is more accurate to denote it by  $q_{\mathbf{z}}(\mathbf{z}; \omega)$ , in which  $\omega$  denote set of variational parameters.

 Sometimes this method is called *meanfield* variational method; this name is because of decomposition of  $q(\cdot)$ . Note that even without decomposition assumptions here, one could follow all of the following derivations. When using the final results in updating, decomposition will help us to find easier update rules.

To approximate the true posterior  $p(\cdot)$  using the decomposed distribution  $q(\cdot)$  we should find a measure of distance between two functions, and also is practical in computational sense<sup>2</sup>.

■ **Lemma 1.1** We have the following fact,

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q_{\mathbf{Z}}) + \text{KL}(q_{\mathbf{Z}}||p_{\mathbf{Z}|\mathbf{X}}). \quad (1.3)$$

In which we defined the following two notations; the first one is KL-*divergence*,

$$\text{KL}(q_{\mathbf{Z}}||p_{\mathbf{Z}|\mathbf{X}}) \triangleq \mathbb{E}_q \left[ \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})} \right] \quad (1.4)$$

And the second one is the lower bound on the likelihood (to be shown)

$$\mathcal{L}(q, \boldsymbol{\theta}) \triangleq \int_{\mathbb{R}^m} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} \quad (1.5)$$

*Proof.* Now using equation 1.1 we have,

$$\begin{aligned} \log p(\mathbf{X}|\boldsymbol{\theta}) &= \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) - \log p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \\ \log p(\mathbf{X}|\boldsymbol{\theta}) &= \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} - \log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})}. \end{aligned}$$

Multiplying two sides in  $q(\mathbf{Z})$  we have,

$$\begin{aligned} q(\mathbf{Z}) \log p(\mathbf{X}|\boldsymbol{\theta}) &= q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} - q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \\ \int_{\mathbb{R}^m} q(\mathbf{Z}) \log p(\mathbf{X}|\boldsymbol{\theta}) d\mathbf{Z} &= \int_{\mathbb{R}^m} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} - \int_{\mathbb{R}^m} q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z}. \end{aligned}$$

<sup>2</sup>In practical sense, this is more important! This is the reason for long domination of the bloodthirsty quadratic error!!

Note that in the left part of the above equation  $p(\mathbf{X}|\boldsymbol{\theta})$  is not a function of  $\mathbf{z}$  and thus,  $\int_{\mathbb{R}^m} q(\mathbf{Z}) \log p_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\theta}) d\mathbf{Z} = \log p(\mathbf{X}|\boldsymbol{\theta})$ . Note that,

$$\text{KL}(q_{\mathbf{Z}}||p_{\mathbf{Z}|\mathbf{X}}) \triangleq - \int_{\mathbb{R}^m} q(\mathbf{Z}) \log \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} \quad (1.6)$$

$$= \int_{\mathbb{R}^m} q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})} d\mathbf{Z} \quad (1.7)$$

$$= \mathbb{E}_q \left[ \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})} \right] \quad (1.8)$$

Putting the results together we have,

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q_{\mathbf{Z}}) + \text{KL}(q_{\mathbf{Z}}||p_{\mathbf{Z}|\mathbf{X}}). \quad (1.9)$$

■

The equation (1.3) is shown in Figure (1.2). Now in equation 1.3 note that left

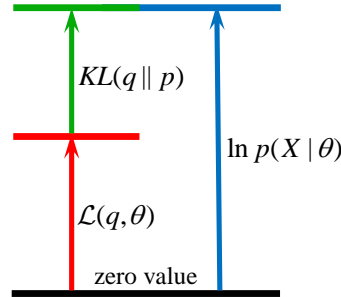


Figure 1.2: Representation of equation 1.3.

side of the equation is not a function of  $q(\cdot)$ . The training consists of increasing the overall likelihood we are doing a two step procedure similar to EM algorithm:

1. Initialize  $\theta$ , and variational parameters of  $q(\cdot)$ .
2. Repeat until convergence
  - (a) E-step:

$$q^{(t+1)} = \arg \max_q \mathcal{L}(q^{(t)}, \theta^{(t)}) = \arg \min_q \text{KL}(q_{\mathbf{Z}}||p_{\mathbf{Z}|\mathbf{X}}) \quad (1.10)$$

- (b) M-step:

$$\theta^{(t+1)} = \arg \max_{\theta} \mathcal{L}(q^{(t+1)}, \theta^{(t)}) \quad (1.11)$$

Visualization of these two steps is shown in Figure 1.3.

**R** Basically we are doing *coordinate ascent* optimization, i.e. given one multivariable objective function, fix all of the variables but one, and maximize with respect to that variable.

### 1.3 Yet another justification

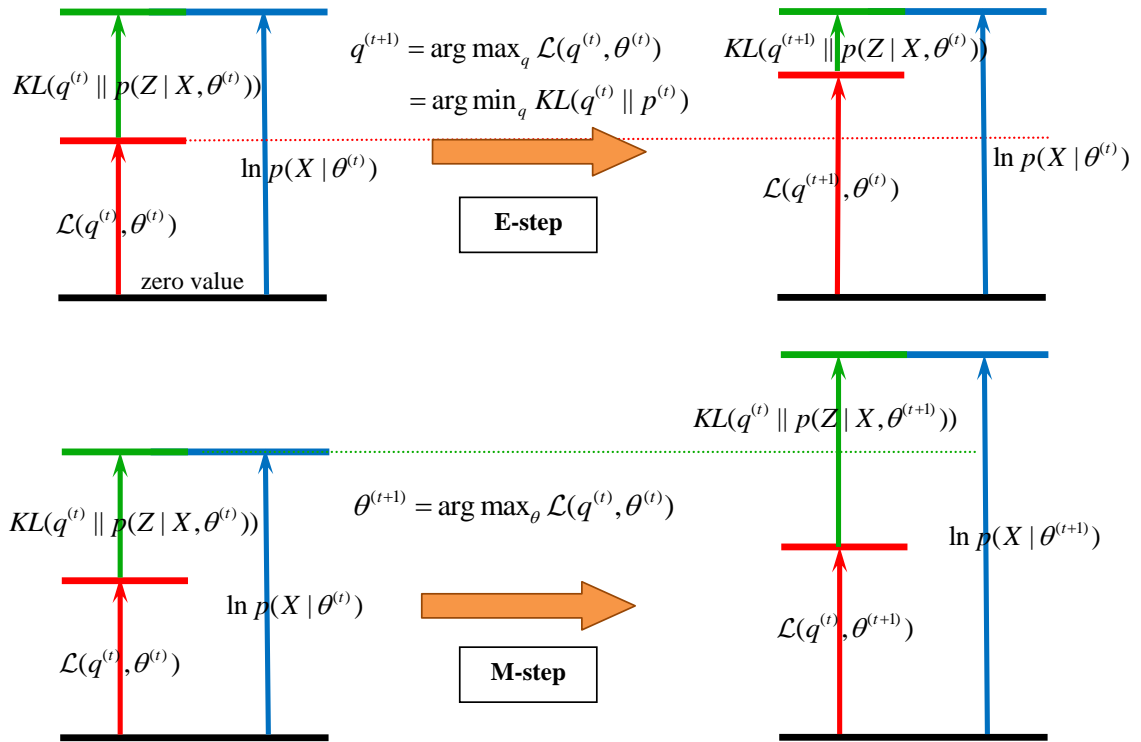


Figure 1.3: Steps in variational learning; the above part is E-step and the other is M-step.

■ **Lemma 1.2** The  $\mathcal{L}(q, \theta)$  is a lower bound on the likelihood distribution. In other words,

$$p(\mathbf{X}|\theta) \geq \exp [\mathcal{L}(q, \theta)]$$

*Proof.* I use Jensen's inequality here to show that  $\mathcal{L}(q, \theta)$  a lower bound for the original likelihood:

$$\begin{aligned} \ln p(\mathbf{X}|\theta) &= \ln \int_{\mathbb{R}^m} p(\mathbf{X}, \mathbf{Z}|\theta) d\mathbf{Z} \\ &= \ln \int_{\mathbb{R}^m} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} d\mathbf{Z} \\ &\geq \int_{\mathbb{R}^m} q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} d\mathbf{Z} = \mathcal{L}(q, \theta) \end{aligned}$$

This shows that  $\exp [\mathcal{L}(q, \theta)]$  a lower bound for likelihood  $p(\mathbf{X}|\theta)$ :

$$\Rightarrow p(\mathbf{X}|\theta) \geq \exp [\mathcal{L}(q, \theta)]$$

■

**Corollary 1.1** Maximizing  $\mathcal{L}(q, \theta)$  (lower bound) will result in maximizing the likelihood  $p(\mathbf{X}|\theta)$ . For this reason,  $\mathcal{L}(q, \theta)$  is sometimes called **ELBO** (Evi-

dence Lower Bound).

Another useful point in the above result is that, one could use the lower bound on likelihood as a good measure of convergence, instead of main likelihood, in case it has a complicated form.

**R** However through lots of examples people have shown effectiveness of maximizing lower bound on the likelihood (instead of maximizing likelihood itself), indeed it is a serious question that when this argument is true and when is not. This topic of some ongoing research in community.

**Proposition 1.1 — Convexity of the lower bound.** The ELBO bound is convex with respect to each of the  $q(Z_i)$  (proof?).

## 1.4 More simplification of updates for mean-field family

Let's consider the equation 1.5,

$$\begin{aligned}\mathcal{L}(q, \boldsymbol{\theta}) &= \int_{\mathbb{R}^m} q(\mathbf{Z}) \log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} d\mathbf{Z} \\ &= \int_{\mathbb{R}^m} q(\mathbf{Z}) \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) d\mathbf{Z} - \int_{\mathbb{R}^m} q(\mathbf{Z}) \log q(\mathbf{Z}) d\mathbf{Z}\end{aligned}$$

**R** The second term in the above final relation is *information entropy* of  $q(\mathbf{Z})$ .

$$\begin{aligned}\mathcal{L}(q, \boldsymbol{\theta}) &= \mathbb{E}_q [\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) - \log q(\mathbf{Z})] \\ &= \mathbb{E}_q \left[ \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) - \log \prod_{k=1}^m q(Z_k) \right] \\ &= \mathbb{E}_q [\log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})] - \sum_{k=1}^m \mathbb{E}_q [\log q(Z_k)]\end{aligned}$$

Now let's say we want to optimize the above expression with respect to one term in approximate posterior, say  $q(Z_j)$ ; thus, we try to take the corresponding term

out of the whole equation,

$$\begin{aligned}
\mathcal{L}(q, \boldsymbol{\theta}) &= \mathbb{E}_{q(\mathbf{Z})} [\log p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta})] - \sum_{k=1}^m \mathbb{E}_{q(Z_k)} [\log q(Z_k)] \\
&= \mathbb{E}_{q(Z_j)} \left[ \mathbb{E}_{\prod_{i=1, i \neq j}^m q(Z_i)} [\log p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta})] \right] - \sum_{k=1}^m \mathbb{E}_{q(Z_k)} [\log q(Z_k)] \\
&= \mathbb{E}_{q(Z_j)} \left[ \mathbb{E}_{\prod_{i=1, i \neq j}^m q(Z_i)} [\log p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta})] \right] - \mathbb{E}_{q(Z_j)} [\log q(Z_j)] - \sum_{\substack{k=1 \\ k \neq j}}^m \mathbb{E}_{q(Z_k)} [\log q(Z_k)] \\
&= \mathbb{E}_{q(Z_j)} \left[ \log \left( \exp \left\{ \mathbb{E}_{\prod_{i=1, i \neq j}^m q(Z_i)} [\log p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta})] \right\} \right) \right] - \mathbb{E}_{q(Z_j)} [\log q(Z_j)] - \sum_{\substack{k=1 \\ k \neq j}}^m \mathbb{E}_{q(Z_k)} [\log q(Z_k)] \\
&= \mathbb{E}_{q(Z_j)} \left[ \log \left( \exp \left\{ \mathbb{E}_{\prod_{i=1, i \neq j}^m q(Z_i)} [\log p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta})] \right\} \right) - \log q(Z_j) \right] - \sum_{\substack{k=1 \\ k \neq j}}^m \mathbb{E}_{q(Z_k)} [\log q(Z_k)] \\
&= \mathbb{E}_{q(Z_j)} \left[ \log \frac{\exp \left\{ \mathbb{E}_{\prod_{i=1, i \neq j}^m q(Z_i)} [\log p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta})] \right\}}{q(Z_j)} \right] - \sum_{\substack{k=1 \\ k \neq j}}^m \mathbb{E}_{q(Z_k)} [\log q(Z_k)] \\
&= -\text{KL} \left( q(Z_j) \parallel \exp \left\{ \mathbb{E}_{\prod_{i=1, i \neq j}^m q(Z_i)} [\log p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta})] \right\} \right) - \sum_{\substack{k=1 \\ k \neq j}}^m \mathbb{E}_{q(Z_k)} [\log q(Z_k)]
\end{aligned}$$

Now only the left term is a function of  $q(Z_j)$ ; to maximize  $\mathcal{L}(q, \boldsymbol{\theta})$  with respect to  $q(Z_j)$ , we should minimize the KL-divergence, which results in the following,

$$q(Z_j) \propto \exp \left\{ \mathbb{E}_{q(Z_{-j})} [\log p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta})] \right\} \quad (1.12)$$

where  $q(Z_{-j})$  stands for  $\prod_{i=1, i \neq j}^m q(Z_i)$ .



There is a very good similarities between Gibbs sampling and mean-field variational learning. In Gibbs sampling we sample from conditionals (function of only one variable); this is similar to coordinate ascent updates in variational learning.

## 1.5 Examples

### 1.5.1 Dirichlet Process Mixture

Just as a brief reminder, the DPM model is defined in the following form:

$$G | \alpha, G_0 \sim \text{DP}(\alpha, G_0)$$

$$\eta_n | G \sim G$$

$$X_n | \eta_n \sim p(x_n | \eta_n)$$

We can concatenate the component parameters which are the same and represent the with  $\eta_1^*$ :

$$(\eta_1, \dots, \eta_n) \rightarrow (\eta_1^*, \dots, \eta_K^*)$$

In order to apply the variational updates we use the stick-breaking representation of the DP:

$$\begin{aligned} \begin{cases} V_i \sim \text{Beta}(1, \alpha) \\ \eta_i^* \sim G_0 \end{cases} &\Rightarrow \pi_i(\mathbf{v}) = v_i \prod_{j=1}^{i-1} (1 - v_j) \\ &\Rightarrow G = \sum_{i=1}^{+\infty} \pi_i(\mathbf{v}) \delta_{\eta_i^*} \end{aligned}$$

The sampling procedure could be explained in the following way:

1.  $V_i \sim \text{Beta}(1, \alpha), i = 1, 2, 3, \dots$
2.  $\eta_i^* \sim G_0, i = 1, 2, 3, \dots$
3. For example data  $n = 1, \dots, N$ 
  - (a)  $Z_n | \{v_1, v_2, \dots\} \sim \text{Multi}(\pi_i(\mathbf{v}))$
  - (b)  $X_n | z_n \sim p(x_n | \eta_{z_n}^*)$

The only assumption that we make is on the distributional form of  $p(x_n | z_n, \{\eta_i^*\})$  and  $p(\eta_i^* | \lambda)$ :

$$p(\eta_i^* | \lambda) = h(\eta_i^*) \exp\{\lambda_1^\top \eta_i^* + \lambda_2 (-a(\eta_i^*)) - a(\lambda)\} \quad (1.13)$$

$$p(x_n | z_n, \{\eta_i^*\}) = \prod_{i=1}^{\infty} \left( h(x_n) \exp\{\eta_i^{*\top} x_n - a(\eta_i^*)\} \right)^{\mathbf{1}_{\{z_n=i\}}} \quad (1.14)$$

In the next step we derive the variational updates for DPM.

■ **Example 1.4 — Variational updates for DPM.** First consider the full likelihood based on the graphical model of the DPM base on stick-breaking representation:

$$p(X, Z, \eta^*, V) = p(X | Z, \eta^*) p(Z | V) p(\eta^* | \lambda) p(V | \alpha) \quad (1.15)$$

In this distribution we are assuming that  $X$  is the only variable observed and we want to estimate the latent parameters of the model  $Z, \eta^*, V$ . Also remember that based on the stick-breaking construction we have:

$$p(Z_n | V) = V_{Z_n} \prod_{i=1}^{i < Z_n} (1 - V_i) \quad (1.16)$$

$$= \prod_{i=1}^{+\infty} (1 - V_i)^{\mathbf{1}_{\{Z_n > i\}}} V_i^{\mathbf{1}_{\{Z_n = i\}}} \quad (1.17)$$

It is easy to show that the Equation 1.16 follows from the Equation 1.17.



$$p(Z_n|V) = \prod_{i=1}^{+\infty} (1 - V_i)^{\mathbf{1}\{Z_n > i\}} V_i^{\mathbf{1}\{Z_n = i\}} \quad (1.18)$$

$$V_i \sim \text{Beta}(1, \alpha) \Rightarrow p(V|\alpha) \propto (1 - V_i)^{\alpha-1} \quad (1.19)$$

We approximate each of these latent variables with a distribution which we denote with  $q(\cdot)$  (to emphasize that it is an approximation):

- $Z_t \rightarrow q_{\phi_n}(z_n), n = 1 \dots N$
- $\eta_t^* \rightarrow q_{\tau_t}(\eta_t^*), t = 1 \dots T$
- $V_t \rightarrow q_{\gamma_t}(v_t), t = 1 \dots T$

Note that each of these approximate family of distributions are characterized by some hyperparameters. For example  $Z$  which is approximated via  $q_{\phi_n}(z_n)$  (the first item) which is characterized via  $\phi_n$  as the parameters of the distribution. Thus inference on the model, basically means finding these latent parameters.

To make the computations tractable, a *truncated* stick breaking process is assumed. In other words, let's say we perform the stick-breaking for  $T$  times, until all of the stick gets used. This means that in the last breaking step, we use the whole remaining stick, i.e.  $q(v_T) = 1$ . The joint distribution of the hyperparameters is

$$q(Z, \eta^*, V) = \prod_{t=1}^{T-1} q_{\gamma_t}(v_t) \prod_{t=1}^T q_{\tau_t}(\eta_t^*) \prod_{n=1}^N q_{\phi_n}(z_n). \quad (1.20)$$

Now let's find the updates for  $V$ :

$$\ln q(v_t) = \mathbb{E}_{q_{-v_t}} [\ln p(X, Z, \eta^*, V)] + C$$

where  $q_{-v_t}$  means the joint approximate distribution in Equation 1.20, except for the  $t$ -th term of  $V$ . We can replace  $p(X, Z, \eta^*, V)$  with its definition in Equation 1.15 and keep only the terms which contain  $V_t$ .

$$\begin{aligned} \ln q(v_t) &= \mathbb{E}_{q_{-v_t}} [\ln p(X, Z, \eta^*, V)] + C \\ &= \mathbb{E}_{q_{-v_t}} [\log p(X|V) + \ln p(V|\alpha)]. \end{aligned}$$

Then we can continue to simplify it using Equation 1.18 and Equation 1.19.

$$\begin{aligned} \ln q(v_t) &= \mathbb{E}_{q_{-v_t}} [\ln p(X|V) + \ln p(V|\alpha)] \\ &= \mathbb{E}_{q_{-v_t}} \left[ \ln \prod_{n=1}^N \prod_{i=1}^{+\infty} (1 - V_i)^{\mathbf{1}\{Z_n > i\}} V_i^{\mathbf{1}\{Z_n = i\}} + \ln(1 - V_t)^{\alpha-1} \right] \\ &= \sum_{n=1}^N \mathbb{E}_{q_{-v_t}} [\mathbf{1}\{Z_n > t\}] \times \ln(1 - V_t) \\ &\quad + \sum_{n=1}^N \mathbb{E}_{q_{-v_t}} [\mathbf{1}\{Z_n = t\}] \times \ln V_t + (\alpha - 1) \ln(1 - V_t) \end{aligned}$$

Also note that,

$$\begin{aligned}\mathbb{E}_{q-v_t} [\mathbf{1} \{z_n > t\}] &= q(z_n > t) = \sum_{i=t+1}^T \phi_{n,i} \\ \mathbb{E}_{q-v_t} [\mathbf{1} \{z_n = i\}] &= q(z_n = i) = \phi_{n,i}\end{aligned}$$

which gives us,

$$\ln q(v_t) = \sum_{n=1}^N \sum_{i=t+1}^T \phi_{n,i} \times \ln(1 - V_t) + \sum_{n=1}^N \phi_{n,t} \ln V_t + (\alpha - 1) \ln(1 - V_t) \quad (1.21)$$

$$(1.22)$$

What is the distributional form of  $q(v_t)$ ? Suppose it has Beta distribution,

$$V_t \sim \text{Beta}(\gamma_{t,1}, \gamma_{t,2})$$

Then,

$$\ln q(v_t) = (\gamma_{t,1} - 1) \ln V_t + (\gamma_{t,2} - 1) \ln(1 - V_t) + C.$$

This matches the form of Equation 1.21. Note that we never made any assumption on the distributional form of  $q(v_t)$ , but it naturally came out. Comparing this with Equation 1.21, we will end up with the following,

$$\begin{cases} \gamma_{t,1} = 1 + \sum_n \phi_{n,t} \\ \gamma_{t,2} = \alpha + \sum_n \sum_{j=t+1}^T \phi_{n,j} \end{cases} \quad (1.23)$$

Now let's derive the factor which contains  $\eta^*$ :

$$\ln q(\eta_t^*) = \mathbb{E}_{q-\eta_t^*} [\ln p(X, Z, \eta^*, V)] + C$$

We replace the full likelihood from Equation 1.15 and only keep the terms which contain  $\eta_t^*$ :

$$\ln q(\eta_t^*) = \mathbb{E}_{q-\eta_t^*} [\ln p(X|Z, \eta^*) + \ln p(\eta^*|\lambda)] + C$$

Based on the assumption on the prior we have,

$$\ln p(\eta^*|\lambda) = \ln h(\eta^*) + \lambda_1^\top \eta^* + \lambda_2 (-a(\eta^*)) - a(\lambda)$$

$$\begin{aligned}\Rightarrow \mathbb{E}_{q-\eta_t^*} [\ln p(\eta^*|\lambda)] &= \mathbb{E}_{q-\eta_t^*} \ln h(\eta^*) + \lambda_1^\top \mathbb{E}_{q-\eta_t^*} \eta^* + \lambda_2 \mathbb{E}_{q-\eta_t^*} (-a(\eta^*)) + C \\ &= \ln h(\eta_t^*) + \lambda_1^\top \eta_t^* - \lambda_2 a(\eta_t^*) + C\end{aligned}$$

And,

$$\begin{aligned}
 \mathbb{E}_{q_{-\eta_t^*}} [\ln p(X|Z, \eta^*)] &= \mathbb{E}_{q_{-\eta_t^*}} \left[ \ln \prod_{n=1}^N p(x_n|z_n, \eta^*) \mathbf{1}_{\{z_n=t\}} \right] \\
 &= \sum_{n=1}^N \mathbb{E}_{q_{-\eta_t^*}} [\mathbf{1}_{\{z_n=t\}}] \left( \eta_t^{*\top} x_n - a(\eta_t^*) + C \right) \\
 &= \sum_{n=1}^N \phi_{n,t} \left( \eta_t^{*\top} x_n - a(\eta_t^*) + C \right)
 \end{aligned}$$

Next we plug-in the distributions

$$\begin{aligned}
 \ln q(\eta_t^*) &= \sum_{n=1}^N \phi_{n,t} \left( \eta_t^{*\top} x_n - a(\eta_t^*) + C \right) + \ln h(\eta_t^*) + \lambda_1^\top \eta_t^* - \lambda_2 a(\eta_t^*) + C \\
 &= \left( \lambda_1 + \sum_{n=1}^N \phi_{n,t} x_n \right)^\top \eta_t^* + \left( \lambda_2 + \sum_{n=1}^N \phi_{n,t} \right) (-a(\eta_t^*)) + \ln h(\eta_t^*) + C
 \end{aligned}$$

This matches the prior distribution over  $\eta^*$  in the Equation 1.13, with the following parameters:

$$\begin{cases} \tau_{t,1} = \lambda_1 + \sum_n \phi_{n,t} x_n \\ \tau_{t,2} = \lambda_2 + \sum_n \phi_{n,t} \end{cases} \quad (1.24)$$

The last update is for the parameters of the distribution corresponding to  $Z$ :

$$\ln q(z_n = t) = \mathbb{E}_{q_{-\eta_t^*}} [\ln p(X, Z, \eta^*, V)] + C$$

Then we replace the definition of the full likelihood in Equation 1.15 and keep the factors that contain  $z_n$ :

$$\ln q(z_n = t) = \mathbb{E}_{q_{-z_n}} [\ln p(X|Z, \eta^*) + \ln p(Z|V)] + C$$

Next we simplify each term in the above equation:

$$\begin{aligned}
 \mathbb{E}_{q_{-z_n}} [\ln p(X|Z, \eta^*)] &= \mathbb{E}_{q_{-z_n}} \left[ \ln \prod_{i=1}^{+\infty} \left( h(x_n) \exp\{\eta_i^{*\top} x_n - a(\eta_i^*)\} \right)^{\mathbf{1}_{\{z_n=i\}}} \right] \\
 &= \mathbb{E}_{q_{-z_n}} \left[ \sum_{i=1}^{+\infty} \mathbf{1}_{\{z_n=i\}} \ln \left( h(x_n) \exp\{\eta_i^{*\top} x_n - a(\eta_i^*)\} \right) \right] \\
 &= \mathbb{E} [\eta_t^*]^\top X_n - \mathbb{E} a(\eta_t^*) + C
 \end{aligned}$$

And,

$$\begin{aligned}\mathbb{E}_{q_{-z}}[\ln p(Z_n = t|V)] &= \mathbb{E}_q \left[ \ln V_t \prod_{i=1}^{t-1} (1 - V_i) \right] \\ &= \mathbb{E}_q \ln V_t + \mathbb{E}_q \ln \sum_{i=1}^{t-1} \ln(1 - V_i)\end{aligned}$$

which would give the following update rule:

$$\begin{cases} \phi_{n,t} \propto \exp(S_t) \\ S_t = \mathbb{E}_q \ln V_t + \sum_{i=1}^{t-1} \mathbb{E}_q \ln(1 - V_i) + \mathbb{E}_q [\eta_t^*]^\top X_n - \mathbb{E}_q \alpha(\eta_t^*) \end{cases} \quad (1.25)$$

The overall algorithm is looping over the Equations 1.23, 1.24, and 1.25.

Since we have assumed the independence between different terms of the posterior distribution we have induces many local optima in the parameter space of the problem. Thus the initialization of the model might play a very important role in training the model. Testing the model with many different random initialization seems to be a reasonable approach. But how to control the convergence to good answer? Remember the goal was to choose the parameters which maximize the posterior. Evaluating the exact posterior is complicated; instead we can evaluate the lower-bound on it:

$$\mathcal{L} \geq \exp \mathbb{E}_q \left[ \ln \frac{p}{q} \right] = \exp \{ \mathbb{E}_q [\ln p] - \mathbb{E}_q [\ln q] \}$$

■

## 1.6 On minimization of divergence measures

Back to the variational M-step in Equation 1.11, it is just a mapping, from distribution  $p(\cdot)$  to a (tractable) family of distributions  $q(\cdot)$ . In addition to the *KL-divergence*, this mapping could be done via many other divergence measures between functions, to see examples and some properties see the Wiki pages for *KL-divergence*, Bregman divergence or *f-divergence*.

There is a nice discussion of divergence minimization in Altun and Smola (2006). Here we briefly mention some of the important results.

## 1.7 Energy minimization justifications

[TBW]

## 1.8 Variational learning with exponential family

Let us first review some properties about the exponential families. If you are already familiar with the exponential family, skip the definition.

**Definition 1.1 — exponential families.** We want to define a probability distribution on  $\mathcal{X}$  domain. Assume the vector of *sufficient estimators*  $\phi(\mathbf{x}) = [\phi(\mathbf{x})_1, \dots, \phi(\mathbf{x})_d]^\top$  in which  $\phi_i(\mathbf{x}) : \mathcal{X} \rightarrow \mathbb{R}$  is a function defined for any  $j = 1, \dots, d$ . This can correspond to the all nodes and connections in a Graphical model. In addition assume the vector of *canonical parameters*  $\boldsymbol{\theta} = [\theta_1, \dots, \theta_d]^\top$ , in which  $\theta_i \in \mathbb{R}$ . Now define the distribution as following,

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = h(\mathbf{x}) \exp(\boldsymbol{\theta}^\top \phi(\mathbf{x}) - A(\boldsymbol{\theta}))$$

where  $h(\mathbf{x})$  is an arbitrary function of  $\mathbf{x}$ . Usually this is  $h(\mathbf{x}) = 1$ .  $A(\boldsymbol{\theta})$  is the *cumulative generating function*, and is defines as  $A(\boldsymbol{\theta}) = \log \int_{\mathbf{x} \in \mathcal{X}} h(\mathbf{x}) \exp(\boldsymbol{\theta}^\top \phi(\mathbf{x})) \nu(d\mathbf{x})$ , where  $\nu(\cdot)$  is the measure defined on  $\mathcal{X}$ . Note that the normalizing function for the distribution is defined as  $\mathcal{Z}(\boldsymbol{\theta}) = \int_{\mathbf{x} \in \mathcal{X}} h(\mathbf{x}) \exp(\boldsymbol{\theta}^\top \phi(\mathbf{x})) \nu(d\mathbf{x}) = \exp(A(\boldsymbol{\theta}))$ .

The distribution is called *minimal* if the set of sufficient estimators are linearly independent. In other words, there is no  $\boldsymbol{\theta} \in \mathbb{R}^n$  for which  $\boldsymbol{\theta}^\top \phi(\mathbf{x}) = 0$  for all  $x \in \mathcal{X}$ . If the distribution is not regular, it is called *over-complete*. The distribution is called *regular* if for any vector coefficients the distribution is normalizable(a valid distribution). In other words, if we define the set  $\Omega = \{\boldsymbol{\theta} \in \mathbb{R}^n \mid |A(\boldsymbol{\theta})| < +\infty\}$ .

This family of distributions has different properties, some of which we are listing here. Many important distributions lie in the exponential family, for example Gaussian, Multinomial and Bernoulli. There is a rich table for transformation of many distributions can be found on the Wiki. In addition to the parametric distributions mentioned in the Wiki, it can be shown that any joint probability distribution on discrete random variables can be transformed into the exponential form.

The *cumulative generating function* has very important properties. The first derivative of commutative generating function is the expectation of the vector of the sufficient statistics,

$$\nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta}) = \mathbb{E}_p(\phi(\mathbf{x})). \quad (1.26)$$

This property is very useful since we can compute the integral of expectation, by differentiation. There is a similar relation for the second derivative,

$$\nabla_{\boldsymbol{\theta}}^2 A(\boldsymbol{\theta}) = \text{Var}_p(\phi(\mathbf{x})).$$

The *sufficiency* is a statistical property, and it means that “having the values of the sufficient statistics, we don’t need the data points to do inference”. In other words, after learning model, we can just through away the training data points. This is helpful because usually the dimension of the data points are much higher than the number of the sufficient statistics. Also if you consider the joint distribution of IID variables (likelihood),  $p_{\boldsymbol{\theta}}(\mathbf{x}_1, \dots, \mathbf{x}_n) = \prod_{i=1}^n h(\mathbf{x}_i) \exp(\boldsymbol{\theta}^\top \sum_{i=1}^n \phi(\mathbf{x}_i) - nA(\boldsymbol{\theta}))$ . This shows the only information visible from the training random variables is the summation  $\sum_{i=1}^n \phi(\mathbf{x}_i)$ . Thus,

have the value of this summation is enough for inference using this exponential model. The exact definition of sufficiency can be explained using the Fisher-Neyman factorization theorem (see Wiki).

Continuing the previous note, let's find the maximum likelihood estimation of the parameter vector  $\boldsymbol{\theta}$  using the likelihood of the IID random variable observations,

$$\boldsymbol{\theta}_{ML} = \arg \max_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\mathbf{x}_1, \dots, \mathbf{x}_n).$$

Putting the derivative of the joint distribution with respect to parameters to zero, we find the maximum-likelihood estimation,  $\nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \phi(\mathbf{x}_i)$ . Using our previous findings, we see that  $\mathbb{E}_p(\phi(\mathbf{x})) = \frac{1}{n} \sum_{i=1}^n \phi(\mathbf{x}_i)$  which says that the expected value of the sufficient statistic matches its average found using IID samples (by having enough data). This is very useful since in finding the expected value of the sufficient statistics by averaging their values from the training data, there is no need for parameter estimation for the original distribution.

Another very important property is the *convexity* of  $A(\boldsymbol{\theta})$  with respect to  $\boldsymbol{\theta}$ . In other words,

$$A(\beta \boldsymbol{\theta}_1 + (1 - \beta) \boldsymbol{\theta}_2) < \beta A(\boldsymbol{\theta}_1) + (1 - \beta) A(\boldsymbol{\theta}_2), \forall \beta \in (0, 1), \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$$

This is easy to prove; just observe that the second derivative (Hessian) of  $A(\boldsymbol{\theta})$  equals to is the covariance function of the sufficient estimators, and is a positive-semi-definite matrix(if you don't know why see Wiki).

Another nice property of the exponential family is the *conjugacy*. This means that, if we choose an exponential prior distribution in a Bayesian model with an exponential likelihood, the posterior distribution will be in the same exponential family.

*Proof.* This is easy to show; say the output observation is  $\mathbf{x}$  with parameter vector of  $\boldsymbol{\theta}$  with exponential distribution,

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x}) \exp(\boldsymbol{\theta}(\boldsymbol{\eta})^\top \phi(\mathbf{x}) - A(\boldsymbol{\theta}(\boldsymbol{\eta}))),$$

which is the definition we used for the exponential family with the difference that the vector of parameters is function of another random variable  $\boldsymbol{\eta}$ . The prior over the  $\boldsymbol{\eta}$  we define to be,

$$p_{\boldsymbol{\lambda}}(\boldsymbol{\eta}) = g(\boldsymbol{\eta}) \exp(\boldsymbol{\theta}(\boldsymbol{\eta})^\top \boldsymbol{\lambda} - B(\boldsymbol{\lambda})).$$

Let's assume we have a matrix of observation  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$ , and the likelihood for the random IID observation variables is,

$$p(\mathbf{X}|\boldsymbol{\eta}) = \prod_{i=1}^n h(\mathbf{x}_i) \exp\left(\boldsymbol{\theta}(\boldsymbol{\eta})^\top \sum_{i=1}^n \phi(\mathbf{x}_i) - nA(\boldsymbol{\theta}(\boldsymbol{\eta}))\right).$$

Using the Bayes formula we know that posterior equals to  $p(\boldsymbol{\eta}|\mathbf{X}; \boldsymbol{\lambda}) \propto p(\mathbf{X}|\boldsymbol{\eta})p(\boldsymbol{\eta}; \boldsymbol{\lambda})$  (Note that here I am assuming that  $\boldsymbol{\eta}$  is a vector of random variables (not constant) but  $\boldsymbol{\lambda}$  is a vector of parameters; that's why I write  $p(\boldsymbol{\eta}; \boldsymbol{\lambda})$  instead of  $p(\boldsymbol{\eta}|\boldsymbol{\lambda})$ ). Using the Bayes formula, the posterior is following,

$$p(\boldsymbol{\eta}|\mathbf{X}; \boldsymbol{\lambda}) \propto \exp \left( \boldsymbol{\theta}(\boldsymbol{\eta})^\top \left( \boldsymbol{\lambda} + \sum_{i=1}^n \phi(\mathbf{x}_i) \right) - (nA(\boldsymbol{\theta}(\boldsymbol{\eta})) + B(\boldsymbol{\lambda})) \right).$$

■

The exponential distribution also arises naturally from the *maximum entropy principle*. The maximum entropy procedure consists of seeking the probability distribution which maximizes information entropy, subject to the constraints of the information. If we constrain the expected values of the sufficient statistics to be mean of the empirical mean of them under the sampled data, the resulting distribution which maximizes the entropy is the exponential family. To make the statement more accurate, let's express it in terms of formula. Given iid random variables,  $\mathbf{x}_1, \dots, \mathbf{x}_n \sim p^*$ , where  $p^*$  is the underlying *unknown* distribution. Define a set of functions  $\{\phi_i(\cdot) : \mathcal{X} \rightarrow \mathbb{R}\}_{i=1}^d$ , and consider the empirical mean of the sampled data under these functions,

$$\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^n \phi_j(\mathbf{x}_i), \quad \forall j \in \{1, \dots, d\}$$

The problem is that we are looking for a distribution,  $p(\cdot)$  which has the same set of expected values for the defined functions as their empirical distribution, i.e.

$$\mathbb{E}_p[\phi_j(\mathbf{x})] = \int_{\mathcal{X}} \phi_j(\mathbf{x}) p(\mathbf{x}) \nu(d\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \phi_j(\mathbf{x}_i) = \hat{\mu}_j,$$

and at the same time maximizes the entropy under this distribution,

$$E = \int_{\mathcal{X}} -p(\mathbf{x}) \log p(\mathbf{x}) \nu(d\mathbf{x}).$$

This constrained optimization will result in the parametric form of the exponential family of distributions defined here.

Looking at the update rule at equation 1.12, one might suspect that we probably can devise easier updates for each  $q(Z_i)$ , when they are from exponential families. In fact, having only a few moments of exponents in exponential families are *sufficient* estimators of this family of distributions. Let's say 1 has the following general form in exponential families,

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = f(\boldsymbol{\chi}, \nu) \exp \{ \boldsymbol{\eta}(\mathbf{X}, \mathbf{Z})^\top \boldsymbol{\chi} - \nu A(\boldsymbol{\eta}(\mathbf{X}, \mathbf{Z})) \}$$

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \ln f(\boldsymbol{\chi}, \nu) + \boldsymbol{\eta}(\mathbf{X}, \mathbf{Z})^\top \boldsymbol{\chi} - \nu A(\boldsymbol{\eta}(\mathbf{X}, \mathbf{Z}))$$

$$\mathbb{E}_{q(Z_{-j})} \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) = \ln f(\boldsymbol{\chi}, \nu) + \mathbb{E}_{q(Z_{-j})} \boldsymbol{\eta}(\mathbf{X}, \mathbf{Z})^\top \boldsymbol{\chi} - \mathbb{E}_{q(Z_{-j})} \nu A(\boldsymbol{\eta}(\mathbf{X}, \mathbf{Z}))$$

then,

$$q(Z_j) \propto f(\boldsymbol{\chi}, \nu) \exp \{ \mathbb{E}_{q(Z_{-j})} \boldsymbol{\eta}(\mathbf{X}, \mathbf{Z})^\top \boldsymbol{\chi} \}$$

This shows that that choosing the approximating functions  $q(\mathbf{Z})$ , from the same exponential family of the conditional  $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$  would crucially help in E-step updates; we only need to calculate expectation of the exponent with respect to approximating family. More on this property and more examples at Wainwright and Jordan (2008).

### 1.8.1 Mean parametrization and marginal polytopes

Let's assume the probability density  $p(\cdot)$  (not necessarily in the exponential family). Assume a set of local functions  $\{\phi_i(\cdot) : \mathcal{X} \rightarrow \mathbb{R}\}_{i \in \mathcal{I}}$ , in which  $\mathcal{I}$  is a set of indices. Also a vector of mean parameters  $\boldsymbol{\mu} = [\mu_1, \dots, \mu_d]^\top$  which are defined as following,

$$\mu_j = \mathbb{E}_p[\phi_j(\mathbf{x})] = \int_{\mathcal{X}} \phi_j(\mathbf{x}) p(\mathbf{x}) \nu(d\mathbf{x})$$

Now we define the whole space of realizable mean vectors by any families of distributions  $\mathcal{P}$  (not limited to exponential family) as *marginal polytope*,

$$\mathcal{M} = \{\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{I}|} | \exists p(\cdot) \in \mathcal{P} \text{ s.t. } \mathbb{E}_p[\phi(\mathbf{x})] = \boldsymbol{\mu}\}$$

If in the above definition we limit the family of the distributions  $\mathcal{P}$  to the exponential family  $\mathcal{E}$ , since the exponential family is a strict subset of the general distributions, the resulting polytope  $\mathcal{M}^0$  is a strict subset of the polytope  $\mathcal{M}$ ,

$$\mathcal{M}^0 = \{\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{I}|} | \exists p(\cdot) \in \mathcal{E} \text{ s.t. } \mathbb{E}_p[\phi(\mathbf{x})] = \boldsymbol{\mu}\} \subset \mathcal{M}. \quad (1.27)$$

**Proposition 1.2** The marginal polytope  $\mathcal{M}$  is a convex subset of  $\mathbb{R}^{|\mathcal{I}|}$  (proof?).

■ **Example 1.5 — A simple Gaussian MRF.** Let's assume  $\phi(x) = [x, x^2]^\top$ , and the vector of means  $[\mu_1, \mu_2]$  realized under any arbitrary distribution  $p(\cdot)$  such that  $[\mu_1, \mu_2] = [\mathbb{E}_p[x], \mathbb{E}_p[x^2]]$ . The only constrain in such realization is that  $\mathbb{V}(x) = \mathbb{E}_p[x^2] - \mathbb{E}_p^2[x] = \mu_2 - \mu_1^2 \geq 0$  ■

■ **Example 1.6 — General Gaussian MRF.** Assume a graph, on which each vertex is associated with a continuous variable on  $\mathbb{R}$  and the variables which are connected by vertices are dependent based on a Gaussian distribution. Let's assume connection between all of the variables (a complete graph,  $k_n$ ). It is more convenient to denote the sufficient statistics in a matrix instead of a vector, since their pairwise interaction (a quadratic model) is important,

$$\phi(\mathbf{x}) = \begin{bmatrix} 1 & x_1 & x_2 & \dots & x_n \\ x_1 & x_1^2 & x_1x_2 & \dots & x_1x_n \\ x_2 & x_2x_1 & x_2^2 & \dots & x_2x_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_n & x_nx_1 & x_nx_2 & \dots & x_n^2 \end{bmatrix}.$$

Given these local functions, we want to find the possible mean values such that



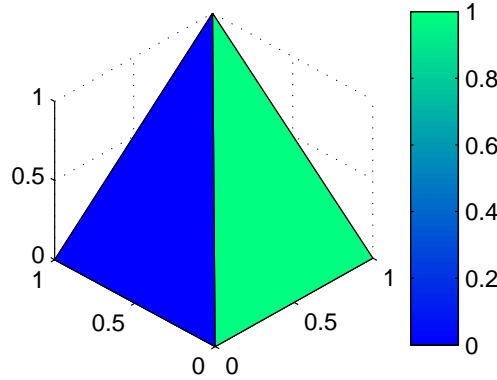


Figure 1.4: The marginal polytope for the Ising model in the example.

$\mathbb{E}_p[\phi(\mathbf{x})] = \mu$ . We show the corresponding mean matrix as follows,

$$U(\mu) = \begin{bmatrix} 1 & \mu_1 & \mu_2 & \dots & \mu_n \\ \mu_1 & \mu_{11} & \mu_{12} & \dots & \mu_{1n} \\ \mu_2 & \mu_{21} & \mu_{22} & \dots & \mu_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu_n & \mu_{n1} & \mu_{n2} & \dots & \mu_{nn} \end{bmatrix}.$$

Now all of the constraints on the mean-space for this model could be found by setting  $U(\mu)$  be positive definite,

$$\mathcal{M}_{\text{Gaussian-MRF}} = \left\{ \mu \in \mathbb{R}^{n+\binom{n}{2}} \mid U(\mu) \succeq 0 \right\}$$

■ **Example 1.7 — A simple Ising model.** Let's assume a very simple Ising models with only two random variables on a graph with two vertices which are connected by an edge. Then the local functions are  $\phi(\mathbf{x}) = [x_1, x_2, x_1x_2]^\top$ . Now we want to find the space of all  $[\mu_1, \mu_2, \mu_3]$  such that,  $[\mu_1, \mu_2, \mu_3] = [\mathbb{E}[x_1], \mathbb{E}[x_2], \mathbb{E}[x_1x_2]]$ . If we simplify this,  $\mathbb{E}[x_1] = 1 \times \mathbb{P}(x_1 = 1) + 0 \times \mathbb{P}(x_1 = 0) = \mathbb{P}(x_1 = 1)$ , similarly  $\mathbb{E}[x_2] = \mathbb{P}(x_2 = 1)$  and  $\mathbb{E}[x_1x_2] = 1 \times 1 \times \mathbb{P}(x_1 = 1, x_2 = 1) + 0 \times 1 \times \dots + \dots = \mathbb{P}(x_1 = 1, x_2 = 1)$ . Each of the individual probabilities can vary between zero and 1. Also the joint distribution  $\mathbb{P}(x_1 = 1, x_2 = 1)$  is strictly less than each of the other individual distributions. Thus the resulting polytope is depicted in Figure 1.4. ■

### 1.8.2 Convex dualities

Here want to provide another view of variational inference. Let us assume an arbitrary function  $f(u)$  which is defined on  $u \in \mathbb{R}^n$ . We show the *conjugate dual*

of  $f(\cdot)$  by  $f^*(\cdot)$  and define this function as following:

$$f^*(v) = \sup_{u \in \mathbb{R}^n} \{\langle u, v \rangle - f(u)\}$$

for any  $v \in \mathbb{R}^n$ . In general the above definition holds in any domain that the Lebesgue measure holds (or on other words the inner product  $\langle u, v \rangle$  makes sense) and is not limited to real numbers.

**R** Re-using the same definition above we find the following,

$$(f^*)^*(u) = \sup_{v \in \mathbb{R}^n} \{\langle u, v \rangle - f^*(v)\}$$

gives the double conjugate dual of the function. In certain conditions, the double conjugate dual of a function equals to itself. If the function  $f(\cdot)$  is *well-behaved* (i.e. *convex and semi-continuous*) then the double conjugate dual of a function equals to itself:

$$\begin{cases} f^*(v) = \sup_{u \in \mathbb{R}^n} \{\langle u, v \rangle - f(u)\} \\ f(u) = \sup_{v \in \mathbb{R}^n} \{\langle u, v \rangle - f^*(v)\} \end{cases}$$

which shows the strong coupling between the functions and its conjugate dual. If the function is *concave* the same duality holds with  $\inf(\cdot)$  operator, instead of  $\sup(\cdot)$ .

$$\begin{cases} f^*(v) = \inf_{u \in \mathbb{R}^n} \{\langle u, v \rangle - f(u)\} \\ f(u) = \inf_{v \in \mathbb{R}^n} \{\langle u, v \rangle - f^*(v)\} \end{cases}$$

This is called **Fenchel's duality theorem**.

The reason for introducing these dualities is that, sometimes the optimization procedure is harder than optimization in its conjugate dual form (or vice versa). Thus we can use these conjugate dualities to find easier optimization schemes, as it will be shown.

**Proposition 1.3** The conjugate dual function is always a convex function (proof?).

### 1.8.3 The log-partition function and conjugate duality

Let's instead of the general functions  $f(\cdot)$ , assume the log-partition function (cumulant function).

$$A^*(\mu) = \sup_{\theta \in \Omega} \{\theta^\top \mu - A(\theta)\}.$$

As in the Equation 1.26 we have shown that  $\nabla_\theta A(\theta) = \mathbb{E}_p(\phi(\mathbf{x}))$ . We can consider  $\nabla_\theta A(\theta) : \Omega \rightarrow \mathcal{M}$  as a function which maps from the parameter space to the mean space, and we call this a *forward mapping*. Thus, because  $A(\theta)$  is defined for the exponential family, this forward mapping, maps all possible parameters to all possible means by these parameters in the exponential family which we called  $\mathcal{M}^0$  in Equation 1.27. Thus the mapping using  $\nabla_\theta A(\theta) : \Omega$  covers the whole  $\mathcal{M}^0$  which is a strict subset of  $\mathcal{M}$ . Thus, there might be some elements in  $\mu \in \mathcal{M} \setminus \mathcal{M}^0$  which are not realizable by any parameter in the exponential family. If we limit the domain to  $\mathcal{M}^0$  it is easy to show that the mapping is one-to-one if the exponentials are minimal (easy to show by contradiction).

**Proposition 1.4** The mapping  $\nabla_\theta A(\theta) : \Omega \rightarrow \mathcal{M}^0$  is one-to-one if and only if the exponential distribution is minimal.

The *backward mapping* is defined in the similar way to the forward mapping; For minimal exponential family, for any  $\mu \in \mathcal{M}^0$ , there exists a  $\theta \in \Omega$  such that  $\mathbb{E}_{p_\theta}(\phi(\mathbf{x})) = \mu$ . Note that among non-exponential families there might be some other other distributions that have the same mean  $\mu$ , but all of them have less entropy than the exponential one, since the exponential family has the maximum entropy, given the means as constraints.

For any  $\mu \in \mathcal{M}^0$ , let  $\theta$  be the corresponding parameter based on the equation  $\mathbb{E}_{p_\theta}(\phi(\mathbf{x})) = \mu$ . We can show that the dual function takes the following form,

$$A^*(\mu) = \begin{cases} -H(p_\theta), & \mu \in \mathcal{M}^0 \\ +\infty & \text{otherwise} \end{cases}$$

in which  $H(p_\theta)$  is the entropy of the distribution  $p_\theta$ . This property is very useful when using the maximum-entropy rule for model-selection, since it gives the direct connection to the entropy of the model. At the same time, given the parameters of the model, calculating the double-conjugate (the cumulant itself) will give the corresponding vector of means which is as if *inference* given a model. This explanation should prove the importance of the variational conjugate dualities and their usefulness in inference and model-selection problems.

■ **Example 1.8 — Conjugate duality on a Bernoulli distribution.** The Bernoulli distribution PMF is defined as  $p(x) = \beta^x(1 - \beta)^{1-x}$ , over random variable  $x \in \{0, 1\}$  where  $\beta$  is the probability of success in one toss. By a little modification we can change the representation to exponential form:

$$\begin{aligned} p(x) &= \beta^x(1 - \beta)^{1-x} \\ &= \exp \left\{ \log [\beta^x(1 - \beta)^{1-x}] \right\} \\ &= \exp (x \log \beta + (1 - x) \log(1 - \beta)) \\ &\propto \exp (x\theta), \quad A(\theta) = \log(1 + \exp(\theta)) \end{aligned}$$

where  $\theta = \log \frac{\beta}{1-\beta}$ . We define  $A^*(\mu) = \sup_{\theta \in \mathbb{R}} \{\theta\mu - \log(1 + \exp(\theta))\}$ . Simplifying this equation for any  $\mu \in (0, 1)$  we find that  $A^*(\mu) = \mu \log \mu + (1 - \mu) \log(1 - \mu)$  which is the entropy for the Bernoulli distribution as mentioned before. ■

## 1.9 Belief Propagation vs. Mean-field approximation

Here is a general tip for when to use BP and when the MF:

- Use MF: When you have an intractable model, which consists of several tractable models, which have a weak coherence with each other.
- Use BP: When you have an intractable model, with a lot of coherence between the substructures. Often adding global structures on models, creates strong coherence between the subgraphs of a graphical model.

And some pros and cons:

- Mean-field method, usually are applicable to simpler models, with less internal correlation, and it is basically doing coordinate-descent on the variables

of the problem. Typically it is fast, and it always converges, though it might converge to a wrong answer.

- BP could be applied to harder problems, but there is no guarantee for convergence and correctness. But since could be applied to a (almost) any configuration, it gives a lot of flexibility in designing the graph.

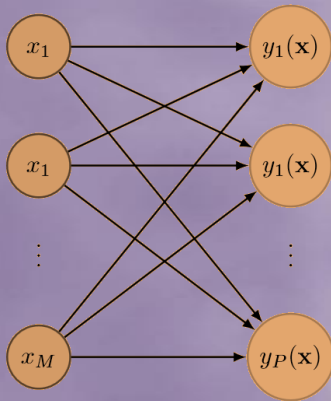
Note that, when the problem has a peaked unimodal posterior (unique optimal answer), the answer are very close to each other. In a good model it usually ends up having peaked near-gaussian posterior.

### 1.10 Bibliographical notes

In preparation of this part I have used Bishop (2006). At some visualizations I've also been inspired by Julia Hockenmaier's slides <sup>3</sup>. Some examples and ideas are also inspired from the lecture notes at Cevher (2008). An important relevant paper is Wainwright and Jordan (2008) which has given the inspiration in many of the notations and examples. David Burkett's ACL tutorial had very good visualizations and points, for a tutorial. Some examples are from David Fosyth's optimization course. Thanks to Jin Wang for discussing the derivation of VB for DPM.

---

<sup>3</sup><http://courses.engr.illinois.edu/cs598jhm/>



## Bibliography

Yasemin Altun and Alex Smola. Unifying divergence minimization and statistical inference via convex duality. In *Learning theory*, pages 139–153. Springer, 2006.

C.M. Bishop. *Pattern recognition and machine learning*, volume 4. springer New York, 2006.

Volkan Cevher. *Variational Bayes Approximation*. Rice University, 2008. Lecture notes of Graphical Models course.

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