
Probabilistic Machine Learning:

9. Variational Inference

ESC 2024 Spring Session 3주차



Contents

- 1. Variational Inference**
- 2. Variational Mixture of Gaussians**
- 3. Variational Linear Regression**
- 4. Exponential Family Distributions**
- 5. Local Variational Methods**
- 6. Variational Logistic Regression**
- 7. Expectation Propagation**

1. Variational Inference

Factorized distribution

Properties of factorized approximation

Example

Model comparison

Variational Inference

Probabilistic model

→ $p(\mathbf{Z}|\mathbf{X})$ 를 사용하여 $p(\mathbf{X}, \mathbf{Z})$ 의 기대값을 구한다.

Problem

- 잠재변수의 차원이 높은 경우 계산이 어렵다.
- 기대값 자체를 구하기 어려운 모델이 존재한다.

Solution

- Approximate Inference (근사추정)

Variational Inference

Approximate Inference

- stochastic (확률적 방법)
- deterministic (결정론적 방법)

Variational Inference: deterministic

- 함수의 형태를 제한 함으로써 posterior를 근사하는 방식을 사용

Variational Inference

Functional(범함수)

→ 함수를 입력 받아 실수를 반환하는 함수

Calculus of Variations(변분법)

→ 적분형태의 식을 최소/최대가 되게 하는 함수를 찾는 방법

$$H[p] = \int p(x) \ln p(x) dx$$

→ 찾아야 하는 함수의 형태를 제한하기 때문에 근사화된 식을 구할 수 있다.

1)quadratic

2)고정된 기저 함수의 선형결합

3)factorization assumption

Variational Inference

ELBO 최대화 in VI cf) EM Algorithm

$$\ln p(\mathbf{X}) = L(q) + KL(q||p)$$

→ EM algorithm에서와 마찬가지로 고정된 주변확률에서 KL-divergence를 최소화 함으로써 ELBO를 최대화 하는 $q_j^*(z_j)$ 를 구하게 된다.

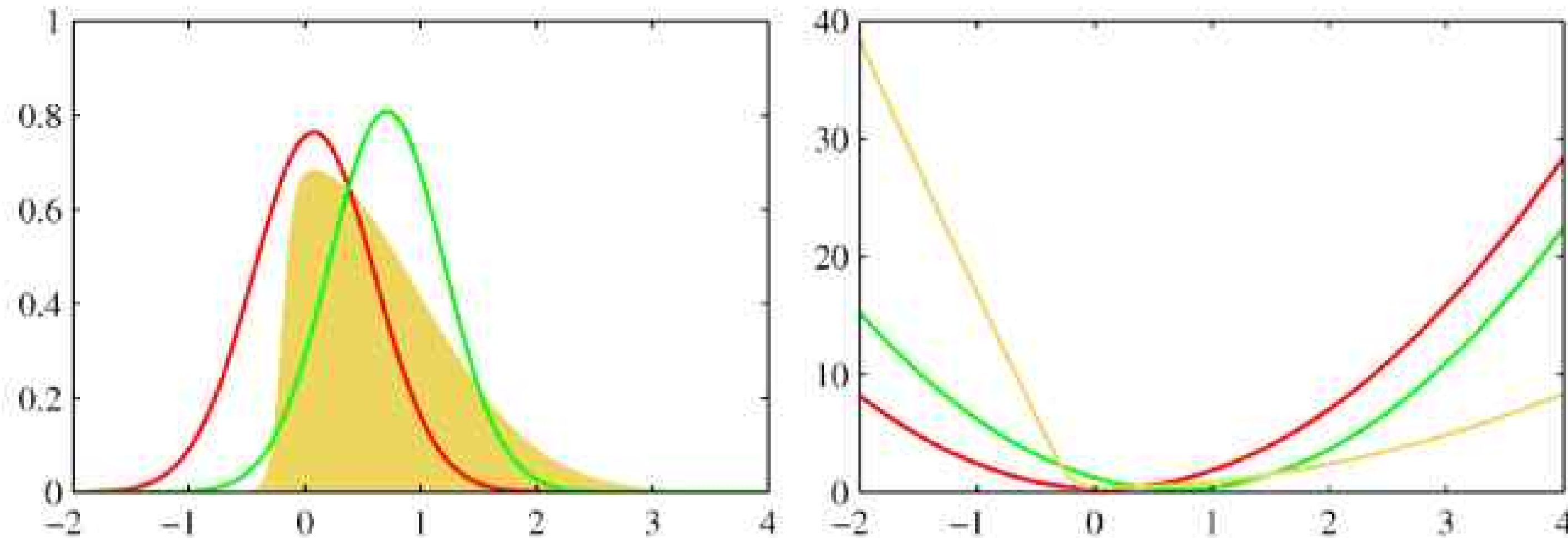
→ EM에서와 다르게 파라미터 theta가 모두 잠재변수로 흡수되고, 잠재변수는 연속형 변수로 처리되어 합산 식이 적분으로 대체된다.

→ EM에서는 $q = p(\mathbf{Z}|\mathbf{X})$ 로 하여 KL을 최소화 하였지만, VI에서는 posterior를 구하기 어렵기 때문에 q를 제한적인 계열의 분포로 생각하고, KL를 최소화하는 파라미터를 구한다.

Variational Inference

Parametric distribution (모수를 사용하는 분포)

→ $q(\mathbf{Z}|\mathbf{w})$ 로 생각하여 w 에 의해 분포 모양이 결정되게 한다.
따라서 ELBO는 w 에 대한 함수로 생각할 수 있다.



Factorized distributions

Factorized distributions

→ $q(\mathbf{Z})$ 의 종류를 인수분해된 형태로 제한

1) 잠재변수가 disjoint group 으로 나누어진다고 가정 => factorization 가정

$$q(\mathbf{Z}) = \prod_i^M q_i(\mathbf{Z}_i)$$

2) ELBO를 가장 크게 하는 $q_j^*(z_j)$ 를 찾는다.

$$L(q) = \int \prod_i q_i \times \left\{ \ln p(\mathbf{X}, \mathbf{Z}) - \sum_i \ln q_i \right\} d\mathbf{Z}$$

전개 과정에서 임의의 특정 j 에 관련 있는 텀과
그렇지 않은 i 들로 분리하여 앞으로 계산에 쓰일 식을 구한다.

$$E_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] = \int \ln p(\mathbf{X}, \mathbf{Z}) \prod_{i \neq j} q_i d\mathbf{Z}_i$$

Factorized distributions

$$L(q) = -KL(q_j || \tilde{p}(\mathbf{X}, \mathbf{Z})) + const$$

위의 식을 기초로 ELBO를 최대화 하기 위해서는 다음 조건을 만족하면 된다. (q와 p가 유사해야 KL이 최소가 되므로 ELBO가 최대가 되기 때문)

$$\ln q_j^*(\mathbf{Z}_j) = E_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})] + const$$

$$q_j^*(\mathbf{Z}) = \frac{\exp(E_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})])}{\int \exp(E_{i \neq j}[\ln p(\mathbf{X}, \mathbf{Z})]) d\mathbf{Z}_j}$$

이 식이 VI를 사용하기 위한 가장 기본적인 식이 된다.

Qi에 대한 해가 qj에 종속이므로 EM과 마찬가지로 수렴할 때까지 update하여 최종해를 구한다.

Properties of factorized approximations

- VI 방식은 실제 사후 분포를 인수분해된 형태로 구하는 형태를 사용한다.
- 인수 분해된 분포로 원래 분포를 근사하게 된다.

Example: Gaussian

두 개의 연관변수 z_1, z_2 에 대한 Gaussian분포를 가정

$$p(\mathbf{z}) = N(\mathbf{z}|\mu, \Lambda^{-1}) \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \Lambda = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}$$

1) 인수분해 형태를 가정

$$q(\mathbf{z}) = q_1(z_1)q_2(z_2)$$

2-1) 최적화 값인 q_1^* 구하기 (z_1 만 종속변수로 여긴다.) **증명

$$\begin{aligned} \ln q_1^*(z_1) &= E_{z_2} [\ln p(\mathbf{z})] + const \\ &= E_{z_2} \left[-\frac{1}{2}(z_1 - \mu_1)^2 \Lambda_{11} - (z_1 - \mu_1) \Lambda_{12} (z_2 - \mu_2) \right] + const \\ &= -\frac{1}{2} z_1^2 \Lambda_{11} + z_1 \mu_1 \Lambda_{11} - z_1 \Lambda_{12} (E[z_2] - \mu_2) + const \quad (10.11) \end{aligned}$$

Example: Gaussian

$$-\frac{1}{2}z_1^2\Lambda_{11} + z_1\mu_1\Lambda_{11} - z_1\Lambda_{12}(E[z_2] - \mu_2) + const$$

→ 위 식은 z_1 에 대하여 quadratic 이기 때문에 $q_1^*(z_1)$ 은 Gaussian임을 알 수 있다.

→ 이는 처음부터 $q_1(z_1)$ 이 Gaussian임을 가정한 것이 아님에도 최적화 결과를 통해 유도된 것

2-2) 완전 제곱식을 이용하여 Gaussian의 평균과 정확도를 계산하여 최적화 값을 구한다.

$$**\text{증명} \quad q_1^*(z_1) = N(z_1|m_1, \Lambda_{11}^{-1}) \quad q_2^*(z_2) = N(z_2|m_2, \Lambda_{22}^{-1})$$

$$m_1 = \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(E[z_2] - \mu_2) \quad m_2 = \mu_2 - \Lambda_{22}^{-1}\Lambda_{12}(E[z_1] - \mu_1)$$

→ 분포가 서로의 기대값에 의존하기 때문에 어떠한 수렴기준을 만족할 때까지 차례로 변수를 업데이트하는 구조를 가지게 된다.

Minimalizing KL(p||q)

Minimalizing KL(p||q)

- 지금까지 본 VI는 KL(q||p)를 최소화 하는 문제였다.
- 인수 분해된 근사분포라는 가정 하에 이번엔 KL(p||q)를 최소화하는 문제를 살펴보자.

$$KL(p||q) = - \int p(x) \ln q(x) dx - \left(- \int p(x) \ln p(x) dx \right) = - \int p(x) \ln \left\{ \frac{q(x)}{p(x)} \right\} dx$$

$$KL(p||q) = - \int p(\mathbf{Z}) \left[\sum_{i=1}^M \ln q_i(\mathbf{Z}_i) \right] d\mathbf{Z} + const \quad ***필기$$

- 위 식의 constant는 $p(\mathbf{Z})$ 에 대한 엔트로피 값으로 $q(\mathbf{Z})$ 와는 무관한 값이 된다.

$$q_j^*(\mathbf{Z}_j) = \int p(\mathbf{Z}) \prod_{i \neq j} d\mathbf{Z}_i = p(\mathbf{Z}_j)$$

- 라그랑주 승수법을 활용하면 위와 같은 최적화 식을 만들어 낼 수 있다.

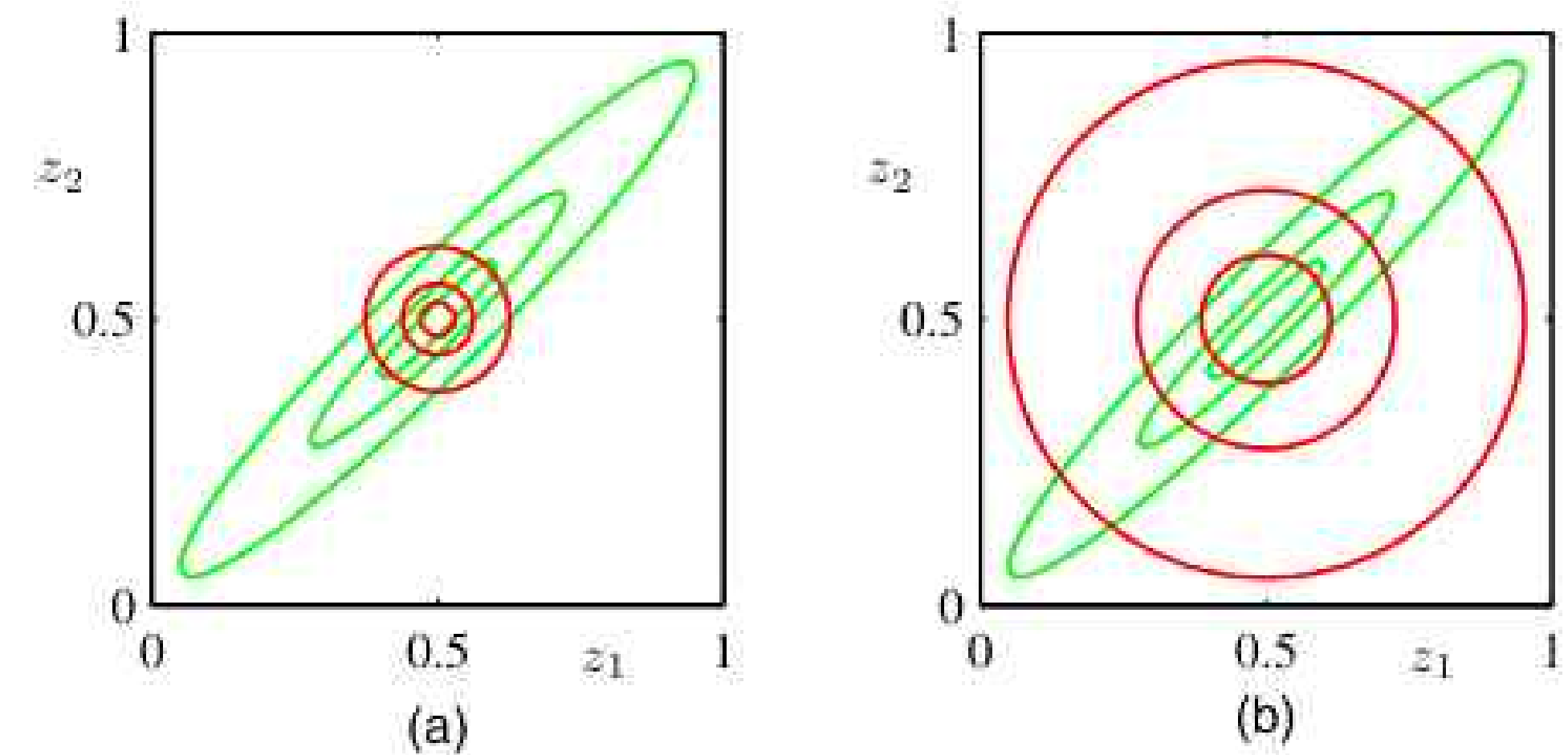
Minimalizing $KL(p||q)$

Green

: z_1, z_2 에 대한 Gaussian $p(z)$ 의 표준편차

Red

: z_1, z_2 에 대한 $q(z)$ 로 근사한 경로



$KL(q||p)$ 최소화에선 $p(z)$ 가 0에 가까울 때 $q(z)$ 를 0에 가깝도록 최적화 하게 되고,
반대로 $KL(p||q)$ 최소화에선 $q(z)$ 가 0에 가까울 때 $p(z)$ 를 0에 가깝도록 최적화 한다.

그림 (a)는 $KL(q||p)$ 를 최소화, (b)는 $KL(p||q)$ 를 최소화 한 것이다.

Example: The Univariate Gaussian

Goal: 평균 μ 와 정밀도 τ 의 posterior를 추론하는 것

Likelihood function

$$p(\mathcal{D}|\mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} \exp\left\{-\frac{\tau}{2} \sum_{n=1}^N (x_n - \mu)^2\right\}.$$

$$\begin{aligned} \ln q_\mu^*(\mu) &= \mathbb{E}_\tau [\ln p(\mathcal{D}|\mu, \tau) + \ln p(\mu|\tau)] + \text{const} \\ &= -\frac{\mathbb{E}[\tau]}{2} \left\{ \lambda_0 (\mu - \mu_0)^2 + \sum_{n=1}^N (x_n - \mu)^2 \right\} + \text{const}. \end{aligned}$$

Mu & tau's prior distribution

$$\begin{aligned} p(\mu|\tau) &= \mathcal{N}(\mu|\mu_0, (\lambda_0\tau)^{-1}) \\ p(\tau) &= \text{Gam}(\tau|a_0, b_0) \end{aligned}$$

→ $q(\mu)$ 가 $\mathcal{N}(\mu|\mu_N, \lambda_N^{-1})$ 의 분포를 따른다.

Posterior에 대한 factorized 가정

$$q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau).$$

$$\begin{aligned} \mu_N &= \frac{\lambda_0\mu_0 + N\bar{x}}{\lambda_0 + N} \\ \lambda_N &= (\lambda_0 + N)\mathbb{E}[\tau]. \end{aligned}$$

Example: The Univariate Gaussian

정밀도 $q(\tau)$ 에 대해서도 최적화를 시켜보자

$$\begin{aligned}\ln q_\tau^*(\tau) &= \mathbb{E}_\mu [\ln p(\mathcal{D}|\mu, \tau) + \ln p(\mu|\tau)] + \ln p(\tau) + \text{const} \\ &= (a_0 - 1) \ln \tau - b_0 \tau + \frac{N}{2} \ln \tau \\ &\quad - \frac{\tau}{2} \mathbb{E}_\mu \left[\sum_{n=1}^N (x_n - \mu)^2 + \lambda_0 (\mu - \mu_0)^2 \right] + \text{const}\end{aligned}$$

→ $q(\tau)$ 는 $\text{Gam}(\tau|a_N, b_N)$ 를 따른다는 알 수 있다.

$$\begin{aligned}a_N &= a_0 + \frac{N}{2} \\ b_N &= b_0 + \frac{1}{2} \mathbb{E}_\mu \left[\sum_{n=1}^N (x_n - \mu)^2 + \lambda_0 (\mu - \mu_0)^2 \right].\end{aligned}$$

이렇게 특정한 functional form에 대한 가정이 없이도, factorization 가정 만으로도 Conjugacy가 도출됨을 알 수 있다.

Model comparison

Variational Inference를 통해 prior probability $p(m)$ 을 가진 후보 모델들을 비교할 수 있다.

Goal: $p(m|X)$ 구하기

→ 모델들은 서로 다른 구조를 가졌을 수 있고, Z 들이 다른 차원수를 가졌을 수도 있기 때문에

$q(\mathbf{Z}, m) = q(\mathbf{Z}|m)q(m)$. 임을 사용해야 한다. 이를 이용해 최적화를 시행하면

$$\ln p(\mathbf{X}) = \mathcal{L}_m - \sum_m \sum_{\mathbf{Z}} q(\mathbf{Z}|m)q(m) \ln \left\{ \frac{p(\mathbf{Z}, m|\mathbf{X})}{q(\mathbf{Z}|m)q(m)} \right\} \quad \mathcal{L}_m = \sum_m \sum_{\mathbf{Z}} q(\mathbf{Z}|m)q(m) \ln \left\{ \frac{p(\mathbf{Z}, \mathbf{X}, m)}{q(\mathbf{Z}|m)q(m)} \right\}.$$

임을 통해 다음과 같은 결과를 얻는다.

$$q(m) \propto p(m) \exp\{\mathcal{L}_m\}.$$

2. Variational Mixture of Gaussian

Variational distribution

Variational lower bound

Predictive density

Determining the number of components

Induced factorization

Variational Mixture of Gaussian

먼저 각 데이터 포인트 \mathbf{x}_n 에 대해 binary latent variable z_n 이 주어진다고 하자.
그러면 \mathbf{Z} 의 conditional distribution을 다음과 같이 정의할 수 있다.

$$p(\mathbf{Z}|\boldsymbol{\pi}) = \prod_{n=1}^N \prod_{k=1}^K \pi_k^{z_{nk}}.$$

또한 latent variable과 모델 파라미터에 대한 observed data의 conditional distribution은 다음과 같이 정의된다.

$$p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{n=1}^N \prod_{k=1}^K \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})^{z_{nk}}$$

마지막으로 prior distribution을 다음과 같이 정의한다.

$$p(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_0) = C(\boldsymbol{\alpha}_0) \prod_{k=1}^K \pi_k^{\alpha_0 - 1} \quad p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda}) \\ = \prod_{k=1}^K \mathcal{N}(\boldsymbol{\mu}_k | \mathbf{m}_0, (\beta_0 \boldsymbol{\Lambda}_k)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_k | \mathbf{W}_0, \nu_0)$$

Variational distribution

1) 변수들에 대한 joint distribution

$$p(\mathbf{X}, \mathbf{Z}, \pi, \mu, \Lambda) = p(\mathbf{X}|\mathbf{Z}, \mu, \Lambda)p(\mathbf{Z}|\pi)p(\pi)p(\mu|\Lambda)p(\Lambda)$$

2) Factorization 가정

$$q(\mathbf{Z}, \pi, \mu, \Lambda) = q(\mathbf{Z})q(\pi, \mu, \Lambda).$$

3) 최적화 식 적용 & 1) 대입

$$\ln q^*(\mathbf{Z}) = \mathbb{E}_{\pi, \mu, \Lambda}[\ln p(\mathbf{X}, \mathbf{Z}, \pi, \mu, \Lambda)] + \text{const.}$$

$$\ln q^*(\mathbf{Z}) = \mathbb{E}_{\pi}[\ln p(\mathbf{Z}|\pi)] + \mathbb{E}_{\mu, \Lambda}[\ln p(\mathbf{X}|\mathbf{Z}, \mu, \Lambda)] + \text{const.}$$

$$\begin{aligned} \ln \rho_{nk} &= \mathbb{E}[\ln \pi_k] + \frac{1}{2} \mathbb{E}[\ln |\Lambda_k|] - \frac{D}{2} \ln(2\pi) \\ &\quad - \frac{1}{2} \mathbb{E}_{\mu_k, \Lambda_k} [(\mathbf{x}_n - \mu_k)^T \Lambda_k (\mathbf{x}_n - \mu_k)] \end{aligned}$$

4) 정의한 식들을 이용하여 solution 도출

$$q^*(\mathbf{Z}) \propto \prod_{n=1}^N \prod_{k=1}^K \rho_{nk}^{z_{nk}}, \quad q^*(\mathbf{Z}) = \prod_{n=1}^N \prod_{k=1}^K r_{nk}^{z_{nk}}$$

$$r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^K \rho_{nj}}.$$

Variational distribution

$$q^*(\mathbf{Z}) = \prod_{n=1}^N \prod_{k=1}^K r_{nk}^{z_{nk}}$$

결과를 보면 prior와 같은 functional form을 가진다는 것을 알 수 있다.

똑같은 방식으로 $q(\pi, \mu, \Lambda)$ 에 대해서도 최적화를 진행하자.

먼저 다음과 같은 변수들을 정의한다.

$$N_k = \sum_{n=1}^N r_{nk}$$

$$\bar{\mathbf{x}}_k = \frac{1}{N_k} \sum_{n=1}^N r_{nk} \mathbf{x}_n$$

$$\mathbf{S}_k = \frac{1}{N_k} \sum_{n=1}^N r_{nk} (\mathbf{x}_n - \bar{\mathbf{x}}_k)(\mathbf{x}_n - \bar{\mathbf{x}}_k)^T.$$

Variational distribution

동일한 과정을 적용하여 다음의 식을 얻을 수 있다.

$$\begin{aligned} \ln q^*(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \ln p(\boldsymbol{\pi}) + \sum_{k=1}^K \ln p(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) + \mathbb{E}_{\mathbf{Z}} [\ln p(\mathbf{Z}|\boldsymbol{\pi})] \\ &+ \sum_{k=1}^K \sum_{n=1}^N \mathbb{E}[z_{nk}] \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}) + \text{const.} \end{aligned} \quad (10.54)$$

→ 위 식에서 $\boldsymbol{\pi}$ 에 관한 term과 나머지 term은 나뉘어져 있으므로 다음 식이 성립한다.

$$q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\boldsymbol{\pi}) \prod_{k=1}^K q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k).$$

(factorization)

Variational distribution

π_i 에 관한 식을 꺼내어 정리하면 다음과 같이 된다.

$$\ln q^*(\boldsymbol{\pi}) = (\alpha_0 - 1) \sum_{k=1}^K \ln \pi_k + \sum_{k=1}^K \sum_{n=1}^N r_{nk} \ln \pi_k + \text{const}$$

이제 양 변에 exponential을 취하면 $q^*(\pi)$ 가 $\alpha_k = \alpha_0 + N_k$ 에 대한 dirichlet 분포임을 알 수 있다.

$$q^*(\boldsymbol{\pi}) = \text{Dir}(\boldsymbol{\pi} | \boldsymbol{\alpha})$$

마찬가지로 $q^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k)$ 을 살펴보면 Gaussian-Wishart distribution이 되는 것을 알 수 있다.

$$q^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \mathcal{N}(\boldsymbol{\mu}_k | \mathbf{m}_k, (\beta_k \boldsymbol{\Lambda}_k)^{-1}) \mathcal{W}(\boldsymbol{\Lambda}_k | \mathbf{W}_k, \nu_k)$$

지금까지 정리한 parameter에 대한 update equation은 EM에서 M step에 해당한다.

이렇게 responsibility를 구하고, 이를 이용하여 parameter에 대한 variational 분포를 업데이트 하는 것을 반복함으로써 variational posterior 분포에 대한 최적화가 이루어진다.

Variational lower bound

모델의 ELBO를 직접 계산하는 것도 가능하다.

$$\begin{aligned}
 \mathcal{L} &= \sum_{\mathbf{Z}} \iiint q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})}{q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})} \right\} d\boldsymbol{\pi} d\boldsymbol{\mu} d\boldsymbol{\Lambda} \\
 &= \mathbb{E}[\ln p(\mathbf{X}, \mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] - \mathbb{E}[\ln q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] \\
 &= \mathbb{E}[\ln p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] + \mathbb{E}[\ln p(\mathbf{Z}|\boldsymbol{\pi})] + \mathbb{E}[\ln p(\boldsymbol{\pi})] + \mathbb{E}[\ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda})] \\
 &\quad - \mathbb{E}[\ln q(\mathbf{Z})] - \mathbb{E}[\ln q(\boldsymbol{\pi})] - \mathbb{E}[\ln q(\boldsymbol{\mu}, \boldsymbol{\Lambda})] \tag{10}
 \end{aligned}$$

$$\begin{aligned}
 \mathbb{E}[\ln p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})] &= \frac{1}{2} \sum_{k=1}^K N_k \left\{ \ln \tilde{\Lambda}_k - D\beta_k^{-1} - \nu_k \text{Tr}(\mathbf{S}_k \mathbf{W}_k) \right. \\
 &\quad \left. - \nu_k (\bar{\mathbf{x}}_k - \mathbf{m}_k)^T \mathbf{W}_k (\bar{\mathbf{x}}_k - \mathbf{m}_k) - D \ln(2\pi) \right\}
 \end{aligned}$$

$$\mathbb{E}[\ln p(\mathbf{Z}|\boldsymbol{\pi})] = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \ln \tilde{\pi}_k$$

$$\mathbb{E}[\ln p(\boldsymbol{\pi})] = \ln C(\boldsymbol{\alpha}_0) + (\alpha_0 - 1) \sum_{k=1}^K \ln \tilde{\pi}_k$$

$$\begin{aligned}
 \mathbb{E}[\ln p(\boldsymbol{\mu}, \boldsymbol{\Lambda})] &= \frac{1}{2} \sum_{k=1}^K \left\{ D \ln(\beta_0/2\pi) + \ln \tilde{\Lambda}_k - \frac{D\beta_0}{\beta_k} \right. \\
 &\quad \left. - \beta_0 \nu_k (\mathbf{m}_k - \mathbf{m}_0)^T \mathbf{W}_k (\mathbf{m}_k - \mathbf{m}_0) \right\} + K \ln B(\mathbf{W}_0, \nu_0) \\
 &\quad + \frac{(\nu_0 - D - 1)}{2} \sum_{k=1}^K \ln \tilde{\Lambda}_k - \frac{1}{2} \sum_{k=1}^K \nu_k \text{Tr}(\mathbf{W}_0^{-1} \mathbf{W}_k)
 \end{aligned}$$

$$\mathbb{E}[\ln q(\mathbf{Z})] = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \ln r_{nk}$$

$$\mathbb{E}[\ln q(\boldsymbol{\pi})] = \sum_{k=1}^K (\alpha_k - 1) \ln \tilde{\pi}_k + \ln C(\boldsymbol{\alpha})$$

$$\mathbb{E}[\ln q(\boldsymbol{\mu}, \boldsymbol{\Lambda})] = \sum_{k=1}^K \left\{ \frac{1}{2} \ln \tilde{\Lambda}_k + \frac{D}{2} \ln \left(\frac{\beta_k}{2\pi} \right) - \frac{D}{2} - \mathbb{H}[q(\boldsymbol{\Lambda}_k)] \right\}$$

Predictive density

Bayesian mixture Gaussian model을 적용할 때 종종 새로운 데이터 포인트에 대한 Predictive density가 목적이 되곤 한다.

그렇다면 Predictive density는 이 데이터 포인트에 대한 latent variable인 z^{\wedge} 에 대해 다음과 같이 표현된다.

$$p(\hat{\mathbf{x}}|\mathbf{X}) = \sum_{\hat{\mathbf{z}}} \iiint p(\hat{\mathbf{x}}|\hat{\mathbf{z}}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\hat{\mathbf{z}}|\boldsymbol{\pi}) p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathbf{X}) d\boldsymbol{\pi} d\boldsymbol{\mu} d\boldsymbol{\Lambda}$$

이전에 사용했던 식을 이용해 z^{\wedge} 을 먼저 계산하면 다음을 얻게 된다.

$$p(\hat{\mathbf{x}}|\mathbf{X}) = \sum_{k=1}^K \iiint \pi_k \mathcal{N}(\hat{\mathbf{x}}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}) p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathbf{X}) d\boldsymbol{\pi} d\boldsymbol{\mu} d\boldsymbol{\Lambda}.$$

$p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathbf{X})$ 를 $q(\boldsymbol{\pi})q(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ 로 근사시켜 대입하게 되면 t분포의 혼합분포를 얻을 수 있다.

$$p(\hat{\mathbf{x}}|\mathbf{X}) = \frac{1}{\hat{\alpha}} \sum_{k=1}^K \alpha_k \text{St}(\hat{\mathbf{x}}|\mathbf{m}_k, \mathbf{L}_k, \nu_k + 1 - D)$$

Determining the number of components

Variational lower bound를 통해 K개의 components로 이루어진 mixture model에 대한 Posterior distribution을 도출해낼 수 있음을 배웠다.

K개의 모델의 배치만 바꾸면 동일한 결과를 가지는 다른 모델을 만들 수 있으므로 K!개의 동일한 설정이 존재한다.

따라서 서로 다른 K값에 대한 비교가 이루어질 때 multi-modality를 반영해야 한다.
간단한 해결책은 모델 비교 시 lower bound term에 $\ln(K!)$ 를 더하는 것이다.
Bayesian 프레임워크에서는 이런 식으로 trade-off가 일어나 비교가 가능해진다.

Induced factorization

Factorization assumption

$$q(\mathbf{Z}, \pi, \mu, \Lambda) = q(\mathbf{Z})q(\pi, \mu, \Lambda).$$

Additional factorization(Induced factorization)

$$q^*(\mathbf{Z}) = \prod_{n=1}^N \prod_{k=1}^K r_{nk}^{z_{nk}} \quad q(\pi, \mu, \Lambda) = q(\pi) \prod_{k=1}^K q(\mu_k, \Lambda_k).$$

- 인수분해 가정과 conditional independence 사이의 상호작용으로 발생한 결과이다.
- 학습 및 추론에 있어서 효율성을 제공한다.
- 두 변수 사이의 dseparation 관계를 이용하여 factorization 성립 여부를 알아낼 수 있다.
(graphical test)

3. Variational Linear Regression

Variational distribution

Predictive distribution

Lower bound

Variational Linear Regression

\mathbf{w} 에 대한 likelihood function과 prior distribution은 다음처럼 주어진다.

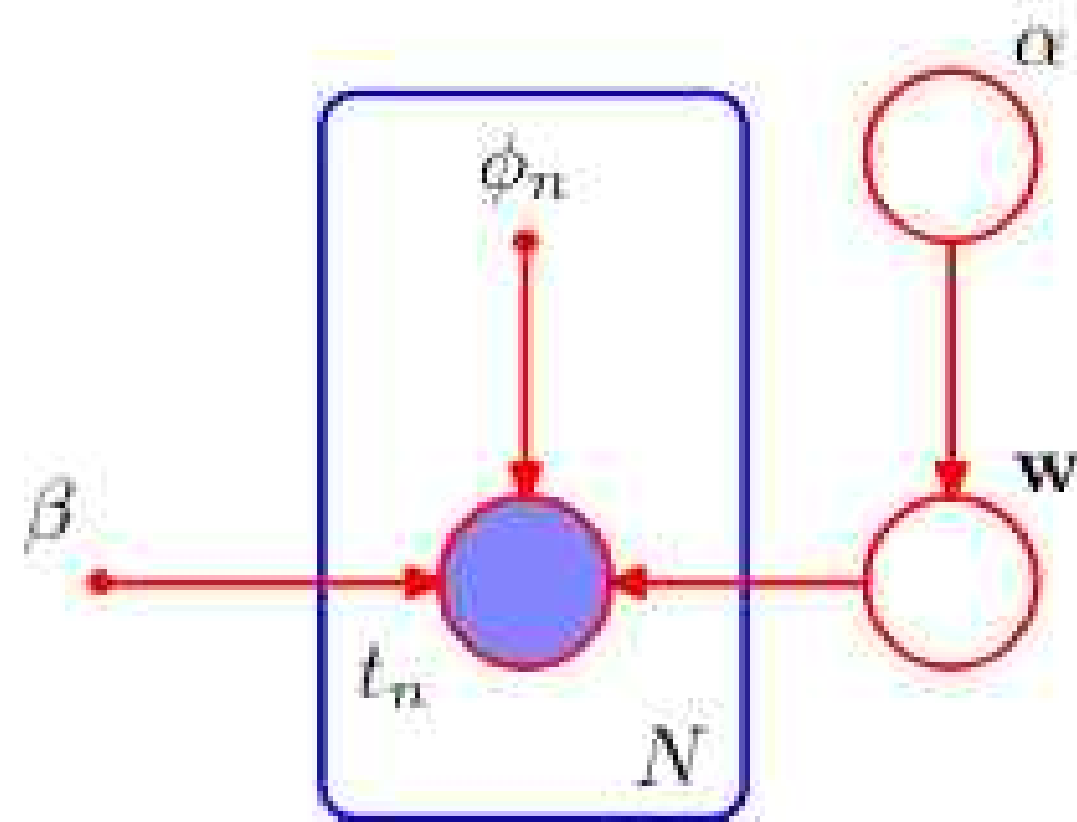
$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^N \mathcal{N}(t_n | \mathbf{w}^T \phi_n, \beta^{-1})$$

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1} \mathbf{I})$$

추가적으로 정밀도의 prior와 모든 variable들의 joint는 다음처럼 주어진다.

$$p(\alpha) = \text{Gam}(\alpha | a_0, b_0)$$

$$p(\mathbf{t}, \mathbf{w}, \alpha) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}|\alpha)p(\alpha).$$



Variational distribution

1st Goal: Posterior $p(\mathbf{w}, \alpha | \mathbf{t})$ 에 대한 근사치 찾기

1) Factorization assumption

$$q(\mathbf{w}, \alpha) = q(\mathbf{w})q(\alpha).$$

2) 각각의 variable에 대한 VI적용

$$\begin{aligned}\ln q^*(\alpha) &= \ln p(\alpha) + \mathbb{E}_{\mathbf{w}} [\ln p(\mathbf{w}|\alpha)] + \text{const} \\ &= (a_0 - 1) \ln \alpha - b_0 \alpha + \frac{M}{2} \ln \alpha - \frac{\alpha}{2} \mathbb{E}[\mathbf{w}^T \mathbf{w}] + \text{const}.\end{aligned}$$

$$q^*(\alpha) = \text{Gam}(\alpha | a_N, b_N)$$

$$\begin{aligned}a_N &= a_0 + \frac{M}{2} \\ b_N &= b_0 + \frac{1}{2} \mathbb{E}[\mathbf{w}^T \mathbf{w}].\end{aligned}$$

$$\begin{aligned}\ln q^*(\mathbf{w}) &= \ln p(\mathbf{t}|\mathbf{w}) + \mathbb{E}_{\alpha} [\ln p(\mathbf{w}|\alpha)] + \text{const} \\ &= -\frac{\beta}{2} \sum_{n=1}^N \{\mathbf{w}^T \phi_n - t_n\}^2 - \frac{1}{2} \mathbb{E}[\alpha] \mathbf{w}^T \mathbf{w} + \text{const} \\ &= -\frac{1}{2} \mathbf{w}^T (\mathbb{E}[\alpha] \mathbf{I} + \beta \Phi^T \Phi) \mathbf{w} + \beta \mathbf{w}^T \Phi^T \mathbf{t} + \text{const}.\end{aligned}$$

$$q^*(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_N, \mathbf{S}_N)$$

$$\begin{aligned}\mathbf{m}_N &= \beta \mathbf{S}_N \Phi^T \mathbf{t} \\ \mathbf{S}_N &= (\mathbb{E}[\alpha] \mathbf{I} + \beta \Phi^T \Phi)^{-1}.\end{aligned}$$

Predictive distribution

새 입력 \mathbf{x} 가 주어졌을 때의 결과값 t 에 대한 predictive distribution은 parameter에 대한 Gaussian variational posterior로 쉽게 계산할 수 있다.

$$\begin{aligned} p(t|\mathbf{x}, \mathbf{t}) &= \int p(t|\mathbf{x}, \mathbf{w})p(\mathbf{w}|\mathbf{t}) d\mathbf{w} \\ &\approx \int p(t|\mathbf{x}, \mathbf{w})q(\mathbf{w}) d\mathbf{w} \\ &= \int \mathcal{N}(t|\mathbf{w}^T \phi(\mathbf{x}), \beta^{-1}) \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) d\mathbf{w} \\ &= \mathcal{N}(t|\mathbf{m}_N^T \phi(\mathbf{x}), \sigma^2(\mathbf{x})) \end{aligned}$$

Where, $\sigma^2(\mathbf{x}) = \frac{1}{\beta} + \phi(\mathbf{x})^T \mathbf{S}_N \phi(\mathbf{x}).$

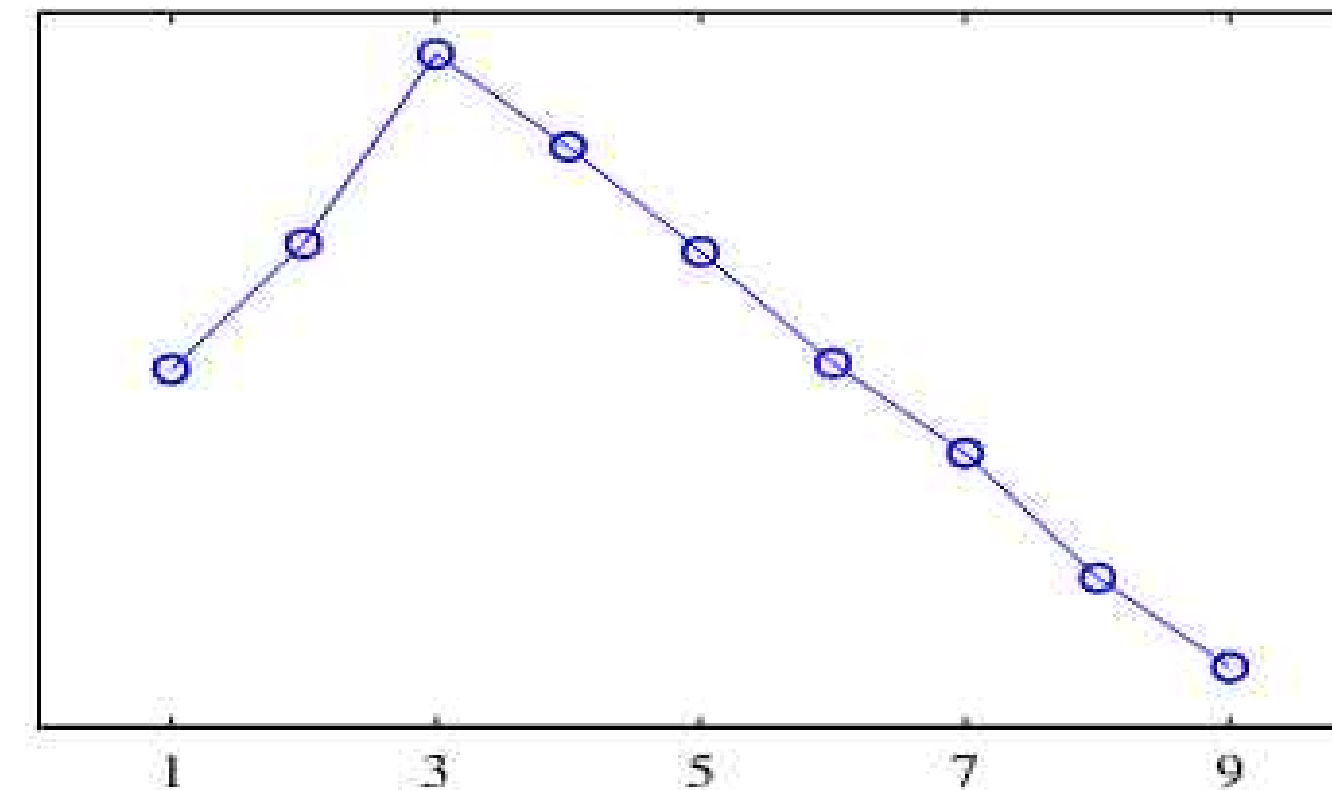
Lower bound

$$\begin{aligned}
 \mathcal{L}(q) &= \mathbb{E}[\ln p(\mathbf{w}, \alpha, \mathbf{t})] - \mathbb{E}[\ln q(\mathbf{w}, \alpha)] \\
 &= \mathbb{E}_{\mathbf{w}}[\ln p(\mathbf{t}|\mathbf{w})] + \mathbb{E}_{\mathbf{w}, \alpha}[\ln p(\mathbf{w}|\alpha)] + \mathbb{E}_{\alpha}[\ln p(\alpha)] \\
 &\quad - \mathbb{E}_{\alpha}[\ln q(\mathbf{w})]_{\mathbf{w}} - \mathbb{E}[\ln q(\alpha)].
 \end{aligned}$$

이 식의 각 항들은 앞 장들에서 구한 결과를 바탕으로 구할 수 있다.

$$\begin{aligned}
 \mathbb{E}[\ln p(\mathbf{t}|\mathbf{w})]_{\mathbf{w}} &= \frac{N}{2} \ln \left(\frac{\beta}{2\pi} \right) - \frac{\beta}{2} \mathbf{t}^T \mathbf{t} + \beta \mathbf{m}_N^T \Phi^T \mathbf{t} \\
 &\quad - \frac{\beta}{2} \text{Tr} [\Phi^T \Phi (\mathbf{m}_N \mathbf{m}_N^T + \mathbf{S}_N)] \\
 \mathbb{E}[\ln p(\mathbf{w}|\alpha)]_{\mathbf{w}, \alpha} &= -\frac{M}{2} \ln(2\pi) + \frac{M}{2} (\psi(a_N) - \ln b_N) \\
 &\quad - \frac{a_N}{2b_N} [\mathbf{m}_N^T \mathbf{m}_N + \text{Tr}(\mathbf{S}_N)] \\
 \mathbb{E}[\ln p(\alpha)]_{\alpha} &= a_0 \ln b_0 + (a_0 - 1) [\psi(a_N) - \ln b_N] \\
 &\quad - b_0 \frac{a_N}{b_N} - \ln \Gamma(a_N)
 \end{aligned}$$

$$\begin{aligned}
 -\mathbb{E}[\ln q(\mathbf{w})]_{\mathbf{w}} &= \frac{1}{2} \ln |\mathbf{S}_N| + \frac{M}{2} [1 + \ln(2\pi)] \\
 -\mathbb{E}[\ln q(\alpha)]_{\alpha} &= \ln \Gamma(a_N) - (a_N - 1) \psi(a_N) - \ln b_N + a_N.
 \end{aligned}$$



4

Exponential Family Distributions

Introduction to Probabilistic Deep Learning

Exponential Family Distribution

$$\mathbf{X}, \mathbf{Z} \rightarrow \mathbf{X}, \mathbf{Z}, \boldsymbol{\eta}$$

이제부터는 hidden variable을 latent variable(\mathbf{Z}), parameter($\boldsymbol{\eta}$)로 구분하여 생각할 것.

GMM을 예를 들면, k th gaussian에 들어있는지를 나타내는 indicator variable z_{kn} 은 latent variable이고 평균과 공분산 행렬, mixing coefficient가 parameter이다.

iid인 데이터에서 \mathbf{X}, \mathbf{Z} 의 joint dist.가 exponential family라고 가정하자.

이때의 likelihood는 다음과 같이 쓸 수 있다.

$$p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\eta}) = \prod_{n=1}^N h(\mathbf{x}_n, \mathbf{z}_n) g(\boldsymbol{\eta}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}_n, \mathbf{z}_n) \}.$$

$\boldsymbol{\eta}$ 의 prior 역시 conjugate 하게 exponential family 형태로 잡으면 다음과 같다.

$$p(\boldsymbol{\eta} | \nu_0, \boldsymbol{\chi}_0) = f(\nu_0, \boldsymbol{\chi}_0) g(\boldsymbol{\eta})^{\nu_0} \exp \{ \nu_0 \boldsymbol{\eta}^T \boldsymbol{\chi}_0 \}$$

Variational Factorization

$$q(\mathbf{Z}, \boldsymbol{\eta}) = q(\mathbf{Z})q(\boldsymbol{\eta})$$

이번엔 latent variable \mathbf{Z} 와 parameter $\boldsymbol{\eta}$ 가 factorize하는 variational distribution을 가정해보자. factorize 된 두개의 항 $q(\mathbf{Z})$ 와 $q(\boldsymbol{\eta})$ 를 각각 근사하여 합쳐 \mathbf{Z} 와 $\boldsymbol{\eta}$ 의 joint dist.를 구할것이다.

1. $q(\mathbf{Z})$

$q(\mathbf{Z})$ 는 앞에서 배운 식을 통해 근사할 수 있다. 아래 식은 앞슬라이드에서 가정한 exponential family 형태의 likelihood를 대입해준 거니까

$$\begin{aligned}\ln q^*(\mathbf{Z}) &= \mathbb{E}_{\boldsymbol{\eta}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\eta})] + \text{const} \\ &= \sum_{n=1}^N \left\{ \ln h(\mathbf{x}_n, \mathbf{z}_n) + \mathbb{E}[\boldsymbol{\eta}^T] \mathbf{u}(\mathbf{x}_n, \mathbf{z}_n) \right\} + \text{const}.\end{aligned}$$

두 번째 줄을 보면 $q^*(\mathbf{Z})$ 가 n 개의 곱으로 쪼개지는 것을 확인할 수 있는데, 이를 통해 $q^*(\mathbf{Z})$ 도 각각의 data로 만든 $q(\mathbf{z}_n)$ 으로 factorize 되는 것을 확인할 수 있다. 즉

$$q^*(\mathbf{Z}) = \prod_n q^*(\mathbf{z}_n) \text{ where } q^*(\mathbf{z}_n) = h(\mathbf{x}_n, \mathbf{z}_n) g(\mathbb{E}[\boldsymbol{\eta}]) \exp \left\{ \mathbb{E}[\boldsymbol{\eta}^T] \mathbf{u}(\mathbf{x}_n, \mathbf{z}_n) \right\}$$

EM algorithm

2. $q(\eta)$

같은 방식으로 $q(\eta)$ 도 근사시킬 수 있다.

$$\ln q^*(\eta) = \ln p(\eta | \nu_0, \chi_0) + \mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z} | \eta)] + \text{const}$$
$$q^*(\eta) = f(\nu_N, \chi_N) g(\eta)^{\nu_N} \exp \{ \eta^T \chi_N \}$$

$$\nu_N = \nu_0 + N$$
$$\chi_N = \chi_0 + \sum_{n=1}^N \mathbb{E}_{\mathbf{z}_n} [\mathbf{u}(\mathbf{x}_n, \mathbf{z}_n)].$$

3. EM

E step


$q(\mathbf{Z})$ 로 $\mathbb{E}[\mathbf{u}(\mathbf{x}_n, \mathbf{z}_n)]$ 를 구하고 이를 이용해 $q(\eta)$ 를 update

M step

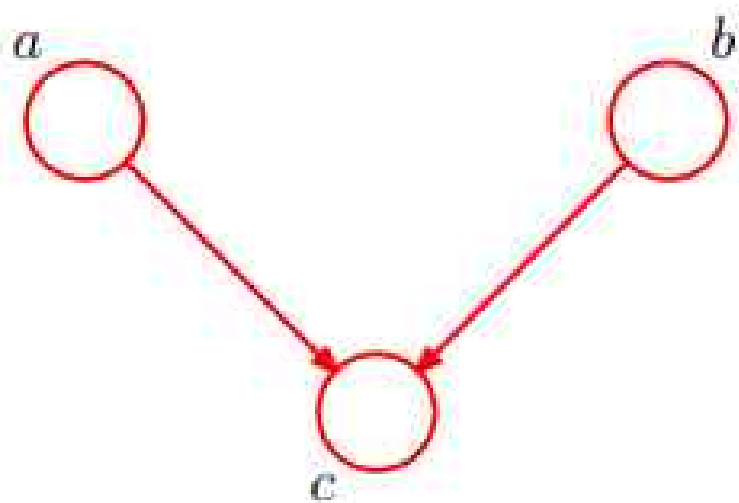
$q(\eta)$ 을 사용해 $\mathbb{E}(\eta^T)$ 를 구하고 이를 이용해 $q(\mathbf{Z})$ 를 update

$$q^*(\mathbf{z}_n) = h(\mathbf{x}_n, \mathbf{z}_n) g(\mathbb{E}[\eta]) \exp \{ \mathbb{E}[\eta^T] \mathbf{u}(\mathbf{x}_n, \mathbf{z}_n) \}$$

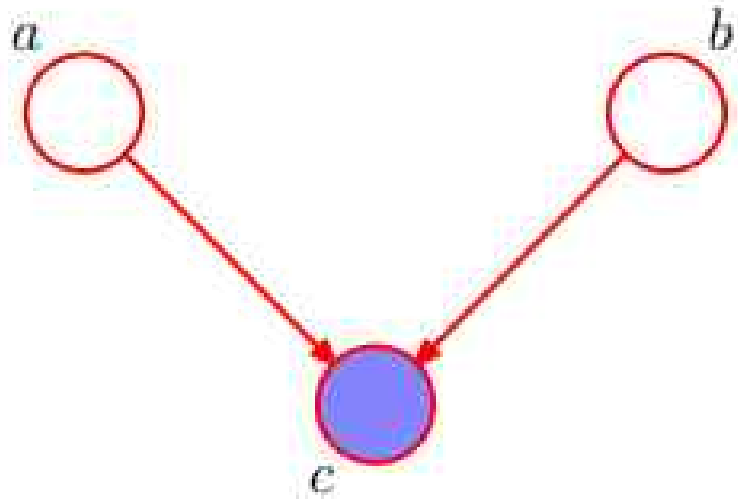
$$q^*(\eta) = f(\nu_N, \chi_N) g(\eta)^{\nu_N} \exp \{ \eta^T \chi_N \}$$

$$\chi_0 + \sum_{n=1}^N \mathbb{E}_{\mathbf{z}_n} [\mathbf{u}(\mathbf{x}_n, \mathbf{z}_n)].$$


Graphs



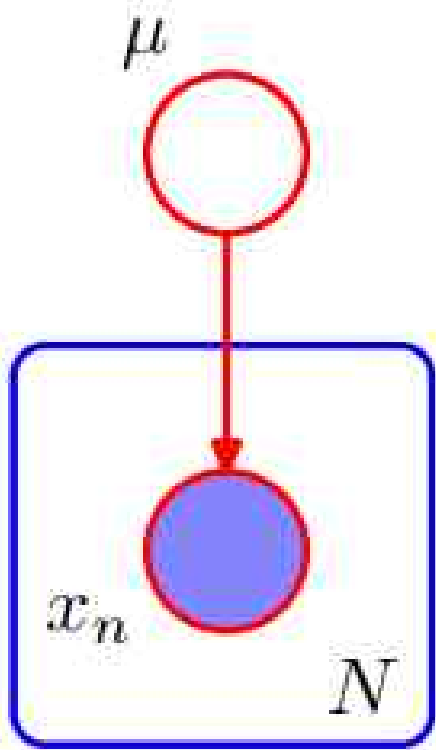
$$p(c|a,b)p(a)p(b), \text{ } a \perp b$$



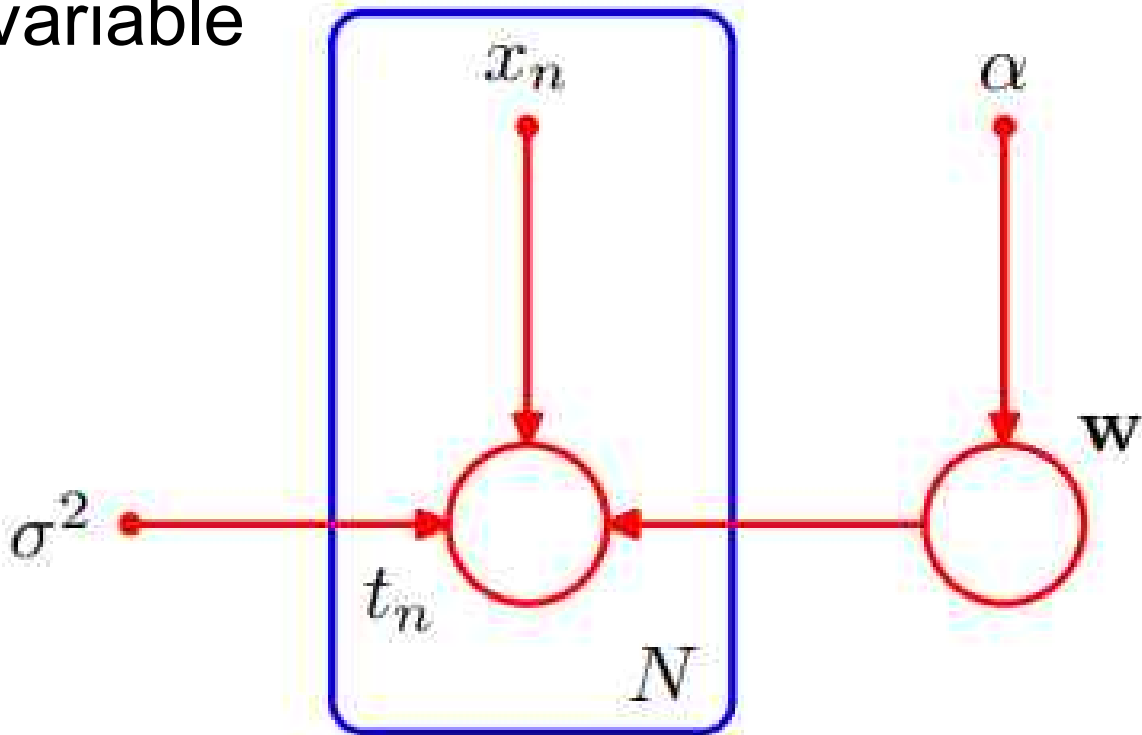
$$p(c|a,b)p(a,b) \text{ } a \not\perp b \mid c.$$

어떤 dependency를 갖는 모델인지 그림
으로 표현하는 방법

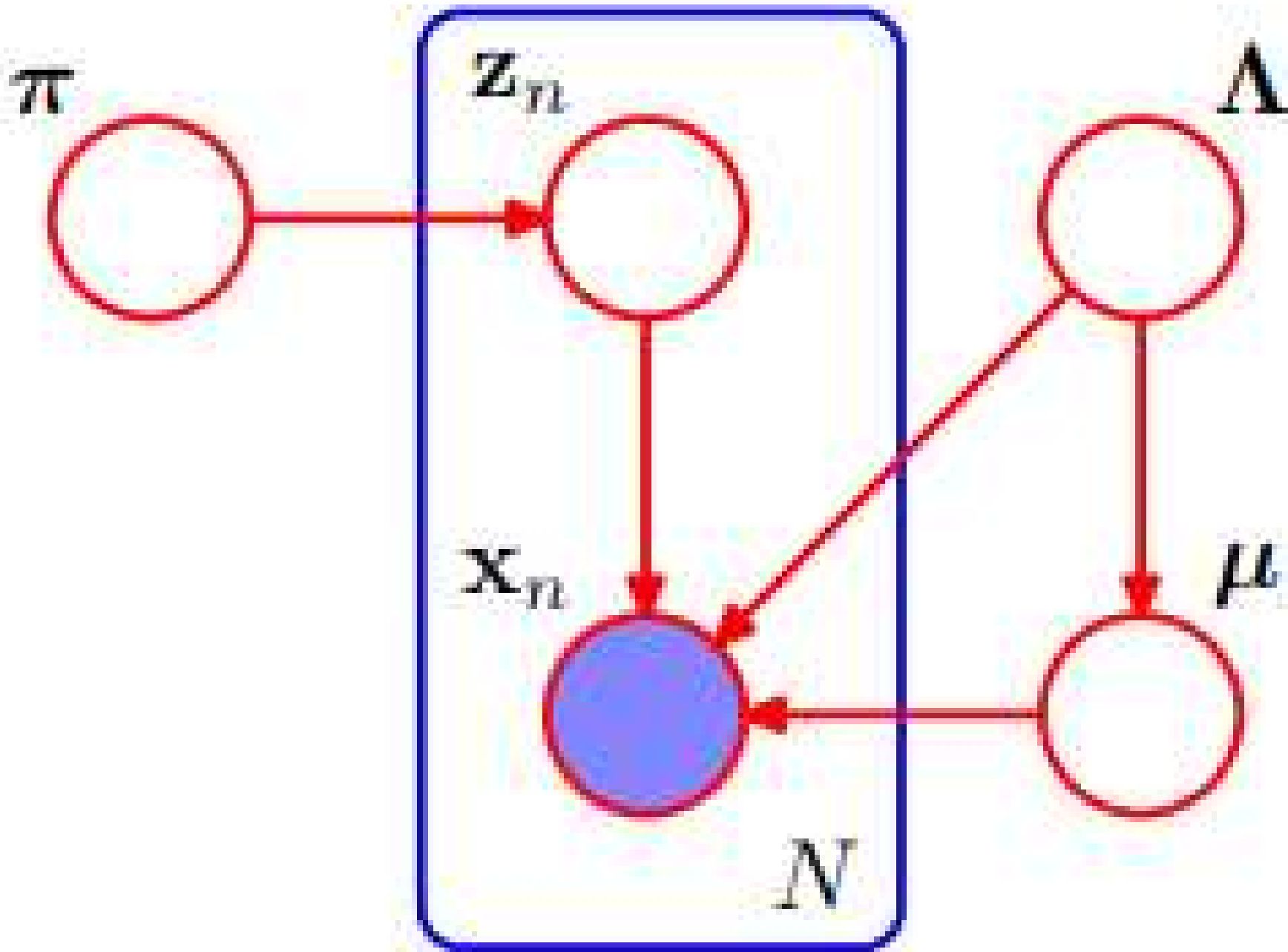
색칠 : observed
상자 밖 : hidden variable



$$p(x_1,x_2,...,x_N|\mu)p(\mu)$$

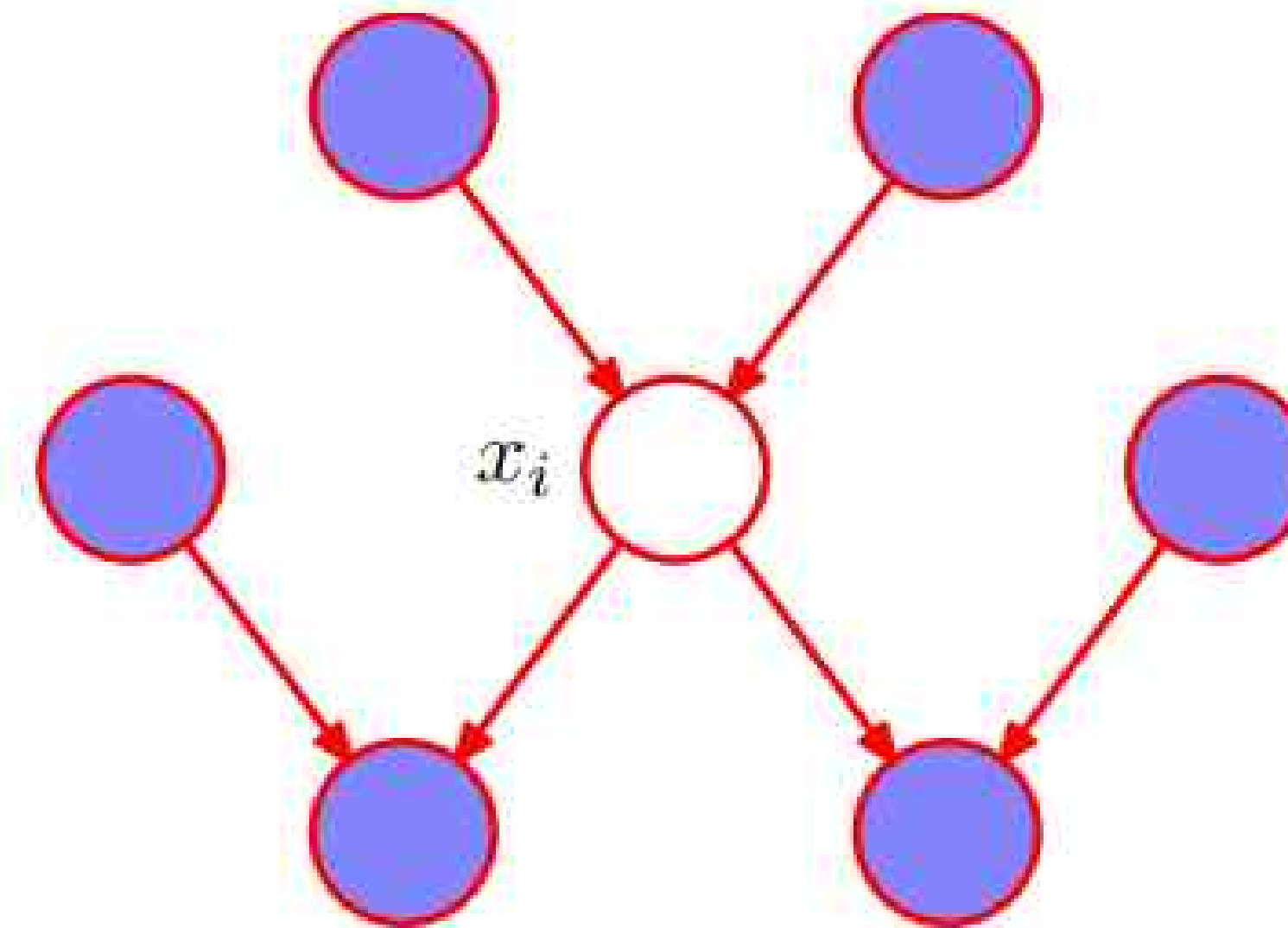


$$p(\mathbf{t}, \mathbf{w}|\mathbf{X}, \alpha, \sigma^2) = p(\mathbf{w}|\alpha) \prod_{n=1}^N p(t_n|\mathbf{w}, x_n, \sigma^2).$$



Graphs - Markov blanket

The Markov blanket of a node x_i comprises the set of parents, children and co-parents of the node. It has the property that the conditional distribution of x_i , conditioned on all the remaining variables in the graph, is dependent only on the variables in the Markov blanket.



x_i 만 주변 node들에서 isolate시킨 모양으로, 그림과 같은 dependent 구조에서 x_i 의 parents, childere, co-parents가 모두 관측된 모양이다.

Variational message passing

$$\ln q_j^*(\mathbf{x}_j) = \mathbb{E}_{i \neq j} \left[\sum_i \ln p(\mathbf{x}_i | \text{pa}_i) \right] + \text{const.}$$

좌변이 x_j 에 대한 식이므로, 우변에서 x_j 와 관련없는 term들은 모두 constant로 흡수된다. 그렇게 되면 살아남은 term들은 x_j 의 markov blanket에 해당하는 node들 뿐이고, 이를 이용해 식을 다음과 같이 쓸 수 있다.

이러한 측면에서 $q_j(x_j)$ 의 update는 local calculation(local message passing)으로 볼 수 있다.

특히 여기서 모든 conditional distribution이 conjugate-exponential structure를 가진 모델로 시선을 좁히면, variational update procedure은 local message passing 알고리즘의 형태로 표현될 수 있다고 한다.

다음 장에서 이 방법에 대해 알아보자

5

Local Variational Methods

Introduction to Probabilistic Deep Learning

Local Approximate by Bounds & Duality

예시) $y = e^x$ 를 tangent line을 lower bound로 하여 근사

$$y(x) = f(\xi) + f'(\xi)(x - \xi) \quad y(x) \leq f(x)$$

$$y(x, \lambda) = \lambda x - \lambda + \lambda \ln(-\lambda). \quad \lambda = -\exp(-\xi)$$

let $g(\lambda) = -\lambda + \lambda \ln(-\lambda)$ (intercept)

$$f(x) = \max_{\lambda} \{ \lambda x - \lambda + \lambda \ln(-\lambda) \} \quad \longleftrightarrow \quad \begin{aligned} g(\lambda) &= -\min_x \{ f(x) - \lambda x \} \\ &= \max_x \{ \lambda x - f(x) \}. \end{aligned}$$

duality

* concave의 경우 min과 max를 바꿔주기만 하면 된다

Approximate sigmoid

sigmoid의 경우는 upper bound를 구하여 approximate하면 간단히 할 수 있다. 더 중요한 lower bound를 구하는 것을 알아보자. lower bound는 $\ln \sigma(x)$ 의 lower bound를 구하게 되는데, 그 이유는 $\sigma(x)$ 는 convex도 concave도 아니지만 log를 취한 $\ln \sigma(x)$ 는 concave하기 때문이다.

$$\begin{aligned}\ln \sigma(x) &= -\ln(1 + e^{-x}) = -\ln \{e^{-x/2}(e^{x/2} + e^{-x/2})\} \\ &= x/2 - \ln(e^{x/2} + e^{-x/2}).\end{aligned}\quad \ln \sigma(x) \leq \lambda x - g(\lambda)$$

$\ln \sigma(x)$ 가 x^2 에 대해 convex이므로 x^2 의 linear function으로 lower bound를 구해보자. Duality 식은 다음과 같다.

$$g(\lambda) = \max_{x^2} \left\{ \lambda x^2 - f(\sqrt{x^2}) \right\}$$

여기서 gradient=0인 λ 를 구하면 $\lambda(\xi) = -\frac{1}{4\xi} \tanh\left(\frac{\xi}{2}\right) = -\frac{1}{2\xi} \left[\sigma(\xi) - \frac{1}{2} \right]$

이를 대입하여 $\sigma(x)$ 의 lower bound를 구하면 다음과 같다.

$$\sigma(x) \geq \sigma(\xi) \exp \left\{ (x - \xi)/2 - \lambda(\xi)(x^2 - \xi^2) \right\}$$

lower bound가 exp(x에 대한 quadratic form)이므로, $\sigma(x)$ 의 lower bound으로 Gaussian을 쓸 수 있다는 것을 알 수 있다.

Use of Approximated sigmoid

- sigmoid 함수는 log odds를 posterior로 바꿔주기 때문에 자주 나오게 된다.
- 그러나 여기서 구한 sigmoid의 lower bound는 multiclass로 확장된 softmax의 경우로 확장되어 쓸 수 없다. 이 경위는 Gibbs(1997)가 구한 방법을 사용한다.
- Bayesian logistic regression을 할때, predictive distribution을 구하는 과정에서 적용할 수 있다.

$I = \int \sigma(a)p(a) da$ 는 계산할 수 없는데, sigmoid의 lower bound를 사용하여 구하면 적분 가능한

꼴로 바꿀 수 있다.

$$I \geq \int f(a, \xi)p(a) da = F(\xi).$$

6

Variational Logistic Regression

Introduction to Probabilistic Deep Learning

Variational posterior distribution

방학세션때 Laplace 근사를 통한 Logistic regression을 진행했다면, 이번 단원에서 variational approximation을 사용할 것이다.

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w}) d\mathbf{w} = \int \left[\prod_{n=1}^N p(t_n|\mathbf{w}) \right] p(\mathbf{w}) d\mathbf{w}.$$

위와 같은 marginal likelihood의 lower bound를 maximize 해보자. 5장에서 구한 sigmoid의 lower bound를 적용하면 likelihood를 다음과 같이 쓸 수 있다.

$$p(t|\mathbf{w}) = e^{at} \sigma(-a) \geq e^{at} \sigma(\xi) \exp \left\{ -(a + \xi)/2 - \lambda(\xi)(a^2 - \xi^2) \right\}$$

이 식의 우변을 $h(\mathbf{w}, \xi)$ 라 하면, joint function에 대한 lower bound는 다음과 같다.

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}) \geq h(\mathbf{w}, \xi)p(\mathbf{w})$$

ξ : variational parameter. 관측치 마다 존재하여 ξ_n 으로 쓴다.

where ξ denotes the set $\{\xi_n\}$ of variational parameters, and

$$h(\mathbf{w}, \xi) = \prod_{n=1}^N \sigma(\xi_n) \exp \left\{ \mathbf{w}^T \phi_n t_n - (\mathbf{w}^T \phi_n + \xi_n)/2 - \lambda(\xi_n)([\mathbf{w}^T \phi_n]^2 - \xi_n^2) \right\}.$$

Variational posterior distribution

앞에서 구한 식의 문제점은, normalize 되지 않아서 pdf로 볼 수 없지만, 그렇다고 normalizing constant를 붙이게 되면 더이상 lower bound가 아닌게 된다는 모순에 빠지게 된다는 것이다. 즉, lower bound는 구했으나 그것이 pdf가 아니라는 것이다.

이는 양변에 \ln 을 취해 식을 다시 유도하면 해결될 뿐더러, posterior도 쉽게 구할 수 있게 된다.

$$\ln \{p(\mathbf{t}|\mathbf{w})p(\mathbf{w})\} \geq \ln p(\mathbf{w}) + \sum_{n=1}^N \left\{ \ln \sigma(\xi_n) + \mathbf{w}^T \phi_n t_n - (\mathbf{w}^T \phi_n + \xi_n)/2 - \lambda(\xi_n)([\mathbf{w}^T \phi_n]^2 - \xi_n^2) \right\}.$$

$$\begin{aligned} \ln\{p(\mathbf{t}|\mathbf{w})p(\mathbf{w})\} - \ln p(\mathbf{w}) = \ln p(\mathbf{t}|\mathbf{w}) \geq & -\frac{1}{2}(\mathbf{w} - \mathbf{m}_0)^T \mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0) \\ & + \sum_{n=1}^N \left\{ \mathbf{w}^T \phi_n (t_n - 1/2) - \lambda(\xi_n) \mathbf{w}^T (\phi_n \phi_n^T) \mathbf{w} \right\} + \text{const.} \end{aligned}$$

log likelihood가 quadratic function of \mathbf{w} 이므로 likelihood는 $\exp(\text{quadratic form of } \mathbf{w})$, 즉 Gaussian이라고 볼 수 있다. prior 역시 conjugate하게 Gaussian으로 잡으면 posterior $q(\mathbf{w})$ 는 Gaussian임을 알 수 있다.

Variational posterior distribution

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_N, \mathbf{S}_N)$$

where

$$\mathbf{m}_N = \mathbf{S}_N \left(\mathbf{S}_0^{-1} \mathbf{m}_0 + \sum_{n=1}^N (t_n - 1/2) \phi_n \right)$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + 2 \sum_{n=1}^N \lambda(\xi_n) \phi_n \phi_n^T.$$

여기서 쓰인 variational method는 Laplace와는 다르게 관측치마다 variational parameter ξ_n 이 있어 좀더 flexible하고, 따라서 정확도도 더욱 좋다.

그러나 sequential하게 data point를 한번 쓰고 버리는 식이기 때문에 Batch learning에 관점에선 적합하지 않다. 또한 5장에서 말했듯 multiclass로 확장되지 않는다는 단점이 있다(이 경우 Gibbs(1997)을 사용)

Optimizing the variational parameter ξ - EM algorithm

EM algorithm을 통해 Expected complete-data log likelihood $Q(\xi, \xi^{\text{old}}) = \mathbb{E} [\ln h(\mathbf{w}, \xi)p(\mathbf{w})]$ 를 maximize 하고, 이를 이용해 다시 ξ 를 구하는 식으로 하여 ξ 를 최적화할 것이다. 우선 $\{\xi_n\}$ 의 초기값을 임의로 배정하자. 이를 ξ -old라 하자.

E step

ξ 를 통해 w 의 posterior를 찾는다.($q(w)$)
이를 이용해 $Q(\xi, \xi\text{-old})$ 를 구한다.

$$Q(\xi, \xi^{\text{old}}) = \mathbb{E} [\ln h(\mathbf{w}, \xi)p(\mathbf{w})]$$

좌변이 $q(w)$ 에 대한 기댓값이기 때문이다.

M-step

$Q(\xi, \xi\text{-old})$ 의 gradient=0을 이용해, $Q(\xi, \xi\text{-old})$ 를 maximize 하는 ξ 를 구한다.

$$(\xi_n^{\text{new}})^2 = \phi_n^T \mathbb{E}[\mathbf{w}\mathbf{w}^T] \phi_n = \phi_n^T (\mathbf{S}_N + \mathbf{m}_N \mathbf{m}_N^T) \phi_n$$

Optimizing the variational parameter ξ - Another approach

$$\ln p(\mathbf{t}) = \ln \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w}) d\mathbf{w} \geq \ln \int h(\mathbf{w}, \xi)p(\mathbf{w}) d\mathbf{w} = \mathcal{L}(\xi).$$

식의 우변의 형태를 직접 구할 수 있다. 적분 안이 $\exp(\text{quadratic form of } w)$ * Gaussian 이므로 gaussian으로 본뒤 normalizing constant를 추가하여 적분 안의 식을 구할 수 있고, 따라서 marginalizing 및 ln 계산까지 하면 $L(\xi)$ 를 구할 수 있다.

이렇게 구한 $L(\xi)$ 를 ξ 에 대해 미분하여 이 값을 maximize하는 ξ 를 구하면 된다.

이는 앞장의 EM algorithm으로 구한값과 비슷하게 나온다.

Inference of hyperparameters

지금까지는 prior에 있는 hyperparameter를 아는 것으로 취급했다(hyperparameter 예시 : w 의 prior을 gaussian으로 가정했을 때 precision). 이제부터는 그 hyper parameter까지도 고려하여 posterior를 찾아보자.

우선 notation 및 prior을 다음과 같다고 하자.

α : hyperparameter, \mathbf{w} : parameter

$p(\mathbf{w}|\alpha) = N(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$ (simple isotropic gaussian prior)

$p(\alpha) = \text{Gam}(\alpha|a, b)$

Marginal likelihood $p(\mathbf{t}) = \iint p(\mathbf{w}, \alpha, \mathbf{t}) d\mathbf{w} d\alpha$,

where $p(\mathbf{w}, \alpha, \mathbf{t}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}|\alpha)p(\alpha)$

Inference of hyperparameters

$$\ln p(\mathbf{t}) = \mathcal{L}(q) + \text{KL}(q||p) \quad (10.169)$$

where the lower bound $\mathcal{L}(q)$ and the Kullback-Leibler divergence $\text{KL}(q||p)$ are defined by

$$\mathcal{L}(q) = \iint q(\mathbf{w}, \alpha) \ln \left\{ \frac{p(\mathbf{w}, \alpha, \mathbf{t})}{q(\mathbf{w}, \alpha)} \right\} d\mathbf{w} d\alpha \quad (10.170)$$

$$\text{KL}(q||p) = - \iint q(\mathbf{w}, \alpha) \ln \left\{ \frac{p(\mathbf{w}, \alpha | \mathbf{t})}{q(\mathbf{w}, \alpha)} \right\} d\mathbf{w} d\alpha. \quad (10.171)$$

$$p(\mathbf{w}, \alpha, \mathbf{t}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}|\alpha)p(\alpha)$$

빨간색으로 칠한 likelihood $p(\mathbf{t}|\mathbf{w})$ (sigmoid)

때문에 $\mathcal{L}(q)$ 계산 불가능

Inference of hyperparameters - marginal likelihood의 lower bound

$$\begin{aligned}\ln p(\mathbf{t}) &\geq \mathcal{L}(q) \geq \tilde{\mathcal{L}}(q, \boldsymbol{\xi}) \\ &= \iint q(\mathbf{w}, \alpha) \ln \left\{ \frac{h(\mathbf{w}, \boldsymbol{\xi}) p(\mathbf{w}|\alpha) p(\alpha)}{q(\mathbf{w}, \alpha)} \right\} d\mathbf{w} d\alpha.\end{aligned}$$

• 참고

$$h(\mathbf{w}, \boldsymbol{\xi}) = \prod_{n=1}^N \sigma(\xi_n) \exp \left\{ \mathbf{w}^T \boldsymbol{\phi}_n t_n - (\mathbf{w}^T \boldsymbol{\phi}_n + \xi_n)/2 \right. \\ \left. - \lambda(\xi_n)([\mathbf{w}^T \boldsymbol{\phi}_n]^2 - \xi_n^2) \right\}.$$

앞에서 sigmoid의 lower bound를 찾은 것을 통해 $L(q)$ 의 lower bound를 구할 수 있다.

이제 $q(\mathbf{w}, \alpha)$ 를 근사해보자.

Inference of hyperparameters - $q(\mathbf{w}, \alpha)$

1. Assume factorization $q(\mathbf{w}, \alpha) = q(\mathbf{w})q(\alpha)$

이러한 factorization이 가능한 variational model을 가정하고 각각의 항 $q(\mathbf{w})$, $q(\alpha)$ 를 expectation으로 나타내어 보자.

$$\begin{aligned}\ln q(\mathbf{w}) &= \mathbb{E}_{\alpha} [\ln \{h(\mathbf{w}, \boldsymbol{\xi})p(\mathbf{w}|\alpha)p(\alpha)\}] + \text{const} \\ &= \ln h(\mathbf{w}, \boldsymbol{\xi}) + \mathbb{E}_{\alpha} [\ln p(\mathbf{w}|\alpha)] + \text{const}.\end{aligned}$$

$h(\mathbf{w}, \boldsymbol{\xi})$ 는 $\exp(\mathbf{w}$ 의 quadratic form), $p(\mathbf{w}|\alpha)$ 는 gaussian이므로 $q(\mathbf{w})$ 는 gaussian임을 알 수 있다.

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$$

where we have defined

$$\begin{aligned}\boldsymbol{\Sigma}_N^{-1} \boldsymbol{\mu}_N &= \sum_{n=1}^N (t_n - 1/2) \boldsymbol{\phi}_n \\ \boldsymbol{\Sigma}_N^{-1} &= \mathbb{E}[\alpha] \mathbf{I} + 2 \sum_{n=1}^N \lambda(\xi_n) \boldsymbol{\phi}_n \boldsymbol{\phi}_n^T.\end{aligned}$$

$$\ln q(\alpha) = \mathbb{E}_{\mathbf{w}} [\ln p(\mathbf{w}|\alpha)] + \ln p(\alpha) + \text{const}.$$

여기서는 Gaussian 과 gamma의 결합이므로 gamma 분포로 나타낼 수 있다.

$$q(\alpha) = \text{Gam}(\alpha | a_N, b_N) = \frac{1}{\Gamma(a_0)} a_0^{b_0} \alpha^{a_0-1} e^{-b_0 \alpha}$$

where

$$\begin{aligned}a_N &= a_0 + \frac{M}{2} \\ b_N &= b_0 + \frac{1}{2} \mathbb{E}_{\mathbf{w}} [\mathbf{w}^T \mathbf{w}].\end{aligned}$$

Inference of hyperparameters - optimize ξ_n

$$\ln p(\mathbf{t}) \geq \mathcal{L}(q) \geq \tilde{\mathcal{L}}(q, \xi)$$

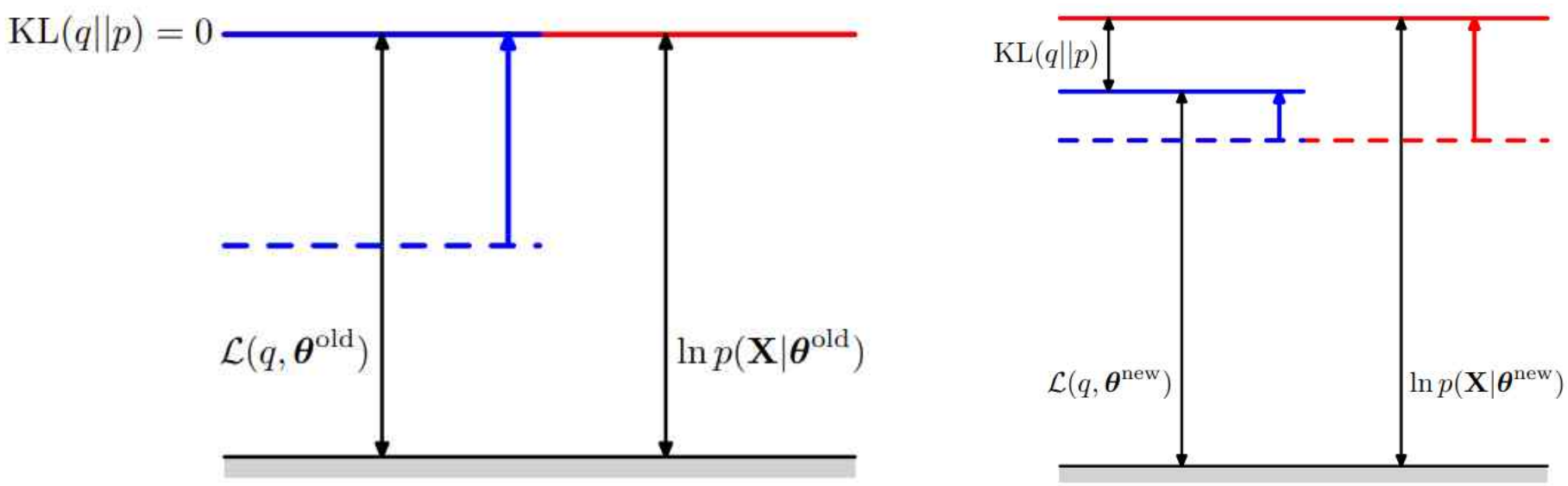
$$= \iint q(\mathbf{w}, \alpha) \ln \left\{ \frac{h(\mathbf{w}, \xi) p(\mathbf{w}|\alpha) p(\alpha)}{q(\mathbf{w}, \alpha)} \right\} d\mathbf{w} d\alpha.$$

$$\tilde{\mathcal{L}}(q, \xi) = \int q(\mathbf{w}) \ln h(\mathbf{w}, \xi) d\mathbf{w} + \text{const.}$$

$$(\xi_n^{\text{new}})^2 = \phi_n^T (\Sigma_N + \mu_N \mu_N^T) \phi_n.$$

EM

$q(\mathbf{w}) \leftrightarrow \xi, q(\alpha)$



1. ξ 와 관련 없는 애들은 const.로 보내버리고 α 에 대해 적분
2. 구한 $L(q, \xi)$ 을 maximize하는 ξ 를 찾음
3. 지금까지 구한 $q(\mathbf{w}), q(\alpha), \xi$ 를 돌아가면서 update

7

Expectation Propagation

Introduction to Probabilistic Deep Learning

EP - Another deterministic approximation inference

minimize $KL(p||q)$ by *moment matching*

moment matching 이란?

q의 모수(mean, covariance matrix 등 moment로 구해지는 모수)를 p와 같게 설정해줌으로써 KL divergence를 줄이는 방법.

Expectation Propagation

We are given a joint distribution over observed data \mathcal{D} and stochastic variables $\boldsymbol{\theta}$ in the form of a product of factors

$$p(\mathcal{D}, \boldsymbol{\theta}) = \prod_i f_i(\boldsymbol{\theta}) \quad (10.202)$$

and we wish to approximate the posterior distribution $p(\boldsymbol{\theta}|\mathcal{D})$ by a distribution of the form

$$q(\boldsymbol{\theta}) = \frac{1}{Z} \prod_i \tilde{f}_i(\boldsymbol{\theta}). \quad (10.203)$$

We also wish to approximate the model evidence $p(\mathcal{D})$.

1. Initialize all of the approximating factors $\tilde{f}_i(\boldsymbol{\theta})$.
2. Initialize the posterior approximation by setting

$$q(\boldsymbol{\theta}) \propto \prod_i \tilde{f}_i(\boldsymbol{\theta}). \quad (10.204)$$

minimize $\text{KL} \left(\frac{f_j(\boldsymbol{\theta}) q^{\setminus j}(\boldsymbol{\theta})}{Z_j} \parallel q^{\text{new}}(\boldsymbol{\theta}) \right)$ by moment matching

3. Until convergence:

- (a) Choose a factor $\tilde{f}_j(\boldsymbol{\theta})$ to refine.
- (b) Remove $\tilde{f}_j(\boldsymbol{\theta})$ from the posterior by division

$$q^{\setminus j}(\boldsymbol{\theta}) = \frac{q(\boldsymbol{\theta})}{\tilde{f}_j(\boldsymbol{\theta})}.$$

- (c) Evaluate the new posterior by setting the sufficient statistics (moments) of $q^{\text{new}}(\boldsymbol{\theta})$ equal to those of $q^{\setminus j}(\boldsymbol{\theta}) f_j(\boldsymbol{\theta})$, including evaluation of the normalization constant

$$Z_j = \int q^{\setminus j}(\boldsymbol{\theta}) f_j(\boldsymbol{\theta}) d\boldsymbol{\theta}. \quad (10.206)$$

- (d) Evaluate and store the new factor

$$\tilde{f}_j(\boldsymbol{\theta}) = Z_j \frac{q^{\text{new}}(\boldsymbol{\theta})}{q^{\setminus j}(\boldsymbol{\theta})}. \quad (10.207)$$

4. Evaluate the approximation to the model evidence

$$p(\mathcal{D}) \simeq \int \prod_i \tilde{f}_i(\boldsymbol{\theta}) d\boldsymbol{\theta}. \quad (10.208)$$

EP vs Variational Bayes

EP	Variational Bayes
<ul style="list-style-type: none">- converge 한다는 보장이 없음- converge한다 해도, 매 iteration마다 energy ft이 감소하지는 않음 (solution은 energy ft의 stationary point긴함)	<ul style="list-style-type: none">-iteration마다 적어도 lower bound가 감소는 안함.
<ul style="list-style-type: none">- minimize $KL(p q)$	<ul style="list-style-type: none">- minimize $KL(q p)$

- p를 q와 비슷하게 하는 방법이므로, p가 multimodal인 경우(ex mixture) p가 sensible 하지 않아 성능이 좋지 않음. posterior의 모든 mode를 캡처하려하기 때문. 그치만 logistic의 경우 EP가 local variational method, Laplace보다도 성능이 좋음.

감사합니다