Variational Inference (VI)

- In Bayesian learning, finding posterior $p(z|x)$ is difficult or finding moments w.r.t. $p(z|x)$ is difficult.

- So we need approximations
  - VI is such an approximation.
  - There are other techniques : (a) Sampling
    (b) Expectation propagation

- Sampling (or MCMC) is quite accurate or infact exact if we can sample many.

- Goal of VI : To find best possible approximation of the exact posterior distribution.

$$q(z) \approx p(z|x)$$

- How to do that ? By some optimality criterion.

For VI, we need some assumption:

- We must need to have (or formulate) $p(X|Z)$.
- We also need to have priors $p(Z)$.
- So we have $p(X,Z) = p(X|Z)\,p(Z)$ expression.

Example: We use GMM case.

prior: $p(\underline{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$ $\quad \underline{z}$ is a 1-of-K representation vector.

conditional distribution: $p(\underline{x}|z_k=1) = \mathcal{N}(\underline{x}|\mu_k, \Sigma_k)$, $z_k \in \{0,1\}$

or $p(\underline{x}|\underline{z}) = \prod_{k=1}^{K} \mathcal{N}(\underline{x}|\mu_k, \Sigma_k)^{z_k}$

cf $p(x) = \sum_z p(x|z)\,p(z)$

$p(x|\mu,\Sigma,\pi) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k,\Sigma_k)$.

$\therefore$ $p(\underline{x},\underline{z}) = p(\underline{x}|\underline{z}) \cdot p(\underline{z})$

$= \prod_{k=1}^{K} \{\pi_k \mathcal{N}(\underline{x}|\mu_k,\Sigma_k)\}^{z_k}$.

$X = [\underline{x}_1\ \underline{x}_2 \cdots \underline{x}_N]^T$ $\quad Z = [\underline{z}_1\ \underline{z}_2 \cdots \underline{z}_N]^T$

$\therefore$ $p(X,Z) = \prod_{n=1}^{N} p(\underline{x}_n, \underline{z}_n)$

$p(X,Z|\pi,\mu,\Sigma) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\underline{x}_n|\mu_k,\Sigma_k)^{z_{n,k}}$

- Now, if we need $p(Z|X,\pi,\mu,\Sigma)$, that is a bit trouble.
- Just for the sake of an example, we want $q(Z)$ that best approximates $p(Z|X, \text{the parameters})$.
- Another example: $\underline{z} = [\underline{z}_1^T\ \underline{z}_2^T \cdots \underline{z}_Q^T]^T$. Then if we want $p(\underline{z}_i|X,\cdots) = \int_{\underline{z}_j, j\neq i} p(Z|X,\pi,\mu,\Sigma)\,dz_j$, this is more difficult due to complex dependencies between $\{\underline{z}_i\}$.

Optimization formulation

Notation: $E_q[h(z)] = \int q(z)\, h(z)\, dz$

$$\ln p(x) = E_q[\ln p(x)] \quad \rightarrow \text{because } p(x) \text{ does not depend on } q(z). \text{ The equality is valid for any } q.$$

$$= E_q\left[\ln \frac{p(x,z)}{p(z|x)}\right]$$

$$= E_q\left[\ln \frac{p(x,z)}{q(z)}\right] - E_q\left[\ln p(z|x)\right]$$

$$= E_q\left[\ln \frac{p(x,z)}{q(z)} \cdot \frac{q(z)}{p(z|x)}\right]$$

$$= \underbrace{E_q\left[\ln \frac{p(x,z)}{q(z)}\right]}_{L(q)} + \underbrace{E_q\left[\ln \frac{q(z)}{p(z|x)}\right]}_{KL(q \| p(z|x))}$$

$$= L(q) + KL(q \| p(z|x)).$$

As $KL(q \| p(z|x)) \geq 0$, $\Rightarrow L(q)$ works as a LB.

- Note $\ln p(x)$ is fixed. So, if we define the optimization problem (easy) as

$$q^* = \arg\max_q L(q) = \arg\max_q E_q\left[\ln \frac{p(x,z)}{q(z)}\right],$$

then the $KL(q^* \| p(z|x))$ will be minimum.

By this logic q^* approaches to $p(z|x)$ or

$$q^* \approx p(z|x).$$

Factorized approximations

$$z = \left\{ z_1, z_2 \cdots z_M \right\}$$

$$p(z|x) \approx q(z) = \prod_{i=1}^{M} q_i(z_i) .$$

assumption : z_is are statistically independent.

Task : find $\{q_i^*\}_{i=1}^{M}$. Then for $q^* = \prod_{i=1}^{M} q_i^*$.

Statement : Optimal density for each group (i.e. q_i^*) can be found iteratively, by choosing

$$\ln q_i^*(z_i) = E_{q_{j\neq i}} \left[\ln p(x,z) \right] + \text{const}$$

$$= E_{q_{j\neq i}} \left[\ln p(x, z_1 z_2 \cdots z_{i-1} z_i z_{i+1} \cdots z_M \right] + c$$

→ Here 'c' is a normalization constant, and $E_{q_{j\neq i}}[\cdot]$ means that the q_js are kept fixed for $j \neq i$, and these q_js are used to calculate expectation over all those other group of variables z_j, except z_i.

→ For $i = 1, 2, \cdots, M$, this is repeated for each q_i, keeping all other $q_{j\neq i}$ fixed.

→ optimality : $$\{q_i^*\} = \underset{\{q_i\}}{\arg\max}\ E_{q_i} \left[\ln \frac{p(x,z)}{\prod_{i=1}^{M} q_i(z_i)} \right]$$

→ $$q_i^*(z_i) = \frac{\exp\left[E_{q_{j\neq i}} \left[\ln p(x,z) \right] \right]}{\int_{z_i} \exp\left[E_{q_{j\neq i}} \left[\ln p(x,z) \right] \right] dz_i} .$$

Theorem : (Theorem 9.2 of Arnie's compendium).

Assuming $Z = [Z_1\ Z_2]$, and given $p(Z,X) = p(Z_1, Z_2, X)$, we want to have $q_r(Z) = q_{r_1}(Z_1)\, q_{r_2}(Z_2) \approx p(Z|X)$. Then, for any fixed q_{r_2}, a density function q_{r_1}, obtained as

$$\ln q_{r_1}(Z_1) = E_{q_{r_2}}\left[\ln p(Z,X)\right] + c$$
$$= E_{q_{r_2}}\left[\ln p(Z_1, Z_2, X)\right] + c \quad \cdots ①$$

maximizes $L(q_r) = E_{q_r}\left[\ln \frac{p(Z,X)}{q_r(Z)}\right]$. $\cdots ②$

proof: • After taking expectation over Z_2 in ①, the remaining expression is a function of Z_1. With proper normalization, this function can be interpreted as the logarithm of a density function, called $\tilde{p}(Z_1)$ here.

$$\text{So, } E_{q_{r_2}}\left[\ln p(Z_1 Z_2 X)\right] = \int_{Z_2} q_{r_2}(Z_2) \ln p(Z_1, Z_2, X)\, dZ_2$$
$$= f(Z_1, X) + \text{const}$$
$$= \ln \tilde{p}(Z_1) + \text{const}$$

(X is fixed and $\tilde{p} \geq 0$.)

$$\text{Now, } L(q_r) = \iint_{Z_1 Z_2} q_{r_1}(Z_1)\, q_{r_2}(Z_2) \ln p(Z,X)\, dZ_1 dZ_2 - \iint_{Z_1 Z_2} q_{r_1}(Z_1) q_{r_2}(Z_2) \left[\ln q_{r_1}(Z_1) + \ln q_{r_2}(Z_2)\right] dZ_1 dZ_2$$

$$= \int_{Z_1} q_{r_1}(Z_1) \underbrace{\left[\int q_{r_2}(Z_2) \ln p(Z,X)\, dZ_2\right]}_{E_{q_{r_2}}[\ln p(Z,X)] = \ln \tilde{p}(Z_1) + \text{const}} dZ_1 - \int_{Z_1} q_{r_1}(Z_1) \ln q_{r_1}(Z_1) \underbrace{\int_{Z_2} q_{r_2}(Z_2)\, dZ_2}_{1} dZ_1$$
$$- \underbrace{\int_{Z_1} q_{r_1}(Z_1)\, dZ_1}_{1} \underbrace{\int_{Z_2} q_{r_2}(Z_2) \ln q_{r_2}(Z_2)\, dZ_2}_{\text{const}}$$

$$= \int_{Z_1} q_{r_1}(Z_1) \ln \tilde{p}(Z_1) + \text{const} \underbrace{\int_{Z_1} q_{r_1}(Z_1)\, dZ_1}_{1} - \int_{Z_1} q_{r_1}(Z_1) \ln q_{r_1}(Z_1)\, dZ_1 - \text{const}$$

$$= \int_{Z_1} q_{r_1}(Z_1) \ln \frac{\tilde{p}(Z_1)}{q_{r_1}(Z_1)} + \text{const} = -\left[-\int_{Z_1} q_{r_1}(Z_1) \ln \frac{\tilde{p}(Z_1)}{q_{r_1}(Z_1)}\right] + \text{const}$$

$$= -KL\left(q_{r_1} \| \tilde{p}\right) + \text{const}.$$

- Note that $q_1(z_1) = \tilde{p}(z_1)$ leads to $KL = 0$, and hence maximizes $L(q)$.
- So the optimal choice is $\ln q_1(z_1) = \ln \tilde{p}(z_1)$

$$= E_{q_2}\left(\ln p(z, x)\right) + \text{const.}$$

(Theorem proved).

- This proof covers general statement:

$$\ln q_i^*(z_i) = E_{q_{j \neq i}}\left[\ln p(x, z)\right] + \text{const.}$$

where we use $q_i = q_1$ and $q_{j \neq i} = q_2$.

Comment on the use of VI for Gaussian mixture (section 10.2 of Bishop Book).

It is shown that the use of VI with factorized distribution leads to two optimization steps analogous to E and M steps of the maximum likelihood EM algorithm.

EM - Special case of VI (from Arne's compendium)

- Let us explore relation between EM and VI.

- First, we recall the EM algo. (General EM)

Given a joint distribution $p(X,Z|\theta)$ over observed variables X and latent variables Z, governed by the parameters θ, the goal is to maximize likelihood $p(X|\theta)$ w.r.t. θ. That is $\theta^* = \arg\max_\theta p(X|\theta)$.

1. Choose an initial θ^{old}.
2. E Step: Evaluate $p(Z|X,\theta^{old})$.
3. M Step: Define $Q(\theta,\theta^{old}) = \sum_Z p(Z|X,\theta^{old}) \ln p(X,Z|\theta)$

$$\theta^{new} = \arg\max_\theta Q(\theta,\theta^{old}).$$

4. Check convergence. If not converged $\theta^{old} \leftarrow \theta^{new}$.

• Now we investigate VI & EM connection

Assumptions: ⓐ $p(X|Z,\theta)$ is known

ⓑ Prior $p(Z|\theta)$ and $p(\theta)$ are known.

$\therefore \ln p(X,Z,\theta) = \ln p(X|Z\theta)\, p(Z|\theta)\, p(\theta)$ is known.

→ Let us assume the goal is to estimate a point estimate $\hat{\theta}$.
(in analogy with the EM)

↓ VI factorization

→ $p(\theta,Z|X) \approx q(\theta,Z) = q_1(\theta|\hat{\theta})\, q_2(Z) = q_1(\theta|\hat{\theta},X)\, q_2(Z|\hat{\theta},X)$

$\hat{\theta}$ is a hyper-parameter

Why $q_1(\theta|\hat{\theta})$? We force the posterior density q_1 to be very sharply peaked at $\hat{\theta}$ so we model

$$q_1(\theta|\hat{\theta}) = \prod_K \frac{1}{e_K}\, \delta\left(\frac{\theta_K - \hat{\theta}_K}{e_K}\right) \quad \rightarrow \text{Dirac Delta}$$

$$\boxed{E_{q_1}[g(\theta)] = \int g(\theta)\, q_1(\theta|\hat{\theta})\, d\theta \rightarrow g(\hat{\theta})}$$

→ Strategy: In each step of VI iteration, we apply the general factorization solution $\ln q_i^*(z_i) = E_{q_{j\neq i}}[\ln p(X,Z,\theta)]$
where we assume $Z = (Z, \theta)$.

ⓐ Let we have $\hat{\theta} = \theta^{old}$.

ⓑ $q_1(\theta|\hat{\theta}) = q_1(\theta|\theta^{old})$.

ⓒ $\ln q_2(Z) = E_{q_1}[\ln p(X,Z,\theta^{old})] + \text{const} = \ln p(X,Z,\theta^{old}) + \text{const}.$

Using proper normalization, $q_2(Z) \propto p(X,Z,\theta^{old}) \propto p(Z|X,\theta^{old})$
compare with E step of EM.

ⓓ To find a new point estimate $\hat{\theta} = \theta'$, we have $q_1(\theta|\hat{\theta}) = q_1(\theta|\theta')$.

Use the relation by lower bound

$$\theta^{new} = \arg\max_{\theta'} L(q) \quad L(q) = E_q\left[\ln \frac{p(X,Z,\theta)}{q_1(\theta|\theta')\cdot q_2(Z)}\right].$$

$$L(q) = E_q\left[\ln \frac{p(x,z,\theta)}{q_1(\theta|\theta')\, q_2(z)}\right]$$

$$= \int_{\theta,z} q_1(\theta|\theta')\, q_2(z) \left[\ln p(x,z,\theta) - \ln q_1(\theta|\theta') - \ln q_2(z)\right] d\theta\, dz$$

$$= \int_z q_2(z) \left[\int_\theta q_1(\theta|\theta') \ln p(x,z,\theta)\, d\theta\right] dz$$

$$- \int_\theta q_1(\theta|\theta') \ln q_1(\theta|\theta')\, d\theta \underbrace{\int_z q_2(z)\, dz}_{1}$$

$$- \underbrace{\int_\theta q_1(\theta|\theta')\, d\theta}_{1} \int_z q_2(z) \ln q_2(z)\, dz$$

$$= \int_z q_2(z) \ln p(x,z,\theta')\, dz - \int_\theta q_1(\theta|\theta') \ln q_1(\theta|\theta')\, d\theta$$

$$- \int_z q_2(z) \ln q_2(z)\, dz$$

Note : (i) The first term depends on θ' and θ^{old} (because $q_2(z)$ is a function of θ^{old}. We express the first term as $\mathcal{C}(\theta', \theta^{old})$.

(ii) The second term is constant. Because the change in θ' only leads to change in peak for Dirac distribution $q_1(\theta|\theta')$.

(iii) The third term is independent of θ'.

So, maximization of $L(q)$ $\equiv$ maximization of $\mathcal{C}(\theta', \theta^{old})$.

From step (c) $q_2(z|x,\theta) = p(x,z,\theta^{old}) \cdot e^{const}$

$$= c_1\, p(x,z,\theta^{old})$$

$$= c_1\, p(z|x,\theta^{old}) \cdot p(x,\theta^{old})$$

Now, x is fixed and θ^{old} is fixed, then $p(x,\theta^{old}) =$ constant

$$\therefore q_2(z|x,\theta) = c_1\, p(z|x,\theta^{old}) \cdot x \text{ constant}$$

$$= c_2\, p(z|x,\theta^{old}).$$

$$\int_z q_2(z|x,\theta)\, dz = 1 \quad \text{and} \quad \int_z p(z|x,\theta^{old})\, dz = 1$$

$$\Rightarrow c_2 = 1 \quad \text{or,} \quad q_2(z|x,\theta) = p(z|x,\theta^{old}).$$

$$\theta^{new} = \arg\max_{\theta'} L(q)$$

$$= \arg\max_{\theta'} c(\theta', \theta^{old})$$

$$c(\theta', \theta^{old}) = \int_z q_2(z|x,\theta) \ln P(x,z,\theta') \, dz$$

$$= \int_z P(z|x,\theta^{old}) \ln P(x,z|\theta') \cdot P(\theta') \, dz$$

$$= \int_z P(z|x,\theta^{old}) \ln P(x,z|\theta') dz + \int_z P(z|x,\theta^{old}) \ln P(\theta') dz$$

Compare with M-step of EM. $\longrightarrow$ $= Q(\theta', \theta^{old}) + \ln P(\theta')$

Points: (i) If $P(\theta')$ is a non-informative prior ~~there~~ such as constant, then $\ln P(\theta')$ is constant. We have the same solution as general EM (M-step) for ML.

(ii) If $P(\theta')$ is non-constant, then we have MAP.

(e) If not converged, $\theta^{old} \leftarrow \theta^{new}$.

Conclusion: (i) VI and EM both lead to same computational algorithm.

(ii) Bayesian VI considers θ as random. But EM assumes θ some unknown non-random quantity.

(iii) So, VI is general than EM and EM can be viewed as a special case of VI.

Some comments of Gibbs sampling and Expectation propagation.