

Variational Bayes and Variational Message Passing

Mohammad Emtiyaz Khan

CS,UBC

Variational Inference

Find a **tractable** distribution $Q(H)$ that closely approximates the true posterior distribution $P(H|V)$.

$$\begin{aligned}\log P(V) &= \sum_H Q(H) \log P(V) \\ &= \sum_H Q(H) \log \frac{P(H, V)}{P(H|V)} \\ &= \sum_H Q(H) \log \left[\frac{P(H, V)}{Q(H)} \frac{Q(H)}{P(H|V)} \right] \\ &= \underbrace{\sum_H Q(H) \log \frac{P(H, V)}{Q(H)}}_{\mathcal{L}(Q)} + \underbrace{\sum_H -Q(H) \log \frac{P(H|V)}{Q(H)}}_{KL(Q||P)}\end{aligned}$$

Variational Inference

$$\log P(V) = \mathcal{L}(Q) + KL(Q||P) \quad (1)$$

$$\mathcal{L}(Q) = \sum_H Q(H) \log \frac{P(H, V)}{Q(H)} \quad (2)$$

$$KL(Q||P) = - \sum_H Q(H) \log \frac{P(H|V)}{Q(H)} \quad (3)$$

- Find $Q(H)$ that maximizes lower bound $\mathcal{L}(Q)$ (and hence minimizes KL divergence).
- For $Q(H) = P(H|V)$, KL vanishes to zero, but $P(H|V)$ is intractable (that's why variational approach).
- Trick : Consider a restricted class of $Q(H)$, and then find the member which minimizes the KL divergence.

Factorized Distributions

$$Q(H) = \prod_i Q_i(H_i) \quad (4)$$

Substituting this in the expression for lower bound,

$$\begin{aligned} \mathcal{L}(Q) &= \sum_H \prod_i Q_i(H_i) \log \frac{P(H, V)}{\prod_i Q_i(H_i)} \quad (\text{Outline}) \\ &= \sum_H \prod_i Q_i(H_i) \log P(H, V) - \sum_H \prod_i Q_i(H_i) \sum_i \log Q_i(H_i) \\ &= \sum_H \prod_i Q_i(H_i) \log P(H, V) - \sum_i \sum_{H_i} \prod_i Q_i(H_i) \log Q_i(H_i) \\ &= \sum_H \prod_i Q_i(H_i) \log P(H, V) + \sum_i \mathbb{H}(Q_i) \end{aligned}$$

Factorized Distributions

Now separate out all the terms in one factor Q_j .

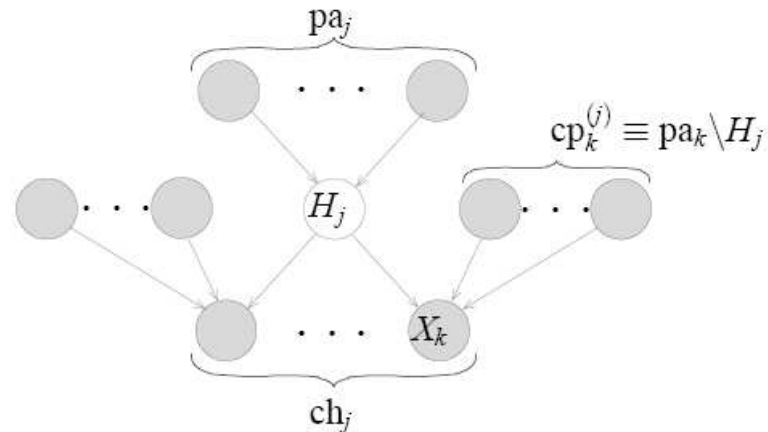
$$\begin{aligned}\mathcal{L}(Q) &= \sum_{H_j} Q_j(H_j) \underbrace{\langle \log P(H, V) \rangle_{\sim Q_j(H_j)}}_{\log Q_j^*(H_j)} + \mathbb{H}(Q_i) + \sum_{i \neq j} \mathbb{H}(Q_i) \\ &= -KL(Q_j \| Q_j^*) + \text{terms not in } Q_j\end{aligned}\quad (5)$$

This bound is maximized wrt Q_j when

$$\log Q_j(H_j) = \log Q_j^*(H_j) = \langle \log P(H, V) \rangle_{\sim Q_j(H_j)} + c \quad (6)$$

Now iterate, guaranteed convergence ...

Variational Bayes for Bayesian Networks



$$\begin{aligned}
 \log Q_j^*(H_j) &= \langle \log P(H, V) \rangle_{\sim Q_j(H_j)} + \mathbf{c} \\
 &= \sum_i \langle \log P(X_i | \mathbf{pa}_i) \rangle_{\sim Q_j(H_j)} + \mathbf{c} \\
 &= \langle \log P(H_j | \mathbf{pa}_j) \rangle_{\sim Q_j(H_j)} \\
 &\quad + \sum_{k \in \text{ch}_j} \langle \log P(X_k | \mathbf{pa}_j) \rangle_{\sim Q_j(H_j)} + \mathbf{c}
 \end{aligned}$$

Exponential-Conjugate Models

$$P(Y|\theta) = \exp[\phi_Y^T(\theta)u(Y) + f(Y) + g(\theta)] \quad (7)$$

$$u(Y) = \text{Natural statistics} \quad (8)$$

$$\phi_Y(\theta) = \text{Natural Parameter vector} \quad (9)$$

$$g(\theta) = \text{Constant of integration} \quad (10)$$

Example I: Bernoulli Distribution

$$p(x|\mu) = \mu^x(1-\mu)^{1-x} \quad (11)$$

$$\log p(x|\mu) = x \log \mu + (1-x) \log(1-\mu) \quad (12)$$

$$= \underbrace{\log \frac{\mu}{(1-\mu)}}_{\phi(\mu)} \underbrace{x}_{u(x)} + \underbrace{\log(1-\mu)}_{g(\mu)} \quad (13)$$

Exponential-Conjugate Models

$$P(Y|\theta) = \exp[\phi_Y^T(\theta)u(Y) + f(Y) + g(\theta)] \quad (14)$$

$$P(Y|\phi) = \exp[\phi^T u(Y) + f(Y) + \tilde{g}(\phi)] \text{ (Re-parametrization)}$$

Property I: $\langle u(Y) \rangle_{P(Y|\theta)} = -\frac{d\tilde{g}(\phi)}{d\phi}$

$$\log p(x|\mu) = \underbrace{\log \frac{\mu}{(1-\mu)}}_{\phi(\mu)} \underbrace{x}_{u(x)} + \underbrace{\log(1-\mu)}_{g(\mu)} \quad (15)$$

$$\phi = \log \frac{\mu}{(1-\mu)} \Rightarrow \mu = \frac{e^\phi}{1+e^\phi} \quad (16)$$

$$g(\mu) = \log(1-\mu) = -\log(1+e^\phi) = \tilde{g}(\phi) \quad (17)$$

$$E(x) = \langle u(Y) \rangle = e^\phi (1+e^\phi)^{-1} = \mu \quad (18)$$

Exponential-Conjugate Models

$$P(Y|\theta) = \exp[\phi_Y^T(\theta)u(Y) + f(Y) + g(\theta)] \quad (19)$$

Example II: Gaussian Distribution $\theta \rightarrow Y \rightarrow X \leftarrow \beta$

$$p(Y|\theta) = (2\pi)^{-1/2} \exp^{-\frac{1}{2}(Y-\theta)^2}$$

$$\log p(Y|\theta) = \underbrace{[\theta, -1/2]}_{\phi_Y(\theta)} \underbrace{\begin{bmatrix} Y \\ Y^2 \end{bmatrix}}_{u_Y(Y)} - \underbrace{\frac{1}{2}\theta^2}_{g_Y(\theta)} - \underbrace{\frac{1}{2}\log(2\pi)}_{f_Y(Y)}$$

$$p(X|Y, \beta) = (2\pi)^{-1/2} \beta^{1/2} \exp^{-\frac{\beta}{2}(X-Y)^2}$$

$$\log p(X|Y, \beta) = \underbrace{[\beta Y, -\beta/2]}_{\phi_X(Y, \beta)} \underbrace{\begin{bmatrix} X \\ X^2 \end{bmatrix}}_{u_X(X)} + \underbrace{\frac{-1}{2}(\beta Y^2 + \log \beta)}_{g_X(Y, \beta)} - \underbrace{\frac{1}{2}\log(2\pi)}_{f_X(X)}$$

Exponential-Conjugate Models

Property II: Multi-linearity $\theta \rightarrow Y \rightarrow X \leftarrow \beta$

$$\begin{aligned}
 \log p(X|Y, \beta) &= \underbrace{[\beta Y, -\beta/2]}_{\phi_X(Y, \beta)} \underbrace{\begin{bmatrix} X \\ X^2 \end{bmatrix}}_{u_X(X)} + \underbrace{\frac{-1}{2}(\beta Y^2 + \log \beta)}_{g_X(Y, \beta)} - \underbrace{\frac{1}{2} \log(2\pi)}_{f_X(X)} \\
 &= \underbrace{[\beta X, -\beta/2]}_{\phi_{XY}(X, \beta)} \underbrace{\begin{bmatrix} Y \\ Y^2 \end{bmatrix}}_{u_Y(Y)} + \underbrace{\frac{-1}{2}(\beta X^2 + \log \beta)}_{g_{XY}(X, \beta)} - \underbrace{\frac{1}{2} \log(2\pi)}_{f_Y(Y)} \\
 \log p(Y|\theta) &= \underbrace{[\theta, -1/2]}_{\phi_Y(\theta)} \underbrace{\begin{bmatrix} Y \\ Y^2 \end{bmatrix}}_{u_Y(Y)} - \underbrace{\frac{1}{2}\theta^2}_{g_Y(\theta)} - \underbrace{\frac{1}{2} \log(2\pi)}_{f_Y(Y)}
 \end{aligned}$$

Exponential-Conjugate Models

Consider Y node and its children in $\theta \rightarrow Y \rightarrow X \leftarrow \beta$,

$$\begin{aligned}\log P(Y|\theta) &= \phi_Y^T(\theta)u_Y(Y) + f_Y(Y) + g_Y(\theta) \\ \log P(X|Y, \beta) &= \phi_X^T(Y, \beta)u_X(X) + f_X(X) + g_X(Y, \beta) \\ &= \phi_{XY}^T(X, \beta)u_Y(Y) + g_{XY}(Y, \beta)\end{aligned}$$

Recall that,

$$\begin{aligned}\log Q_Y^*(Y) &= \langle \log P(Y|\theta) \rangle_{\sim Q_Y(Y)} + \langle \log P(X|Y, \beta) \rangle_{\sim Q_Y(Y)} + \mathbf{c} \\ &= \langle \phi_Y^T(\theta)u_Y(Y) + f_Y(Y) + g_Y(\theta) \rangle_{\sim Q_Y(Y)} \\ &\quad + \langle \phi_{XY}^T(X, \beta)u_Y(Y) + g_{XY}(Y, \beta) \rangle_{\sim Q_Y(Y)} + \mathbf{c} \\ &= \langle \phi_Y^T(\theta) + \phi_{XY}^T(X, \beta) \rangle_{\sim Q_Y(Y)} u_Y(Y) + f_Y(Y) + c_1\end{aligned}$$

Exponential-Conjugate Models

$$\log Q_Y^*(Y) = \langle \phi_Y^T(\theta) + \phi_{XY}^T(X, \beta) \rangle_{\sim Q_Y(Y)} u_Y(Y) + f_Y(Y) + c_1$$

Finally,

$$\begin{aligned}\langle \phi_Y^T(\theta) \rangle &= [\theta, -1/2] \\ \langle \phi_{XY}^T(X, \beta) \rangle &= \langle [\beta X, -\beta/2] \rangle\end{aligned}$$

Later is found using the property I **(explain)**.

Back to Bayesian Networks

Take each node, write the expression as a function of natural statistics of that node.

$$\begin{aligned} & \log Q_Y^*(Y) \\ &= \langle \log P(Y | \text{pa}_Y) \rangle_{\sim Q_Y(Y)} + \sum_{k \in \text{ch}_j} \langle \log P(X_k | \text{pa}_j) \rangle_{\sim Q_Y(Y)} + c \\ &= \left[\langle \phi_Y^T(\theta) + \sum_{k \in \text{ch}_j} \phi_{X_k Y}^T(X, \beta) \rangle_{\sim Q_Y(Y)} \right] u_Y(Y) + f_Y(Y) + c_1 \end{aligned}$$

The compute the expectation of natural statistics of each children node, and use that to find the quantity in bracket.

Variational Message Passing

Message from a parent node Y to a child node X :

$$m_{Y \rightarrow X} = \langle u_Y \rangle \quad (20)$$

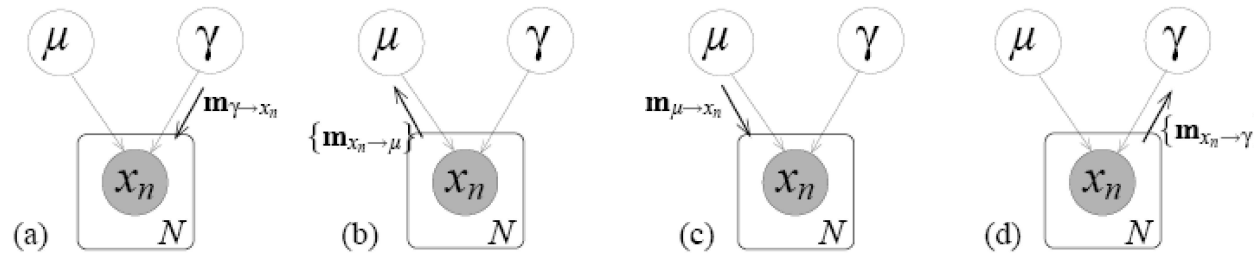
Message from a child node X to a parent node Y :

$$m_{X \rightarrow Y} = \tilde{\phi}_{XY}(\langle u_X \rangle, \{m_{i \rightarrow X}\}_{i \in cp_Y}) \quad (21)$$

Node Y update it's posterior Q_Y^* :

$$\phi_Y^* = \tilde{\phi}_Y(\{m_{i \rightarrow Y}\}_{i \in pa_Y}) + \sum_{j \in ch_Y} m_{j \rightarrow Y} \quad (22)$$

Variational Message Passing



Algorithm 1 The variational message passing algorithm

1. Initialise each factor distribution Q_j by initialising the corresponding moment vector $\langle \mathbf{u}_j(X_j) \rangle$.
2. For each node X_j in turn,
 - Retrieve messages from all parent and child nodes, as defined in (18) and (19). This will require child nodes to retrieve messages from the co-parents of X_j .
 - Compute updated natural parameter vector ϕ_j^* using (20).
 - Compute updated moment vector $\langle \mathbf{u}_j(X_j) \rangle$ given the new setting of the parameter vector.
3. Calculate the new value of the lower bound $\mathcal{L}(Q)$ (if required).
4. If the increase in the bound is negligible or a specified number of iterations has been reached, stop. Otherwise repeat from step 2.

Discussion

- Initialization and message passing schedule.
- Calculation of Lower Bound
- Allowable Model
- VIBES