Probabilistic Graphical Models

Lecture 27,28,29

Variational Inference
Mean-field algorithm
Evidence Lower Bound

Remember: Inference

$$P(X_t \mid X_e = x_e) \Longrightarrow \boxed{P(X_c)} \quad \text{marginalization} \quad \boxed{\phi_c(X_c)}$$

$$x_1^*, \dots, x_n^* = \arg\max_{x_1 \dots x_n} P(X_1, X_2, \dots, X_n)$$

$$P(X_c) = \sum_{X \setminus X_c} P(X_1, \dots, X_n)$$

How to solve this using an optimization problem?

Remember: Latent Variable Models

K. N. Toosi
University of Technology

- Latent variable models

  $P_\theta(X, Z) = ✓$

  data $\underline{x^1, x^2, \ldots, x^n}$

  $\underline{p(Z|X)} = ✓$ Hard to compute

Situations where we need normalization

$$\boxed{P(X) = \frac{1}{Z}\underbrace{\tilde{P}(X)}_{\text{product of factors}} = \frac{1}{Z}\, e^{F(X)}}$$

$F(X) \rightarrow$ sum of

MRF

BN or MRF with evidence

$$P(X) = \underbrace{P(X_t, X_e)}_{X_e \rightarrow \text{evidence (known)}} \Longrightarrow \text{need } p(X_t \mid X_e) = \frac{p(X_t, x_e)}{\underbrace{\sum_{X_t} p(X_t, x_e)}_{Z(X_e)}}$$

$$p(X) = p(y, \underbrace{z}_{\text{latent}}) \qquad p(z \mid y) = \frac{p(z,y)}{\underbrace{\sum_{z} p(z,y)}_{\rightarrow Z(y)}}$$

Variational Inference on MRFs

$$P(X) = \frac{1}{Z}\tilde{P}(X)$$

complex

(Hard to do inference on)

$$\text{find } Q(X) \begin{cases} Q(X) \text{ is easy to handle} \\ Q(X) \text{ is close to } P(X) \end{cases}$$

Kullback–Leibler (KL) Divergence

pgm 27 (II)

Measure distance between $Q(X), P(X)$

KL-divergece

$$KL(Q \| P) = \overset{\int}{\sum_{X}} Q(X) \log \frac{Q(X)}{P(X)} = E_{Q}\left\{\log \frac{Q(X)}{P(X)}\right\}$$

$$= \sum_{X} Q(X) \left[\log Q(X) - \log P(X) \right]$$

$P(X) = Q(X)$ for all $X \Longrightarrow KL(Q \| P) = 0$

$KL(Q \| P) = 0 \Longrightarrow P(X) = Q(X)$ for all X

almost all (continuous)

$KL(Q \| P) \geq 0$

$KL(Q \| P) \neq KL(P \| Q)$ in general

KL Divergence and Entropy

$$KL(Q\|P) = \sum_x Q(x) \log Q(x) - \sum_x Q(x) \log P(X)$$

what choice of Q maximizes $\sum_x Q(x) \log P(X)$?

$$\sum_x Q(X) = 1$$

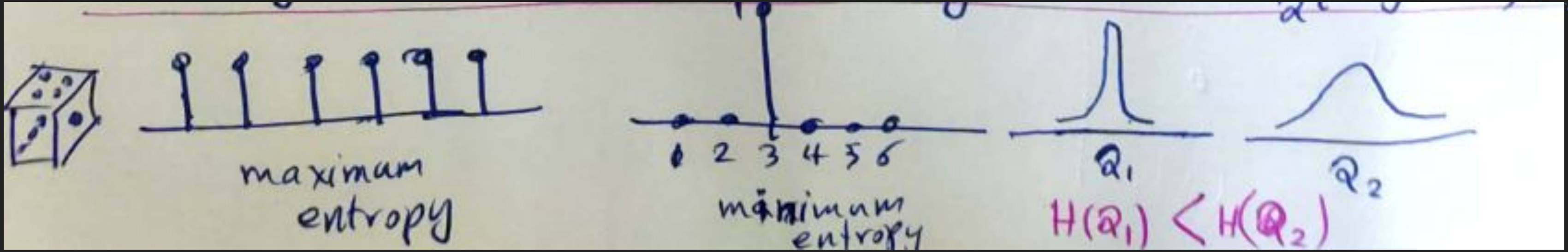
$$\int_x Q(x) = 1$$

$$\text{Entropy} = \sum_x Q(x) \log \frac{1}{Q(x)} = -\sum_x Q(x) \log Q(x) = H(Q)$$

$$KL(Q\|P) = -H(Q) - E_Q\{\log P(x)\}$$

minimizing $KL(Q\|P) \equiv$ maximizing $H(Q) + E_Q\{\log P(x)\}$

KL Divergence and Entropy



pgm27 III

$$KL(Q\|P) = -H(Q) - E_Q\{\log P(x)\}$$

$$\underset{Q \in \mathcal{Q}}{\text{minimize}}\ KL(Q\|P)$$

Example: Fully Factorized Q

Simple case: $Q(X) = Q(X_1, X_2, \ldots, X_n) = Q_1(X_1)\, Q_2(X_2) \cdots Q_n(X_n)$

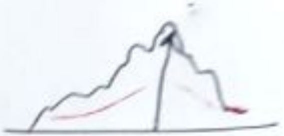
fully factorized case

KL Divergence is not Symmetric

K. N. Toosi
University of Technology

$$P(X) = \frac{1}{Z}\tilde{P}(X) = \frac{1}{Z}e^{F(X)}$$

$$\min_{Q} KL(Q\|P) = \sum_{X} Q(X)\log\frac{Q(X)}{P(X)}$$



$$= \sum_{x} P(x)\log P(x) - \sum_{x} P(x)\boxed{\log Q(x)}$$

why not minimize $KL(P\|Q) = \sum_{X} \underline{P(X)}\log\frac{P(X)}{Q(X)}$?

Variational Lower Bound

K. N. Toosi
University of Technology

$$KL(Q\|P) = \sum_x Q(x)\left[\log Q(x) - \log P(x)\right]$$

$$= \sum_x Q(x)\left[\log Q(x) - \log \frac{1}{Z}\tilde{P}(x)\right]$$

$$= \sum_x Q(x)\log Q(x) - \sum_x Q(x)\log \tilde{P}(x) + \underbrace{\sum_x Q(x)\log Z}$$

$$= \sum_x Q(x)\log Q(x) - \sum_x Q(x)\underbrace{\log \tilde{P}(x)}_{F(x)} + \log Z$$

$$KL(Q\|P) = \underbrace{\sum_x Q(x)\log\frac{Q(x)}{\tilde{P}(x)}}_{-L(Q)} + \log Z$$

$$\boxed{L(Q) = E_Q\left\{\frac{\tilde{P}(x)}{Q(x)}\right\}}$$

$$L(Q) = +\log Z - KL(Q\|P) \Rightarrow L(Q) \leq \log(Z)$$

$KL \geq 0$

variational lower bound!

The Mean-field Algorithm

Meanfield inference $Q(X) = Q(X_1 \ldots X_n) = \prod_{i=1}^{n} Q_i(X_i)$

Fully factorized

$$Q^* = \arg\min_{Q} \sum_{x} Q(x) \log \frac{Q(x)}{\tilde{P}(x)} + \log Z$$

$$= \arg\min_{Q} \sum_{x} Q(x) \log \frac{Q(x)}{\tilde{P}(x)}$$

$$\sum_{x} Q(x) \log \frac{Q(x)}{\tilde{P}(x)} = \boxed{\sum_{x} Q(x) \log Q(x)} - \sum_{x} Q(x) \underbrace{\log \tilde{P}(x)}_{F(x)}$$

$$\sum_{x} Q(x) \log Q(x) = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_n} Q_1(x_1)\, Q_2(x_2) \cdots Q_n(x_n) \left[\sum_{i=1}^{n} \log Q_i(x_i) \right]$$

$$= \sum_{i=1}^{n} Q_i(x_i) \log Q_i(x_i)$$

The Mean-field Algorithm

$$\sum_x Q(x) \log \tilde{P}(X) = \sum_x Q(x)\, F(x) = \sum_x Q(x) \sum_c F_c(X_c)$$

$$= \sum_c \sum_X Q(x)\, F_c(X_c) = \sum_c \sum_{X_c} \left[\sum_{X \setminus X_c} Q(X_c, X \setminus X_c) \right] F_c(X_c)$$

$$= \sum_c \sum_{X_c} Q(X_c)\, F_c(X_c)$$

$$\sum_i Q_i(X_i) \log Q_i(X_i) - \sum_c \sum_{X_c} Q(X_c)\, F_c(X_c)$$

The Mean-field Algorithm for Pairwise MRF

$$P(x) = \frac{1}{Z} e^{F(x)} \quad , \quad F(x) = \sum_{i=1}^{n} F_i(x_i) + \sum_{(i,j)\in\mathcal{E}} F_{ij}(x_i, x_j)$$

$$-L(Q,\tilde{P}) = \sum_{i} \sum_{x_i} Q_i(x_i)\log Q_i(x_i) - \sum_{i=1}^{n} Q_i(x_i)\, F_i(x_i) - \sum_{(i,j)\in\mathcal{E}} Q_i(x_i)\, Q_j(x_j)\, F_{ij}(x_i, x_j)$$

$$L(Q,\tilde{P}) = -\sum_{i} \sum_{x_i} Q_i(x_i)\log Q_i(x_i) + \sum_{i=1}^{n} \sum_{x_i} Q_i(x_i)\, F_i(x_i) + \sum_{(i,j)\in\mathcal{E}} \sum_{x_i} \sum_{x_j} Q_i(x_i)\, Q_j(x_j)\, F_{ij}(x_i, x_j)$$

The Mean-field Algorithm for Pairwise MRF

pgm 28 II

$$Q(X) = Q(X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} Q_i(X_i)$$

mean-field

$$P(X) = \frac{1}{Z}\tilde{P}(X) = \frac{1}{Z}e^{F(X)}$$

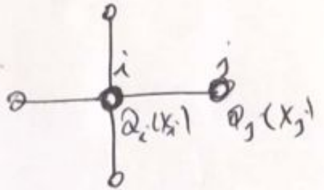
$$\sum_{i=1}^{n} F_i(X_i) + \sum_{(i,j)\in\mathcal{E}} F_{ij}(X_i, X_j)$$

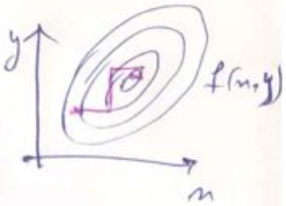
$$\mathcal{L}(Q,\tilde{P}) = -\sum_{i=1}^{n}\sum_{X_i} \underline{Q_i(X_i)}\log Q_i(X_i) + \sum_{i=1}^{n}\sum_{X_i} Q_i(X_i)F_i(X_i)$$

$$+\sum_{(i,j)\in\mathcal{E}}\sum_{X_i}\sum_{X_j} Q_i(X_i)Q_j(X_j)F_{ij}(X_i,X_j)$$

The Mean-field Algorithm for Pairwise MRF

$$\max_{\substack{Q \\ q_{ik}}} \mathcal{L}(Q,P) \quad \text{subject to} \quad \sum_{k=1}^{L} q_{ik} = 1 \quad \text{for } i = 1 \cdots n$$





coordinate ascent

Loop

for $i = 1 \cdots n$

$\max_{q_i:}$

$$\max_{q_{i1}, q_{i2}, \cdots q_{iL}} L(Q,P) \Longrightarrow \frac{\partial}{q_{i\ell}} L(Q,P) + \lambda\left(\sum_{k=1}^{L} q_{ik} - 1\right)$$

$$= -\log q_{i\ell} - 1 + F_i(\ell) + \sum_{j \in N_i} \sum_{k'=1}^{L} q_{jk'} F_{ij}(\ell, k') + \underline{\lambda} = 0$$

for $\ell = 1, 2, \cdots, L$

$$\sum_{k=1}^{\ell} q_{ik} = 1$$

The Mean-field Algorithm for Pairwise MRF

pgm28 III

$$\Rightarrow \log q_{il} = F_i(l) + \sum_{j \in N_i} \sum_{k=1}^{L} q_{jk} F_{ij}(l,k) + \lambda - 1$$

$$q_{il} = \exp\Big(F_i(l) + \sum_{j \in N_i} \sum_{k=1}^{L} q_{jk} F_{ij}(l,k)\Big) e^{\lambda - 1}$$

$l = 1, 2, \dots, L$ $\qquad \sum_{k=1}^{L} q_{il} = 1$

$$q'_{il} = \exp\Big(F_i(l) + \sum_{j = N_i} \sum_{k=1}^{L} q_{jk} F_{ij}(l,k)\Big)$$

$$q_{il} \leftarrow \frac{q'_{il}}{\sum_{k=1}^{L} q'_{ik}}$$

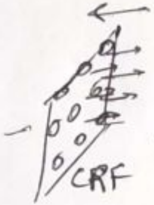
mean-field message passing

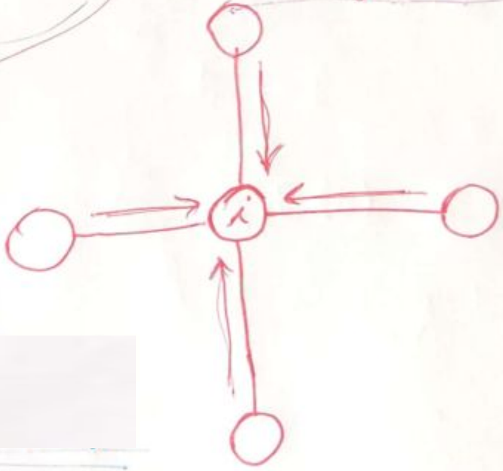
Meanfield message passing

$$q'_{il} = \exp\left(F_i(l) + \sum_{j=N_i}\sum_{k=1}^{L} q_{jk}\, F_{ij}(l,k)\right)$$

$$q_{il} \leftarrow \frac{q'_{il}}{\sum_{k=1}^{L} q'_{ik}}$$

mean-field message passing





Variational Inference: Continuous Case

Case 2: Continuous $Q_i(X_i)$

Among all $Q_i(X)$ with $Q_i(X) \geq 0$, $\int Q_i(X)\,dx = 1$

$i = 1, \dots, n$

find the one that maximizes $\mathcal{L}(Q, \tilde{P})$

minimizes $KL(Q \| P)$

Variational Calculus $f : \mathbb{R}^m \longrightarrow \mathbb{R}^n$ is a function

optimize $J(f)$ s.t. so some constraint

Latent Variable Models and Variational Learning

Variational Learning

$$p_\theta(x) = \frac{1}{Z(\theta)} \tilde{P}_\theta(x) = \frac{1}{Z(\theta)} e^{F_\theta(x)}$$

Latent Variable Models

$$P_\theta(x) = \int P_\theta(x, z)\, dz$$

Data $x^1, x^2, \dots, x^m$

$$\max_\theta \sum_i \log P_\theta(x^i)$$

$$\max_\theta ll(\theta) = \max_\theta \sum_{i=1}^{m} \log P_\theta(x^i)$$

$$= \max \sum_{i=1}^{m} \log \int P_\theta(x^i, z)\, dz$$

Classic Method fail!

Solution 1

$$\frac{\partial \ell\ell(\theta)}{\partial \theta_k} = \frac{\partial}{\partial \theta_k} \sum_{i=1}^{m} \log \underbrace{\int P_\theta(x^i, z)\, dz}_{\text{hard to compute}}$$

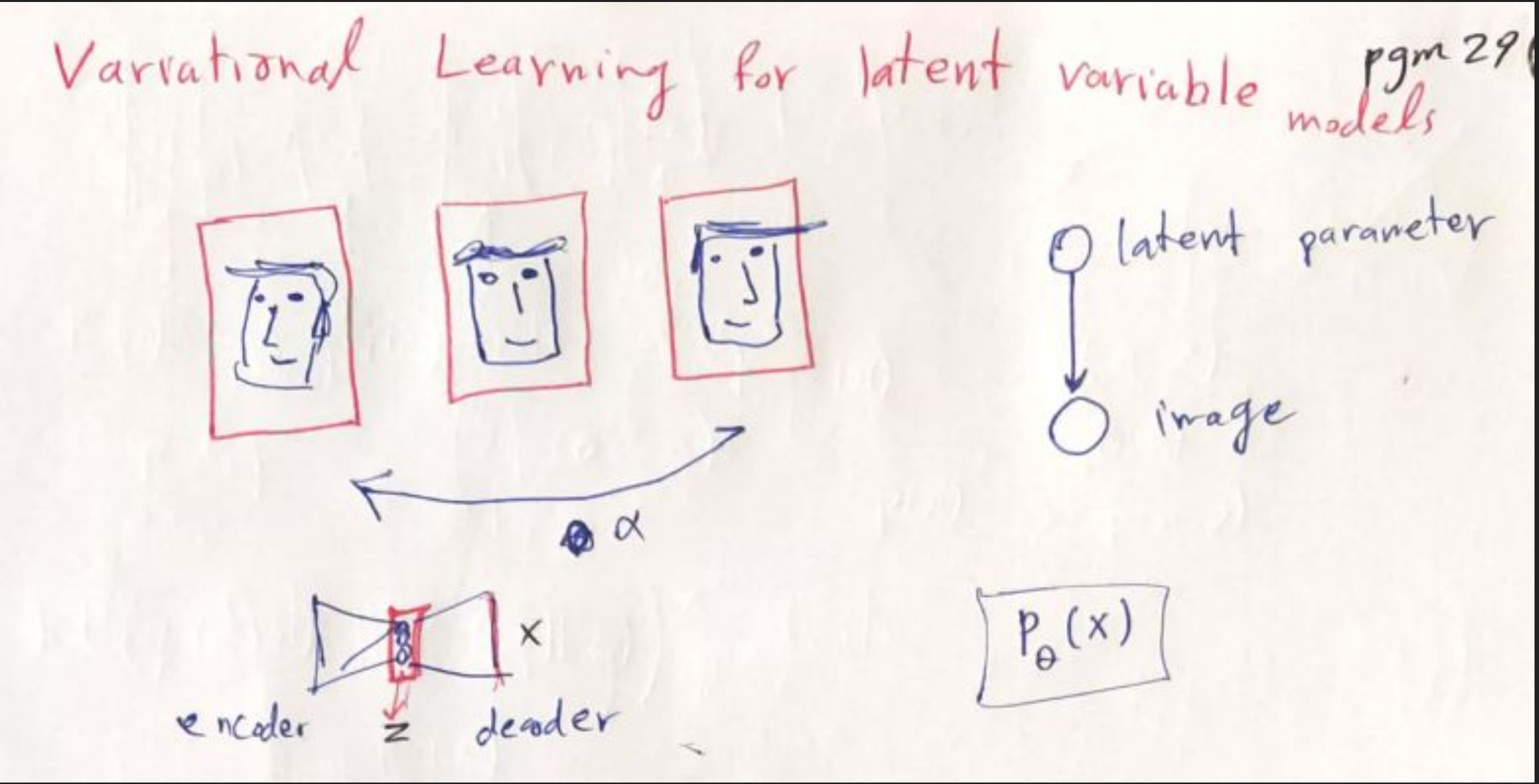
$$= \sum_{i=1}^{m} \frac{\int \frac{\partial}{\partial \theta_k} P_\theta(x^i; z)\, dz}{\int \underbrace{P_\theta(x^i; z)}\, dz}$$

Solution 2: Expectation - Maximization

$\Longrightarrow$ Need posterior $\underbrace{P_\theta(z|x^i)}$ for $i=1, \dots, m$

hard to compute

Variational Learning



Remember: Latent Variable Models

model $P_\theta(x) = \sum_z P_\theta(x,z)$ latent variable model

Data $= \{x^1, x^2, \dots, x^m\} = D$

$$\max_\theta \; ll(\theta, D) = \sum_{i=1}^{m} \log P_\theta(x^i) = \sum_{i=1}^{m} \log \sum_z P_\theta(x^i, z)$$

Usually $P_\theta(x,z) = \boxed{P_\theta(x|z)}\, p(z)$

$\boxed{\text{or } P_{\theta_1}(x|z)\, P_{\theta_2}(z)}$

Example

Remember: Latent Variable Models

K. N. Toosi
University of Technology

Example

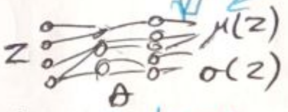
$P_\theta(x,z) = P_\theta(x|z)\,p(z)$

$P_\theta(x|z) = N(x; \mu(z), I\sigma(z))$

1 $z \sim p(z)$

2 compute $\boxed{\mu(z), \sigma(z)}$

3 $x \sim P_\theta(x|z) = N(x; \mu(z), I\sigma(z))$



a deep Neural Net

How to train the neural network

Need to comput $\frac{\partial}{\partial \theta_i} \sum_{i=1}^{m} \log \sum_{z} P_\theta(x^i, z)$

Intractible

$= \frac{\partial}{\partial \theta_i} \sum_{i=1}^{m} \log \sum_{z} P_\theta(x^i|z)\,p(z)$

for EM Need $P_\theta(z|x)$ → posterior → intractable!

VI on MRFs vs Latent Variables Models

K. N. Toosi
University of Technology

MRF

$$P_\theta(y) = \frac{1}{Z(\theta)} \tilde{P}(y)$$

$$Z(\theta) = \sum_{y} \tilde{p}(y)$$

minimize over q

$$KL(q \,\|\, P_\theta(y))$$

latent variable model

$$P_\theta(z|x) = \frac{1}{P_\theta(x)} P_\theta(x,z)$$

$$P_\theta(x) = \sum_{z} P_\theta(x,z)$$

minimize over $q(z|x)$

$$KL(\underbrace{q(z|x)}_{\text{approximate posterior}} \,\|\, P_\theta(z|x))$$

Example: $q(z|x) = \prod q_i(z_i|x)$

KL-divergence

K. N. Toosi
University of Technology

$$KL\left(q(z|x) \,\|\, P_\theta(z|x)\right) = \sum_z q(z|x) \log \frac{q(z|x)}{P_\theta(z|x)}$$

$$= \sum_z q(z|x) \log \frac{q(z|x)}{\dfrac{P_\theta(x,z) \longrightarrow \text{easy}}{P_\theta(x) \longrightarrow \text{hard}}}$$

$$= \sum_z q(z|x) \left(\log \frac{q(z|x)}{P_\theta(x,z)} + \log P_\theta(x) \right)$$

$$= \sum_z q(z|x) \log \frac{q(z|x)}{P_\theta(x,z)} + \sum_z q(z|x) \overbrace{\log P_\theta(x)}^{\text{independent of } z}$$

$$= -\sum_z q(z|x) \log \frac{P_\theta(x,z)}{q(z|x)} + \log P_\theta(x)$$

Evidence Lower Bound (ELBO)

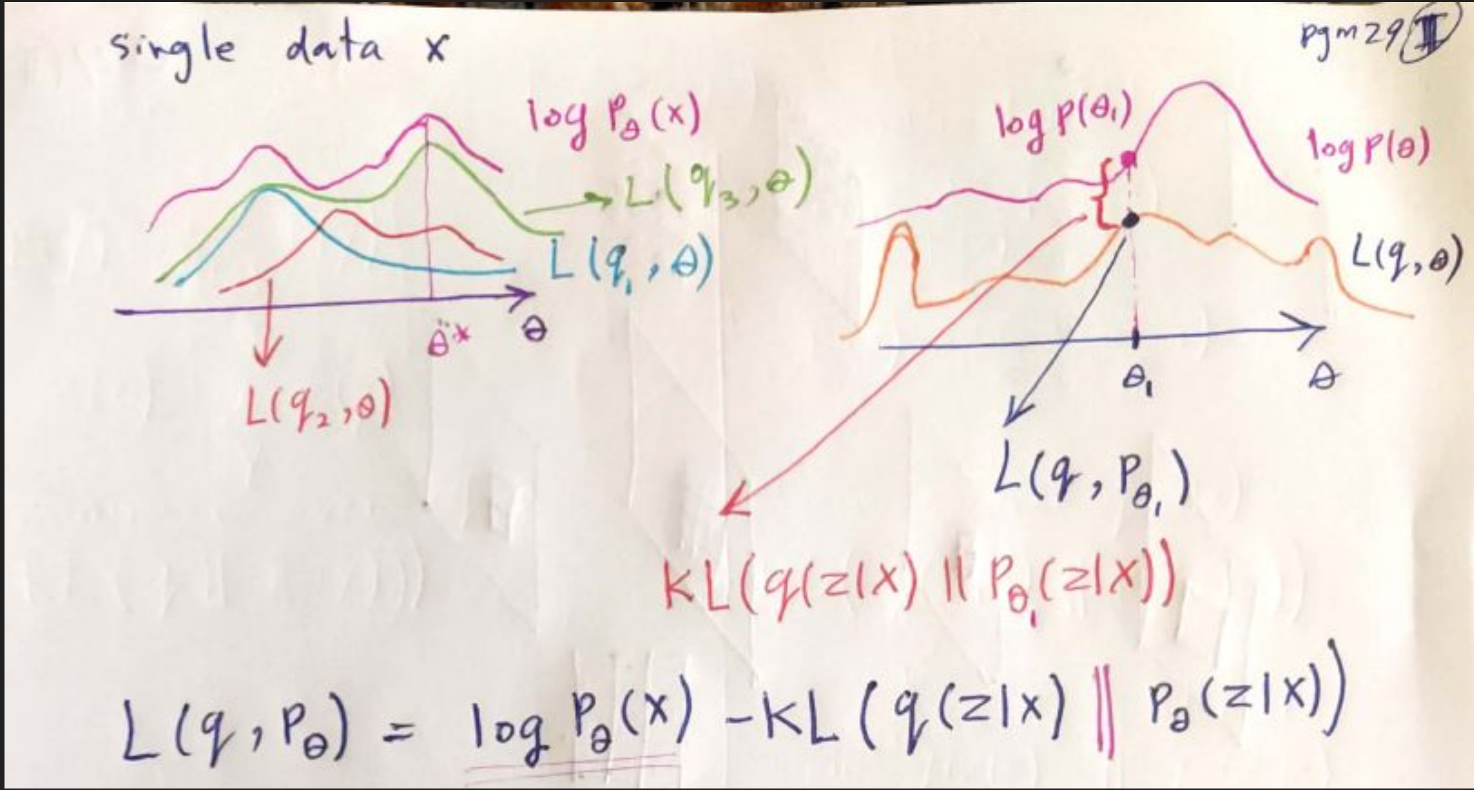
$$\underbrace{\sum_z q(z|x) \log \frac{P_\theta(x,z)}{q(z|x)}}_{\mathcal{L}(q,\, P_\theta)} = \log P_\theta(x) - KL\big(q(z|x) \,\|\, P_\theta(z|x)\big)$$

$\mathcal{L}(q, P_\theta)$: variational lower bound

Evidence Lower BOund (ELBO)

$\mathcal{L}(q, P_\theta) \leqslant \log P_\theta(x)$ for any choice of q

Evidence Lower Bound (ELBO)



Evidence Lower Bound (ELBO)

$$L(q, P_\theta) = \log P_\theta(x) - KL\left(q(z|x) \,\|\, P_\theta(z|x)\right)$$

1: Maximizing $L(q, P_\theta)$ with respect to θ pushes $\log P_\theta(x)$ up.

2: Maximizing $L(q, P_\theta)$ w.r.t θ works better when $q(z|x)$ is closer to $P_\theta(z|x)$ $\left[KL\left(q(z|x) \| P_\theta(z|x)\right) \text{ is smaller}\right]$

3: Maximizing $L(q, P_\theta)$ w.r.t. q minimizes $KL\left(q(z|x) \,\|\, P_\theta(z|x)\right)$

K. N. Toosi University of Technology

Variational Learning

$$\max_{q,\theta} \mathcal{L}(q, P_\theta) = \sum_{z} q(z|x) \log \frac{P_\theta(z,x)}{q(z|x)}$$

How to compute (approximate) $\mathcal{L}(q, P_\theta) = \sum_{z} \sim$?

$$\mathcal{L}(q, P_\theta) = E_{q(z|x)} \left\{ \log \frac{P_\theta(z,x)}{q(z|x)} \right\}$$

$$\simeq \frac{1}{p} \sum_{i=1}^{p} \log \frac{P_\theta(z^i, x)}{q(z^i|x)}$$ where $z^1, z^2, \dots, z^p \sim q(z|x)$

Variational Learning

K. N. Toosi University of Technology

Data $x^1, x^2, x^3, \dots, x^m$

$$\max_\theta \ \ell\ell(\theta) = \sum_{i=1}^{m} \log P_\theta(x^i) = \sum_{i=1}^{m} \log \sum_z \log P_\theta(x^i; z)$$

instead

$$\max_{\theta, q} \mathcal{L}(q, P_\theta) = \sum_{i=1}^{m} \sum_z q_i(z|x^i) \log \frac{P_\theta(x^i; z)}{q_i(z|x^i)}$$

$$= \mathcal{L}(q, \theta) = \sum_{i=1}^{m} E_{q(z|x^i)} \{\log P_\theta(x^i; z)\} + H\{q(z|x^i)\}$$

For each x^i a different q might be optimum.

$\Longrightarrow$ A different q_i for each x^i

Variational Learning Algorithm

Start from some $\theta \leftarrow \theta_0, q_1, q_2, \dots, q_m$

optimize w.r.t. θ

$$\frac{\partial}{\partial\theta} L(q, P_\theta) = \sum_{i=1}^{m} E_q\left\{\frac{\partial}{\partial\theta}\log \underbrace{P_\theta(x^i;z)}_{P_\theta(x^i|z)\,p(z)}\right\}$$

$$= \sum_{i=1}^{m} E_q\left\{\frac{\partial}{\partial\theta}\log P_\theta(x^i|z)\right\}$$

$$\sim \sum_{i=1}^{m}\sum_{j=1}^{m'} \frac{\partial}{\partial\theta}\log P_\theta(x^i|z_j^i)$$

$z_1^i, z_2^i, \dots, z_{m'}^i \sim q_i(z|x^i)$

usually $m'=1$

$$\theta \leftarrow \theta + \lambda \frac{\partial}{\partial\theta} L(q, P_\theta)$$

maximize w.r.t. $q_1, q_2, \dots, q_m$

can do independently for each q_i.