Lecture: Variational inference

We have $p(x)$. We want „good" $q(x) \approx p(x)$

Entropy

x, y - observations

$h(x)$ - information from observ.

$h((x,y)) = h(x) + h(y)$, if x, y - ind. r. v.

$p(x,y) = p(x) \cdot p(y)$, - || -

$\log p(x,y) = \log p(x) + \log p(y)$

$h(x) = \underbrace{-\log p(x)}$ - information

$$\boxed{\begin{aligned} H(p) &= -\sum_{x \in \mathcal{X}} p(x) \log p(x) \\ &= \mathbb{E}_p - \log p(x) \end{aligned}}$$ - entropy

$$H(p) \to \max_p$$

Ex. $x_1, \dots, x_k$ - possible values for r.v. X

$\vec{p} = \{p_1, \dots, p_k\}$ - probabilities for $x_1, \dots, x_k$

$$\sum_{i=1}^{k} p_i = 1, \quad p_i > 0, \quad i = 1, \dots, k$$

$$-\sum_{i=1}^{k} p_i \log p_i + \lambda\left(\sum p_i - 1\right) \to \max_{\vec{p}}$$

$$-\log p_i - 1 + \lambda = 0, \quad i = 1, \dots, k$$

$$\log p_i = (1 - \lambda) \Rightarrow \text{All } p_i \text{ are equal}$$

$$p_i = \exp(1 - \lambda)$$

$$H(p) = \int p(x) \log p(x)\,dx = \mathbb{E}_p \log p(x)$$

Ex. $\int p(x)\,dx = 1 \qquad \int x\,p(x)\,dx = \mu, \quad \int (x-\mu)^2 p(x)\,dx = \sigma^2$

$$L(p) = -\int p(x) \log p(x)\,dx + \lambda_1\left(\int p(x)\,dx - 1\right) + \lambda_2\left(\int x\,p(x)\,dx - \mu\right) +$$
$$+ \lambda_3\left(\int (x-\mu)^2 p(x)\,dx - \sigma^2\right) \to \max_p$$

$$\frac{\partial L(p)}{\partial p} - \frac{d}{\,} \frac{\partial L(p)}{\,} = 0$$ - Euler-Lagrange eq.

$\partial p \quad dx \quad \partial p$

Solution is Gaussian $N(\mu, \sigma^2)$

$$L(p) = \int p(x) \left[-\log p(x) + \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 \right]$$

Mutual entropy

$$H(\vec{y} \mid \vec{x}) = -\iint p(\vec{x}, \vec{y}) \ln p(\vec{y} \mid \vec{x}) \, d\vec{y} \, d\vec{x}$$

$$H(\vec{y} \mid \vec{x}) + H(\vec{x}) = H(\vec{x}, \vec{y})$$

Kullback - Liebler divergence

p, q – two distributions

with densities $p(x)$, $q(x)$

$KL(p \| q)$ measures how different

are they

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

$$= H(p|q) - H(p)$$

1. $KL(p|q) \geq 0$

$$-\int p(x) \ln \frac{q(x)}{p(x)} dx \geq -\int p(x)\left(\frac{q(x)}{p(x)} - 1\right) dx =$$

$$= -\int q(x) dx + \int p(x) dx = 1 - 1 = 0$$

2. Non-symmetric

Let's establish a connection between KL divergence & MLE (maximum likelihood estimate)

$p(x)$ is true

$q_\theta(x)$ - want to be close to $p(x)$,

$\theta \in \Theta \subseteq \mathbb{R}^d$

$$KL(p \| q_\theta) \to \min_\theta$$

$$KL(p \| q_\theta) = -\int p(x) \ln \frac{q(x)}{p(x)} dx = -\mathbb{E}_p \ln \frac{q(x)}{p(x)} \approx$$

$$\approx \mathbb{E}_{p_e} \ln \frac{q(x)}{p(x)} = \qquad (p_e \text{ - empirical})$$

$$= -\frac{1}{n} \sum_{i=1}^{n} [\ln q_\theta(x_i) - \ln p(x_i)] \to \min_\theta \iff$$

$$L_\theta(D) = \sum_{i=1}^{n} \ln q_\theta(x_i) \to \max_\theta$$

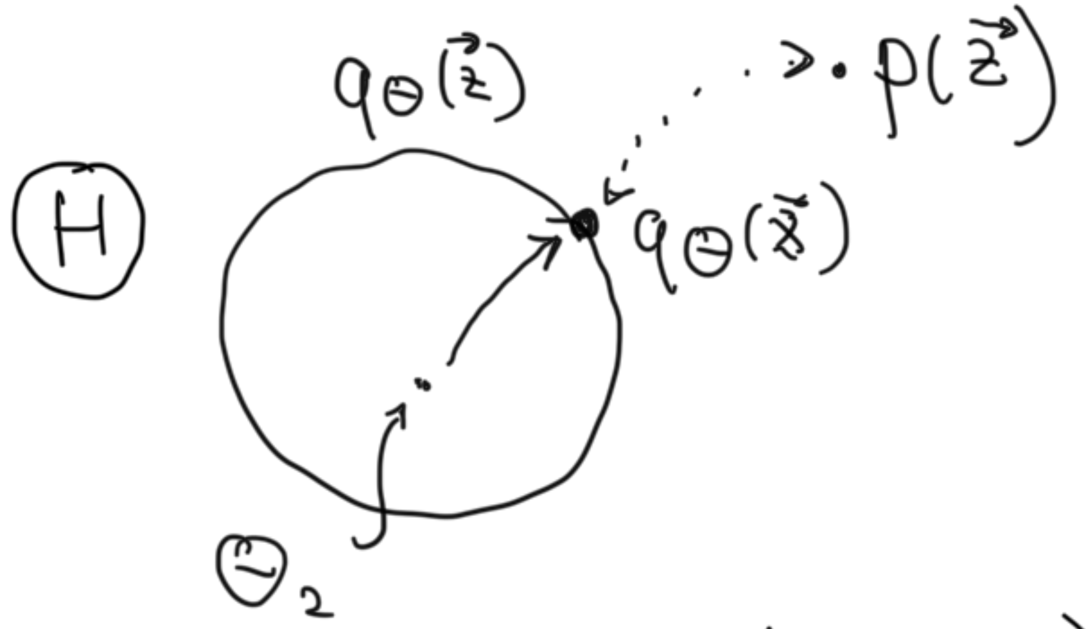
Mutual information

Def. $I(x,y) = KL(p(x,y) \| p(x)p(y))$

$$I(x,y) \geq 0$$

$$I(x,y) = H(x) - H(x|y) = H(y) - H(y|x)$$

Variational inference – approximate Bayesian inference



$$q^* = \underset{\Theta \in H}{\operatorname{argmin}}\ d(q,p)$$

Let us use $d(q,p) = KL(q\|p)$

$$q(\vec{\Theta}) \approx p(\vec{\Theta}|\mathcal{D}) \leftarrow \text{posterior}$$

$$KL(q(\vec\theta) \,|\, p(\vec\theta|D)) = \int q(\vec\theta) \ln \frac{q(\vec\theta)}{p(\vec\theta|D)} d\vec\theta =$$

$$= \int q(\vec\theta) \ln \frac{q(\vec\theta)\, p(D)}{p(\vec\theta, D)} d\vec\theta =$$

$$= \int q(\vec\theta) \ln p(D)\, d\vec\theta + \int q(\vec\theta) \ln q(\vec\theta)\, d\vec\theta -$$

$$- \int q(\vec\theta) \ln p(\vec\theta, D)\, d\vec\theta =$$

$$= \ln p(D) - \int q(\vec\theta) \ln \frac{p(\vec\theta, D)}{q(\vec\theta)} d\vec\theta$$

fit

$$ELBO(q) = \int q(\vec\theta) \ln \frac{p(\vec\theta, D)}{q(\vec\theta)} d\vec\theta$$

regularization

1. We maximize $ELBO(q)$ w.r.t. q

2. $\ln p(D) = ELBO(q) + KL(q|p) \Rightarrow$

$\quad\quad KL(q|p) \geq 0$

$$ELBO(q) \leq \ln p(D)$$

– evidence lower bound

3. To calculate $\mathrm{ELBO}(q)$ we don't need $p(\mathcal{D})$, only $p(\mathcal{D}|\Theta)$ and $p(\Theta)$ $\Rightarrow$ we look at unnormalized density (c.t. Laplace approximation)

Mean - Field Variational Bayes

We factorize $\vec{\Theta}$ to $_\tau\{\vec{\Theta}_1, \dots, \vec{\Theta}_\tau)$

$$Q_{MFVB} = \{ q : q(\vec{\Theta}) = \prod_{j=1} q_j(\Theta_j) \}$$

$$\mathrm{ELBO}(q) = \int q(\vec{\Theta}) \log p(\mathcal{D}|\vec{\Theta}) + \int q(\vec{\Theta}) \log \frac{1}{q(\vec{\Theta})} d\vec{\Theta} =$$

$$= \int q_j(\theta_j) \left[\log p(\mathcal{D},\theta) \prod_{i \neq j} q_i(\theta_i) \right] d\theta +$$

$$+ \sum_{i=1}^{I} \int q_i(\theta_i) \log \frac{1}{q_i(\theta_i)} d\theta_i =$$

$$= \int q_j(\theta_j) \left[\int \log p(\mathcal{D},\vec{\theta}) \prod_{i \neq j} q_i(\theta_i) \, d\vec{\theta}_{/j} \right] d\theta_j +$$

$$+ \int q_j(\theta_j) \log \frac{1}{q_j(\theta_j)} d\theta_j + \text{const} =$$

independent of q_j

q_i - fixed, if $i \neq j$

q_j - what we optimize now

$$= \int q_j(\theta_j) \, \mathbb{E}_{i \neq j} \log p(\mathcal{D},\vec{\theta}) \, d\theta_j +$$

$$+ \int q_j(\theta_j) \log \frac{1}{q_j(\theta_j)} d\theta_j + C =$$

$$= KL\left(q_j(\theta_j) \,|\, \exp\left(\mathbb{E}_{i \neq j} \log p(\mathcal{D}|\vec{\theta})\right)\right)$$

$$\boxed{q_j^*(\theta_j) = \exp\left(\mathbb{E}_{i \neq j} \log p(\mathcal{D}|\vec{\theta})\right)}$$

Algorithm : 1. Initialize $q_j^0(\theta_j), j = 1 \dots \tau$

2. Update factors one by one

$$q_j^t(\theta_j) \longleftarrow \vec{q}^{\,t}(\vec{\theta})$$

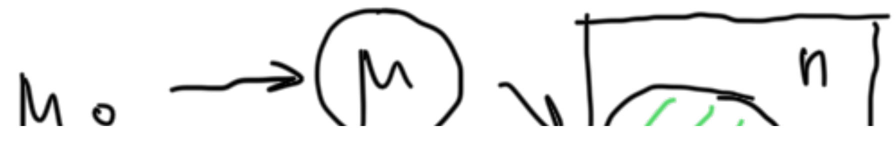
Ex. $x_i \sim \mathcal{N}(\mu, 1/\tau)$

$$\mathcal{D} = \{x_i\}_{i=1}^{n}, \quad x_i - \text{i.i.d.}$$

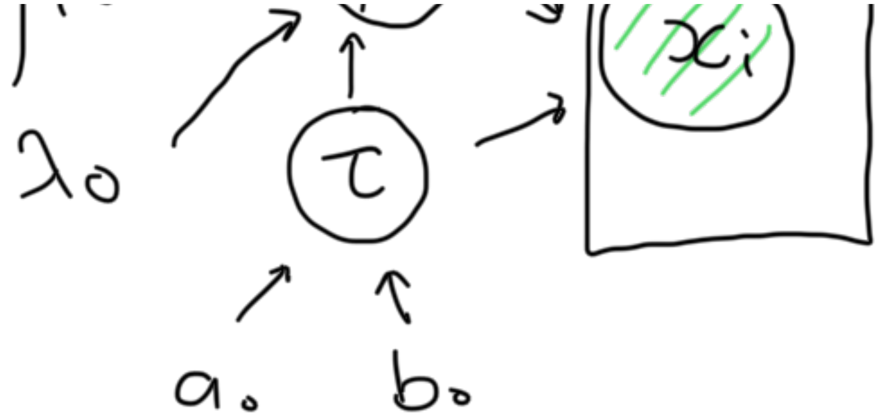
Likelihood $$p(\mathcal{D} \mid \mu, \tau) = \left(\tau/2\pi\right)^{n/2} \exp\left(-\tau/2 \sum_{i=1}^{n} (x_i - \mu)^2\right)$$

Priors $$p(\mu \mid \tau) = \mathcal{N}(\mu \mid \mu_0, (\lambda_0 \tau)^{-1})$$

$$p(\tau) = \Gamma(\tau \mid a_0, b_0)$$



graphical model



$\Theta = \{\mu, \tau\}$ - ?

$p(\mu, \tau \mid \mathcal{D})$ - not tractable

$q(\mu, \tau) = q_\mu(\mu)\, q_\tau(\tau)$ - MFVB factor.

1. $\log q^*_\mu(\mu) = \mathbb{E}_\tau\left[\log p(\mathcal{D} \mid \mu, \tau) + \log p(\mu \mid \tau)\right] + \text{const} =$

$$= -\frac{\mathbb{E}[\tau]}{2}\left\{\lambda_0(\mu-\mu_0)^2 + \sum_{i=1}(x_i-\mu)^2\right\} + \text{const}$$

$$q^*_\mu(\mu) = \mathcal{N}(\mu \mid \mu_n, \lambda_n^{-1})$$

$\mu_n = \frac{\lambda_0\mu_0 + n\bar{x}}{}$, $\lambda_n = (\lambda_0 + n)\,\mathbb{E}[\tau]$,

$$\lambda_0 + n$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \mathbb{E}_{q^*_\tau}[\tau]$$

2. Fix $q_\mu(\mu)$ - find $q^*_\tau(\tau)$

$$\log q^*_\tau(\tau) = (a_0 - 1)\log\tau - b_0\tau + \frac{n}{2}\log\tau -$$
$$- \frac{\tau}{2}\mathbb{E}_\mu\left[\sum_{i=1}^{n}(x_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right]$$

$$q^*_\tau(\tau) = \Gamma(\tau \mid a_n, b_n)$$

$$a_n = a_0 + n/2$$

$$b_n = b_0 + 1/2\,\mathbb{E}_\mu\left[\sum_{i=1}^{n}(x_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right]$$

Alternative approach, SVI

$$q(\vec{\theta}) = N(\vec{\theta} \mid \vec{\mu}, \Sigma)$$

$$q(\theta) = \mathcal{N}(\theta \mid \mu, \Sigma)$$

$$KL(q \mid p(\Theta \mid D)) \to \min_{\mu, \Sigma}$$

... more on this later

Expectation propagation

VI: $KL(q \mid p) \to \min$

EP: $KL(p \mid q) \to \min_q$, $q \in Q$ - exp. family

$$q_\theta(\vec{z}) = h(\vec{z}) \exp(\vec{\theta}\, T(\vec{z}) + A(\vec{\theta}))$$

$$KL(p \mid q) = \int p(\vec{z}) \ln \frac{p(\vec{z})}{q_\theta(\vec{z})} d\vec{z} =$$

$$= \text{const} - \int p(\vec{z}) \left[\ln h(\vec{z}) + A(\vec{\theta}) + \vec{\theta}^T T(\vec{z}) \right] d\vec{z} =$$

$$= \text{const} - A(\vec{\theta}) - \vec{\theta}^T \mathbb{E}_p T(\vec{z})$$

$$-\nabla_\theta A(\vec{\theta}) = \mathbb{E}_p T(\vec{z}) \approx \frac{1}{n} \sum_{i=1}^{n} T(\vec{z}_i), \quad \vec{z}_i \sim p(\vec{z})$$

$\mathbb{E}_q [T(\vec{z})] = \mathbb{E}_p T(\vec{z})$ - method of moments

$$q(\vec{z}) \sim N(z \mid \vec{\mu}, \Sigma),$$

$\vec{\mu}, \Sigma$ - from p