Lecture 8 - Variational Inference

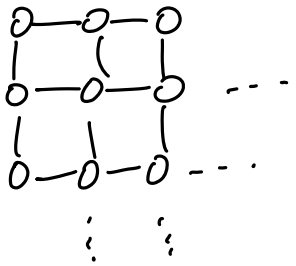
1. Variational lower bounds

- **Motivation**

→ Inference is often intractable !

Ex: • inference in non-tree graphical model

e.g. Ising model



$$p(x) \propto \exp\left\{ \sum_j \theta_j x_j + \sum_{i<j} \theta_{ij} x_i x_j \right\} \qquad x_j \in \{\pm 1\}$$

$$\mu_j = \mathbb{E}_\theta[x_j] \qquad \mu_{ij} = \mathbb{E}_\theta[x_i x_j]$$

→ BP does not converge to correct marginals

→ exact marginals are hard to compute

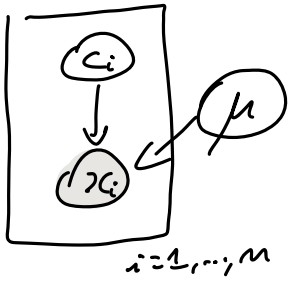
• Complicated posteriors in Bayesian inference

e.g. Bayesian mixture model

$$c_i \sim \text{Unif}(\{1,\dots,K\})$$

$$\mu_k \sim \mathcal{N}(0,\sigma^2) \quad \text{(prior)}$$

$$x_i \mid c_i = k \sim \mathcal{N}(\mu_k, 1)$$



Latent variables are $z = (c, \mu)$

posterior : $p(z|x) = \dfrac{p(c,\mu,x)}{p(x)}$ ← ok

← Not ok !!

requires computing intractable integral

$$p(x) = \sum_c p(c) \int \cdots \int p(\mu) \prod_{i=1}^{n} p(x_i|c_i,\mu)\, d\mu_1 \ldots d\mu_k$$

$c = (c_1, \ldots, c_n)$ take K^n possible values !

Variational lower bound in Exponential Families

$$p(x;\theta) = \exp\left(\langle \theta, \phi(x) \rangle - A(\theta)\right)$$

$$A(\theta) = \log \int \exp(\langle \theta, \phi(x) \rangle)\, d\nu(x)$$

- recall:
  - θ : natural parameter
  - $\mu(\theta) = \mathbb{E}_\theta[\phi(x)] = \nabla A(\theta)$: mean parameter
  - $\theta \mapsto \mu$: inference

- $\mu(\theta)$ achieves maximum in Legendre dual:

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \langle \theta, \mu \rangle - A^*(\mu)$$

A^* ≈ neg-entropy

$$\mathcal{M} = \{\mu \in \mathbb{R}^d : \exists\, q,\ \mathbb{E}_q[\phi(x)] = \mu\}$$

(convex set, polytope if $\mathcal{X}$ is discrete)

→ convex problem in μ

- Lower bound on $A(\theta)$

$$A(\theta) = \log \int \exp\left(\langle \theta, \phi(x) \rangle\right) d\nu(x)$$

$$= \log \int q(x) \frac{\exp(\langle \theta, \phi(x) \rangle)}{q(x)} d\nu(x) \qquad \text{for any } q(x)$$

$$\geq \mathbb{E}_q\left[\log \frac{\exp(\langle \theta, \phi(x) \rangle)}{q(x)}\right] \qquad \text{(Jensen's ineq.)}$$

$$= \mathbb{E}_q\left[\langle \theta, \phi(x) \rangle\right] - \mathbb{E}_q\left[\log q\right]$$

$$= \langle \theta, \mu_q \rangle + H(q) \qquad \left(\mu_q := \mathbb{E}_q\{\phi(x)\}\right)$$

→ Equality iff $q(x) \propto \exp(\langle \theta, \phi(x) \rangle)$, i.e. $\begin{cases} q(x) = p(x;\theta) \\ \mu_q = \mu(\theta) \end{cases}$ (exact inference)

⇒ "Inference as optimization" / Variational inference (V.I.)

$$A(\theta) = \sup_{q(x)} \langle \theta, \mathbb{E}_q[\phi(x)] \rangle + H(q)$$

V.I.: - consider tractable families of distributions $q \in Q$ amenable to tractable optimization

- usually, pick Q s.t. $\mu_q = \mathbb{E}_q\{\phi(x)\}$ is easy to compute for $q \in Q$
  e.g. q factorizes over disconnected graph
- parameterize by μ_q, express $H(q) = \bar{H}(\mu_q)$
- not always convex! $\lambda q_1 + (1-\lambda) q_2 \notin Q$

- Link with KL divergence

Gap in the lower bound:

$$A(\theta) - \mathbb{E}_q\left[\log \frac{\exp(\langle\theta, \phi(x)\rangle)}{q(x)}\right] = A(\theta) - \mathbb{E}_q\left[\log \frac{\exp(\langle\theta, \phi(x)\rangle - A(\theta))\exp(A(\theta))}{q(x)}\right]$$

$$= \mathbb{E}_q\left[\log \frac{q(x)}{p(x;\theta)}\right] = KL(q \| p_\theta)$$

> Maximize L.B. on $A(\theta)$ $\Leftrightarrow$ Minimize $KL(q \| p_\theta)$

Variational lower bound for posterior inference

$\rightarrow$ $p(x) = \int p(x, z)\, dz$

$\uparrow$ "evidence"

x: observed

z: latent variables / parameters in Bayesian inference

- Evidence lower bound (ELBO)

$$\log p(x) = \log \int p(x, z)\, dz$$

$$= \log \mathbb{E}_q\left[\frac{p(x, z)}{q(z)}\right]$$

$$\geq \mathbb{E}_q[\log p(x, z)] + H(q) =: ELBO(q)$$

Note:

$\rightarrow$ optimal q: $q(z) = p(z|x)$

$\rightarrow$ $\max_q ELBO(q) \Leftrightarrow \min_q KL(q \| p(\cdot|x))$

$\rightarrow$ similar lower bound used for EM algorithm

2. Mean-field Variational Inference

Key idea : use product distributions q as tractable families
i.e. q factorizes on the disconnected graph

- **Ex : Ising model**

$G = (V, E)$ $\quad x_i \in \{0, 1\}$

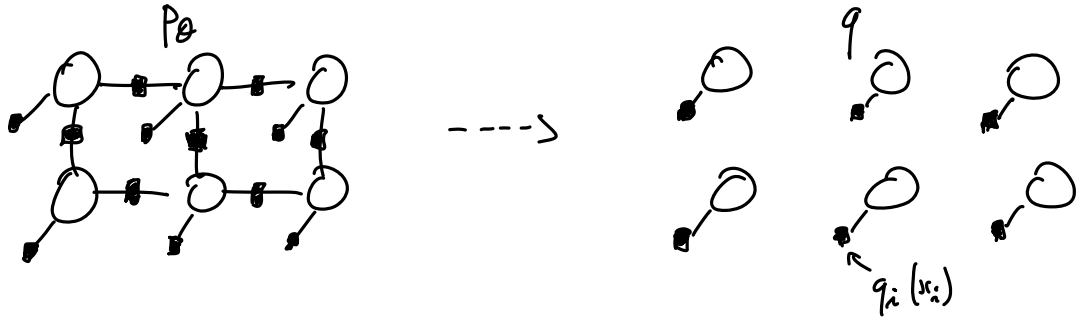
$$p(x;\theta) \propto \exp\left\{ \sum_{i \in V} \theta_i x_i + \sum_{(i,j) \in E} \theta_{ij} x_i x_j \right\}$$

mean params : $\mu_i = \mathbb{E}_\theta [X_i]$ $\quad \mu_{ij} = \mathbb{E}_\theta [X_i X_j]$ $\quad \phi(x) = \begin{pmatrix} (x_i)_i \\ (x_i x_j)_{ij} \end{pmatrix}$

→ Consider a fully factorized (product) family $\mathcal{Q}$

$$q(x) = \prod_{i \in V} q_i(x_i)$$

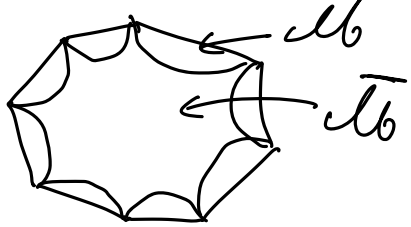
parameterized by $\mu_i := q_i(x_i = 1) = \mathbb{E}_q [X_i]$



Note :- This choice imposes the constraint $\mu_{ij} = \mu_i \cdot \mu_j$ on the set of realizable mean params :

$$\bar{\mathcal{M}} = \{ \mu \in \mathcal{M} : \mu_{ij} = \mu_i \cdot \mu_j \} \subset \mathcal{M}$$

- $\bar{\mathcal{M}}$ may be non-convex !



- Variational Lower bound

$$A(\theta) = \sup_q \; \langle \theta, \mathbb{E}_q[\phi(x)] \rangle + H(q)$$

$$\geq \sup_{q \in \mathcal{Q}} \; \langle \theta, \mathbb{E}_q[\phi(x)] \rangle + H(q)$$

$$= \max_{(\mu_i)_{i \in V}} \; \sum_i \theta_i \mu_i + \sum_{(ij) \in E} \theta_{ij} \mu_i \mu_j + \mathbb{E}_q\Big[\sum_i \log q_i(x_i)\Big]$$

$$= \max_{(\mu_i)_{i \in V}} \; \sum_i \theta_i \mu_i + \sum_{ij} \theta_{ij} \mu_i \mu_j + \sum_i \bar{H}(\mu_i) =: L(\mu)$$

with $\bar{H}(\mu_i) = \mu_i \log \mu_i + (1-\mu_i) \log (1-\mu_i)$

- Mean Field V.I. algorithm
  - → coordinate ascent on $L(\mu)$, i.e., maximize w.r.t. each μ_i separately.

$$\frac{\partial L}{\partial \mu_i} = \theta_i + \sum_{j \in N(i)} \theta_{ij} \mu_j + \log \mu_i + 1 - \log(1-\mu_i) - 1$$

$$\frac{\partial L}{\partial \mu_i} = 0 \Rightarrow \quad \log \frac{\mu_i}{1-\mu_i} = \theta_i + \sum_{j \in N(i)} \theta_{ij} \mu_j$$

$$\frac{\mu_i}{1-\mu_i} = \exp\Big(\theta_i + \sum_j \theta_{ij} \mu_j\Big)$$

$$\mu_i = \sigma\Big(\theta_i + \sum_{j \in N(i)} \theta_{ij} \mu_j\Big)$$

$$\sigma(z) = \frac{1}{1+\exp(-z)}$$

Algorithm: (MFVI for Ising model)

While $L(\mu)$ has not converged:

For $i \in V$:

$$\mu_i \leftarrow \sigma\left(\theta_i + \sum_{j \in N(i)} \theta_{ij} \mu_j\right)$$

- **Ex: Bayesian Inference**

$$p(x) = \int p(x, z)\, dz \quad \text{where} \quad z \rightarrow \begin{cases} \text{hidden variables} \\ \text{parameters} \end{cases}$$

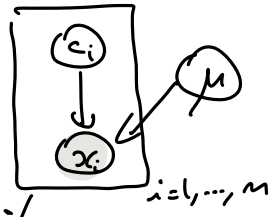
Ex: Bayesian mixture model (BMM): $z = ((c_i)_i, (\mu_k)_k)$

($(c_i)_i$: cluster assignments; $(\mu_k)_k$: cluster parameters)

$c_i \sim \text{Unif}(\{1, \dots, K\})$

$\mu_k \sim \mathcal{N}(0, \sigma^2)$

$x_i \mid c_i = k \sim \mathcal{N}(\mu_k, 1)$



- **Factorized variational family:**

$$q(z) = \prod_{j=1}^{m} q_j(z_j)$$

- **Generic coordinate ascent algorithm:**

$$\text{ELBO}(q) = \mathbb{E}_q[\log p(x, z)] + \sum_j \bar{H}(q_j)$$

As a function of q_j:

$$\text{ELBO}(q) = \mathbb{E}_q[\log p(z_j \mid z_{-j}, x)] + \bar{H}(q_j) + \text{const}(q_{-j})$$

max over q_j $\Rightarrow$ $q_j(z_j) \propto \exp\left(\mathbb{E}_{q_{-j}}[\log p(z_j | z_{-j}, x)]\right)$

Algorithm: (Coordinate Ascent Variational Inference)

While ELBO(q) has not converged:

For $j \in \{1, \dots, m\}$:

$q_j(z_j) \propto \exp\{\mathbb{E}_{-j}[\log p(z_j | z_{-j}, x)]\}$

Remark:

- usually, take $q_j(z_j)$ to have same form as $p(z_j | z_{-j}, x)$
- $p(z_j | z_{-j}, x)$ is often available in closed form when
  - $\rightarrow$ z_j is a discrete latent variable
  - $\rightarrow$ z_j is a parameter with conjugate prior to $p(x | z_j)$

Def: (conjugate prior)

$p(z)$ is a conjugate prior to the likelihood $p(x|z)$ when the posterior $p(z|x) \propto p(x|z)p(z)$ belongs to the same family

Ex:

- Beta-Bernoulli

$p(z) = \text{Beta}(z, \alpha, \beta) \propto z^{\alpha-1}(1-z)^{\beta-1}$

$p(x|z) = \text{Ber}(x, z) \propto z^{x}(1-z)^{1-x}$

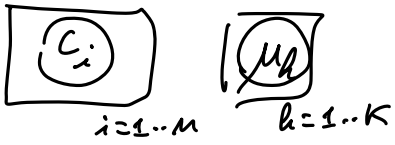
$\Rightarrow p(z|x) = \text{Beta}(z, \alpha+x, \beta+(1-x))$

- other: Dirichlet-Multinomial, Gaussian-Gaussian (fixed covariance)

- **BMM example**: (see [Blei, Kucukelbir, McAuliffe '17])

→ $q(z) = q(\mu, c) = \prod_{k=1}^{K} q(\mu_k; m_k, s_k^2) \prod_{i=1}^{m} q(c_i; \varphi_i)$

Gaussian — categorical



We have:

$$ECBO(q) = \sum_{i=1}^{m} \mathbb{E}_q[\log p(x_i, c_i | \mu)] + \sum_i \bar{H}_i(\varphi_i)$$

$$+ \sum_{k=1}^{K} \mathbb{E}_q[\log p(\mu_k)] + \sum_k \bar{H}_k(m_k, s_k^2)$$

→ max over $\{\varphi_i\}$ is similar to an E-step:

$$q(c_i) \propto \exp\{\mathbb{E}_\mu[\log p(x_i, c_i | \mu)]\}$$

$$\varphi_{ik} \propto \exp\{\mathbb{E}_\mu[\mu_k] x_i - \tfrac{1}{2}\mathbb{E}_\mu[\mu_k^2]\}$$

→ max over $\{m_k, s_k\}$ is similar to M-step:

$$m_k = \frac{\sum_i \varphi_{ik} x_i}{\frac{1}{\sigma^2} + \sum_i \varphi_{ik}}, \quad s_k^2 = \frac{1}{\frac{1}{\sigma^2} + \sum_i \varphi_{ik}}$$

Algorithm: ("Variational Bayes" for BMM)

While ELBO has not converged:

(E-step)

For $i = 1 \dots n$

$$\varphi_{ik} \propto \exp\left\{ \mathbb{E}_\mu[\mu_k] x_i - \frac{1}{2} \mathbb{E}_\mu[\mu_k^2] \right\}$$

(M-step)

for $k = 1, \dots, K$

$$m_k = \frac{\sum_i \varphi_{ik} x_i}{\frac{1}{\sigma^2} + \sum_i \varphi_{ik}}, \quad s_k^2 = \frac{1}{\frac{1}{\sigma^2} + \sum_i \varphi_{ik}}$$

3. Stochastic VI and Online EM

→ In EM and (Bayesian) VI, the E-step often requires computing expected sufficient statistics across the entire dataset

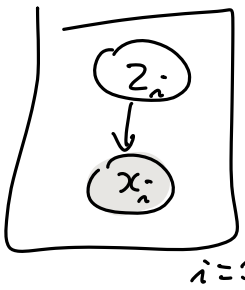
Q: Can we go faster for large datasets?

→ Yes, using stochastic approximation!

Ex: Online EM (Stochastic VI is similar)

[Neal & Hinton, '98], [Cappé & Moulines '09]

[Hoffman et al. '13]



n i.i.d. observations x_i

$$\log p(x;\theta) \geq \mathbb{E}_q\left[\log p(x,z;\theta)\right] + H(q)$$
$$= \sum_i \mathbb{E}_{q_i}\left[\log p(x_i,z_i;\theta)\right] + H(q_i)$$

→ In exponential families, the M-step ($\arg\max_\theta$) usually depends on "Expected sufficient statistics"

$$S_q = \sum_i \mathbb{E}_q\left[\phi(x_i,z_i)\right]$$

e.g. for Gaussian Mixtures : $S_q = \left(S_q^{1,k}, S_q^{2,k}\right)_k$

$$\begin{cases} S_q^{0,k} = \sum_i \mathbb{E}_q\left[\mathbb{1}\{z_i=k\}\right] \\ S_q^{1,k} = \sum_i \mathbb{E}_q\left[\mathbb{1}\{z_i=k\}\right] x_i \\ S_q^{2,k} = \sum_i \mathbb{E}_q\left[\mathbb{1}\{z_i=k\}\right] x_i x_i^T \end{cases}$$

⇒ M-step : $\mu_k = \dfrac{S_q^{1,k}}{S_q^{0,k}} \qquad \Sigma_k = \dfrac{S_q^{2,k}}{S_q^{0,k}}$

→ Instead of updating each q_i at every E-step, Online EM update only one :

(Initialize $S^{(0)} = \sum_i S_i^{(0)}$)

- E-step : pick random $i \in [n]$

$$q_i(z_i) \propto p(z_i | x_i ; \theta_t)$$
$$S_i^{(t+1)} = \mathbb{E}_{q_i}\left[\phi(x_i,z_i)\right]$$
$$S^{(t+1)} = S^{(t)} + \left(S_i^{(t+1)} - S_i^{(t)}\right)$$

- M-step : update θ_{t+1} using $S^{(t+1)}$

→ Can also use stochastic approximation on infinite data stream $(x_t)_{t=1,\dots}$

$$S^{(t+1)} = (1-\alpha_t)\, S^{(t)} + \alpha_t\, S_t^{(t)}$$