Variational Inference

\[ \text{Variational Lower Bound} \]

\[ \log p(x) = \log \int p(x, \theta) \, d\theta = \log \int q(\theta) \frac{p(x, \theta)}{q(\theta)} \, d\theta = \log \mathbb{E}_{q}[\frac{p(x, \theta)}{q(\theta)}] \]
\[ \geq \mathbb{E}_{q}[\log \frac{p(x, \theta)}{q(\theta)}] \quad \text{(Jensen's inequality)} \]
\[ = -\text{KL}(q \parallel p(x, \theta)) \]

\[ \Rightarrow \text{Variational optimization finds tightest lower bound:} \]
\[ \log p(x) \geq \max_{q \in \mathcal{Q}} -\text{KL}(q(\theta) \parallel p(x, \theta)) \]
for some class of dist's \( \mathcal{Q} \)
- Bound is tight iff \( q(\theta) = p(\theta | x) \)
- Key Ideas: VI turns statistical inference into optimization

\[ \Rightarrow \text{Different VI methods amount to different } \mathcal{Q} \]
- Mean Field = \( \mathcal{Q}_{MF} \) marginal posterior independence
- Belief Propagation = \( \mathcal{Q}_{BP} \) "Tree-like" posterior
- Expectation Propagation = \( \mathcal{Q}_{EP} \) "Tree-Like" in moments

\[ \Rightarrow \text{Aside: Info-Theoretic interpretation suggests we should minimize} \]
\[ \min_{q} \text{KL}(q(\theta | x) \parallel q(\theta)) = \mathbb{E}_{q(\theta | x)} \left[ \log \frac{p(\theta | x)}{q(\theta)} \right] \]

\[ \text{Can't compute this, don't have this} \]
Mean Field Variational

⇒ "Mean field" assumes d-dimensional $\Theta$, marginally independent:

\[ q(\Theta) = \prod_{i=1}^{d} q_i(\Theta_i) \]

⇒ Originates from many-body problem in stat. mechanics:

\[ \xi_3 = \text{"microstates" (e.g., spin, velocity, ...)} \]

\[ p(\xi) \approx \prod_{i=1}^{d} \frac{1}{Z_i} e^{-\beta H_i(\xi_i)} \]

GIBBS' DIST'N
BOLTZMANN DIST'N

⇒ Mean field variational lower bound:

\[ \log p(x) \geq \max_q \mathbb{E}_q \left[ \log p(\Theta, x) \right] + \sum_i H(q_i) \triangleq \mathcal{L}(q) \]

- Sometimes called "Evidence Lower Bound" (ELBO)

⇒ How do we optimize $\mathcal{L}(q)$? Coordinate ascent.

- Take derivative w.r.t. each $q_i(\Theta_i)$, set to 0, solve:

\[ q^*_i(\Theta_i) \propto \exp \left\{ \mathbb{E}_{q_{\neq i}} \left[ \log p(\Theta, x) \right] \right\} \]

(*)

Where $\mathcal{T}_i = \{ \Theta_j \mid j \neq i \}$  

**KNOWN AS "COMPLETE CONDITIONAL"**
COORDINATE ASCENT VIS:

\( \Rightarrow \) Monotonically increases \( \mathcal{L}(\psi) \)

\( \Rightarrow \) Fixed-points of (**) \( \Leftrightarrow \) Local optima of \( \mathcal{L}(\psi) \)

**Ex. Gaussian Mixture Model**

\[ m_k \sim N(0, \sigma^2) \text{ for } k=1, \ldots, K \]

\[ z_n \sim \text{CAT}(\pi_1, \ldots, \pi_K) \text{ for } n=1, \ldots, N \]

\[ x_n | z_n \sim N(m_{z_n}, 1) \]

- Variational distributions:
  \[ q_n(z_n | \pi_n) = \text{CAT}(z_n | \pi_n) \]
  \[ q_n(m_k | \mu_k, S_k) = N(m_k | \mu_k, S_k) \]

where \( \{ \pi_n, \mu_k, S_k \} \) are variational parameters

---

1. **Update Cluster Assignments**

\[ q_n(z_n | \pi_n) \propto \exp \left\{ \log p(z_n) + \mathbf{E}_{q(m)} \left[ \log p(x_n | z_n, m) \right] \right\} \tag{**} \]

   \[ \text{Prior} = \pi_k \]

   \[ \text{Expected Log-Likelihood} \]

- Compute expected log-likelihood:

\[ \mathbf{E}_{q(m)} \left[ \log p(x_n | z_n, m) \right] = \mathbf{E} \left[ \sum_k \mathbb{I}(z_n = k) \log N(x_n | m_k, 1) \right] \]

\[ = \sum_k \mathbb{I}(z_n = k) \mathbf{E} \left[ -\frac{1}{2} \log 2\pi 1 - \frac{1}{2} x_n^2 + \frac{x_n m_k^2}{2 m_k^2} \right] \]

\[ = \sum_k \mathbb{I}(z_n = k) \left( \mathbf{E} \left[ q_{m_k}(m_k) \right] m_k \right) x_n - \frac{1}{2} \left( \mathbf{E} \left[ q_{m_k}(m_k) m_k^2 \right] \right) + \text{const.} \]

- Plug back into (**)

\[ q_n(z_n = k) = \gamma_n(k) \propto \exp \left\{ m_k x_n - \frac{1}{2} (S_n + m_k^2) \right\} \]

2. **Update Component means** \( q_n(\mu_k, \mu_k, S_k) \) (we will skip this)
CONDITIONALLY CONJUGATE MODELS

⇒ In the GMM example our update was:

\[ q_n(z_n) \propto p(z_n) \exp \left[ \mathbb{E} \left[ \log p(x_n | z_n, m) \right] \right] \propto \text{Cat}(\cdot) \]

which is same family as prior \( p(z_n) \)

- We call this conditional conjugacy

⇒ In general this holds when:

1) Complete conditionals are expfamily

2) \( q(\theta) \propto p(\theta) \exp \left[ \mathbb{E} \left[ \log p(x, z | \theta) \right] \right] \)

\[ = p(\theta) f(x | \theta) \]

is conjugate ⇒ lies in same family as \( p(\theta) \)

⇒ Conditional conjugacy ⇒ Closed-form updates

VARIATIONAL OPTIMIZATION

⇒ How did we get fixed-point condition?

\[ q_i(\theta_i) \propto \exp \left[ \mathbb{E} \left[ \log p(\theta_i | \theta_{-i}, x) \right] \right] \]

⇒ Form Lagrangian of variational problem:

\[ \max_{q} \mathcal{L}(q) \triangleq \mathbb{E}_q \left[ \log p(x, \theta) \right] + \sum_i H(q_i(\theta_i)) \]

s.t. \( \int q_i(\theta_i) = 1 \quad \forall i = 1, \ldots, N \)

Lagrangian:

\[ \mathcal{L}(q) + \sum_i \lambda_i (1 - \int q_i(\theta_i) d\theta_i) \triangleq \mathcal{J}(q, \lambda) \]

⇒ Calculus of variations defines \textit{derivative} of functional

\[ \mathcal{J}(f) = \int L(f(x)) \, dx \]

As:

\[ \frac{\partial \mathcal{J}}{\partial f(x)} = \frac{\partial L}{\partial f} \]

\textbf{Note:} This is a simplification of the Euler-Lagrange equations.
Take functional derivative wrt $q_i(\theta_i)$:

$$\frac{\partial}{\partial q_i(\theta_i)} L(q) + \frac{\partial}{\partial q_i(\theta_i)} \lambda_i (1 - \int q_i(\theta_i) d\theta_i)$$

$$= \frac{\partial}{\partial q_i(\theta_i)} \mathbb{E}_{q_i} \left[ \mathbb{E}_{q_{\theta_i}} \left[ \log p(\theta_i, \theta_{\theta_i}, x_i) \right] \right]$$

$$+ \frac{\partial}{\partial q_i(\theta_i)} H(q_i(\theta_i)) - \lambda_i$$

$$= \mathbb{E}_{q_i} \left[ \log p(\theta_i, \theta_{\theta_i}, x_i) \right] - \log q_i(\theta_i) - 1 - \lambda_i = 0 \Rightarrow$$

$$q_i(\theta_i) = \exp \left\{ \mathbb{E}_{q_i} \left[ \log p(\theta_i, \theta_{\theta_i}, x_i) \right] \right\} \frac{1}{\lambda_i \mathbb{Z}_i(\lambda_i)}$$
Beste Variational Problem:

⇒ Key Ideas: Optimize over dist'n \( q(\theta) \) consistent w/ tree-structured MRFs.

⇒ Any tree-structured dist'n \( q(\theta) \) can be written as:

\[
q(\theta) = \prod_{s \in V} q_s(\theta_s) \prod_{(s,t) \in E} \frac{q_{st}(\theta_s, \theta_t)}{q_s(\theta_s) q_t(\theta_t)}
\]

(\*)

where \( q_s(\theta_s) \) are marginals:

\[
q_s(\theta_s) = \int q(\theta_s) \, d\theta_{-s}
\]

⇒ Ex:

\[
q(\theta) = (q_{11}(\theta_1) q_{12}(\theta_2) q_{13}(\theta_3) q_{14}(\theta_4) \prod_{s} q_s(\theta_s) q_s(\theta_s))
\]

\[
\frac{q_{12}(\theta_1, \theta_2)}{q_{11}(\theta_1) q_{12}(\theta_2)} \frac{q_{13}(\theta_1, \theta_3)}{q_{11}(\theta_1) q_{13}(\theta_3)} \frac{q_{14}(\theta_1, \theta_4)}{q_{11}(\theta_1) q_{14}(\theta_4)}
\]

\[
= q_{11}(\theta_1) q_{12}(\theta_2) q_{13}(\theta_3) q_{14}(\theta_4)
\]

⇒ Dist'ns of the form (\*) necessarily satisfy Local Marginal Consistency:

\[
q_s(\theta_s) = \int q_{st}(\theta_s, \theta_t) \, d\theta_t \quad \forall \; t \in \Gamma(s) \quad \forall \; s \; \text{neighbors of } s
\]

⇒ Beste Variational Problem:

\[
\max_{q_\theta} E_q[\log p(\theta, x)] + H(q_\theta)
\]

s.t.

\[
\int q_{st}(\theta_s, \theta_t) \, d\theta_t = q_s(\theta_s) \quad \forall \; t \in \Gamma(s) \quad \text{(Local Consistency)}
\]

\[
\int q_s(\theta_s) \, d\theta_s = 1 \quad \forall \; s \in V \quad \text{(Normalization)}
\]