Approximate Inference

Variational Inference

Task: eval the posterior distribution
\[ p(\mathcal{Z} | X) \]

\[ p(\mathcal{Z} | X, \phi) \]

latent variables / params

Often the case that

- space in which \( \mathcal{Z} \) lives is very, eval'ing all possible (enumerating all) \( \mathcal{Z}'s \)
- posterior doesn't have a nice analytic form
  - continuous vars: integrations might not be closed for
  - discrete

alternative MCMC

Variational Inference

Applied to the Bayesian inference prob. (section 10.1 of PRML)

\( \mathcal{Z} \) latent vars + params (set)

\( X \) set of observed vars \( \mathcal{Z} \) might be \( \mathcal{N} \) iid obs

\( X = \{ x_1, \ldots, x_N \} \)

\( \mathcal{Z} = \{ z_1, \ldots, z_N \} \)

Probabilistic model

\[ p(X, \mathcal{Z}) \]

Goal:

Find posterior dist \( p(\mathcal{Z} | X) \) and evidence \( p(X) \)

\[ \ln p(X) = \mathcal{L}(q) + KL(q || p) \]

where

\[ \mathcal{L}(q) = \int q(\mathcal{Z}) \cdot \ln \left\{ \frac{p(X, \mathcal{Z})}{q(\mathcal{Z})} \right\} d\mathcal{Z} \]

\[ KL(q || p) = - \int q(\mathcal{Z}) \ln \left\{ \frac{p(X | \mathcal{Z})}{q(\mathcal{Z})} \right\} d\mathcal{Z} \]

multi-dimensional ints dimensionally

dim is # \( \mathcal{Z} \) vars and dimensionality of each
\[
\ln p(x) = \mathbb{L}(q) + \text{KL}(q || p)
\]

\[
= \int q(z) \ln \left\{ \frac{p(x, z)}{q(z)} \right\} dz - \int q(z) \ln q(z) dz
\]

\[
= \int q(z) \ln \left\{ \frac{p(z|x) \frac{p(x)}{q(z)}}{q(z)} \right\} dz - \int q(z) \ln p(z|x) dz + \int q(z) \ln q(z) dz
\]

\[
= \int q(z) \ln p(x) dz = \ln p(x) \int q(z) dz = \ln p(x) \cdot 1
\]

- Differs EM step only in that \( \theta \) no longer appears

- Case: maximize \( \mathbb{L}(q) \) (lower bound on the evidence) - equivalent to minimizing the KL div.

If all \( q \)'s possible then \( q(z) = p(z|x) \) is minimax but \( p(z|x) \) is complicated.

**Approach**

Restrict the family of distribution \( q(z) \) to "simple" distributions, and then to seek the member of this family that most closely approx \( p(x) \)

**Note:**

- choice of \( q(z) \) all about tractability
- more complex \( q(z) \)'s are limited by computation
- no overfitting
Choices for $Q(z)$
- Parameterized $Q(z|\omega)$ is governed by
- Factorized $Q(z) = \prod_{i=1}^{M} Q_i(z_i)$ family of approximately

$Q(z) = \prod_{i=1}^{M} Q_i(z_i)$

no restrictions on the
form of ind. $q_i$'s dists

Amongst all dists in this family, which
makes $L(q)$ the largest?

$L(q) = \int q(z) \ln \left\{ \frac{p(x,z)}{q(z)} \right\} dz$

Use family $Q(z) = \prod_{i=1}^{M} Q_i(z_i)$ where $Z_0$ is a subset
of the latest vars

cell $Q_i(z_i) = q_i$

$L(q) = \int \left( \prod_{i} q_i \right) \ln \left\{ \frac{p(x,z)}{\prod_{i} q_i} \right\} dz$

EM-like objective, find conditions at optimal $L(q)$
for each $q_i$. Split out a single term $q_i$. Max $L(q)$
\[ L(q) = \int \left( \prod_{i} q_{i} \right) \left( \ln P(X, Z) - \frac{1}{Z} \sum_{i} \ln q_{i} \right) dZ \]

Extract single factor \( q_{j} \)

\[ = \frac{\int q_{j} \left( \prod_{i \neq j} q_{i} \right) \ln P(X, Z) dZ}{\int \left( \prod_{i} q_{i} \right) \left( \sum_{i} \ln q_{i} \right) dZ} \]

\( \text{where} \ q_{j} = q_{j}(Z_{j}), \ Z_{j} \in Z \)

\[ = \int q_{j}(Z_{j}) \prod_{i \neq j} q_{i} \ln P(X, Z) dZ_{1} \cdots dZ_{j} \cdots dZ_{m} dZ_{j} \]

\[ = -\int q_{j} \left( \prod_{i \neq j} q_{i} \right) \left( \ln q_{j} + \sum_{i \neq j} \ln q_{i} \right) dZ \]

Opening \( \text{a boxed quantity} \)

\[ = -\int q_{j} \ln q_{j} dZ_{j} + \int q_{j} \left( \prod_{i \neq j} q_{i} \right) \left( \sum_{i \neq j} \ln q_{i} \right) dZ \]

Integrate \( dZ_{j} \)

\[ \text{Mess - clean it up} \]

\[ L(q) = \int q_{j} \left( \prod_{i \neq j} \ln P(X, Z) dZ_{1} \cdots dZ_{j} \cdots dZ_{m} \right) dZ_{j} \]

\[ - \int q_{j} \ln q_{j} dZ_{j} \]

If we define

\[ \ln \hat{p}(X, Z_{j}) = \mathbb{E}_{q_{j}} \left[ \ln p(X, Z) \right] \]

\[ = \int \cdots \int \ln P(X, Z) dZ_{1} \cdots dZ_{m} q_{j} dZ_{j} \]

\[ L(q) = \int q_{j} \ln \hat{p}(X, Z_{j}) dZ_{j} - \int q_{j} \ln q_{j} dZ_{j} + \text{const} \]
\[ L(q) = \sum_{z_j} \ln \tilde{p}(x, z_j) dZ_j - \sum_{z_j} \ln q_j(z_j) dZ_j \quad \text{const} \]

Goal: coordinate-wise maximization of \( L(q) \)

- In particular, right now we aim \( L(q) \) with \( q_j \)

Recognize the \( L(q) \) is a coordinate KL div. between \( \tilde{p}(x, z_j) \) and \( q_j(z_j) \)

Minimize KL divergence between \( \tilde{p}(x, z_j) \) and \( q_j(z_j) \).

The optimal \( q_j^*(z_j) \) is given by

\[
\ln q_j^*(z_j) = \mathbb{E}_{\tilde{p}} \left[ \ln \tilde{p}(x, z) \right] + const
\]

- Consider this form a second the "coordinate component" now is a function of \( z_j \) (subject of poster lat. vars.). The right hand side is the joint dist. of all obs. & latent variables, but with all latent vars. besides \( z_j \) integrated out - this leaves a function of some vars. \( z_j \) on right hand side as well.

- The specific dist. form of \( q_j^*(z_j) \) will often emerge from this rule.

- \( \tilde{p}(x, z) \) probably has interesting cond. dependencies to exploit.
- In expectation (actually \( \log \) norm) many/worst of the terms will be absorbed into the constant but not all
- Coupling between approximating factors, e.g.

\[ r_k \in \mathbb{R} \]

- No closed form sol'n in general.
\[ \ln q^*_j(z_j) = \mathcal{E}_{i,j} \left[ \ln p(x, z_i) \right] + \text{const} \]

Lacks normalization (const)

- sole yields \( q^*_j \) up to a multiplicative factor

- one can normalize this distribution by either

- (usually)
  - inspection (will be become clear)
  - or by explicit normalization

\[ q^*_j(z_j) = \frac{\exp \left( \mathcal{E}_{i,j} \left[ \ln p(x, z_i) \right] \right)}{\int \exp \left( \mathcal{E}_{i,j} \left[ \ln p(x, z_i) \right] \right) dz_j} \]

- Set of the equals for all \( q^*_i, i = 1 \ldots M \) is a set of "consistencies" conditions for the max. They are not an explicit soln, nor in general closed form, and have to be cycled through until numerical convergence.

**State without proof**

Convergence of these interdependent updates is guaranteed because the objective is convex.

**Teaching Example**

Variational approximation to a full covariance Gaussian (\( \mathbb{R}^D \))

Recall, in general \( z \in \mathbb{R}^D \) \( \sim \mathcal{N}(m, \Sigma) \)

\[ p(z) \neq p(z_1) p(z_2) \] unless \( \Sigma = \mathcal{D}_m \)

and \( z_1, \ldots, z_2 \)

Goal: find the independent Gaussian dist (diagonal Gaussian) that best approximates \( p(z) \).

Factorization \( q(z_1, z_2) \)