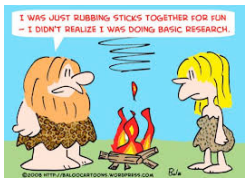
BTRY 6790, Probabilistic Graphical Models



Nov. 14, 2013

Plan for Today

- Introduction to variational inference

Variational Inference

- As discussed, intractable inference problems can also be approached with deterministic approximation schemes
- Examples include:
  - Methods based on the ***Laplace approximation***
  - ***Variational inference***
  - ***Expectation propagation***
- We will focus on variational inference, giving a general intro and overview

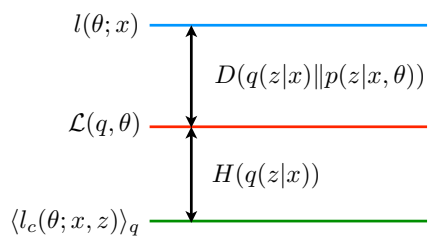
Calculus of Variations

- Variational methods are named for the ***calculus of variations***, developed in the 1700s by Euler, Lagrange, and others
- The COV is a kind of "meta-calculus" concerned with ***functionals*** (which take functions as arguments) rather than ***functions*** (which take numbers).
- Examples include:
  - Entropy: $H[p] = \int p(x) \log p(x) dx$
  - KLD wrt some $q(x)$: $D[p] = \int p(x) \log \frac{p(x)}{q(x)} dx$
  - Arc length: $A[f] = \int_{x_1}^{x_2} \sqrt{1 + |f'(x)|^2}\, dx$

Calc. of Var., cont.

- The COV is often useful for finding an extremal value of a function *f*. For example,
  - What function *f* defines the curve (on a particular surface) of shortest length between two points?
  - What pdf *f* (subject to certain constraints) minimizes the KL divergence with respect to a distribution of interest?
- In variational inference, it is used to find a tractable (factorized) distribution that is as close as possible to the desired intractable distribution, in terms of the KL divergence

Recall: Key Quantities in EM



Recall: Bounds

- Consider the quantity:

$$\mathcal{L}(q,\theta) = \sum_z q(z|x) \log \frac{p(x,z|\theta)}{q(z|x)}$$

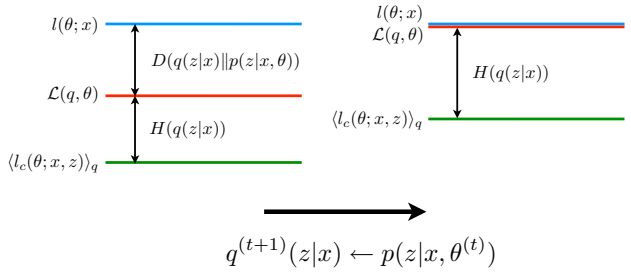
- $\mathcal{L}(q, \theta)$ is an *upper bound* on the ECLL:

$$\begin{aligned}\mathcal{L}(q,\theta) &= \sum_z q(z|x) \log p(x,z|\theta) - \sum_z q(z|x) \log q(z|x) \\ &= \langle l_c(\theta;x,z)\rangle_q + H(q(z|x))\end{aligned}$$

- $\mathcal{L}(q, \theta)$ is a *lower bound* on the ILL:

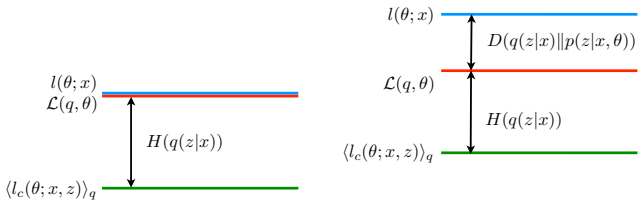
$$\begin{aligned}\mathcal{L}(q,\theta) &= \sum_z q(z|x) \log \frac{p(z|x,\theta)p(x|\theta)}{q(z|x)} \\ &= \sum_z q(z|x) \log \frac{p(z|x,\theta)}{q(z|x)} + \sum_z q(z|x) \log p(x|\theta) \\ &= -D(q(z|x)\|p(z|x,\theta)) + \log p(x|\theta)\end{aligned}$$

E-Step



$$q^{(t+1)}(z|x) \leftarrow p(z|x,\theta^{(t)})$$

M-Step



$$\theta^{(t+1)} \leftarrow \arg\max_\theta \langle l_c(\theta;x,z)\rangle_{q^{(t+1)}}$$

Variational Inference Idea

- Instead of using the posterior for the averaging distribution $q(z)$, assume it has a general form that is easy to work with
- Using variational principles, adjust $q(z)$ to maximize the value of $\mathcal{L}$, or equivalently, to minimize the KLD of $q(z)$ wrt the posterior
- While $\mathcal{L}$ will not in general equal the log likelihood, it will remain a valid lower bound
- Thus, the procedure will both "fit" an approximate posterior and produce a lower bound to the log likelihood

Fully Bayesian Model

- In a fully Bayesian setting, the parameters are also random. Assume θ is included with the latent variables Z
- In addition, allow the latent variables to be either discrete or continuous. Thus,

$$\mathcal{L}(q) = \int q(Z) \log \frac{p(X,Z)}{q(Z)} dZ$$

$$D(q\|p) = \int q(Z) \log \frac{p(Z|X)}{q(Z)} dZ$$

$$\log p(X) = \log \left\{ \int p(X|Z)p(Z)\, dZ \right\}$$

Factorized $q(Z)$

- Assume Z is partitioned into M disjoint groups, Z_1, ..., Z_M, and $q(Z)$ factorizes as:

$$q(Z) = \prod_{i=1}^{M} q_i(Z_i)$$

- Thus, $\mathcal{L}$ can be written:

$$\begin{aligned}\mathcal{L}(q) &= \int q(Z) \log \frac{p(X,Z)}{q(Z)} dZ \\ &= \int \left(\prod_i q_i(Z_i) \right) \left(\log p(X,Z) - \sum_i \log q_i(Z_i) \right) dZ\end{aligned}$$

Dependence on $q_j(Z_j)$

$$\begin{aligned}
\mathcal{L}(q) &= \int \left(\prod_i q_i(Z_i)\right)\left(\log p(X,Z) - \sum_i \log q_i(Z_i)\right) dZ \\
&= \int\int q_j(Z_j)\left(\prod_{i\neq j} q_i(Z_i)\right)\left(\log p(X,Z) - \log q_j(Z_j) - \sum_{i\neq j}\log q_i(Z_i)\right) dZ_{-j}\, dZ_j \\
&= \left\{\int q_j(Z_j)\int\left(\prod_{i\neq j} q_i(Z_i)\right)\log p(X,Z)\, dZ_{-j}\, dZ_j\right\} - \left\{\int q_j(Z_j)\log q_j(Z_j)\left(\int\prod_{i\neq j} q_i(Z_i)\, dZ_{-j}\right) dZ_j\right\} \\
&\quad - \left\{\int\left(\prod_{i\neq j} q_i(Z_i)\right)\sum_{i\neq j}\log q_i(Z_i)\left(\int q_j(Z_j)\, dZ_j\right) dZ_{-j}\right\} \\
&= \int q_j(Z_j)\log\tilde{p}(X,Z_j)\, dZ_j - \int q_j(Z_j)\log q_j(Z_j)\, dZ_j + C \\
&= \int q_j(Z_j)\log\frac{\tilde{p}(Z_j|X)}{q_j(Z_j)} + C' = -D(q_j\|\tilde{p}) + C'
\end{aligned}$$

where $\log\tilde{p}(X,Z_j) = \int\left(\prod_{i\neq j} q_i(Z_i)\right)\log p(X,Z)\, dZ_{-j}$ and $\tilde{p}(Z_j|X) = \frac{\tilde{p}(X,Z_j)}{\int\tilde{p}(X,Z'_j)\, dZ'_j}$

Maximization

- Thus, maximization of $\mathcal{L}$ with respect to $q_j(Z_j)$ is accomplished by setting $\log q_j(Z_j)$ equal to a normalized version of:

$$\begin{aligned}
\log\tilde{p}(X,Z_j) &= \int\left(\prod_{i\neq j} q_i(Z_i)\right)\log p(X,Z)\, dZ_{-j} \\
&= E_{q_{-j}}[\log p(X,Z)]
\end{aligned}$$

- That is:

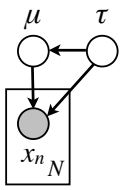
$$q_j^*(Z_j) = \frac{\exp\left(E_{q_{-j}}[\log p(X,Z)]\right)}{\int\exp\left(E_{q_{-j}}[\log p(X,Z)]\, dZ_j\right)}$$

- Maximization can be done for each j in turn; convergence is guaranteed

Ex #1: Univariate Gaussian

- Assume a fully Bayesian univariate Gaussian model with:

$$\begin{aligned}
\mathcal{D} &= \{x_1,\ldots,x_N\} \\
p(\mathcal{D}|\mu,\tau) &= \left(\frac{\tau}{2\pi}\right)^{N/2}\exp\left\{-\frac{\tau}{2}\sum_{n=1}^N(x_n-\mu)^2\right\} \\
p(\mu|\tau) &= \mathcal{N}(\mu|\mu_0,(\lambda_0\tau)^{-1}) \\
p(\tau) &= \mathrm{Gam}(\tau|a_0,b_0)
\end{aligned}$$



- Therefore,

$$\begin{aligned}
\log p(\mathcal{D},\mu,\tau) &= \log p(\mathcal{D}|\mu,\tau) + \log p(\mu|\tau) + \log p(\tau) \\
&= \frac{N}{2}\log\tau - \frac{\tau}{2}\left\{\sum_{n=1}^N(x_n-\mu)^2\right\} + \frac{1}{2}\log\tau \\
&\quad - \frac{\lambda_0\tau}{2}(\mu-\mu_0)^2 + (a_0-1)\log\tau - b_0\tau + C
\end{aligned}$$

Variational Approach

- Assume the averaging distrib. factors as:

$$q(\mu,\tau) = q_\mu(\mu)q_\tau(\tau)$$

- Applying the general result from above,

$$\begin{aligned}
\log q_\mu^*(\mu) &= E_\tau[\log p(\mathcal{D},\mu,\tau)] + C \\
&= E_\tau\left[-\frac{\tau}{2}\left\{\sum_{n=1}^N(x_n-\mu)^2\right\} - \frac{\lambda_0\tau}{2}(\mu-\mu_0)^2\right] + C \\
&= -\frac{E[\tau]}{2}\left\{\lambda_0(\mu-\mu_0)^2 + \sum_{n=1}^N(x_n-\mu)^2\right\} + C
\end{aligned}$$

- It can be shown, by completing the square, that:

$$\begin{aligned}
q_\mu^*(\mu) &= \mathcal{N}(\mu|\mu_N,\lambda_N^{-1}) \\
\mu_N &= \frac{\lambda_0\mu_0 + N\bar{x}}{\lambda_0+N} \\
\lambda_N &= (\lambda_0+N)E[\tau]
\end{aligned}$$

Variational Approach, cont.

- Similarly,

$$\begin{aligned}
\log q_\tau^*(\tau) &= E_\mu[\log p(\mathcal{D},\mu,\tau)] + C \\
&= \frac{N}{2}\log\tau + (a_0-1)\log\tau - b_0\tau \\
&\quad - \frac{\tau}{2}E_\mu\left\{\lambda_0(\mu-\mu_0)^2 + \sum_{n=1}^N(x_n-\mu)^2\right\} + C
\end{aligned}$$

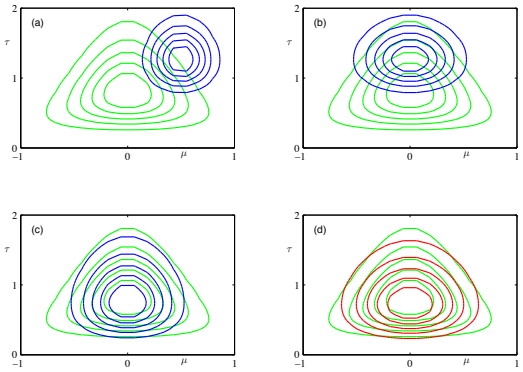
- Therefore,

$$\begin{aligned}
q_\tau(\tau) &= \mathrm{Gam}(\tau|a_N,b_N) \\
a_N &= a_0 + \frac{N+1}{2} \\
b_N &= b_0 + \frac{1}{2}E_\mu\left\{\lambda_0(\mu-\mu_0)^2 + \sum_{n=1}^N(x_n-\mu)^2\right\}
\end{aligned}$$

Estimation

- These expressions give optimal values of q_μ and q_τ, but each is defined in terms of moments computed under the other
- In this case, it turns out to be possible to solve for both simultaneously (see Bishop)
- However, in general it will be necessary to start with an initial guess and iterate. As noted, the problem is convex and convergence is guaranteed

Iterative Solution



Ex #2: Gaussian Mixture

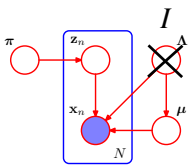
- Assume a fully Bayesian version of the Gaussian mixture model, but with identity covariances (see Bishop for full version)

$$p(X|Z,\mu) = \prod_{n=1}^{N}\prod_{k=1}^{K}\mathcal{N}(x_n|\mu_k, I)^{z_{nk}}$$

$$p(Z|\pi) = \prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{z_{nk}}$$

$$p(\pi) = \mathrm{Dir}(\pi|\alpha_0) \propto \prod_{k=1}^{K}\pi_k^{\alpha_0-1}$$

$$p(\mu) = \prod_{k=1}^{K}\mathcal{N}(\mu_k|0, I)$$



Variational Approach

- The complete log likelihood is:

$$\begin{aligned}\log p(X,Z,\mu,\pi) &= \log p(X|Z,\mu) + \log p(Z|\pi) + \log p(\pi) + \log p(\mu)\\ &= \sum_n\sum_k z_{nk}\left\{-\frac{1}{2}(x_n-\mu_k)^T(x_n-\mu_k) + \log\pi_k\right\}\\ &\quad + \sum_k\left\{(\alpha_0-1)\log\pi_k - \frac{1}{2}\mu_k^T\mu_k\right\} + C\end{aligned}$$

- Assume the factorization:

$$q(Z,\mu,\pi) = q(Z)q(\mu)q(\pi)$$

- Starting with $q(Z)$,

$$\begin{aligned}\log q^*(Z) &= E_{\mu\pi}[\log p(X,Z,\mu,\pi)] + C\\ &= E_{\mu\pi}\left[\sum_n\sum_k z_{nk}\left\{-\frac{1}{2}(x_n-\mu_k)^T(x_n-\mu_k) + \log\pi_k\right\}\right] + C\\ &= \sum_n\sum_k z_{nk}\left\{-\frac{1}{2}E_\mu\left[(x_n-\mu_k)^T(x_n-\mu_k)\right] + E[\log\pi_k]\right\} + C\end{aligned}$$

Responsibilities

- Re-express $\log q^*(Z)$ as:

$$\begin{aligned}\log q^*(Z) &= \sum_n\sum_k z_{nk}\left\{-\frac{1}{2}E_\mu\left[(x_n-\mu_k)^T(x_n-\mu_k)\right] + E[\log\pi_k]\right\} + C\\ &= \sum_n\sum_k z_{nk}\log\rho_{nk}\end{aligned}$$

$$\text{where } \log\rho_{nk} = -\frac{1}{2}E_\mu\left[(x_n-\mu_k)^T(x_n-\mu_k)\right] + E[\log\pi_k]$$

- Thus,

$$q^*(Z) = \frac{\prod_n\prod_k\rho_{nk}^{z_{nk}}}{\sum_{z'}\prod_n\prod_k\rho_{nk}^{z'_{nk}}} = \prod_n\prod_k r_{nk}^{z_{nk}}$$

$$\text{where } r_{nk} = \frac{\rho_{nk}}{\sum_{j=1}^{K}\rho_{nj}}$$

- For discrete Z, these "responsibilities" are the expected z_{nk}s under q: $E[z_{nk}] = r_{nk}$

Part Two

- Now,

$$\begin{aligned}\log q^*(\pi) &= E_{Z\mu}[\log p(X,Z,\mu,\pi)] + C\\ &= E_{Z\mu}\left[\sum_n\sum_k z_{nk}\log\pi_k + \sum_k(\alpha_0-1)\log\pi_k\right] + C\\ &= \sum_n\sum_k r_{nk}\log\pi_k + (\alpha_0-1)\sum_k\log\pi_k + C\\ &= \sum_k(N_k+\alpha_0-1)\log\pi_k + C\end{aligned}$$

$$\text{where } N_k = \sum_n r_{nk}$$

- This has the form of a Dirichlet distribution:

$$q^*(\pi) = \mathrm{Dir}(\pi|\alpha_1,\ldots,\alpha_K), \quad \alpha_k = \alpha_0 + N_k$$

Part Three

- Finally,

$$\begin{aligned}\log q^*(\mu) &= E_{Z\pi}[\log p(X,Z,\mu,\pi)] + C\\ &= E_{Z\pi}\left[\sum_n\sum_k z_{nk}\left\{-\frac{1}{2}(x_n-\mu_k)^T(x_n-\mu_k)\right\} - \frac{1}{2}\sum_k\mu_k^T\mu_k\right] + C\\ &= \sum_n\sum_k r_{nk}\left\{-\frac{1}{2}(x_n-\mu_k)^T(x_n-\mu_k)\right\} - \frac{1}{2}\sum_k\mu_k^T\mu_k + C\end{aligned}$$

- Again completing the square,

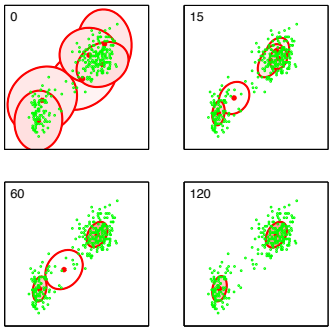
$$q^*(\mu_k) = \mathcal{N}(\mu_k|m_k, P_k^{-1})$$

$$m_k = \frac{N_k}{1+N_k}\bar{x}_k$$

$$P_k^{-1} = (1+N_k)\mathbf{I}$$

$$\bar{x}_k = \frac{1}{N_k}\sum_n r_{nk}x_n$$

Illustration



$$E[\pi_k] = \frac{\alpha_k + N_k}{K\alpha_0 + N}$$

Summary

- The procedure is reminiscent of EM
  - First “responsibilities” are computed based on current averaging distribution
  - Then the averaging distribution is updated based on the responsibilities
- In this case, however, there is no fixed parameter update (*M* step). The “maximization” is actually operating on the averaging distribution
- See discussion by Bishop

That’s All

- Read Bishop 10.1–10.6