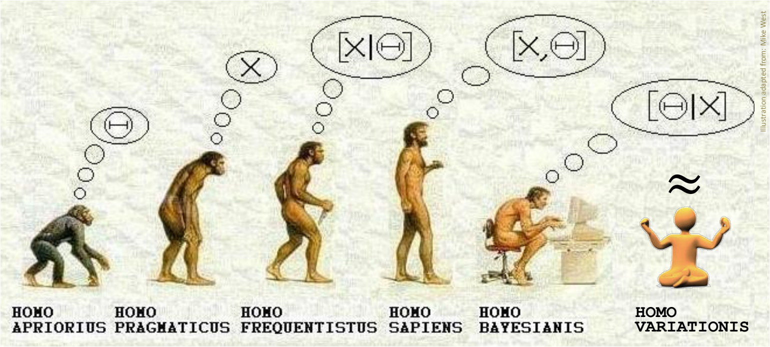
Variational Inference

Parameswaran Raman

October 19, 2015



Topics

- Bayesian Inference
  - The problem and challenges
  - Why approximate inference?
  - Sampling vs Variational Methods
- Variational Inference
  - Basic Idea, Mean Field Approximation
  - ELBO, KL Divergence diagram and formulation
  - Algorithm (using coord descent)
  - Connections: EP, EM algorithm, Mix of Exp Families
- Other Extensions
  - Stochastic Variational Inference [Hoffmann, Blei et al]
  - Streaming Variational Inference [Tamara, Jordan et al]
- References

Quick Recap - Exponential Family

Broad umbrella of distributions that can be expressed in the form,

$$p(x; \theta) = p_0(x) \exp(\langle \phi(x), \theta \rangle - g(\theta))$$

- $p_0(x)$ - base measure
- $\phi(x)$ - sufficient statistics
- θ - natural parameter
- $g(\theta) = \log \int_x \exp(\langle \phi(x), \theta \rangle) dx$ - log-partition function

Examples: Gaussian, Multinomial, Exponential, Dirichlet, Poisson, Gamma, ...

Quick Recap - Exponential Family

Key Properties:

- $g(\theta)$ is convex
- Derivatives of $g(\theta)$ generate moments of $\phi(x)$
  - $\partial_\theta g(\theta) = \mathbb{E}_{p(x;\theta)}[\phi(x)]$
  - $\partial^2_\theta g(\theta) = \mathrm{Var}_{p(x;\theta)}[\phi(x)]$
- Every exponential family distribution has a conjugate prior

Bayesian Inference

Given data $x = \{x_1, x_2, \ldots, x_n\}$,

$$\underbrace{p(\theta|x)}_{posterior} = \frac{\overbrace{p(x|\theta)}^{likelihood} \cdot \overbrace{p(\theta)}^{prior}}{\underbrace{\int p(x,\theta)d\theta}_{\text{marginal likelihood (model evidence)}}}$$

Modeling Assumptions:

- $p(x|\theta) \sim \exp(\langle \phi(x), \theta \rangle) - g(\theta))$
- x are iid

Most inference problems will be one of:

- **Marginalization** $\quad p(x) = \int p(x,\theta)d\theta$
- **Expectation** $\quad \mathbb{E}[f(x|z)] = \int f(x)p(x|z)dz$
- **Prediction** $\quad p(y|x) = \int p(y|\theta, x)p(\theta|x)d\theta$

Computational Challenges

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{\int p(x,\theta) d\theta}$$

- Computing the log-partition function

Solution: Approximate Inference techniques!

Approaches: Type I (sampling based), Type II (variational approximation based), or mix of both

Approximate Inference - Type I

Sampling based methods

1. Design an algorithm that draws samples $\theta_1, \ldots, \theta_k$ from $p(\theta|x)$
2. Use the samples for further computations (say: approximating an expectation originally intractable)

Pros and Cons:

- asymptotically exact
- slow and computationally expensive

Examples: Gibbs Sampling, Importance Sampling, Rejection Sampling

Approximate Inference - Type II

Variational methods

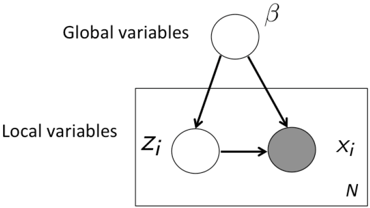
1. Find an analytical proxy $q(\theta)$ that is maximally similar to $p(\theta|x)$
2. Inspect distribution statistics of $q(\theta)$

Pros and Cons:

- more interpretable - and can be much faster
- deterministic algorithm
- hard to derive variational updates
- prone to local minima

Examples: Variational Bayes, Expectation Propagation

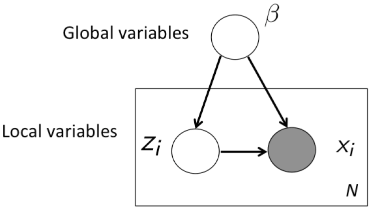
Generic Model



$$p(\beta, z_{1:N}, x_{1:N}) = p(\beta) \prod_{i=1}^{N} p(z_i|\beta) p(x_i|z_i, \beta)$$

- Observations are $x = x_{1:N}$
- Local variables are $z = z_{1:N}$
- Global variables are β
- The ith data point x_i only depends on z_i and β
- Goal: To compute $p(\beta, z|x)$

Generic Model

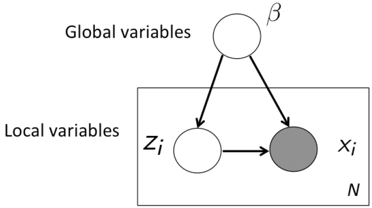


$$p(\beta, z_{1:N}, x_{1:N}) = p(\beta) \prod_{i=1}^{N} p(z_i|\beta) p(x_i|z_i, \beta)$$

- A **complete conditional** is the conditional of a latent variable given the observations and other latent variables
- Assume each complete conditional $\sim$ exp family,

$$p(z_i|\beta, x_i) \propto \exp\{\langle z_i, \eta_l(\beta, x_i)\rangle - g(\eta_l(\beta, x_i))\}$$
$$p(\beta|z, x) \propto \exp\{\langle \beta, \eta_g(z, x)\rangle - g(\eta_g(z, x))\}$$

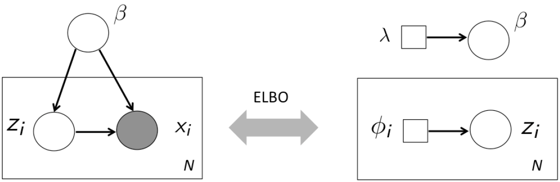
Generic Model



$$p(\beta, z_{1:N}, x_{1:N}) = p(\beta) \prod_{i=1}^{N} p(z_i|\beta) p(x_i|z_i, \beta)$$

- Bayesian Mixture Models
- Time series models (HMMs, Kalman Filters)
- Factorial Models
- Matrix Factorization (PCA, etc)
- Dirichlet Process Mixtures, HDPs
- Multilevel regression
- Mixed-membership models (LDA and variants)

Evidence Lower Bound (ELBO)



- Introduce a **variational distribution** over the latent variables $q(\beta, z)$
- We optimize the **evidence lower bound** (ELBO) with respect to q

$$\log p(x) \geq \mathbb{E}_q \Big[\log p(\beta, Z, x) \Big] - \mathbb{E}_q \Big[\log q(\beta, Z) \Big]$$

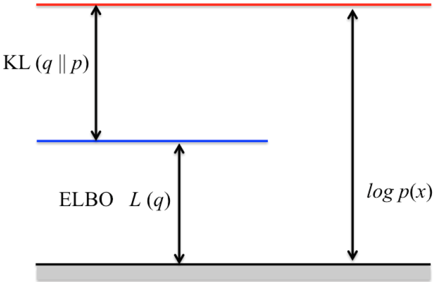
- Up to a constant, this is the negative KL divergence between q and the posterior
- *The ELBO links the observations/model to the variational distribution*

Evidence Lower Bound (ELBO)

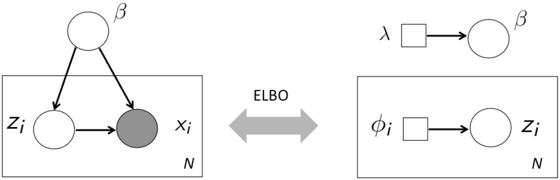
Jensen's Inequality

When f is concave, $f(\mathbb{E}[x]) \geq \mathbb{E}[f(x)]$

$$
\begin{aligned}
\log p(x) &= \log \int_z p(x, z) dz \\
&= \log \int_z p(x, z) \frac{q(z)}{q(z)} dz \\
&= \log \int_z \frac{p(x, z)}{q(z)} q(z) dz \\
&= \log(\mathbb{E}_q[\frac{p(x, z)}{q(z)}]) \\
&\geq \mathbb{E}_q[\log \frac{p(x, z)}{q(z)}]
\end{aligned}
$$



Mean Field Approximation



- We specify $q(\beta, z)$ to be a fully factored variational distribution,

$$q(\beta, z) = q(\beta|\lambda) \prod_{i=1}^{N} q(z_i|\phi_i)$$

- Each instance of each variable has its own distribution
- Each component (factor) is in the same family as the model conditional,

$$q(\beta|z, x) \propto \exp\{\langle \beta, \eta_g(z, x) \rangle - g(\eta_g(z, x))\}$$
$$q(\beta|\lambda) \propto \exp\{\langle \beta, \lambda \rangle - g(\lambda)\}$$

(And same for the local variational parameters)

Optimization of ELBO

- Objective function: We optimize the ELBO wrt variational parameters

$$\mathcal{L}(\lambda, \phi_{1:N}) = \mathbb{E}_q[\log p(\beta, x, z)] - \mathbb{E}_q[\log q(\beta, z)]$$

- Same as finding the $q(\beta, z)$ that is closest in KL Divergence to $p(\beta, z|x)$
- Coordinate Ascent: Iteratively update each variational parameter, holding others fixed
  - **Local Step** (*Var E-Step*):

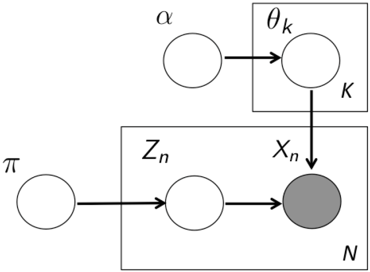
$$\phi_i = \mathbb{E}_q\left[\eta_l(x_i, \beta)\right]$$

  - **Global Step** (*Var M-Step*):

$$\lambda = \mathbb{E}_q\left[\eta_g(x, z)\right]$$

Example: Mix of Exponential Families

Setup:



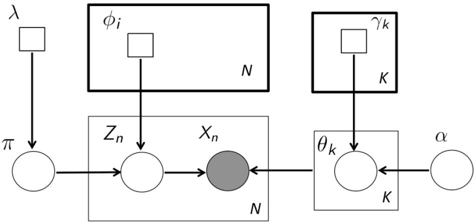
Joint Distribution:

$$p(\pi, \theta|\alpha, Z, X) = p(\pi) \prod_{k=1}^{K} p(\theta_k|\alpha) \left\{ \prod_{i=1}^{N} \sum_{k=1}^{K} p(Z_i = k|\pi_k) p(X_i|\theta_k) \right\}$$

Goal of inference: Estimate posteriors $p(\theta|Z, X, \alpha)$ and $p(Z|X, \theta, \alpha)$

Example: Mixture of Exponential Families

Mean Field Approximation:



$$\mathcal{Q}(\pi, \theta, Z|\lambda, \gamma, \phi) = \mathcal{Q}(\pi|\lambda) \prod_{k=1}^{K} \mathcal{Q}(\theta_k|\gamma_k) \prod_{i=1}^{N} \mathcal{Q}(Z_i|\phi_i)$$

- λ, γ_k and ϕ_i are variational parameters for π, θ_k, Z_i respectively
- *Assumption*: Each factor of the variational distribution above $\sim$ exp family

Example: Mixture of Exponential Families

Variational EM updates can be derived by maximizing $\mathcal{L}(q)$ (ELBO) and cyclically updating variational parameters - λ, γ_k, ϕ_i

- **Var E-Step** (local variational parameters):
  - for $i = 1, 2, \ldots N$

$$\phi_i = \mathbb{E}_q\left[\hat{\eta}(Z_i, X_i, \theta)\right]$$

- **Var M-Step** (global variational parameters):

$$\lambda = \mathbb{E}_q\left[\hat{\eta}(Z_{1:N}, X_{1:N})\right]$$

$$\gamma_k = \mathbb{E}_q\left[\hat{\eta}(Z_{1:N}, X_{1:N})\right]$$

where, $\hat{\eta}$ denotes the natural parameter of exp family for the corresponding model conditional

(S)tochastic (V)ariational (I)nference

- With local and global variables, we decompose the ELBO

$$\mathcal{L} = \mathbb{E}\Big[\log p(\beta)\Big] - \mathbb{E}\Big[\log q(\beta)\Big] + \sum_{i=1}^{N} \mathbb{E}\Big[\log p(z_i, x_i|\beta)\Big] - \mathbb{E}\Big[\log q(z_i)\Big]$$

- Sample a single data point t uniformly from the data and define

$$\mathcal{L}_t = \mathbb{E}\Big[\log p(\beta)\Big] - \mathbb{E}\Big[\log q(\beta)\Big] + N\left(\mathbb{E}\Big[\log p(z_t, x_t|\beta)\Big] - \mathbb{E}\Big[\log q(z_t)\Big]\right)$$

Observations:

1. The ELBO is the expectation of $\mathcal{L}_t$ with respect to the sample
2. The gradient of the t-ELBO is a noisy gradient of the ELBO
3. The t-ELBO is like an ELBO where we saw x_t repeatedly (N times)

SVI Algo

Stochastic Variational EM updates can be derived as below:

- **Var E-Step** (local variational parameters):

  - ~~for i = 1, 2, . . . N~~ Sample t from 1:N uniformly

$$\phi_t = \mathbb{E}_q\left[\hat{\eta}(Z_t, X_t, \theta)\right]$$

- **Var M-Step** (global variational parameters):

$$\hat{\lambda} = \mathbb{E}_q\left[\hat{\eta}(Z_T, X_T)\right]$$

$$\hat{\gamma}_k = \mathbb{E}_q\left[\hat{\eta}(Z_T, X_T)\right]$$

$$\lambda = (1 - \rho)\lambda + \rho\hat{\lambda}$$

where, X_T denotes t-th sample repeated N times, ρ is a decaying step size

SVI Algo: Discussion

- Converges to local optimum
- Local update step (E-Step) can be parallelized for better performance
- In practice, mini-batch sampling is used instead of sampling a single data point
- Not suited for a streaming setting as the number of observations N has to be known (algorithm assumes the sampled data point will be replicated N times). *SDA Bayes [Tamara, Jordan et al] tries to address this*.

(S)treaming (D)istributed (A)synchronous Bayes

- Data $x_1, x_2, \ldots$ generated iid from $p(x|\theta)$ given parameter θ, prior is $p(\theta)$.
- Posterior of θ given a collection of S data points, $C_1 = (x_1, \ldots, x_S)$:

$$p(\theta|C_1) = \frac{p(C_1|\theta)p(\theta)}{p(C_1)}$$

where $p(C_1|\theta) = p(x_1, \ldots, x_S|\theta) = \prod_{s=1}^{S} p(x_s|\theta)$

- Given posterior $p(\theta|C_1, \ldots, C_{b-1})$, we can calculate the posterior after b-th mini batch:

$$p(\theta|C_1, \ldots, C_b) \propto p(C_b|\theta)p(\theta|C_1, \ldots, C_{b-1})$$

- In complex models, posterior cannot be calculated exactly and so we assume a black box approximation algorithm $\mathcal{A}$ that calculates an approximate posterior q: $q(\theta) = \mathcal{A}(C, p(\theta))$.

(S)treaming (D)istributed (A)synchronous Bayes

- Data $x_1, x_2, \dots$ generated iid from $p(x|\theta)$ given parameter θ, prior is $p(\theta)$.
- Posterior of θ given a collection of S data points, $C_1 = (x_1, \dots, x_S)$:

$$p(\theta|C_1) = \frac{p(C_1|\theta)p(\theta)}{p(C_1)}$$

  where $p(C_1|\theta) = p(x_1, \dots, x_S|\theta) = \prod_{s=1}^{S} p(x_s|\theta)$

- Given posterior $p(\theta|C_1, \dots, C_{b-1})$, we can calculate the posterior after b-th mini batch:

$$p(\theta|C_1, \dots, C_b) \propto p(C_b|\theta)p(\theta|C_1, \dots, C_{b-1})$$

- In complex models, posterior cannot be calculated exactly and so we assume a black box approximation algorithm $\mathcal{A}$ that calculates an approximate posterior q: $q(\theta) = \mathcal{A}(C, p(\theta))$.

(S)treaming (D)istributed (A)synchronous Bayes

- Data $x_1, x_2, \dots$ generated iid from $p(x|\theta)$ given parameter θ, prior is $p(\theta)$.
- Posterior of θ given a collection of S data points, $C_1 = (x_1, \dots, x_S)$:

$$p(\theta|C_1) = \frac{p(C_1|\theta)p(\theta)}{p(C_1)}$$

  where $p(C_1|\theta) = p(x_1, \dots, x_S|\theta) = \prod_{s=1}^{S} p(x_s|\theta)$

- Given posterior $p(\theta|C_1, \dots, C_{b-1})$, we can calculate the posterior after b-th mini batch:

$$p(\theta|C_1, \dots, C_b) \propto p(C_b|\theta)p(\theta|C_1, \dots, C_{b-1})$$

- In complex models, posterior cannot be calculated exactly and so we assume a black box approximation algorithm $\mathcal{A}$ that calculates an approximate posterior q: $q(\theta) = \mathcal{A}(C, p(\theta))$.

(S)treaming (D)istributed (A)synchronous Bayes

- Data $x_1, x_2, \ldots$ generated iid from $p(x|\theta)$ given parameter θ, prior is $p(\theta)$.
- Posterior of θ given a collection of S data points, $C_1 = (x_1, \ldots, x_S)$:

$$p(\theta|C_1) = \frac{p(C_1|\theta)p(\theta)}{p(C_1)}$$

  where $p(C_1|\theta) = p(x_1, \ldots, x_S|\theta) = \prod_{s=1}^{S} p(x_s|\theta)$

- Given posterior $p(\theta|C_1, \ldots, C_{b-1})$, we can calculate the posterior after b-th mini batch:

$$p(\theta|C_1, \ldots, C_b) \propto p(C_b|\theta)p(\theta|C_1, \ldots, C_{b-1})$$

- In complex models, posterior cannot be calculated exactly and so we assume a black box approximation algorithm $\mathcal{A}$ that calculates an approximate posterior q: $q(\theta) = \mathcal{A}(C, p(\theta))$.

SDA Bayes - Contd

- Setting $q_0(\theta) = p(\theta)$, one way to recursively calculate an approximation to the posterior is:

$$p(\theta | C_1, \ldots, C_b) \approx q_b(\theta) = \mathcal{A}(C_b, q_{b-1}(\theta))$$

Issue:

Calculating $\mathcal{A}$ might take longer than time interval between mini batch arrivals!

Question:

Can posterior calculations be parallelized?

SDA Bayes - Contd

- Setting $q_0(\theta) = p(\theta)$, one way to recursively calculate an approximation to the posterior is:

$$p(\theta | C_1, \ldots, C_b) \approx q_b(\theta) = \mathcal{A}(C_b, q_{b-1}(\theta))$$

Issue:

Calculating $\mathcal{A}$ might take longer than time interval between mini batch arrivals!

Question:

Can posterior calculations be parallelized?

SDA Bayes - Contd

- Setting $q_0(\theta) = p(\theta)$, one way to recursively calculate an approximation to the posterior is:

$$p(\theta|C_1, \ldots, C_b) \approx q_b(\theta) = \mathcal{A}(C_b, q_{b-1}(\theta))$$

Issue:

Calculating $\mathcal{A}$ might take longer than time interval between mini batch arrivals!

Question:

Can posterior calculations be parallelized?

SDA Bayes - Contd

- Rewriting the posterior updates using Bayes Theorem:

$$p(\theta|C_1,\ldots,C_b) \propto \left[\prod_{b=1}^{B} p(C_b|\theta)\right] p(\theta) \propto \left[\prod_{b=1}^{B} \overbrace{\frac{p(\theta|C_b)}{p(\theta)}}^{\textit{parallelizable}}\right] p(\theta)$$

- Now plugging in the black box approximating algorithm $\mathcal{A}$:

$$p(\theta|C_1,\ldots,C_b) \propto \left[\prod_{b=1}^{B} \overbrace{\frac{\mathcal{A}(C_b, p(\theta))}{p(\theta)}}^{\textit{parallelizable}}\right] p(\theta)$$

- Assuming, $p(\theta)$ and $\mathcal{A}(C_b, p(\theta)) \sim$ exp family with sufficient statistic $\phi(\theta)$, and natural parameters ξ_0 and ξ_b respectively:

$$p(\theta|C_1,\ldots,C_b) \propto \exp\left\{\left[\xi_0 + \sum_{b=1}^{B} \overbrace{(\xi_b - \xi_0)}^{\textit{parallelizable}}\right] \cdot \phi(\theta)\right\}$$

SDA Bayes - Contd

- Rewriting the posterior updates using Bayes Theorem:

$$p(\theta|C_1,\ldots,C_b) \propto \left[\prod_{b=1}^{B} p(C_b|\theta)\right] p(\theta) \propto \left[\prod_{b=1}^{B} \frac{\overbrace{p(\theta|C_b)}^{\text{parallelizable}}}{p(\theta)}\right] p(\theta)$$

- Now plugging in the black box approximating algorithm $\mathcal{A}$:

$$p(\theta|C_1,\ldots,C_b) \propto \left[\prod_{b=1}^{B} \frac{\overbrace{\mathcal{A}(C_b, p(\theta))}^{\text{parallelizable}}}{p(\theta)}\right] p(\theta)$$

- Assuming, $p(\theta)$ and $\mathcal{A}(C_b, p(\theta)) \sim$ exp family with sufficient statistic $\phi(\theta)$, and natural parameters ξ_0 and ξ_b respectively:

$$p(\theta|C_1,\ldots,C_b) \propto \exp\left\{\left[\xi_0 + \sum_{b=1}^{B} \overbrace{(\xi_b - \xi_0)}^{\text{parallelizable}}\right] \cdot \phi(\theta)\right\}$$

SDA Bayes - Contd

- Rewriting the posterior updates using Bayes Theorem:

$$p(\theta|C_1,\ldots,C_b) \propto \left[\prod_{b=1}^{B} p(C_b|\theta)\right] p(\theta) \propto \left[\prod_{b=1}^{B} \frac{\overbrace{p(\theta|C_b)}^{parallelizable}}{p(\theta)}\right] p(\theta)$$

- Now plugging in the black box approximating algorithm $\mathcal{A}$:

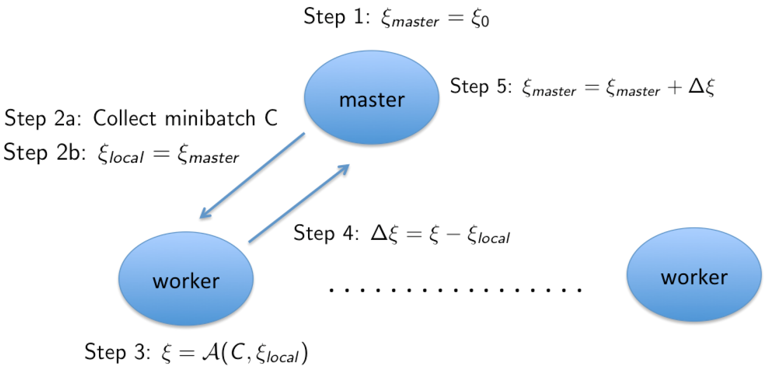
$$p(\theta|C_1,\ldots,C_b) \propto \left[\prod_{b=1}^{B} \frac{\overbrace{\mathcal{A}(C_b, p(\theta))}^{parallelizable}}{p(\theta)}\right] p(\theta)$$

- Assuming, $p(\theta)$ and $\mathcal{A}(C_b, p(\theta)) \sim$ exp family with sufficient statistic $\phi(\theta)$, and natural parameters ξ_0 and ξ_b respectively:

$$p(\theta|C_1,\ldots,C_b) \propto \exp\left\{\left[\xi_0 + \sum_{b=1}^{B} \overbrace{(\xi_b - \xi_0)}^{parallelizable}\right] \cdot \phi(\theta)\right\}$$

SDA Bayes - Contd

Streaming Distributed Asynchronous Algorithm:

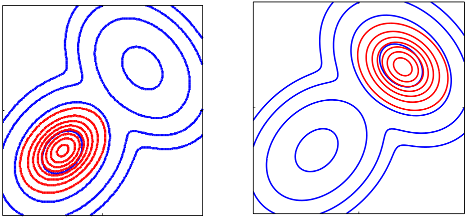


In practice, substitute $\mathcal{A}$ with any variational method like Variational Bayes or EP

Connections to Expectation Propagation (EP)

Variational Bayes

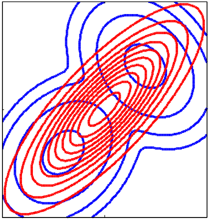
min $KL(q(\theta)||p(\theta|y))$



- $q(\theta)$ will tend to be zero where $p(\theta|x)$ is zero
- may lead to a local minimum

Expectation Propagation

min $KL(p(\theta|y)||q(\theta))$



- $q(\theta)$ will tend to be non-zero where $p(\theta|x)$ is non-zero
- averaging across modes may lead to poor predictive performance

References

[1] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[2] Tamara Broderick, Nicholas Boyd, Andre Wibisono, Ashia C Wilson, and Michael I Jordan. Streaming variational bayes. In *Advances in Neural Information Processing Systems 26*. 2013.

[3] Kay H. Brodersen. Variational Inference. `http://people.inf.ethz.ch/bkay/talks/Brodersen_2013_03_22.pdf`.

[4] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[5] Shakir Mohammed. Variational Inference for Machine Learning. `http://shakirm.com/papers/VITutorial.pdf`.

[6] Alex Tank, Nicholas J Foti, and Emily B Fox. Streaming variational inference for bayesian nonparametric mixture models. *arXiv preprint arXiv:1412.0694*, 2014.