Approximate Inference

Kenji Fukumizu
The Institute of Statistical Mathematics

Computational Methodology in Statistical Inference II

Bayesinan Inference Revisited

- Computation of Integral is required in many Bayeisn inference

$$p(\theta | X) = \frac{p(X | \theta)\pi(\theta)}{\int p(X | \theta)\pi(\theta)d\theta}$$

- Posterior distribution with hidden variables

  Complete model $p(X, Z | \theta)$

  Likelihood $p(X | \theta) = \int \prod_{i=1}^{n} p(X_i, Z_i | \theta) dZ$

$$p(\theta | X) = \frac{\int \prod_{i=1}^{n} p(X_i, Z_i | \theta) dZ \pi(\theta)}{\int \int \prod_{i=1}^{n} p(X_i, Z_i | \theta) \pi(\theta) dZ d\theta}$$

- Marginal likelihood / ABIC

  m: model

$$P(X | m) = \int P(X | \theta, m) P(\theta | m) d\theta$$

Approximation Methods

- Various methods for approximation
  - Laplace approximation

    Quadratic (Gaussian) approximation around the maximum point of the integrand.

  - Variational method (explained here)
  - Expectation propagation (a method similar to belief propagation)
  - Sampling: importance-sampling, MCMC, …

    (Take Iba-san's course!)

  etc.…

The above list is not at all complete.

See *Pattern Recognition and Machine Learning*. C.M. Bishop (2006) Chap. 10 & 11.

Variational Bayesian Learning

Bayesian Learning with Hidden Variables

- Model:

complete data $p(X, Z \mid \theta, m)$

prior of parameter $p(\theta \mid m)$

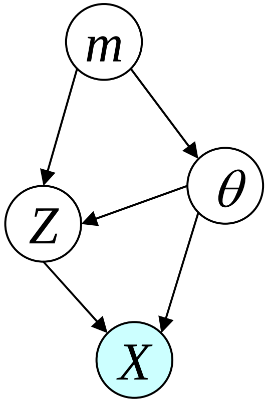
prior of model $p(m)$

X: observable variable

Z: hidden variable

θ: parameter

m: model



- Posterior given data:

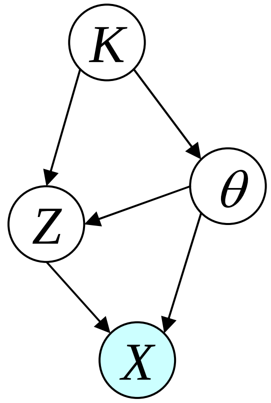
$\mathbf{D} = (X_1, \ldots, X_N), \quad \mathbf{Z} = (Z_1, \ldots, Z_N)$ i.i.d. data

$$p(\mathbf{Z}, \theta, m \mid \mathbf{D}) = \frac{p(\mathbf{D}, \mathbf{Z} \mid \theta, m)\, p(\theta \mid m)\, p(m)}{\sum_m \sum_{\mathbf{Z}} \int p(\mathbf{D}, \mathbf{Z} \mid \theta, m)\, p(\theta \mid m)\, p(m)\, d\theta}$$

where $p(\mathbf{D}, \mathbf{Z} \mid \theta, m) = \prod_{i=1}^{n} p(X_i, Z_i \mid \theta, m)$

- Example: Gaussian mixture in Bayesian viewpoint
  - Model (K components)

$$p(x \mid \theta, K) = \sum_{a=1}^{K} \pi_a \phi(x \mid \mu_a, \Sigma_a) \qquad \theta = (\pi_a, \mu_a, \Sigma_a)_{a=1}^{K}$$



  - Complete model with hidden variable

$$p(X, Z \mid \theta, K) = \prod_{a=1}^{K} \{\pi_a \phi(x \mid \mu_a, \Sigma_a)\}^{Z_a}$$

$Z = (Z_1, \ldots, Z_K)$ takes values in

$$\{ (1,0,0,\ldots,0), \; (0,1,0,\ldots,0), \; \ldots \; (0,0,0,\ldots,1) \}$$

K class

  - Priors on θ

$$p(\pi \mid K) = Dir(\pi \mid \alpha_0^1, \ldots, \alpha_0^K)$$

$$p(\mu_a \mid \Sigma_a, K) = N(\mu \mid \nu_0, (\xi_0 S_a)^{-1})$$

$$p(S_a) = W(S \mid \eta_0, B_0, K) \propto |S|^{\frac{1}{2}(\eta_0 - d - 1)} \exp\left\{ -\frac{1}{2} Tr[S B_0] \right\}$$

where $S = \Sigma^{-1}$

Wishart distribution

Variational Method

- Goal of variational method
  - Direct computation of the posterior $p(\mathbf{Z}, \theta, m \mid \mathbf{D})$ requires the computation of

$$p(\mathbf{D}) = \sum_{m} \sum_{\mathbf{Z}} \int p(\mathbf{D}, \mathbf{Z} \mid \theta, m) p(\theta \mid m) p(m) d\theta$$

The integral and sum are not easy to compute in general.

  - Variational method:

Approximate $p(\mathbf{Z}, \theta, m \mid \mathbf{D})$

by using variational representation of this posterior.

Variational Representation

- Lower bound of marginal likelihood

$$\log p(\mathbf{D}) = \log\left\{\sum_m \sum_{\mathbf{Z}} \int p(\mathbf{D},\mathbf{Z},\theta,m)\,d\theta\right\}$$

$$= \log\left\{\sum_m \sum_{\mathbf{Z}} \int q(\mathbf{Z},\theta,m) \frac{p(\mathbf{D},\mathbf{Z},\theta,m)}{q(\mathbf{Z},\theta,m)}\,d\theta\right\}$$

$q(\mathbf{Z},\theta,m) = q(\mathbf{Z},\theta,m \mid D)$: arbitrary probability

$$= \log E_q\left[\frac{p(\mathbf{D},\mathbf{Z},\theta,m)}{q(\mathbf{Z},\theta,m)}\right]$$

$$\geq E_q\left[\log \frac{p(\mathbf{D},\mathbf{Z},\theta,m)}{q(\mathbf{Z},\theta,m)}\right]$$

(Jensen's inequality)

$$= \sum_m \sum_{\mathbf{Z}} \int q(\mathbf{Z},\theta,m) \log \frac{p(\mathbf{D},\mathbf{Z},\theta,m)}{q(\mathbf{Z},\theta,m)}\,d\theta \qquad \equiv \mathcal{F}[q]$$

Variational Representation

<u>Proposition</u>

$$\log p(\mathbf{D}) = \mathcal{F}[q] + KL(q \,\|\, p)$$

where $$\mathcal{F}[q] = \sum_m \sum_{\mathbf{Z}} \int q(\mathbf{Z},\theta,m) \log \frac{p(\mathbf{D},\mathbf{Z},\theta,m)}{q(\mathbf{Z},\theta,m)} d\theta$$

$$KL(q \,\|\, p) \equiv \sum_m \sum_{\mathbf{Z}} \int q(\mathbf{Z},\theta,m) \log \frac{q(\mathbf{Z},\theta,m)}{p(\mathbf{Z},\theta,m \mid \mathbf{D})} d\theta$$

Proof)

$$\mathcal{F}[q] = -\sum_{Z,m} \int q(\mathbf{Z},\theta,m) \log \frac{q(\mathbf{Z},\theta,m)}{p(\mathbf{D},\mathbf{Z},\theta,m)} d\theta$$

$$= -\sum_{Z,m} \int q(\mathbf{Z},\theta,m) \left\{ \log \frac{q(\mathbf{Z},\theta,m)}{p(\mathbf{Z},\theta,m \mid \mathbf{D})} + \log \frac{p(\mathbf{Z},\theta,m \mid \mathbf{D})}{p(\mathbf{D},\mathbf{Z},\theta,m)} \right\} d\theta$$

$$= -KL(q \,\|\, p) + \log p(\mathbf{D})$$

Variational Representation

- Variational representation of posterior

$$\underbrace{\log p(\mathbf{D})}_{\text{independent of } q} = \mathcal{F}[q] + KL(q \,\|\, p)$$

maximizer of $\mathcal{F}[q]$ $\Leftrightarrow$ minimizer of $KL(q\|p)$

$\Leftrightarrow$ $q(\mathbf{Z}, \theta, m) = p(\mathbf{Z}, \theta, m \mid \mathbf{D})$

$$p(\mathbf{Z}, \theta, m \mid \mathbf{D}) = \arg\max_{q} \mathcal{F}[q]$$

Since q is a function of $(\mathbf{Z},\theta,m)$, the solution is given by the variational method or calculus of variations.

Approximation: Factorization

- Approximation by factorization

  The exact maximization of $\mathcal{F}[q]$ is usually intractable.

  Factorization for tractability

$$q(\mathbf{Z},\theta,m) \approx q(\mathbf{Z}\,|\,m)q(\theta\,|\,m)q(m)$$

The factorization restricts the space of q, and thus the maximization of $\mathcal{F}[q]$ under this restriction gives the posterior only approximately.

But, this gives an EM-like tractable algorithm!

Derivation of VB Method

$$\mathcal{F}[q] = \sum_m \sum_{\mathbf{Z}} \int q(\mathbf{Z}|m) q(\theta|m) q(m) \log \frac{p(\mathbf{D},\mathbf{Z},\theta,m)}{q(\mathbf{Z},\theta,m)} d\theta$$

$$= \sum_m q(m) \sum_{\mathbf{Z}} \int q(\mathbf{Z}|m) q(\theta|m) \log \frac{p(\mathbf{D},\mathbf{Z}|\theta,m)\, p(\theta|m)\, p(m)}{q(\mathbf{Z}|m) q(\theta|m) q(m)} d\theta$$

$$= \sum_m q(m) \left\{ \sum_{\mathbf{Z}} \int q(\mathbf{Z}|m) q(\theta|m) \log \frac{p(\mathbf{D},\mathbf{Z}|\theta,m)}{q(\mathbf{Z}|m)} d\theta + \int q(\theta|m) \log \frac{p(\theta|m)}{q(\theta|m)} d\theta \right\} + \sum_m q(m) \log \frac{p(m)}{q(m)}$$

Derivation of VB Method: Fixed Model

- Fixed model

  Suppose we have a fixed model m.

  Maximize

$$\mathcal{F}_m[q] = \sum_{\mathbf{Z}} \int q(\mathbf{Z} \mid m) q(\theta \mid m) \log \frac{p(\mathbf{D}, \mathbf{Z} \mid \theta, m)}{q(\mathbf{Z} \mid m)} d\theta + \sum_{\mathbf{Z}} \int q(\theta \mid m) \log \frac{p(\theta \mid m)}{q(\theta \mid m)} d\theta$$

  Lagrange functional

$$J[q(\mathbf{Z} \mid m), q(\theta \mid m), \lambda, \nu] = \mathcal{F}_m[q] + \lambda(\textstyle\sum_{\mathbf{Z}} q(\mathbf{Z} \mid m) - 1) + \nu(\int q(\theta \mid m) d\theta - 1)$$

  Euler-Lagrange equations

$$\frac{\partial J[q(\mathbf{Z} \mid m), q(\theta \mid m), \lambda, \nu]}{\partial q(\mathbf{Z} \mid m)} = 0, \quad \frac{\partial J[q(\mathbf{Z} \mid m), q(\theta \mid m), \lambda, \nu]}{\partial \lambda} = 0$$

$$\frac{\partial J[q(\mathbf{Z} \mid m), q(\theta \mid m), \lambda, \nu]}{\partial q(\theta \mid m)} = 0, \quad \frac{\partial J[q(\mathbf{Z} \mid m), q(\theta \mid m), \lambda, \nu]}{\partial \nu} = 0$$

Derivation of VB Method: Fixed Model

- VB-E step

$$\frac{\partial J[q(\mathbf{Z}\mid m), q(\theta\mid m), \lambda, \nu]}{\partial q(\mathbf{Z}\mid m)} = 0$$

$$\Rightarrow \int q(\theta\mid m)\log\frac{p(\mathbf{D},\mathbf{Z}\mid\theta,m)}{q(\mathbf{Z}\mid m)}d\theta - \int q(\mathbf{Z}\mid m)q(\theta\mid m)\frac{1}{q(\mathbf{Z}\mid m)}d\theta + \lambda = 0$$

$$\int q(\theta\mid m)\log p(\mathbf{D},\mathbf{Z}\mid\theta,m)d\theta - \log q(\mathbf{Z}\mid m) - 1 + \lambda = 0$$

$$q(\mathbf{Z}\mid m) = C\exp\left\{\int q(\theta\mid m)\log p(\mathbf{D},\mathbf{Z}\mid\theta,m)d\theta\right\}$$

$$\frac{\partial J[q(\mathbf{Z}\mid m), q(\theta\mid m), \lambda, \nu]}{\partial \lambda} = 0 \quad \Rightarrow \quad \sum_{\mathbf{Z}} q(\mathbf{Z}\mid m) = 1$$

$$\Rightarrow q(\mathbf{Z}\mid m) = C\exp\left\{\int q(\theta\mid m)\log p(\mathbf{D},\mathbf{Z}\mid\theta,m)d\theta\right\}$$

C: normalization constant

Derivation of VB Method: Fixed Model

- VB-M step

  Similarly,

$$\frac{\partial J[q(\mathbf{Z}\,|\,m), q(\theta\,|\,m), \lambda, \nu]}{\partial q(\theta\,|\,m)} = 0 \quad \Longrightarrow \quad q(\theta\,|\,m) = C'\, p(\theta\,|\,m) \exp\left\{ \sum_{\mathbf{Z}} q(\mathbf{Z}\,|\,m) \log p(\mathbf{D}, \mathbf{Z}\,|\,\theta, m) \right\}$$

$$\frac{\partial J[q(\mathbf{Z}\,|\,m), q(\theta\,|\,m), \lambda, \nu]}{\partial \nu} = 0 \quad \Longrightarrow \quad \int q(\theta\,|\,m) d\theta = 1$$

$$\begin{cases} q(\mathbf{Z}\,|\,m) = C \exp\left\{ \int q(\theta\,|\,m) \log p(\mathbf{D}, \mathbf{Z}\,|\,\theta, m) d\theta \right\} \\ q(\theta\,|\,m) = C'\, p(\theta\,|\,m) \exp\left\{ \sum_{\mathbf{Z}} q(\mathbf{Z}\,|\,m) \log p(\mathbf{D}, \mathbf{Z}\,|\,\theta, m) \right\} \end{cases}$$

The two equations are not closed form, and iterations are needed.

Variational Bayes: Fixed Model

- Algorithm (fixed model m)
  1. Intialization
     $$q(\theta \mid m)^{(0)},\ \ q(\mathbf{Z} \mid m)^{(0)}$$
  2. Repeat until some convergence criterion is satisfied.

  VB-E step

  $$q(\mathbf{Z} \mid m)^{(t+1)} = C \exp\left\{\int q(\theta \mid m)^{(t)} \log p(\mathbf{D}, \mathbf{Z} \mid \theta, m) d\theta\right\}$$

  VB-M step

  $$q(\theta \mid m)^{(t+1)} = C' \, p(\theta \mid m) \exp\left\{\sum_{\mathbf{Z}} q(\mathbf{Z} \mid m)^{(t)} \log p(\mathbf{D}, \mathbf{Z} \mid \theta, m)\right\}$$

  t = t+1

Variational Bayes: Fixed Model

- VB-E step for exponential family

  Assume the complete model is given by an exponential family

$$p(X, Z \mid \theta, m) = \exp\left(\theta^T u(x, z) - \psi(\theta)\right)$$

Then,

$$\int q(\theta \mid m)^{(t)} \log p(\mathbf{D}, \mathbf{Z} \mid \theta, m) d\theta = \int q(\theta \mid m)^{(t)} \left\{\theta^T \sum_i u(X_i, Z_i) - N\psi(\theta)\right\} d\theta$$

$$= \bar{\theta}^{(t)T} \sum_i u(X_i, Z_i) - N E_{q(\theta)^{(t)}}[\psi(\theta)]$$

where $\bar{\theta}^{(t)T} = \int \theta q(\theta \mid m)^{(t)} d\theta$

$$\Rightarrow \quad q(\mathbf{Z} \mid m)^{(t+1)} \propto \exp\left\{\int q(\theta \mid m)^{(t)} \log p(\mathbf{D}, \mathbf{Z} \mid \theta, m) d\theta\right\}$$

$$\propto \exp\left(\bar{\theta}^{(t)T} \sum_i u(X_i, Z_i) - N E_{q(\theta)^{(t)}}[\psi(\theta)]\right) = p(\mathbf{D}, \mathbf{Z} \mid \bar{\theta}^{(t)}, m)$$

$$\Rightarrow \quad q(\mathbf{Z}|m)^{(t+1)} = p(\mathbf{Z} \mid \mathbf{D}, \bar{\theta}^{(t)}, m)$$

Comparison with EM Algorithm

Exponential family is assumed for the complete model.

- **EM**

  Goal:

  Maximize $\log p(\mathbf{D} \mid \theta)$ w.r.t. θ

  E-step

  $$q(\mathbf{Z})^{(t+1)} = p(\mathbf{Z} \mid \mathbf{D}, \hat{\theta}^{(t)})$$

  M-step

  $$\hat{\theta}^{(t+1)} = \arg\max_{\theta} \sum_{\mathbf{Z}} q(\mathbf{Z})^{(t+1)} \log p(\mathbf{D}, \mathbf{Z} \mid \theta)$$

- **Variational Bayes**

  Goal:

  Approximate $p(\mathbf{Z}, \theta \mid \mathbf{D})$

  E-step

  $$q(\mathbf{Z})^{(t+1)} = p(\mathbf{Z} \mid \mathbf{D}, \overline{\theta}^{(t)})$$

  M-step

  $$q(\theta)^{(t+1)} \propto p(\theta) \exp\left\{ \sum_{\mathbf{Z}} q(\mathbf{Z})^{(t)} \log p(\mathbf{D}, \mathbf{Z} \mid \theta) \right\}$$

- $\mathcal{F}_m[q]$ increases monotonically.
- VB-E step is computationally as demanding as EM-E step.
- Normalization in VB-M step may be intractable in general.

Model Selection

- MAP approach to model selection

$$\mathcal{F}[q] = \sum_m \sum_{\mathbf{Z}} \int q(\mathbf{Z} \mid m) q(\theta \mid m) q(m) \log \frac{p(\mathbf{D}, \mathbf{Z}, \theta, m)}{q(\mathbf{Z}, \theta, m)} d\theta$$

$$= \sum_m q(m) \mathcal{F}_m[q(\mathbf{Z} \mid m), q(\theta \mid m)] - KL(q(m) \| p(m))$$

Optimize $q(\mathbf{Z}|m)$, $q(\theta|m)$, $q(m)$, and choose m such that

$$m^* = \arg\max_m q(m)$$

- Method (A)

  Compute optimum $q^*(\mathbf{Z}|m)$ and $q^*(\theta|m)$ for all m by VB-EM steps, and optimize $\mathcal{F}[q^*(\mathbf{Z}|m), q^*(\theta|m), q(m)]$ w.r.t. $q(m)$. Choose m as above.
- Method (B)

  Simultaneous optimization of $q(\mathbf{Z}|m)$, $q(\theta|m)$, $q(m)$.

VB for Gaussian Mixture Model

- Model

$$p(x \mid \theta) = \sum_{a=1}^{K} \pi_a \phi(x \mid \mu_a, \Sigma_a)$$

x: d-dimensional

  - Complete model

$$p(X, Z \mid \theta) = \prod_{a=1}^{K} \{\pi_a \phi(x \mid \mu_a, \Sigma_a)\}^{Z_a}$$

$Z = (Z_1,\ldots,Z_K)$ takes values in

$$\left.\begin{array}{l} \{ (1,0,0,\ldots,0), \\ (0,1,0,\ldots,0), \\ \cdots \\ (0,0,0,\ldots,1) \} \end{array}\right\} K \text{ class}$$

  - Conjugate Prior

$$p(\pi) = Dir(\pi \mid \alpha_0^1, \ldots, \alpha_0^K)$$

$$p(\mu_a \mid \Sigma_a) = N(\mu \mid \nu_0, (\xi_0 S_a)^{-1})$$

$$p(S_a) = W(S \mid \eta_0, B_0) \propto |S|^{\frac{1}{2}(\eta_0 - d - 1)} \exp\left\{-\frac{1}{2} Tr[SB_0]\right\}$$

where $S = \Sigma^{-1}$

Wishart distribution

VB for Gaussian Mixture Model

- Update of $q(\theta)^{(t)}$ can be done by update of the parameters.
  By using the conjugate priors,
  $q(\pi)^{(t)}$ is always Dirichlet,
  $$q(\pi)^{(t)} = Dir(\pi \mid \alpha_1^{(t)},\ldots,\alpha_K^{(t)})$$
  $q(\mu|\Sigma)^{(t)}$ is always Gaussian,
  $$q(\mu_a \mid \Sigma_a)^{(t)} = N\left(\mu \mid \bar{\mu}_a^{(t)}, \left((\bar{N}_a^{(t)} + \xi_0) S_a^{(t)}\right)^{-1}\right)$$
  $q(S)^{(t)}$ is always Wishart.
  $$q(S_a)^{(t)} = W(S \mid \eta_a^{(t)}, B_a^{(t)})$$

- The marginal probability of the mean $q(\mu)^{(t)}$ is t-distribution.

$$q(\mu_a)^{(t)} \propto T(\mu \mid \bar{\mu}_a^{(t)}, \Sigma_{\mu_a}^{(t)}, f_{\mu_a}^{(t)})$$

$$f_{\mu_i}^{(t)} = \eta_0 + \bar{N}_i^{(t)} + 1 - d, \quad \Sigma_{\mu_i}^{(t)} = \frac{B_i^{(t)}}{(\bar{N}_i^{(t)} + \xi_0) f_{\mu_i}^{(t)}}$$

t-distribution: $T(x \mid \mu, \Sigma, f) \propto \left\{1 + (x-\mu)^T (f\Sigma)^{-1} (x-\mu)\right\}^{-\frac{f+d}{2}}$

VB for Gaussian Mixture Model

- Algorithm (see Bishop (2006) or 樺島・上田(2003) for the details.)
  1. Initialization of parameters
  2. Repeat

  VB-E step (update of $q^*(\mathbf{Z})$)

$$\tau_i^{n(t+1)} = q(Z_i^n = 1) = \frac{\exp(\gamma_i^{n(t+1)})}{\sum_{j=1}^{K} \exp(\gamma_j^{n(t+1)})} \qquad i = 1,\ldots,K \quad n = 1,\ldots,N$$

where

$$\gamma_i^{n(t+1)} = \Psi(\alpha_0 + \overline{N}_i^{(t)}) - \Psi(K\alpha_0 + \sum_{j=1}^{K} \overline{N}_j^{(t)}) + \frac{1}{2}\sum_{j=1}^{d} \Psi\left(\frac{\eta_0 + \overline{N}_i^{(t)} + 1 - j}{2}\right) - \frac{1}{2}\log|B_i^{(t)}| - \frac{1}{2} Tr\left[(\eta_0 + \overline{N}_i^{(t)}) B_i^{(t)-1}\left(\frac{f_{\mu_i}^{(t)}}{f_{\mu_i}^{(t)} - 2}\Sigma_{\mu_i}^{(t)} + (X_n - \overline{\mu}_i^{(t)})(X_n - \overline{\mu}_i^{(t)})^T\right)\right]$$

Ψ: digamma

VB for Gaussian Mixture Model

VB-M step (update of $q(\theta)^{(t)}$ by updating the parameters)

$$\overline{N}_i^{(n)} = \sum_{n=1}^{N} \tau_i^{n(t)}$$

$$\overline{X}_i^{(t)} = \sum_{n=1}^{N} \tau_i^{n(t)} X_n \qquad \overline{C}_i^{(t)} = \sum_{n=1}^{N} \tau_i^{n(t)} (X_n - \overline{X}_i^{(t)})(X_n - \overline{X}_i^{(t)})^T$$

$$\alpha_i^{(t)} = \alpha_0 + \overline{N}_i^{(t)}$$

$$\eta_i^{(t)} = \eta_0 + \overline{N}_i^{(t)}$$

$$\overline{\mu}_i^{(t)} = \frac{\overline{N}_i^{(t)} \overline{X}_i^{(t)} + \xi_0 \nu_0}{\overline{N}_i^{(t)} + \xi_0}$$

$$B_i^{(t)} = B_0 + \overline{C}_i^{(t)} + \frac{\overline{N}_i^{(t)} \xi_0}{\overline{N}_i^{(t)} + \xi_0} (\overline{X}_i^{(t)} - \nu_0)(\overline{X}_i^{(t)} - \nu_0)^T$$

Demo: VB for Gaussian Mixture

- Matlab demo

Appendix: Jensen's inequality

$f(x)$: convex function on $\mathbf{R}^n$

Q: probability on $\mathbf{R}^n$

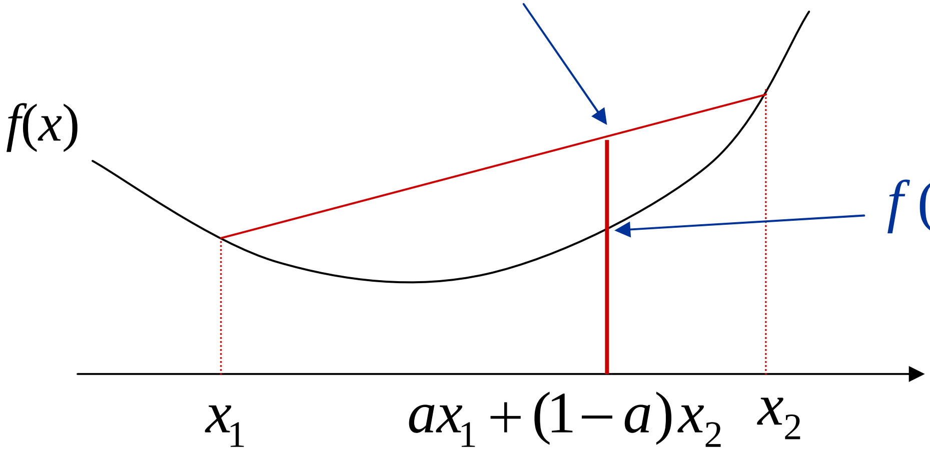
Jensen's inequality

$$E_Q[f(X)] \geq f\left(E_Q[X]\right)$$

$$E_q[f(X)] = af(x_1) + (1-a)f(x_2)$$

$$Q = a\delta_{x_1} + (1-a)\delta_{x_2}$$

$$f(E_q[X]) = f(ax_1 + (1-a)x_2)$$



Appendix: Calculus of Variations

- Optimization of functional

maximize $F[q] = \int \varphi(q(x), x)dx$

For any function $u(x)$

$$0 = \frac{d}{dt} F[q + tu]\Big|_{t=0} = \frac{d}{dt} \int \varphi(q(x) + tu(x), x)dx\Big|_{t=0} = \int \frac{\partial \varphi(q(x), x)}{\partial q} u(x)dx$$

$$\Rightarrow \quad \delta F[q] \equiv \frac{\partial \varphi(q, x)}{\partial q} = 0$$

Euler equation

If there is a constraint on q, e.g. $\int q(x)dx = 1$

$$J[q, \lambda] = F[q] + \lambda(\int q dx - 1)$$

λ: Lagrange multiplier

$$\frac{\partial J[q, \lambda]}{\partial q} = 0, \quad \frac{\partial J[q, \lambda]}{\partial \lambda} = 0$$

Euler-Lagrange equation

Summary

- Various methods for approximate inference
  - Laplace method
  - Variational method
  - Sampling (MCMC, importance sampling, etc)
  - Expectation propagation etc., ...
- Variational method
  - Approximation of posterior for hidden variables, parameter, and model.
  - EM-like algorithm can be used.
  - References on VB

    C.M. Bishop. *Pattern Recognition and Machine Learning*. (2006) Springer.

    樺島・上田. 平均場近似・EM法・変分ベイズ法. 「統計科学のフロンティア11, 計算統計I」岩波書店(2003)

    See also http://www.variational-bayes.org/