MCMC and Variational inference

Department of Statistics

December 2, 2023

University of Seoul

Distribution and random numbers

Generating random number from Binomial distribution

- Bernoulli distribution
- Definition of Binomial distribution

Using uniform distribution, generate numbers from Bernoulli distribution.

Generating random number from Binomial distribution

Let $U \sim [0,1]$ then $\Pr(U \leq p) = p$. That is

$$I(U \leq p) =_d \text{Bernoulli}(p)$$

(1) Set $X = 0$

(2) Generate $U \sim (0,1)$

(3) If $U < p$ then $X \leftarrow X + 1$

(4) Iterate (2)-(3) n times

Inversion method

Let $X \sim F$, $U \sim U[0,1]$ and $X \perp U$, then

$$F^{-1}(U) =_d X$$

(proof) $\Pr(F^{-1}(U) \leq x) = \Pr(U \leq F(x)) = F(x)$. Thus, the CDF of a random variable $F^{-1}(U)$ is F.

Exponential distribution

$X \sim \exp(\lambda)$

- $F(x) = 1 - \exp(-\lambda x)$
- $F^{-1}(u) = -\log(1-u)/\lambda$

Gamma (Erlang) distribution

Let $X \sim \text{Gamma}(n, \beta)$ for $n \in \mathbb{N}$ and $\beta > 0$.

- If $Y_i \sim_{iid} \exp(\beta^{-1})$ for $i = 1, \cdots, n$ $X =_d \sum_{i=1}^{n} Y_i$
- $Y_i \sim \beta \exp(1)$

Normal distribution (Box-Müller method)

The pdf of $X \sim N(0, \sigma^2)$ is

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$$

- $U_1, U_2 \sim U(0, 1)$
- $Z = \sqrt{-2 \log U_2} \cos(2\pi U_1) \sim N(0, 1)$

(Idea of Box-Müller method)

- Let Z_1 and Z_2 be independent random variables following $N(0,1)$.
- Since this density is radially symmetric, it is natural to consider the polar coordinate random variables (R, θ), defined by $0 \leq \theta < 2\pi$ and $Z_1 = R\cos(\theta)$ and $Z_2 = R\sin(\theta)$.
- Clearly, $\theta \sim U[0, 2\pi] =_d 2\pi U_1$ where $U_1 \sim U(0,1)$.
- Intuitively, $R \perp \theta$.

(Idea of Box-Müller method)

$$\begin{aligned} R^2 &= R^2\cos(\theta)^2 + R^2\sin(\theta)^2 \\ &= Z_1^2 + Z_2^2 \\ &=_d \chi^2(2) \\ &=_d \text{Gamma}(1,2) =_d 2\exp(1) \\ &=_d -2\log(1-U_2) =_d -2\log(U_2) \end{aligned}$$

where $U_2 \sim U(0,1)$. So, $R =_d \sqrt{-2\log(U_2)}$ Therefore,

$$Z_1 = R\cos\theta =_d \sqrt{-2\log(U_2)}\cos(2\pi U_1).$$

Multivariate normal distribution

Let $\mathbf{X} \sim N_p(0, \Sigma)$ and suppose that Σ is positive definite.

- Find A satisfying $A^2 = \Sigma$
- Generate $y \sim N_p(0, I)$. Note that $y = (y_1, \cdots, y_p)$ where $y_i \sim_{iid} N(0, 1)$.
- Obtain $x = Ay$.

Rejection sampling

Our object is to obtain random samples from f.

- f is density function of X
- g is density function of Y
- Assume that there exists $k > 0$ such that

$$kg(x) \geq f(x)$$

for all x for which $f(x) > 0$.

Rejection sampling algorithm

1. Sample $Y \sim g$
2. Sample $U \sim U(0,1)$
3. Reject Y if $U > kg(y)/f(y)$, and return to step 1.
4. Otherwise, keep the value of Y. Set $X = Y$ and return to step 1.

Proposition 1

Let $X \sim f$ *and* $Y \sim g$ *and assume that there exists* $k > 0$ *such that* $f(x) \leq kg(x)$ *for all* $x \in \{x : f(x) > 0\}$.

(proof)

$$\begin{aligned}\Pr(U \le f(Y)/(kg(Y))) &= \int \Pr(U \le f(Y)/(kg(Y))|Y=y)g(y)dy \\ &= \int \frac{f(y)}{kg(y)}g(y)dy = 1/k.\end{aligned}$$

$$\begin{aligned}\Pr(Y \le y, (U \le f(Y)/(kg(Y))) &= \int_{-\infty}^{y} d\Pr(Y=y, (U \le f(Y)/(kg(Y))) \\ &= \int_{-\infty}^{y} g(y)\frac{f(y)}{kg(y)}dy = F(y)/k\end{aligned}$$

Therefore, $\Pr(Y \le y|U \le f(Y)/(kg(Y))) = F(y)$, which completes the proof.

Example 1 (Gamma distribution for $\alpha < 1$)

Let $X \sim \text{Gamma}(\alpha, 1)$ then, $f(x) = \frac{x^{\alpha-1}\exp(-x)}{\Gamma(\alpha)}$

- For $0 < x \leq 1$, $f(x) \leq x^{\alpha-1}/\Gamma(\alpha)$
- For $x \geq 1$, $f(x) \leq \exp(-x)/\Gamma(\alpha)$

From the above relation, we can set

$$e(x) = \begin{cases} x^{\alpha-1}/\Gamma(\alpha) & 0 < x < 1 \\ \exp(-x)/\Gamma(\alpha) & x \geq 1 \end{cases}$$

Let

$$K = \int_0^\infty e(x)dx = (\alpha^{-1} + e^{-1})/\Gamma(\alpha)$$

then $g(x) = K^{-1}e(x)$ is pdf.

The cdf of $g(x)$ is given by

$$G(x) = \begin{cases} \frac{x^\alpha}{\Gamma(\alpha)K\alpha} & 0 < x < 1 \\ \frac{e^{-1} - e^{-x}}{\Gamma(\alpha)K} + \frac{1}{\Gamma(\alpha)K\alpha} & x \geq 1. \end{cases}$$

We can easily obtain $G^{-1}(u)$.

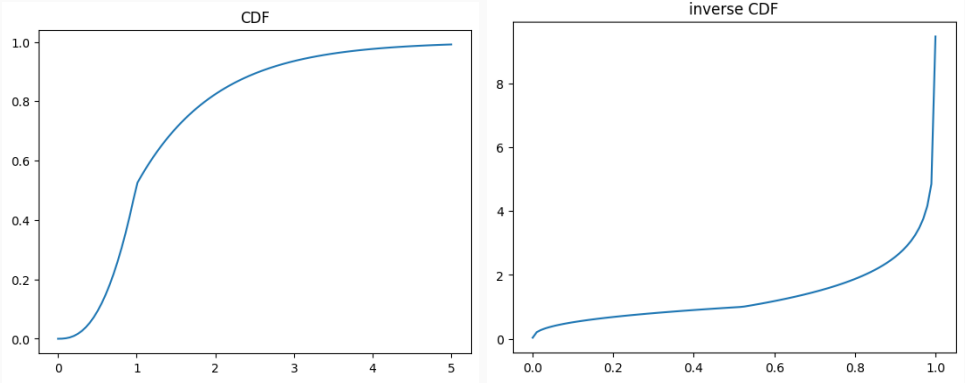


Figure 1: $G(x)$ and $G^{-1}(x)$

Algorithm

1. Sample $Y \sim g$ (inversion method through G^{-1})
2. Sample $U \sim U(0,1)$
3. Reject Y if $U > f(Y)/e(Y)$, and return to step 1.
4. Otherwise, keep the value of Y. Set $X = Y$ and return to step 1.

Example 2 (Gamma distribution for $\alpha > 1$)

Let $X \sim$ Gamma$(\alpha, 1)$. Let $h(x) = d(1 + cx)^3$, and we consider a generating rs defined by $h(X)$.

The pdf of $h(X)$, f_h, is proportional to

$$\exp(g(x)) = h(x)^{\alpha-1} \exp(-h(x))h'(x), \quad (x > -1/c)$$

where $g(x) = (\alpha - 1/3)\log(1 + cx)^3 - d(1 + cx)^3 + d$

Let $d = \alpha - 1/3$ and $c = 1/\sqrt{9d}$ then $\exp(g(x)) \leq \exp(-x^2/2)$.

(continued) Let $f_h(x) = K \exp(g(x))$ then

$$f_h(x) = K \exp(g(x)) \leq K\sqrt{2\pi}\phi(x)$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$

Note that

$$f_h(x)/K\sqrt{2\pi}\phi(x) = \exp(g(x))/(\sqrt{2\pi}\phi(x)).$$

We need not to know the value of K in rejection sampling.

Algorithm

1. Sample $Z \sim \phi$
2. Sample $U \sim U(0,1)$
3. Reject Z if $U > \exp(g(Z))/\exp(-Z^2/2)$, and return to step 1.
4. Otherwise, keep the value of Y. Set $Y = Z$ and return to step 1.

Note that $Y = H(X)$. By inversion of Y, we complete the sampling algorithm.

we may obtain random sample of $y = h(X)$. By letting $x = h^{-1}(y)$, we obtain rs of X.

$$\exp(h(x)) \leq \exp(-x^2/2)$$

We let $e(x) = \exp(-x^2/2)$ then

Example 3 (Hit or Miss method)

Our object is to compute

$$I = \int_a^b g(x)dx$$

for $g(x) \geq 0$.

Assume that $g(x) \in [0, c]$ for all $x \in (a, b)$.

- Set $N_H = 0$
- For $i = 1, \cdots, N$
  - Generate u_i and v_i
  - $x_i = a + u_i(b - a)$
  - If $g(x_i) \geq cv_i$ then $N_H \to N_H + 1$.
- $\hat{I}_H = c(b - a)N_H/N$

- Let p be the probability that a random point falls in S where $S = \{(x, y) : y \leq g(x), y \geq 0\}$
- $\hat{p} = N_H/N$
- $N_H \sim bin(N, p)$

We know that

- $\hat{I}_H$ is unbiased estimator for I, since

$$\mathrm{E}\hat{I}_H = c(b-a)\mathrm{E}(\hat{p}) = I$$

- $\mathrm{Var}(I_H) = c^2(b-a)^2\mathrm{Var}(\hat{p}) = I(c(b-a) - I)/N$

We can compute the confidence interval of $\hat{I}_H$.

$$\hat{I}_H \pm c(b-a) z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

Sample Mean method

Our object is to compute

$$I = \int_a^b g(x)dx$$

for $g(x) \geq 0$.

Note that

$$I = \int_a^b g(x)dx = \int_a^b \frac{g(x)}{f(x)} f(x)dx$$

That is

$$I = E_X[\frac{g(X)}{f(X)}]$$

Algorithm

- For $i = 1, \cdots, N$
  - $x_i \sim U(a, b)$
  - Compute $g(x_i)$
- $\hat{I}_{SM} = \sum_{i=1}^{N} g(x_i) \frac{1}{(b-a)} / N$

Note that $1/(b-a)$ is pdf of X.

- $X \sim U(a, b)$ and $I = (b-a)\mathrm{E}(g(X))$
- $\hat{I}_{SM}$ is unbiased.

$$\begin{aligned} \mathrm{E}(\hat{I}_{SM}) &= (b-a)\frac{1}{N}\sum_{i=1}^{N}\mathrm{E}g(X_i)) \\ &= (b-a)\frac{1}{N}\sum_{i=1}^{N}\int_a^b g(x)\frac{1}{(b-a)}dx = I \end{aligned}$$

- $\mathrm{Var}(\hat{I}_{SM}) = \frac{1}{N}\{(b-a)\int_a^b g(x)^2 dx - I^2\}$

Note that $\mathrm{Var}(\hat{I}_{SM}) \leq \mathrm{Var}(\hat{I}_H)$

Discussion

- Both of $\hat{I}_{SM}$ and $\hat{I}_H$ is unbiased estimator.
- $\mathrm{Var}(\hat{I}_{SM}) \leq \mathrm{Var}(\hat{I}_H)$

Then, we conclude that.

Definition 4 (Markov Chain)

Let X_n be a discrete random variable having finite states. $\{X_n\}$ is Markov chain if

$$P(X_n|X_{n-1}, \cdots, X_1) = P(X_n|X_{n-1}).$$

Theorem 5

Let x_i for $i = 1, \cdots, m$ be a state of X_n and let P be a transition matrix where

$$(P)_{ij} = \Pr(X_{n+1} = x_j | X_n = x_i)$$

If $P \in \mathbb{R}^{m \times m}$ is irreducible then there exists the unique $\pi \in \mathcal{S}^{m-1}$ (m-dimensional simplex) such that

$$\pi = \pi P$$

(Note: π is a row vector! It is a conventional notation)

Theorem 6

If $P \in \mathbb{R}^{m \times m}$ is irreducible and aperiodic then there exists the unique $\pi \in \mathcal{S}^{m-1}$ (m-dimensional simplex) such that

$$\lim_{n \to \infty} \pi_0(P^n) = \pi$$

for any $\pi_0 \in \mathcal{S}^{m-1}$.

Definition 7 (Detailed Balance Condition)

Let P_{ij} be the transition prob from the state i to j and π be the state probability. If $\pi_i P_{ij} = \pi_j P_{ji}$ for all i and j we call that the transition probability matrix P satisfies the detailed balance condition wrt π.

Detailed Balance condition and stationary distribution

If P satisfies the detailed balance condition wrt π, then π is the stationary distribution of P under regular conditions.

(why?) $\sum_i \pi_i P_{ij} = \sum_i \pi_j P_{ji} = \pi_j \sum_i P_{ji} = \pi_j$. That is, π is the solution of $\pi P = \pi$.

If we find P satisfying the detailed balance condition, we can generate a random sample following π by restoring samples from $\pi_0(P^n)$ for a large n.

Markov Chain Monte Carlo

Monte Carlo method

- Evaluating

$$E_\pi[h(X)] = \int h(x)\pi(x)dx$$

  is difficult.

- However, if we can draw independent samples

$$X^{(1)}, X^{(2)}, ..., X^{(n)} \sim \pi(x),$$

  then we can approximate

$$E_\pi[h(X)] \approx \bar{h}_n = \frac{1}{n}\sum_{t=1}^{n} h(X^{(t)}).$$

- This is Monte Carlo integration.

- For independent samples, by Law of Large numbers,

$$\bar{h}_n \to E_\pi[h(X)] \tag{1}$$

  as $n \to \infty$.
- But, generating independent samples from $\pi(x)$ may be difficult.
- It turns out that (1) still applies if we generate samples using a Markov chain. That is, the sequence $X^{(1)}, X^{(2)}, ..., X^{(n)}$ constitutes a certain Markov chain.
- This is the main idea of MCMC.
- Consider X as $\boldsymbol{\theta}$.

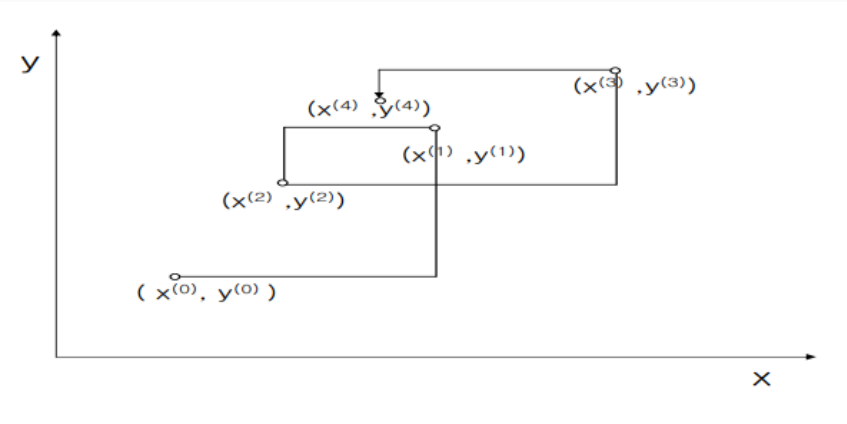
Gibbs sampler

- Let $(X, Y) \sim \pi(x, y)$.
- Generating (X, Y) jointly from $\pi(x, y)$ is difficult.
- However, generating $X|Y = y \sim \pi(x|y)$ and $Y|X = x \sim \pi(y|x)$ is easy.
  (Note that the conditional probability $\pi(x|y)$ is the transition probability that the state y moves to the state x in the next step)
- Under this situation, the Gibbs sampler is an algorithm to construct a Markov chain whose stationary distribution is π.

Gibbs sampler algorithm

1. Initialization: Set $X^{(0)} = x^{(0)}$ and $Y^{(0)} = y^{(0)}$.
2. For $i = 1$ to n,
   - Generate $X^{(i)} \sim \pi(x|y^{(i-1)})$.
   - Generate $Y^{(i)} \sim \pi(y|x^{(i)})$.

- $(X^{(1)}, Y^{(1)}), (X^{(2)}, Y^{(2)}), \ldots$ is a Markov chain with stationary distribution $\pi(x, y)$.
- The sample path of the Gibbs sampler will look something like



Example

- Let $Y_i \sim^{i.i.d} N(\mu, \sigma^2)$ and $\pi(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$.
- We had

$$\pi(\mu, \sigma^2|y) \propto (\frac{1}{\sigma^2})^{n/2+1} \exp\{-\frac{\sum(y_i - \mu)^2}{2\sigma^2}\}$$

- Let $\tau = 1/\sigma^2$. Then, it is easy to derive
  - $\pi(\mu|\sigma^2, y) = N(\bar{y}, \sigma^2/n)$
  - $\pi(\tau|\mu, y) = Gamma(n/2, \sum(y_i - \mu)^2/2)$

Let $x = (x_1, \cdots, x_d)$ and $y = (y_1, \cdots, y_d)$ and let $x \sim_j y$ if $x_i = y_i$ for all $i = j$.

Gibbs sampling and detailed balance condition
Markov chain constructed by Gibbs sampling satisfies the detail balance condition wrt π.

(why?) $P_{xy} = \frac{1}{d} \frac{\pi(y)}{\sum_{z:z\sim_j x} \pi(z)}$ So,

$$\pi(x)P_{xy} = \frac{1}{d}\frac{\pi(x)\pi(y)}{\sum_{z:z\sim_j x}\pi(z)} = \frac{1}{d}\frac{\pi(y)\pi(x)}{\sum_{z:z\sim_j x}\pi(z)} = \pi(y)P_{yx}.$$

Gibbs sampler algorithm for general cases

1. Initialization: Set $X_1^{(0)} = x_1^{(0)}, ..., X_p^{(0)} = x_p^{(0)}$.
2. For $i = 1$ to n,
   - Generate $X_1^{(i)} \sim \pi(x_1 | x_2^{(i-1)}, ..., x_p^{(i-1)})$.
   - Generate $X_2^{(i)} \sim \pi(x_2 | x_1^{(i)}, x_3^{(i-1)}, ..., x_p^{(i-1)})$.
   - Generate $X_3^{(i)} \sim \pi(x_3 | x_1^{(i)}, x_2^{(i)}, x_4^{(i-1)} ..., x_p^{(i-1)})$.
   - .....
   - Generate $X_p^{(i)} \sim \pi(x_p | x_1^{(i)}, ..., x_{p-1}^{(i)})$.

Example 8 (Censored data)

- Let $X_i \sim_{i.i.d}$ Exp(λ).
- Observations are $T_i = \min\{X_i, C_i\}$ and $\delta_i = I(X_i \leq C_i)$ where C_i's are censoring times.
- Prior : $\lambda \sim Gamma(\alpha, \beta)$.
- Objective : Obtain the posterior distribution of λ given $(T_1, \delta_1), ..., (T_n, \delta_n)$.

(continue with the example)

- Note that if we observe $X_1, ..., X_n$, we have

$$\pi(\lambda|X_1, ..., X_n) = Gamma(\alpha + n, \beta + \sum_{i=1}^{n} X_i).$$

- The main idea of the Gibbs sampler is to consider the joint posterior distribution of λ and $(X_1, ..., X_n)$ given the observations.
- That is, the Gibbs sampler generate λ and $(X_1, ..., X_n)$ successively from $\pi(\lambda|X_1, ..., X_n, \text{Data})$ and $\pi(X_1, ..., X_n|\lambda, \text{Data})$.

(continue with the example)

Gibbs sampler algorithm

1. Initialization : $\lambda^{(0)}$ and $X_1^{(0)}, ..., X_n^{(0)}$.
2. For $i = 1$ to n,
   - $\lambda^{(i)} \sim Gamma(\alpha + n, \beta + \sum_{k=1}^{n} X_k^{(i-1)})$.
   - $X_1^{(i)}, ..., X_n^{(i)} \sim \prod_{k=1}^{n} \pi(x_k|\lambda^{(i)}, (T_k, \delta_k))$ where
     - If $\delta_k = 1$, $\pi(x_k = T_k|\lambda^{(i)}, (T_k, \delta_k)) = 1$,
     - If $\delta_k = 0$, $\pi(x_k|\lambda^{(i)}, (T_k, \delta_k)) = \mathsf{Exp}(\lambda^{(i)})|x_k \geq T_k$.

Metropolis-Hastings algorithm

- Let $\pi(x)$ be a distribution of $\mathbb{R}^k$ known except possibly for the normalizing constant.
- The aim is to generate $X \sim \pi$.

Metropolis-Hastings algorithm

1. Choose a transition function $q(y|x)$ of a certain Markov chain.
2. Initialize $x^{(0)}$.
3. For $i = 1$ to n,
   - Generate $\tilde{x} \sim q(x|x^{(i-1)})$.
   - With probability

$$\alpha(x^{(i-1)}, \tilde{x}) = \min\left\{ \frac{\pi(\tilde{x})q(x^{(i-1)}|\tilde{x})}{\pi(x^{(i-1)})q(\tilde{x}|x^{(i-1)})} \right\},$$

   set $x^{(i)} = \tilde{x}$ (acceptance) else set $x^{(i)} = x^{(i-1)}$ (rejection).

MCMC

Markov chain constructed by MH algorithm has the stationary distribution π

(proof) It suffices to prove that the chain satisfies detailed balance condition wrt π.

- The normalizing constant in $\pi(x)$ is not required in the MH algorithm since we only need the ratio $\pi(\tilde{x})/\pi(x^{(i-1)})$.
- If $q(y|x) = \pi(y)$, then we obtain independent samples.
- Usually, q is chosen so that $q(y|x)$ is easy to sample from.
- Theoretically, any density $q(\cdot|x)$ having the same support as $\pi(\cdot)$ should work. However, the choice of q strongly depends on the problem in hand.

Choice of q

- The basic idea of the MH algorithm is
  - from the current position x, move to y according to $q(y|x)$, and
  - we decide to stay at y, roughly speaking, with probability $\pi(y)/\pi(x)$.
- Hence, $q(y|x)$ having more mass when $\pi(y)$ is larger and vice versa is a good candidate.
- Definitely, the best choice of q is π, which is impossible.
- The following three methods are popular:
  - Random walk
  - Independence sampler
  - Utilizing π

Choice of q : Random walk

- $q(y|x) = f(|y - x|)$.
- Then, $y = x + z$ where $z \sim f(|z|)$ (random walk).
- Possible choices of f include the multivariate normal density and the multivariate t density.
- With this q,

$$\alpha(x, y) = \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}.$$

Choice of q : Independence sampler

- $q(y|x) = f(y)$.
- Usual choices of f include the multivariate normal density and the multivariate t density.
- Tails of $f(y)$ must be heavier than tails of $\pi(x)$ for good performance.
- Hence, typically, the variance of f is set to be much larger than the (guestimated) variance of π.
- **Be aware:** The more similar f is to π, the better the MH algorithm performs.

Choice of q : Utilizing π

- Exploit the known form of π to specify q.
- If $\pi(x) \propto \psi(x)h(x)$ where $h(x)$ is an easy-to-generate density and $\psi(x)$ is uniformly bounded. Then, let $q(y|x) = h(y)$.
- Example: Normal-Cauchy model
  - Let $Y_1, ..., Y_n \sim_{i.i.d.} N(\theta, 1)$.
  - $\pi_0(\theta) = \frac{1}{\pi(1+\theta^2)}$.
  - Posterior :

$$\begin{aligned} \pi(\theta|y) &\propto \exp\left(-\frac{\sum_{i=1}^n (y_i - \theta)^2}{2} \right) \times \frac{1}{1+\theta^2} \\ &\propto \exp\left(-\frac{n(\theta - \bar{y})^2}{2} \right) \times \frac{1}{1+\theta^2}. \end{aligned}$$

  - A possibly good choice for $q(y|x)$ is $N(\bar{y}, \tau/n)$ for some $\tau > 1$.

MH algorithm as an optimization algorithm

- Suppose we want to find the maximum of a given function $\pi(x)$.
- Usual numerical methods such as Newton-Raphson or Gradient descent algorithms fails when $\pi(x)$ is not concave.
- The MH algorithm (with random walk q) can be considered as a randomized optimization algorithm:
  - From x, generate y.
  - If $\pi(y) \geq \pi(x)$, move to y.
  - Even if $\pi(y) < \pi(x)$, move to y with positive probability to avoid being trapped at a local maxima.
- Similar optimization algorithms are simulated annealing, genetic algorithm, ...

Convergence diagnostic

- Must do :
  - Plot the times series for each quantity of interest.
  - Plot the auto-correlation functions.
  - Determine the burn-in period and the step size.
- But, realize that you cannot prove that you have converged using any of those.

Variational inference

Approximate Bayesian Inference

- Latent variable $\boldsymbol{\theta} = (\theta_1, ..., \theta_m)$,
- Observations : $\mathbf{x} = (x_1, ..., x_n)$.
- Prior : $p(\boldsymbol{\theta})$.
- Likelihood : $p(\mathbf{x}|\boldsymbol{\theta})$.
- Posterior :

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}.$$

Example : Normal mixture model

$$
\begin{aligned}
\pi = (\pi_1, ..., \pi_K) &\sim \mathcal{D}(\beta, ..., \beta), \\
\mu_k &\sim \mathcal{N}(0, \tau^2), \quad \text{for } k = 1, ..., K, \\
z_i &\sim Multinomial(\pi), \quad \text{for } i = 1, ..., n, \\
x_i &\sim \mathcal{N}(\mu_{z_i}, \sigma^2), \quad \text{for } i = 1, ..., n.
\end{aligned}
$$

- $\boldsymbol{\theta} = (\pi, \boldsymbol{\mu}, \mathbf{z})$.
- Posterior :

$$
p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\pi) \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i|\pi) p(x_i|z_i, \boldsymbol{\mu})}{\int_{\pi} p(\pi) \int_{\boldsymbol{\mu}} \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} \sum_{z_i} p(z_i|\pi) p(x_i|z_i, \boldsymbol{\mu}) d\boldsymbol{\mu} d\pi}
$$

Variational inference

- Variational method is to choose ν where the variational distribution $q(\boldsymbol{\theta}|\nu)$ is well-approximated to the posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$.
- ν : variational parameter
- $q(\boldsymbol{\theta}|\nu)$: variational distribution

Kullback-Leibler Divergence

- Similarity measure : Kullback-Leibler(KL) divergence

$$KL(q||p) = E_q\left[\log \frac{q(\boldsymbol{\theta}|\nu)}{p(\boldsymbol{\theta}|\mathbf{x})}\right]$$

- It is not a "distance" since $KL(q||p) \neq KL(p||q)$.
- $KL(p||p) = 0$.

Evidence Lower Bound(ELBO)

- Minimizing $KL(q||p)$ is equivalent to maximizing ELBO, which will be defined below.
- By Jensen inequality, $f(E[X]) \geq E[f(X)]$ when f is concave.
- Definition of ELBO :

$$\begin{aligned}
\log p(\mathbf{x}) &= \log \int p(\mathbf{x}, \boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \log \int \frac{p(\mathbf{x}, \boldsymbol{\theta})}{q(\boldsymbol{\theta}|\nu)} q(\boldsymbol{\theta}|\nu) d\boldsymbol{\theta} \\
&= \log \left(E_q \left[\frac{p(\mathbf{x}, \boldsymbol{\theta})}{q(\boldsymbol{\theta}|\nu)} \right] \right) \\
&\geq E_q[\log p(\mathbf{x}, \boldsymbol{\theta})] - E_q[\log q(\boldsymbol{\theta}|\nu)] := \mathcal{L}
\end{aligned}$$

- We have to choose the variational distribution where ELBO can be calculated.

KL divergence and ELBO

$$
\begin{aligned}
KL(q||p) &= E_q[\log q(\boldsymbol{\theta}|\nu)] - E_q[\log p(\boldsymbol{\theta}|\mathbf{x})] \\
&= E_q[\log q(\boldsymbol{\theta}|\nu)] - E_q[\log p(\boldsymbol{\theta}, \mathbf{x})] + \log p(\mathbf{x}) \\
&= -\mathcal{L} + \log p(\mathbf{x})
\end{aligned}
$$

Mean-field variational inference

- For the variational distribution, assume that all latent variables are independent:

$$q(\boldsymbol{\theta}|\nu) = \prod_{i=1}^{m} q(\theta_j|\nu_j).$$

- In fact, the latent variables are dependent in view of the posterior distribution.
- $p(\mathbf{x}, \boldsymbol{\theta})$ can be decomposed as follows by the property of conditional distribution:

$$p(\mathbf{x}, \boldsymbol{\theta}) = p(\mathbf{x})p(\boldsymbol{\theta}_{-k}|\mathbf{x})p(\theta_k|\boldsymbol{\theta}_{-k}, \mathbf{x})$$

- We update the variational parameter by the coordinate ascent algorithm.
- For updating ν_k, we write ELBO as follows:

$$\mathcal{L} = \log p(\mathbf{x}) + E_q[\log p(\boldsymbol{\theta}_{-k}|\mathbf{x})] + E_q[\log p(\theta_k|\boldsymbol{\theta}_{-k}, \mathbf{x})] - \sum_{j=1}^{m} E_q[\log q(\theta_j|\nu_j)]$$

Mean-field variational inference

- $\mathcal{L}_k$ is defined as the function of ν_k:

$$\begin{aligned}\mathcal{L}_k &:= E_q[\log p(\theta_k|\boldsymbol{\theta}_{-k}, \mathbf{x})] - E_q[\log q(\theta_k|\nu_k)] \\ &= \int q(\theta_k|\nu_k) E_{-k}[\log p(\theta_k|\boldsymbol{\theta}_{-k}, \mathbf{x})] d\theta_k - \int q(\theta_k|\nu_k) \log q(\theta_k|\nu_k) d\theta_k.\end{aligned}$$

  where E_{-k} is the expectation with respect to $\prod_{j \neq k} q(\theta_j|\nu_j)$.

- Under $\int q(\theta_k|\nu_k)d\theta_k = 1$, $q^*(\theta_k|\nu_k)$ which maximizes $\mathcal{L}_k$ is as follows:

$$q^*(\theta_k|\nu_k) \propto \exp\{E_{-k}[\log p(\theta_k|\boldsymbol{\theta}_{-k}, \mathbf{x})]\}$$

Mean-field variational inference

- Assume that $p(\theta_j|\boldsymbol{\theta}_{-j}, \mathbf{x})$ belongs to an exponential family.

$$p(\theta_j|\boldsymbol{\theta}_{-j}, \mathbf{x}) = h(\theta_j) \exp\{\eta(\boldsymbol{\theta}_{-j}, \mathbf{x})^T t(\theta_j) - a(\eta(\boldsymbol{\theta}_{-j}, \mathbf{x}))\}$$

- This assumption is satisfied in many complicated models:
  - Bayesian mixtures of exponential families with conjugate priors
  - Switching Kalman filters
  - Hierarchical HMMs
  - Mixed-membership models of exponential families
  - Factorial mixtures/HMMs of exponential families
  - Bayesian linear regression
- We choose the variational distribution with the same exponential family.

$$\begin{aligned} q(\theta_j|\nu_j) &= h(\theta_j) \exp\{\nu_j^T t(\theta_j) - a(\nu_j)\} \\ q(\boldsymbol{\theta}|\nu) &= \prod_{j=1}^{m} q(\theta_j|\nu_j) \end{aligned}$$

Mean-field variational inference

- We can calculate $E_{-k}[\log p(\theta_k|\boldsymbol{\theta}_{-k}, \mathbf{x})]$ as follows:

$$\begin{aligned} \log p(\theta_k|\boldsymbol{\theta}_{-k}, \mathbf{x}) &= \log h(\theta_k) + \eta(\boldsymbol{\theta}_{-k}, \mathbf{x})^T t(\theta_k) - a(\eta(\boldsymbol{\theta}_{-k}, \mathbf{x})) \\ E_{-k}[\log p(\theta_k|\boldsymbol{\theta}_{-k}, \mathbf{x})] &= \log h(\theta_k) + E_{-k}[\eta(\boldsymbol{\theta}_{-k}, \mathbf{x})]^T t(\theta_k) \\ &\quad - E_{-k}[a(\eta(\boldsymbol{\theta}_{-k}, \mathbf{x}))] \end{aligned}$$

- We can rewrite $q^*(\theta_k|\nu_k)$ as follows:

$$\begin{aligned} q^*(\theta_k|\nu_k) &\propto h(\theta_k) \exp\{E_{-k}[\eta(\boldsymbol{\theta}_{-k}, \mathbf{x})]^T t(\theta_k)\} \\ q^*(\theta_k|\nu_k) &= q(\theta_k|\nu_k^*), \end{aligned}$$

  where

$$\nu_k^* = E_{-k}[\eta(\boldsymbol{\theta}_{-k}, \mathbf{x})]$$

- We update all ν_k^*'s by the above equation until they converge.

Example : Normal mixture model(revisited)

$$\begin{aligned}
\pi = (\pi_1, ..., \pi_K) &\sim \mathcal{D}(\beta, ..., \beta), \\
\mu_k &\sim \mathcal{N}(0, \tau^2), \quad \text{for } k = 1, ..., K, \\
z_i &\sim Multinomial(\pi), \quad \text{for } i = 1, ..., n, \\
x_i &\sim \mathcal{N}(\mu_{z_i}, \sigma^2), \quad \text{for } i = 1, ..., n.
\end{aligned}$$

- We choose the variational distribution as follows:

$$\begin{aligned}
\pi &\sim \mathcal{D}(b_1, ..., b_K), \\
\mu_k &\sim \mathcal{N}(m_k, s_k^2), \quad \text{for } k = 1, ..., K, \\
z_i &\sim Multinomial(p_{i1}, ..., p_{iK}), \quad \text{for } i = 1, ..., n.
\end{aligned}$$

Example : Normal mixture model(revisited)

- The variational method iteratively update the below equations until the variational parameters converge.

$$
\begin{aligned}
p_{ik}^{*} &\propto \exp\left\{\psi(b_k) - \psi(b_1 + \cdots + b_K) + \frac{x_i m_k}{\sigma^2} - \frac{m_k^2 + s_k^2}{2\sigma^2}\right\}, \\
m_k^{*} &= \frac{\sum_{i=1}^{n} p_{ik} x_i}{\frac{\sigma^2}{\tau^2} + \sum_{i=1}^{n} p_{ik}}, \quad s_k^{*2} = \left(\frac{1}{\tau^2} + \frac{\sum_{i=1}^{n} p_{ik}}{\sigma^2}\right)^{-1}, \\
b_k^{*} &= \beta + \sum_{i=1}^{k} p_{ik}.
\end{aligned}
$$

HW

(It is allowed to use a generating code for only uniform and normal distributions)

- Generalized Extreme distribution
- Gamma distribution $\alpha = 3$, $\beta = 3/2$