PGM . Homework 3. Haekyu Park

Problem 1. Markov Chain Monte Carlo

1.

① Show that x follows F.

$$Pr(X \le x) = Pr(F^{-1}(U) \le x) = Pr(U \le F(x)) = F(x)$$

↑ since we have $X \sim F^{-1}(U)$

↑ $F^{-1}(U) \le x \Leftrightarrow F \circ F^{-1}(U) \le F(x) \Leftrightarrow U \le F(x)$

Thus, x follows F.

② The drawback of this method?

F^{-1} should be explicitly valid to use this method.

Scanned with CamScanner

2.

① Given that K_1 and K_2 have $p(\cdot)$ as stationary density, we have

$$\int p(x) \cdot K_1(x,y)\, dx = p(y) \quad \text{and} \quad \text{---} (*)$$

$$\int p(x) \cdot K_2(x,y)\, dx = p(y). \quad \text{---} (**)$$

② Show that $K_1 \circ K_2$ has $p(\cdot)$ as stationary density.

$$\int p(x) \cdot K_1 \circ K_2(x,z)\, dx$$

$$= \int p(x) \int K_2(x,y) K_1(y,z)\, dy\, dx \quad \text{by (1)}$$

$$= \int p(y) K_1(y,z)\, dy. \quad \text{by } (**)$$

$$= p(z). \quad \text{by } (*)$$

$\therefore K_1 \circ K_2$ has $p(\cdot)$ as stationary density.

Scanned with CamScanner

② . Show that $\lambda K_1 + (1-\lambda) K_2$ for any $\lambda \in [0,1]$ has $p(\cdot)$ as stationary density.

$$\int p(x) \left(\lambda K_1(x,y) + (1-\lambda) K_2(x,y)\right) dx$$

$$= \lambda \cdot \underbrace{\int p(x) K_1(x,y)\, dx}_{\text{by } (*)} + (1-\lambda) \underbrace{\int p(x) K_2(x,y)\, dx}_{\text{by } (**)}$$

$$= \lambda \underbrace{p(y)} + (1-\lambda) \cdot \underbrace{p(y)}$$

$$= p(y)$$

$\therefore \lambda K_1 + (1-\lambda) K_2$ has $p(\cdot)$ as stationary density.

Scanned with CamScanner

3.

We can construct the transition kernel K as:

$$K(s,t) = q(t|s)\,A(s,t) + (1 - r(s))\,\delta_s(t),$$

where $A(s,t) = \min\left(1, \frac{\tilde{p}(t)\,q(s|t)}{\tilde{p}(s)\,q(t|s)}\right) = \min\left(1, \frac{p(t)\,q(s|t)}{p(s)\,q(t|s)}\right)$ is the acceptance probability of a move from state s to state t,

$r(s) = \int q(t|s)\,A(s,t)\,dt$, and $\delta_s(t)$ is the Dirac mass at s.

We want to show that p satisfies detailed balance for all s and t.

That is, we want to show that

$$p(s)\,K(s,t) \overset{?}{=} p(t)\,K(t,s).$$

① LHS $= p(s)\,K(s,t)$

$$= p(s)\left[q(t|s)\,A(s,t) + (1-r(s))\,\delta_s(t)\right]$$

$$= p(s)\,q(t|s)\,A(s,t) + p(s)(1-r(s))\,\delta_s(t)$$

$$= \left(p(s)\,q(t|s) \text{ or } p(t)\,q(s|t)\right) + p(s)(1-r(s))\,\delta_s(t)$$

↑ when $A(s,t) = 1$ ↑ when $A(s,t) \neq 1$

Scanned with CamScanner

② RHS $= p(t)\,K(t,s)$

$= p(t)\left[q(s|t)\,A(t,s) + (1-r(t))\,\delta_t(s)\right]$

$= \left(p(t)\,q(s|t) \text{ or } p(s)\,q(t|s)\right) + p(t)(1-r(t))\,\delta_t(s)$

↑ when $A(t,s)=1$; ↑ when $A(t,s)\neq 1$

③ LHS − RHS

$= \left(p(s)\,q(t|s) \text{ or } p(t)\,q(s|t)\right) - \left(p(t)\,q(s|t) \text{ or } p(s)\,q(t|s)\right)$

↑ when $A(s,t)=1$; ↑ when $A(s,t)\neq 1$; ↑ when $A(t,s)=1$; ↑ when $A(t,s)\neq 1$

$+ p(s)(1-r(s))\,\delta_s(t) - p(t)(1-r(t))\,\delta_t(s)$

↳ if $s=t$, this term is 0.

otherwise, $p(s)(1-r(s))\,\delta_s(t) = p(t)(1-r(t))\,\delta_t(s) = 0$

thus this term is 0 as well.

$= 0 + 0 = 0.$

∴ LHS = RHS, p satisfies detailed balance.

Thus, p is the stationary distribution.

Scanned with CamScanner

4.

The transition kernel of Gibbs chain is

$$K(\mathbf{x}, \mathbf{x}') = p_1(x_1' \mid x_2, \cdots, x_d) \times p_2(x_2' \mid x_1', x_3, \cdots, x_d) \times \cdots \times p_d(x_d' \mid x_1', \cdots, x_{d-1}').$$

To show that $p(\mathbf{x})$ is the stationary distribution of the Markov chain, we want to show that

$$\int K(\mathbf{x}, \mathbf{x}') p(\mathbf{x}) \, d\mathbf{x} = p(\mathbf{x}').$$

Let $\hat{p}_i(\cdot)$ be the marginal over all variables except for x_i.

$$\int K(\mathbf{x}, \mathbf{x}') p(\mathbf{x}) \, d\mathbf{x}$$

$$= \int p_1(x_1' \mid x_2, \cdots, x_d) \, p_2(x_2' \mid x_1', x_3, \cdots, x_d) \times \cdots \times p_d(x_d' \mid x_1', \cdots, x_{d-1}') \times \hat{p}_1(x_2, \cdots, x_d) \, p(x_1 \mid x_2, \cdots, x_d) \, dx_1 \cdots dx_d$$

$$= \int p_2(x_2' \mid x_1', x_3, \cdots, x_d) \times \cdots \times p_d(x_d' \mid x_1', \cdots, x_{d-1}') \times p(x_1', x_2, \cdots, x_d) \, dx_2 \cdots dx_d$$

Scanned with CamScanner

$$= \int P_2(x_2' \mid x_1', x_3, \cdots, x_d) \times \cdots \times P_d(x_d' \mid x_1', \cdots, x_{d-1})$$
$$\times \hat{P}_2(x_1', x_3, \cdots, x_d)\, P(x_2 \mid x_1', x_3, \cdots, x_d)\, dx_2 \cdots dx_d$$

$$= \cdots$$

$$= P(x_1', \cdots, x_d')$$

$$= P(x').$$

Thus, $P(x)$ is the stationary distribution.

Scanned with CamScanner

Problem 2 Box-Muller method

I refer David Barber's text book (Bayesian and Machine Learning).

(Change of Variable)

For a univariate continuous random variable x with distribution $p(x)$ the transformation $y=f(x)$, where $f(x)$ is a monotonic function, has distribution

$$p(y) = p(x)\left(\frac{df}{dx}\right)^{-1}, \quad x = f^{-1}(y).$$

Let's denote $\mathbb{y} = (y_1, y_2)$ and $\mathbb{x} = (x_1, x_2)$.

For multivariate $\mathbb{x}$ and bijection $f(\mathbb{x})$, then $\mathbb{y} = f(\mathbb{x})$ has distribution

$$p(\mathbb{y}) = p(\mathbb{x} = f^{-1}(\mathbb{y}))\left|\det\left(\frac{\partial f}{\partial \mathbb{x}}\right)\right|^{-1}$$

where the Jacobian matrix has elements

$$\left[\frac{\partial f}{\partial \mathbb{x}}\right]_{ij} = \frac{\partial f_i(\mathbb{x})}{\partial x_j}.$$

Scanned with CamScanner

① $\left| \det \frac{\partial f}{\partial x} \right| = \left| \det \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{bmatrix} \right|$

- $\frac{\partial y_1}{\partial x_1} = \frac{\partial}{\partial x_1} \left(\sqrt{-2 \log x_1} \cos 2\pi x_2 \right)$

$= \frac{-\cancel{2} \cdot \frac{1}{x_1}}{\cancel{2}\sqrt{-2\log x_1}} \cdot \cos 2\pi x_2 = \frac{-1}{\sqrt{-2\log x_1}} \cdot \frac{\cos 2\pi x_2}{x_1}$

- $\frac{\partial y_1}{\partial x_2} = \frac{\partial}{\partial x_2} \left(\sqrt{-2 \log x_1} \cos 2\pi x_2 \right)$

$= \sqrt{-2\log x_1} \cdot \left(-\sin(2\pi x_2)\right) \cdot 2\pi = -2\pi \cdot \sqrt{-2\log x_1} \cdot \sin 2\pi x_2$

- $\frac{\partial y_2}{\partial x_1} = \frac{\partial}{\partial x_1} \left(\sqrt{-2\log x_1} \sin 2\pi x_2 \right)$

$= \frac{-2 \cdot \frac{1}{x_1}}{2\sqrt{-2\log x_1}} \sin 2\pi x_2 = \frac{-1}{\sqrt{-2\log x_1}} \cdot \frac{\sin 2\pi x_2}{x_1}$

- $\frac{\partial y_2}{\partial x_2} = \frac{\partial}{\partial x_2} \left(\sqrt{-2 \log x_1} \sin 2\pi x_2 \right)$

$= \sqrt{-2\log x_1} \cdot \cos 2\pi x_2 \cdot 2\pi$

$\therefore \left| \det \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{bmatrix} \right| = \left| -\frac{2\pi}{x_1} \cos^2 2\pi x_2 - \frac{2\pi}{x_1} \sin^2 2\pi x_2 \right| = \frac{2\pi}{x_1}$

Scanned with CamScanner

② $y_1 = \sqrt{-2\log x_1}\cos 2\pi x_2$, $y_2 = \sqrt{-2\log x_1}\sin 2\pi x_2$

$\downarrow$

$$\cos 2\pi x_2 = \frac{y_1}{\sqrt{-2\log x_1}} \qquad \sin 2\pi x_2 = \frac{y_2}{\sqrt{-2\log x_1}}$$

$$\cos^2 2\pi x_2 + \sin^2 2\pi x_2 = 1 = \frac{y_1^2 + y_2^2}{-2\log x_1}$$

$$\therefore x_1 = \exp\left(-\frac{y_1^2 + y_2^2}{2}\right)$$

③ $$p(y) = p(x = f^{-1}(y)) \left| \det\left(\frac{\partial f}{\partial x}\right)\right|^{-1}$$ ← by change of variable

$$= p(x = f^{-1}(y)) \cdot \frac{x_1}{2\pi}$$ ← by ①

$$= p(x = f^{-1}(y)) \cdot \frac{1}{2\pi} \cdot e^{-\frac{1}{2}(y_1^2 + y_2^2)}$$ ← by ②

$$= 1 \cdot \frac{1}{2\pi} \cdot e^{-\frac{1}{2}(y_1^2 + y_2^2)}$$ ← $p(x_1) = p(x_2) = 1$

$$= \mathcal{N}(y_1 \mid 0,1)\ \mathcal{N}(y_2 \mid 0,1)$$

$\therefore$ We arrive at $p(y_1, y_2) = \mathcal{N}(y_1 \mid 0,1)\ \mathcal{N}(y_2 \mid 0,1)$

Scanned with CamScanner

Problem 3

We know that all neighbors of any white variables are all black, and also that all neighbors of any black variables are all white. When we condition on all black variables, the white variables become independent. Similarly, when we condition on all black variables, the black variables become independent.

Thus, $P(b_1, b_2, \cdots \mid w_1, w_2, \cdots) = P(b_1 \mid w_1, \cdots) \cdot P(b_2 \mid w_1, \cdots) \cdot \cdots$

and $P(w_1, w_2, \cdots \mid b_1, b_2, \cdots) = P(w_1 \mid b_1, \cdots) \cdot P(w_2 \mid b_1, \cdots) \cdots$.

In Gibbs sampling procedure, we can draw white samples from their black neighbors conditioned on the white variables by using the above joint distribution. Also, we can draw black samples from their white neighbors conditioned on the black variables by using the above joint distribution.

Scanned with CamScanner

Problem 4

The code is in Problem-4/symptomDisease.m. In the code, I used 500 burn-in, and 10000 sub-sampling. The vector computed is:

==========

Columns 1 through 10

0.0029 0.9979 0.0214 1.0000 0.6472 0.0115 0.0149 0.0008 0.0053 0.9999

Columns 11 through 20

0.0051 1.0000 1.0000 1.0000 0.9907 0.9677 0.9898 0.9055 0.9998 0.9933

Columns 21 through 30

0.0938 0.7473 1.0000 0.9935 0.0056 0.9996 0.0173 0 0.9986 0

Columns 31 through 40

0 0.0996 0.0115 1.0000 0 0.0031 0.9938 0.0015 0 0.0002

Columns 41 through 50

0.9927 0.9985 1.0000 0.9997 0.0001 0.0167 0.9959 0.9966 0.0002 0.9993

Problem 5

1) Derive $p(d \mid s, D) = \int_{W, b, p} p(d \mid s, W, b, p)\, p(W, b, p \mid D)$

$$p(d \mid s, D) = \int_{W, b, p} p(d \mid s, D, W, b, p)\, p(W, b, p \mid D)$$

$$= \int_{W, b, p} p(d \mid s, W, b, p)\, p(W, b, p \mid D)$$

↑ since the prediction is independent to D.

2). Explain how to estimate $p(d_i = 1 \mid s, D)$ using sampling.

We can use Gibbs sampling. By fixing b and p, we can perform Gibbs sampling on $(d \mid s, D)$. By fixing W, p, we can perform Gibbs sampling on $(d \mid s, D)$. By fixing W, b, we can perform Gibbs sampling on $(d \mid s, D)$. Through this controlled Gibbs sampling, we can estimate $p(d_i = 1 \mid s, D)$.

Scanned with CamScanner

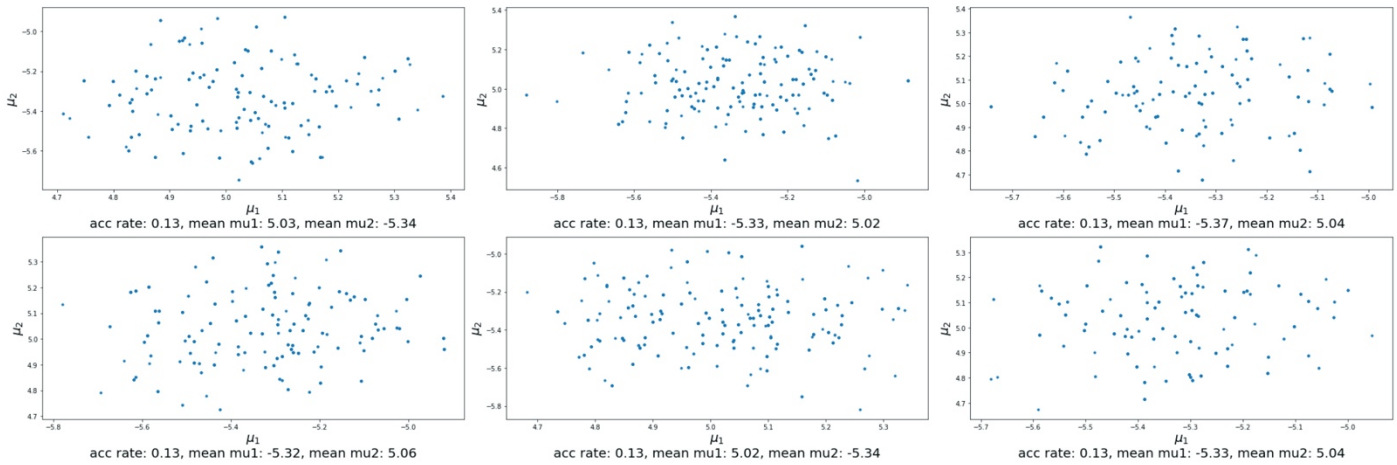
Problem 6

The code is in Problem-6/problem_6.ipynb. The pdf version of the notebook is Problem-6/problem_6.pdf.

a) See problem_6.ipynb, Problem a).

b)

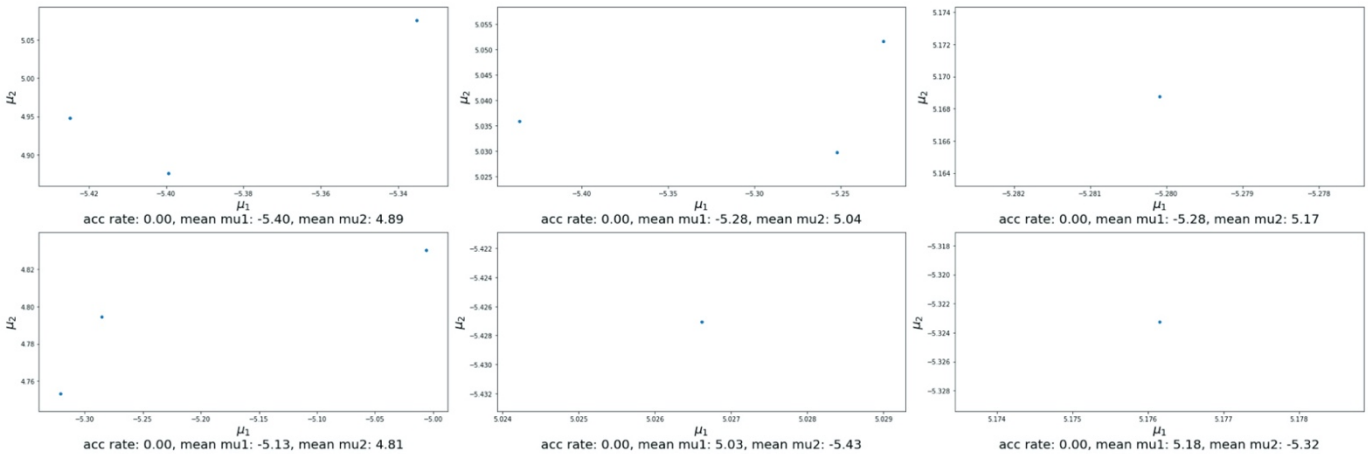
When sigma = 0.5, the plot of the last 1000 mu1 and mu2 is:



The estimated mean of mu1 and mu2, and the acceptance rate are:

mean_mu1	mean_mu2	acceptance rate
5.025064	-5.340499	0.131000
-5.332442	5.021185	0.131091
-5.365892	5.035718	0.130727
-5.316841	5.056698	0.127909
5.019107	-5.335424	0.134545
-5.333710	5.039708	0.127909

When sigma = 5, the plot of the last 1000 mu1 and mu2 is:



The estimated mean of mu1 and mu2, and the acceptance rate are:

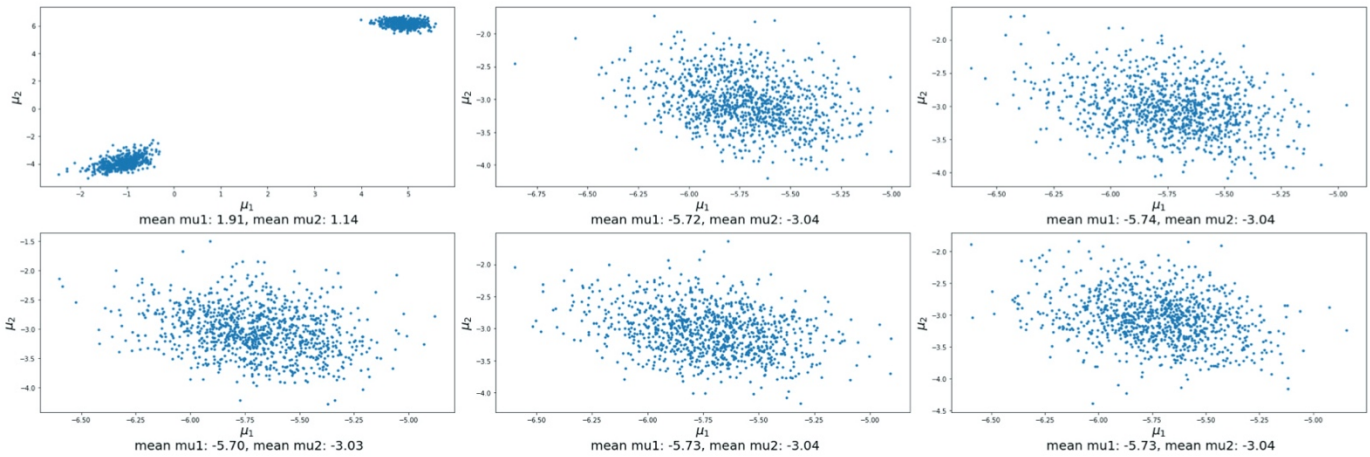
mean_mu1	mean_mu2	acceptance rate
-5.401915	4.894947	0.002273
-5.277091	5.037582	0.001455
-5.280091	5.168751	0.002182
-5.128976	4.807459	0.003273
5.026610	-5.427057	0.001909
5.176153	-5.323255	0.002182

The differences between sigma = 0.5 and sigma = 5 are:

- When sigma = 5, the acceptance rate is lower, compared to when sigma = 0.5.
- When sigma = 0.5, the mu1 and mu2 are more scattered, compared to when sigma = 5.

c)

The plot of the last 1000 mu1 and mu2 is:



The estimated mean of mu1 and mu2 are:

mean_mu1	mean_mu2
1.912066	1.142613
-5.724099	-3.042464
-5.738514	-3.035543
-5.704386	-3.032615
-5.725242	-3.040966
-5.728205	-3.044541

Problem 7

Run Problem-7/KL.m. We can get the value of the minimal KL Divergence for the optimal q as 0.0159647.

Problem 8

Run Problem-8/BPMF.m.

1. The result of loopy belief propagation

	j =1	j =2
i=1	0.0364	0.9636
i=2	0.7064	0.2936
i=3	0.4510	0.5490
i=4	0.8302	0.1698

2. The result of variational mean-field equation

	j =1	j =2
i=1	0.0021	0.9979
i=2	0.9216	0.0784
i=3	0.5668	0.4332
i=4	0.8779	0.1221

3. The marginals of exact inference

	j =1	j =2
i=1	0.0325	0.9675
i=2	0.7104	0.2896
i=3	0.4509	0.5491
i=4	0.8292	0.1708

4. Mean expected deviation in the marginals:

BP = 0.00218917
MF = 0.101594

Belief propagation works better since it has lower mean error.

Problem 9.

1.

Let's denote z be the normalization coefficient for $r(x)$. Then,

$$r(x) = \frac{p(x)f(x)}{z}$$

By construction, $J = \log z$.

Taking $KL(r|q)$ gives

$$\begin{aligned}\log z = J &\geq -\langle \log q(x) \rangle_{q(x)} + \langle \log p(x) \rangle_{q(x)} + \langle \log f(x) \rangle_{q(x)} \\ &= -KL(q(x) \mid p(x)) + \langle \log f(x) \rangle_{q(x)}\end{aligned}$$

2.

$$J \geq -KL(q(x) \mid p(x)) + \langle \log f(x) \rangle_{q(x)} \quad \leftarrow \text{by 1}$$

$$= -KL(q(x) \mid p(x)) + \underbrace{\langle \log f(x) \rangle_{q(x)} - \langle \log q(x) \rangle_{q(x)}} + \langle \log q(x) \rangle_{q(x)}$$

$$= -KL(q(x) \mid p(x)) \underbrace{- KL(q(x) \mid f(x))} + \langle \log q(x) \rangle_{q(x)}$$

$$= -KL(q(x) \mid p(x)) - KL(q(x) \mid f(x)) - H(q(x)).$$

Scanned with CamScanner