Probabilistic Graphical Models
Solution to
HWK#3 (Part A)
Assigned Sunday, March 8, 2020
Due: Tuesday, March 31, 2020

Problem 1 Markov Chain Monte Carlo (15 Points)

1. A simple sampling method adopted by many of the standard math libraries is the *inverse probability transform*: draw $u \sim \text{Unif}(0,1)$, then draw $x \sim F^{-1}(u)$, where F^{-1} is the inverse of the cdf. Show that x generated by this procedure follows distribution F. What is the drawback of this method?

Solution:

1. We have $U \sim \text{Unif}(0,1)$ and $X \sim F^{-1}(U)$. Evaluate the cdf of X at t:

$$\mathbb{P}(X \leq t) = \mathbb{P}\left(F^{-1}(U) \leq t\right) = \mathbb{P}(U \leq F(t)) = F(t).$$

This method is only applicable if the explicit form of F^{-1} is known.

Same result holds for *generalized inverse cdf* defined as

$$F^{-1}(t) = \inf\{x : F(x) \geq t\}.$$

2. We have learned transition matrix in class which only applies to discrete state Markov chains. A more general concept is the transition kernel, which can apply to both discrete and continuous state spaces. The transition kernel $k(x, x')$ is defined as $k(x, x') = P(\theta^{(t+1)} = x' | \theta^{(t)} = x)$, where $\theta^{(t)}$ is the state at time t. Show that if both transition kernels K_1 and K_2 have $p(\cdot)$ as stationary density, so do $K_1 K_2$ and $\lambda K_1 + (1-\lambda)K_2$ for any $\lambda \in [0,1]$. In practice, the former corresponds to sampling from K_1 and K_2 *cyclically* and the latter draws either K_1 with probability λ or K_2 otherwise. Note that in the continuous case, the cyclic kernel can be defined as composition of functions:

$$(K_1 \circ K_2)(x, z) = \int K_2(x, y) K_1(y, z)\, \mathrm{d}y. \tag{1}$$

Solution:

2. We have

$$\int p(x)K_1(x,y)\,\mathrm{d}x = p(y) \quad \text{and} \quad \int p(x)K_2(x,y)\,\mathrm{d}x = p(y).$$

Cyclic kernel $K = K_1 \circ K_2$:

$$\begin{aligned}\int p(x)K(x,z)\,\mathrm{d}x &= \int p(x)\int K_2(x,y)K_1(y,z)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int K_1(y,z)p(y)\,\mathrm{d}y \\ &= p(z).\end{aligned}$$

Mixture of kernels $K = \lambda K_1 + (1-\lambda)K_2$:

$$\begin{aligned}\int p(x)K(x,y)\,\mathrm{d}x &= \int p(x)\left(\lambda K_1(x,y) + (1-\lambda)K_2(x,y)\right)\,\mathrm{d}x \\ &= \lambda p(y) + (1-\lambda)p(y) \\ &= p(y).\end{aligned}$$

Discrete case follows similarly.

3. Recall MH sampling for target distribution $p(x)$ using proposal $q(x|y)$: at state s, first draw $t \sim q(t|s)$, then accept t with probability

$$A = \min\left(1, \frac{\widetilde{p}(t)q(s|t)}{\widetilde{p}(s)q(t|s)}\right),$$

where $\widetilde{p}(x)$ is the unnormalized target distribution. Show that $p(x)$ is the stationary distribution of the Markov chain defined by this procedure. Consider both continuous and discrete cases.

Solution:

3. The transition kernel of Metropolis-Hastings for target density $f(x)$ with proposal $q(y|x)$:

$$K(x,y) = q(y|x)A(x,y) + (1 - r(x))\,\delta_x(y),$$

where $r(x) = \int q(y|x)A(x,y)\,\mathrm{d}y$ and δ_x is the Dirac mass at x.
We only need to show detailed balance holds:

$$f(x)\Big[q(y|x)A(x,y) + (1-r(x))\,\delta_x(y)\Big] \stackrel{?}{=} f(y)\Big[q(x|y)A(y,x) + (1-r(y))\,\delta_y(x)\Big].$$

The first term on the LHS:

$$f(x)q(y|x)A(x,y) = f(x)q(y|x) \wedge f(y)q(x|y).$$

Similarly, the first term on the RHS:

$$f(y)q(x|y)A(y,x) = f(y)q(x|y) \wedge f(x)q(y|x).$$

Thus above two equations are equal. Also by inspecting the second term, we can notice that they are both zero if x≠ y or are both 1 − r(x) if x = y. Thus detailed balance holds for K, which means f(x) is the stationary density of the Markov chain.

4. Recall Gibbs sampling for target distribution $p(\mathbf{x}) = p(x_1, \ldots, x_d)$: for each $j \in \{1, \ldots, d\}$, draw $t \sim p(x_j|\text{rest})$ and set $x_j = t$. Show that $p(\mathbf{x})$ is the stationary distribution of the Markov chain defined by this procedure.

Solution:

4. The transition kernel of Gibbs chain is

$$\begin{aligned} K(\mathbf{x}, \mathbf{x}') = & \, p_1(x_1'|x_2, \ldots, x_d) \\ & \times p_2(x_2'|x_1', x_3, \ldots, x_d) \cdots \times p_d(x_d'|x_1', \ldots, x_{d-1}'). \end{aligned}$$

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

$$\begin{aligned} \int K(\mathbf{x}, \mathbf{x}')p(\mathbf{x}) \, \mathrm{d}\mathbf{x} = & \int \underbrace{p_1(x_1'|x_2, \ldots, x_d) \cdots}_{K(\mathbf{x}, \mathbf{x}')} \\ & \times \underbrace{p^{(1)}(x_2, \ldots, x_d)p(x_1|x_2, \ldots, x_d)}_{p(\mathbf{x})} \, \mathrm{d}x_1 \cdots \mathrm{d}x_d \\ = & \int p_2(x_2'|x_1', x_3, \ldots, x_d) \cdots \\ & \times p(x_1', x_2, \ldots, x_d) \underbrace{p(x_1|x_2, \ldots, x_d) \, \mathrm{d}x_1}_{\text{vanish}} \cdots \mathrm{d}x_d \\ = & \int p_2(x_2'|x_1', x_3, \ldots, x_d) \cdots \\ & \times p^{(2)}(x_1', x_3, \ldots, x_d)p(x_2|x_1', x_3, \ldots, x_d) \, \mathrm{d}x_2 \cdots \mathrm{d}x_d. \end{aligned}$$

Apply recursively, eventually we will arrive at

$$\int K(\mathbf{x}, \mathbf{x}')p(\mathbf{x}) \, \mathrm{d}\mathbf{x} = p(x_1', \ldots, x_d') = p(\mathbf{x}').$$

Alternatively, we can make use of the fact that each Gibbs sampling subroutine is a special case of Metropolis-Hastings with acceptance rate always equal to 1. However note that the transition kernel (current conditional) is changing within each step, so we need the result on the stationarity of cyclic kernels to complete the proof.

Problem 2 Box-Muller method (10 points)

Let $x_1 \sim U\left(x_1 | [0,1]\right)$, $x_2 \sim U\left(x_2 | [0,1]\right)$ *and*

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

Show that

$$p(y_1, y_2) = \int p(y_1|x_1, x_2)p(y_2|x_1, x_2)p(x_1)p(x_2)dx_1dx_2 = \mathcal{N}\left(y_1|0,1\right)\mathcal{N}\left(y_2|0,1\right)$$

and suggest an algorithm to sample from a univariate normal distribution. Hint: Use the change of variable result for vectors $y = (y_1, y_2)$ and $x = (x_1, x_2)$.

Solution:

Writing $\mathbf{y} = (y_1, y_2)$ and $\mathbf{x} = (x_1, x_2)$, we have

$$p(\mathbf{y}) = \int \delta(\mathbf{y} - f(\mathbf{x}))d\mathbf{x} = \int \delta(\mathbf{x} - f^{-1}(\mathbf{y}))\left|\frac{d}{d\mathbf{y}}\mathbf{x}\right| d\mathbf{y}$$

From exercise(8.10), we need to find the absolute value of the determinant of the Jacobian:

$$\begin{vmatrix} \frac{\partial}{\partial x_1}y_1 & \frac{\partial}{\partial x_2}y_1 \\ \frac{\partial}{\partial x_1}y_2 & \frac{\partial}{\partial x_2}y_2 \end{vmatrix} = \begin{vmatrix} -(-2\log x_1)^{-\frac{1}{2}}\frac{\cos 2\pi x_2}{x_1} & -2\pi(-2\log x_1)^{\frac{1}{2}}\sin 2\pi x_2 \\ -(-2\log x_1)^{-\frac{1}{2}}\frac{\sin 2\pi x_2}{x_1} & 2\pi(-2\log x_1)^{\frac{1}{2}}\cos 2\pi x_2 \end{vmatrix} = -\frac{2\pi}{x_1}\left(\cos^2 2\pi x_2 + \sin^2 2\pi x_2\right) = -\frac{2\pi}{x_1}$$

From the transformation, we also have

$$x_1 = e^{-\frac{1}{2}(y_1^2+y_2^2)}$$

Hence, since $p(x_1) = p(x_2) = 1$,

$$p(\mathbf{y}) = \frac{1}{2\pi}e^{-\frac{1}{2}(y_1^2+y_2^2)} = \mathcal{N}\left(y_1|0,1\right)\mathcal{N}\left(y_2|0,1\right)$$

Problem 3 (6 points)

Consider an Ising model on an $M \times M$ square lattice with nearest neighbour interactions:

$$p(x) \propto \exp \beta \sum_{i \sim j} \mathbb{I}\left[x_i = x_j\right]$$

Now consider the $M \times M$ grid as a checkerboard, and give each white square a label w_i, and each black square a label b_j, so that each square is associated with a particular variable. Show that

$$p(b_1, b_2, \ldots, |w_1, w_2, \ldots) = p(b_1|w_1, w_2, \ldots)p(b_2|w_1, w_2, \ldots)\ldots$$

That is, conditioned on the white variables, the black variables are independent. The converse is also true, that conditioned on the black variables, the white variables are independent. Explain how this can be exploited by a Gibbs sampling procedure. This procedure is known as checkerboard *or* black and white sampling.

Solution:

Consider a 4 by 4 example:

$$\begin{pmatrix} w_1 & b_1 & w_2 & b_2 \\ b_3 & w_3 & b_4 & w_4 \\ w_5 & b_5 & w_6 & b_6 \\ b_7 & w_7 & b_8 & w_8 \end{pmatrix}$$

Since all neighbours of any white variable all all black, and similarly, all neighbours of any black variable are all white, when we condition on all the black variables, the white variables become independent, and vice versa. This means that in the Gibbs step to sample from

$$p(b_1, b_2, \ldots, |w_1, w_2, \ldots) = p(b_1|w_1, w_2, \ldots)p(b_2|w_1, w_2, \ldots)\ldots$$

we can draw a sample from the joint distribution by independently drawing a sample from each of the black variables conditioned on its neighbouring white variables. The converse is true when we sample from

$$p(w_1, e_2, \ldots, |b_1, b_2, \ldots) = p(w_1|b_1, b_2, \ldots)p(w_2|b_1, b_2, \ldots)\ldots$$

This is advantageous since we are able to truly draw independent samples from all black variables conditioned on all white variables, and vice versa.

Problem 4 **Approximate Inference via Sampling (10 points)**

Consider a setting in which there are **D** diseases and a patient can either have or not have a particular disease $\mathbf{d_i} \in \{0, 1\}$, for $\mathbf{i = 1,...,D}$. Here $\mathbf{d_i = 1}$ means that the patient has disease '**i**'; $\mathbf{d_i = 0}$ means the patient does not have disease '**i**'. A patient may have more than one disease. There are a set of **S** symptoms the hospital can measure; if $\mathbf{s_j = 1}$, for $\mathbf{j = 1,...,S}$ then the patient has symptom '**j**'; otherwise $\mathbf{s_j = 0}$ means that the patient does not have symptom '**j**'. A simple disease-symptom network is given by

$$p(s_1, \ldots, s_S, d_1, \ldots, d_D) = \prod_{j=1}^{S} p(s_j|\mathbf{d}) \prod_{i=1}^{D} p(d_i)$$

where $\mathbf{d} = (d_1, \ldots, d_D)^T$ *and*

$$p(s_j = 1|\mathbf{d}) = \sigma\left(\mathbf{w}_j^T \mathbf{d} + b_j\right)$$

where $\sigma(x) = 1/(1 + \exp(-x))$.

In the above $\mathbf{w_j}$ is a vector of parameters relating symptom **j** to the diseases and $\mathbf{b_j}$ is related to the prevalence of the symptom. The hospital provides the collection of parameters **W** and **b**, the prior disease probabilities **p** (with $\mathbf{p(d_i = 1) = p_i}$) and a vector **s** of symptoms for the patient, see SymptomDiseasePars.mat.

Use Gibbs sampling (using a reasonable amount of burn-in and sub-sampling) to estimate the vector

$$[p(d_1 = 1|\mathbf{s}), \ldots, p(d_D = 1|\mathbf{s})]$$

Solution:

Running the code "symptom-disease-sampling.m" gives the following output:

ans =

```
%   Columns 1 through 10
%     0.0026    0.9982    0.0210    1.0000    0.6495    0.0111    0.0157    0.0006    0.0070    0.9999
%
%   Columns 11 through 20
%
%     0.0049    1.0000    1.0000    1.0000    0.9883    0.9698    0.9857    0.9045    0.9999    0.9936
%
%   Columns 21 through 30
%
%     0.0843    0.7478    0.9998    0.9948    0.0039    0.9996    0.0158         0    0.9989         0
%
%   Columns 31 through 40
%
%     0.0001    0.0885    0.0106    1.0000         0    0.0023    0.9941    0.0013         0    0.0003
%
%   Columns 41 through 50
%
%     0.9923    0.9977    1.0000    0.9998    0.0001    0.0157    0.9967    0.9934         0    0.9998
```

Nearly all the marginals are almost deterministic, except for disease 5, which is relatively uncertain.

Problem 5 **Approximate Inference via Sampling (10 points)**

For a disease-symptom network similar to the previous question and a collection of N patient records $D = (s^n, d^n),\ n = 1, ..., N$, derive the following for the prediction of the diseases for a new patient with symptoms s:.

$$p(\mathbf{d}|\mathbf{s}, \mathcal{D}) = \int_{\mathbf{W},\mathbf{b},\mathbf{p}} p(\mathbf{d}|\mathbf{s}, \mathbf{W}, \mathbf{b}, \mathbf{p}) p(\mathbf{W}, \mathbf{b}, \mathbf{p}|\mathcal{D})$$

where

$$p(\mathbf{W}, \mathbf{b}, \mathbf{p}|\mathcal{D}) \propto p(\mathbf{W}, \mathbf{b}, \mathbf{p}) \prod_{n=1}^{N} p(\mathbf{s}^n|\mathbf{d}^n, \mathbf{W}, \mathbf{b}) p(\mathbf{d}^n|\mathbf{p})$$

and explain how you could estimate

$$p(d_i = 1|\mathbf{s}, \mathcal{D})$$

using sampling.

Solution:

This follows directly from Bayes rule and the iid assumption. We would need to choose a sensible prior for the parameters, for example independent priors for each parameter block $\mathbf{W}$, $\mathbf{b}$, $\mathbf{p}$. One could then use Gibbs sampling to draw parameter samples from $p(\mathbf{W}, \mathbf{b}, \mathbf{p}|\mathcal{D})$ and use these samples to draw samples:

$$p(\mathbf{d}|\mathbf{s}, \mathcal{D}) \approx \frac{1}{L} \sum_{l=1}^{L} p(\mathbf{d}|\mathbf{s}, \mathbf{W}^l, \mathbf{b}^l, \mathbf{p}^l)$$

where here l is the sample index. For each parameter sample l, we then need to run another sampling scheme to draw samples $\mathbf{d}$.

Problem 6 MCMC Sampling (10 points)

Assume we have a mixture of Gaussians model to generate $x \in R$ given parameters μ_1, μ_2:

$$p(x|\mu_1, \mu_2) \propto \frac{1}{2}\exp(-\frac{1}{2}(x-\mu_1)^2) + \frac{1}{2}\exp(-\frac{1}{2}(x-\mu_2)^2)$$

Assume we have N samples $x_1, \cdots, x_N$ drawn from the model. We also impose a Gaussian prior on μ_1 and μ_2:

$$p(\mu_1, \mu_2) \propto \exp(-\frac{1}{2}\frac{\mu_1^2}{100})\exp(-\frac{1}{2}\frac{\mu_2^2}{100})$$

In the below questions, you will be asked to sample from the posterior using Metropolis–Hastings sampling methods, although we can directly sample from it. We have picked the simplest possible model we could think of so that the mechanics of the samplers are incredibly basic. Of course, one would never actually use Metropolis–Hastings for this posterior, but we want to make the behavior of the sampler very easy to understand.

a) Generate 100 samples $x_1, \cdots, x_{100}$ from $0.5\mathcal{N}(-5, 1^2) + 0.5\mathcal{N}(5, 1^2)$ where $\mathcal{N}(\mu, \sigma^2)$ denotes normal distribution with mean μ and variance σ^2. You can use any package to sample from a Gaussian. You do not need to visualize the samples, just submit your code to get full credit.

b) Use Metropolis Hastings Sampling to sample from $P(\mu_1, \mu_2 | x_1, \cdots, x_{100})$. Let the proposal distribution $q(\mu'_1, \mu'_2 | \mu_1, \mu_2)$ be $\mathcal{N}(\mu_1, \sigma^2)\mathcal{N}(\mu_2, \sigma^2)$. Experiment with $\sigma = 0.5$ and $\sigma = 5$ respectively, let the initial point be $(0, 0)$, discard the first 10000 samples and plot the next 1000 samples. Calculate the acceptance rate during the whole process. Report the estimated mean of μ_1 and μ_2. Repeat the experiment 6 times. What's the difference between $\sigma = 0.5$ and $\sigma = 5$?

c) Use Gibbs Sampling to sample from $P(\mu_1, \mu_2 | x_1, \cdots, x_{100})$. In order to make the sampling simple, introduce latent variables $z_1, \cdots, z_{100}$ for $x_1, \cdots, x_{100}$ indicating the Gaussian component where the data comes from. Discard the first 10000 samples and plot the next 1000 samples. Report the estimated mean of μ_1 and μ_2. Repeat the experiment 6 times.

Solution:

When σ gets larger, the acceptance rate goes down.Metropolis Sampling with $\sigma = 0.5$: Refer to Figure 1 and Table 1.

Metropolis Sampling with σ = 5: Refer to Figure 2 and Table 2.

Gibbs Sampling: Refer to Figure 3 and Table 3.

μ_1	μ_2	acceptance rate
5.00006072478	-4.93338418457	0.129818181818
5.0010081317	-4.95831554185	0.127909090909
-4.92284810472	5.00680816166	0.133545454545
5.04564639596	-4.92087198573	0.136818181818
-4.93499411707	5.01781537163	0.128909090909
5.03688216314	-4.91294615155	0.138454545455

Table 1: Metropolis Sampling with $\sigma = 0.5$

μ_1	μ_2	acceptance rate
-5.01025804393	5.18150739467	0.00254545454545
-4.87431212679	5.019837823	0.00309090909091
-4.9652526189	5.08654509263	0.00236363636364
-5.17325113525	5.18626585913	0.00281818181818
-4.91334451461	4.90938607396	0.00190909090909
5.24078037238	-4.91140730237	0.00154545454545

Table 2: Metropolis Sampling with $\sigma = 5$

μ_1	μ_2
-4.94264733014	5.01718753031
-4.93836794507	5.02427160241
5.01609162106	-4.95018273368
5.02024871795	-4.94013878062
-4.94012845973	5.018491535
5.02560504958	-4.93581447745

Table 3: Gibbs sampling

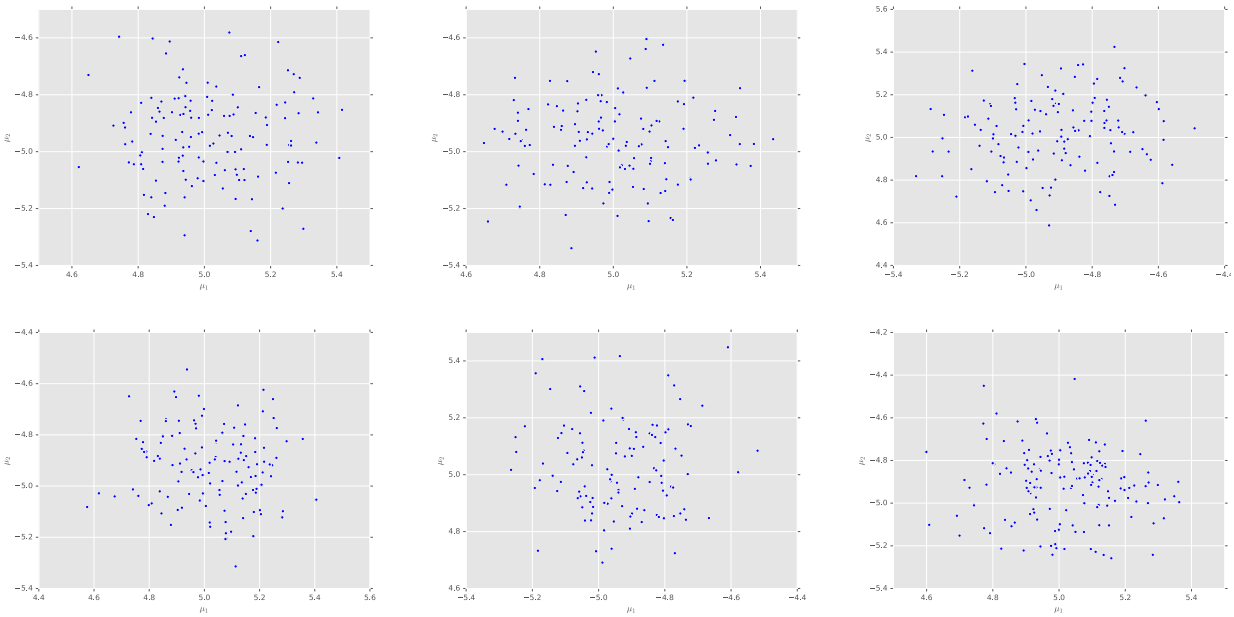

Figure 1: Metropolis Sampling with $\sigma = 0.5$

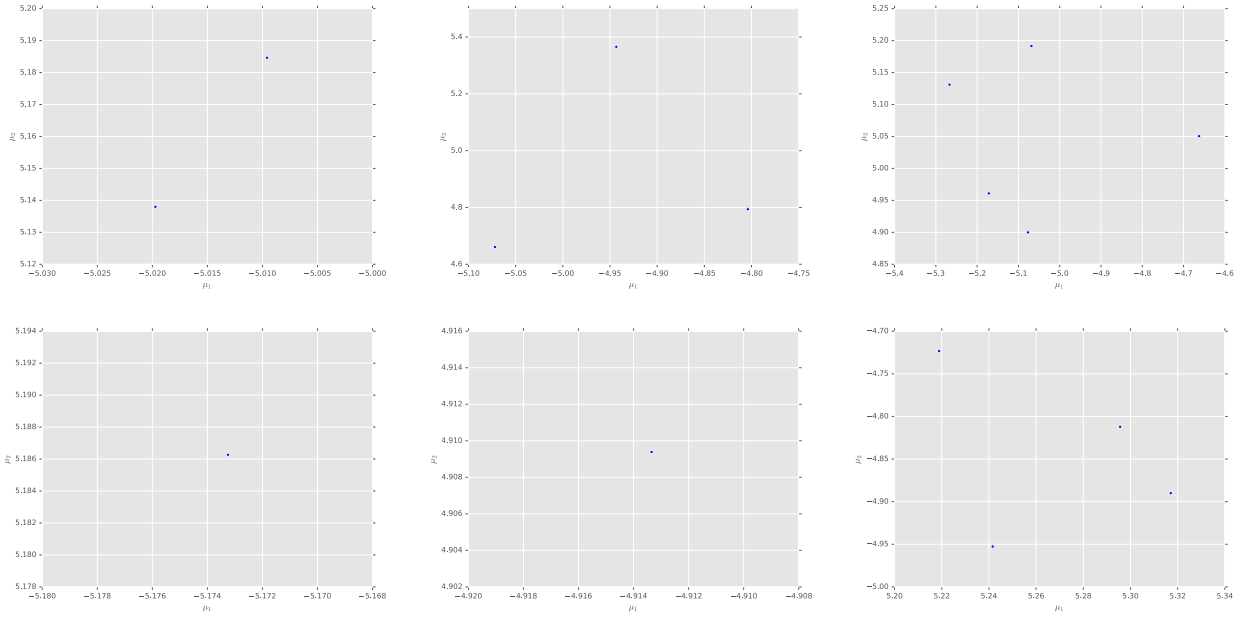

Figure 2: Metropolis Sampling with $\sigma = 5$

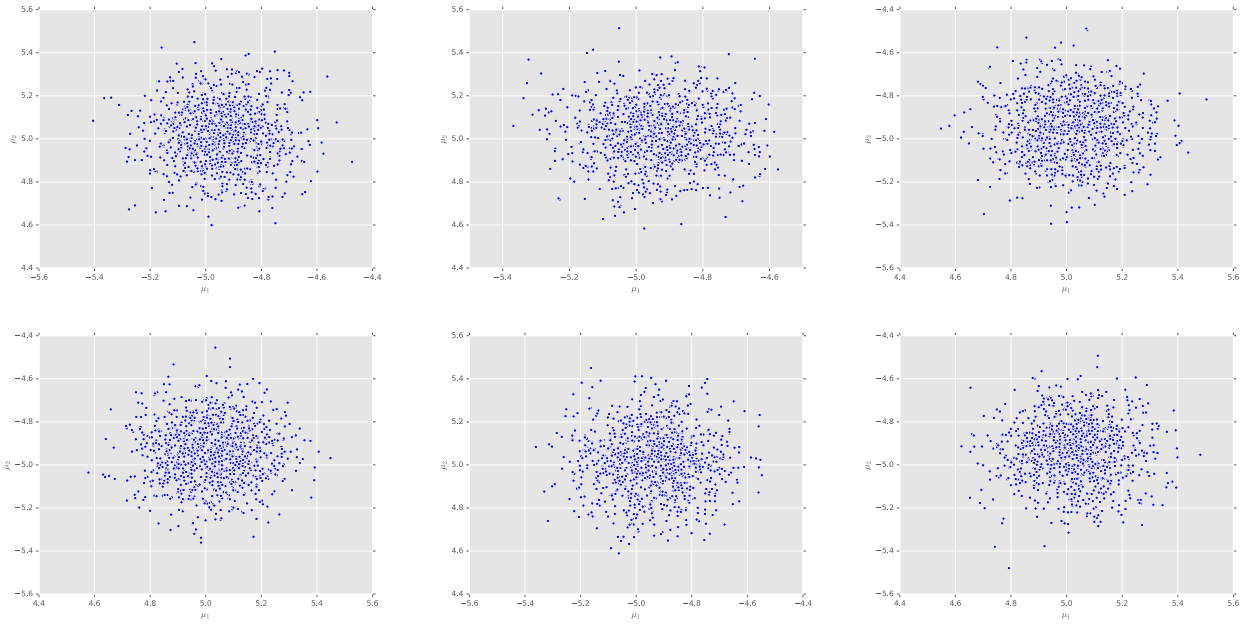

Figure 3: Gibbs sampling

Problem 7 Variational Inference (15 points)

The file p.mat contains a distribution $p(x, y, z)$ on ternary state variables. Using BRML-toolbox, find the best approximation $q(x, y)q(z)$ that minimizes the Kullback-Leibler divergence $KL(q|p)$ and state the value of the minimal Kullback-Leibler divergence for the optimal q.

Solution:

The best approximation of q(x,y)q(z) that minimizes the KL divergence is equal to 0.0160.

See the code "exerciseKLfit.m'' under Problem 7 folder.

Problem 8 Variational Inference (10 points)

Consider the pairwise Markov network defined on a 2 × 2 lattice, as given in the file pMRF.mat.

1. Find the optimal fully factorized approximation $\prod_{i=1}^{4} q_i^{BP}$ by loopy belief propagation, based on the factor graph formalism.

2. Find the optimal fully factorized approximation $\prod_{i=1}^{4} q_i^{MF}$ by solving the variational mean-field equations.

3. By pure enumeration, compute the exact marginals p_i.

4. Averaged over all 4 variables, compute the mean expected deviation in the marginals

$$\frac{1}{4}\sum_{i=1}^{4}\frac{1}{2}\sum_{j=1}^{2}|q_i(x=j)-p_i(x=j)|$$

for both the BP and MF approximations, and comment on your results.

Solution:

See the code "exerciseMFBP.M" under folder Problem 8.

variable(1)

Exact	Loopy BP	MF
0.1374	0.1980	0.1271
0.8626	0.8020	0.8729

variable(2)

Exact	Loopy BP	MF
0.5907	0.6213	0.5392
0.4093	0.3787	0.4608

variable(3)

Exact	Loopy BP	MF
0.1705	0.2265	0.0331
0.8295	0.7735	0.9669

variable(4)

Exact	Loopy BP	MF
0.5196	0.5057	0.5210
0.4804	0.4943	0.4790

mean error BP = 0.0402797
mean error MF = 0.0501128

Both work quite well for this small problem, with BP working slightly better.

Problem 9 Variational Inference (14 points)

Note that in this question, the notation $< \cdot >_q$ denotes averaging with respect to distribution q. *Consider the average of a positive function* $f(x)$ *with respect to a distribution* $p(x)$*:*

$$J = \log \int_x p(x) f(x)$$

where $f(x) \geq 0$*. The simplest version of Jensen's inequality states that*

$$J \geq \int_x p(x) \log f(x)$$

1. *By considering a distribution* $r(x) \propto p(x)f(x)$*, and* $KL(q|r)$*, for some variational distribution* $q(x)$*, show that*

$$J \geq -KL(q(x)|p(x)) + \langle \log f(x) \rangle_{q(x)}$$

The bound saturates when $q(x) \propto p(x)f(x)$*. This shows that if we wish to approximate the average* J*, the optimal choice for the approximating distribution* $q(x)$ *depends on both the distribution* $p(x)$ *and integrand* $f(x)$*.*

2. *Furthermore, show that*

$$J \geq -KL(q(x)|p(x)) - KL(q(x)|f(x)) - H(q(x))$$

where $H(q(x))$ *is the entropy of* $q(x)$*. The first term encourages* q *to be close to* p*. The second encourages* q *to be close to* f*, and the third encourages* q *to be sharply peaked.*

Solution:

1.

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

where, by construction, therefore, J = log Z. Taking KL(r | q) gives the bound

$$\log\ Z = J \geq -\left\langle \log q \right\rangle_q + \left\langle \log p(x) \right\rangle_q + \left\langle \log f(x)) \right\rangle_q = -KL(q||p) + \left\langle \log f(x)) \right\rangle_q$$

2. The result follows simply from adding and subtracting the entropy H(q). The result is particularly interpretable for a function f(x) that is positive and integrates to 1, although the result holds more generally, even when f(x) is not a distribution (on using the standard expression for the KL divergence). Essentially the three requirements translate into q being equal, optimally, to p(x)f(x).