Problem 1

a)

$P(X \leq y)$
$= P(F^{-1}(U) \leq y)$ (NOTE: $X \sim F^{-1}(U)$)
$= P(U \leq F(y))$
$= F(y)$

As shown above, x follows the distribution F. The drawback of this method is that it only works if we know the true distribution of F^{-1}.

b) There are two cases we must show, the cyclical and the mixture. First, let $K(x, z) = (K_1 \circ K_2)(x, z)$. Now, we will show the cyclical kernel has a stationary density:

$\int \int p(x)K_2(x, y)K_1(y, z)dydx$
$= \int p(y)K_1(y, z)dy$ (NOTE: $\int p(x)K_2(x, y)dx = p(y)$)
$= p(z)$

And for the mixture:

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

Despite both of these results being for the continuous case, it should be pretty obvious how they expand to the discrete case, as the integral is just an infinite summation, where the discrete would have a finite summation.

c) The transition probability of MH is as follows:

$p(x \to x') = q(x'|x)A(x', x)$

$A(x, x_t) = min(1, \frac{\tilde{p}(x)q(x_t|x)}{\tilde{p}(x_t)q(x|x_t)})$

So, we have the following:

$p(x)p(x \to x')$
$= p(x)q(x'|x)A(x', x)$
$= min(p(x)q(x'|x), p(x', q(x|x'))$
$= min(p(x')q(x|x'), p(x)q(x', x))$
$= p(x')q(x|x')A(x, x')$
$= p(x')p(x' \to x)$

As shown above, the transition kernel satisfies the detailed balance property.

d)

For notation purposes, let $p(x_{-i})$ be the joint distribution over all variables except for x_i (i.e. $p(x_{-1}) = p(x_2, \ldots, x_d)$.

We know the transition kernel is $K(x, x') = p(x'_1|x_2, \ldots, x_d)p(x'_2|x'_1, x_3, \ldots, x_d) \ldots p(x'_d|x'_1, x'_2, \ldots, x'_{d-1})$. So, we can begin to show p(x) is the stationary distribution of the Markov chain as follows:

$\int K(x, x')p(x)dx$
$= \int p(x'_1|x_2, \ldots, x_d)p(x'_2|x'_1, x_3, \ldots, x_d) \ldots p(x'_d|x'_1, x'_2, \ldots, x'_{d-1})p(x_{-1})p(x_1|x_{-1})dx_1 \cdots dx_d$

We can see $p(x_{-1})p(x_1|x_{-1})dx_1 = p(x)$ and $\int p(x_1|x_{-1})dx_1 = 1$, so we can remove that term, similar to how in variable elimination we could do the same thing (in a conditional distribution, if summing over the input variable, the summation is 1). Additionally, please note we know $p(x'_1|x_2, \ldots, x_d)p(x_{-1}) = p(x'_1, x_2, \ldots, x_d)$. We can proceed as follows:

$= \int p(x'_2|x'_1, x_3, \ldots, x_d) \ldots p(x'_d|x'_1, x'_2, \ldots, x'_{d-1})p(x'_1, x_2, \ldots, x_d)p(x_1|x_{-1})dx_1 \cdots dx_d$
$= \int p(x'_2|x'_1, x_3, \ldots, x_d) \ldots p(x'_d|x'_1, x'_2, \ldots, x'_{d-1})p(x'_1, x_3, \ldots, x_d)p(x_2|x'_1, x_3, \ldots, x_d)dx_2 \cdots dx_d$
$= \int p(x'_3|x'_1, x'_2, x_4, \ldots, x_d) \ldots p(x'_d|x'_1, x'_2, \ldots, x'_{d-1})p(x'_1, x'_2, x_4, \ldots, x_d)p(x_2|x'_1, x_3, \ldots, x_d)dx_2 \cdots dx_d$
$= \int p(x'_3|x'_1, x'_2, x_4, \ldots, x_d) \ldots p(x'_d|x'_1, x'_2, \ldots, x'_{d-1})p(x'_1, x'_2, x_4, \ldots, x_d)p(x_3|x'_1, x'_2, x_4, \ldots, x_d)dx_3 \cdots dx_d$
$\ldots$
$= \int p(x'_1, x'_2, \ldots, x_d)$
$= \int p(x')$

We can see from the above, p(x) is the stationary distribution of the Markov chain.

Problem 2

Let $y_1 = \cos(2\pi x_2)\sqrt{-2\log(x_1)}$.

Let $y_2 = \sin(2\pi x_2)\sqrt{-2\log(x_1)}$.

We execute change of variables as follows:

$$\frac{y_1}{y_2} = \frac{\cos(2\pi x_2)\sqrt{-2\log(x_1)}}{\sin(2\pi x_2)\sqrt{-2\log(x_1)}}$$
$$\implies \frac{y_1}{y_2} = \frac{\cos(2\pi x_2)}{\sin(2\pi x_2)}$$
$$\implies \frac{y_1}{y_2} = \tan(2\pi x_2)$$
$$\implies \arctan(\frac{y_2}{y_1}) = 2\pi x_2$$
$$\implies \frac{1}{2\pi}\arctan(\frac{y_2}{y_1}) = x_2$$

$$y_1^2 + y_2^2 = (\cos(2\pi x_2)\sqrt{-2\log(x_1)})^2 + (\sin(2\pi x_2)\sqrt{-2\log(x_1)})^2$$
$$\implies y_1^2 + y_2^2 = -2\log(x_1)(\cos^2(2\pi x_2) + \sin^2(2\pi x_2))$$
$$\implies y_1^2 + y_2^2 = -2\log(x_1)$$
$$\implies -\frac{1}{2}(y_1^2 + y_2^2) = \log(x_1)$$
$$\implies \exp(-\frac{1}{2}(y_1^2 + y_2^2)) = x_1$$

So, we have found $x_1 = \exp(-\frac{1}{2}(y_1^2 + y_2^2))$ and $x_2 = \frac{1}{2\pi}\arctan(\frac{y_2}{y_1})$.

We can calculate partial derivatives for the jacobian as follows:

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

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

$$\frac{\delta x_2}{\delta y_1} = -y_2 \frac{1}{2\pi(y_1^2+y_2^2)}$$

$$\frac{\delta x_2}{\delta y_2} = y_1 \frac{1}{2\pi(y_1^2+y_2^2)}$$

Then, the jacobian is computed as follows:

$$p(y_1, y_2) \implies J = \left| det \begin{bmatrix} \frac{\delta x_1}{\delta y_1} & \frac{\delta x_1}{\delta y_2} \\ \frac{\delta x_2}{\delta y_1} & \frac{\delta x_2}{\delta y_2} \end{bmatrix} \right|$$
$$= \left| det \begin{bmatrix} -y_1 \exp(-\frac{1}{2}(y_1^2 + y_2^2)) & -y_2 \exp(-\frac{1}{2}(y_1^2 + y_2^2)) \\ -y_2 \frac{1}{2\pi(y_1^2+y_2^2)} & y_1 \frac{1}{2\pi(y_1^2+y_2^2)} \end{bmatrix} \right|$$
$$= \left| -y_1^2 \frac{\exp(-\frac{1}{2}(y_1^2+y_2^2))}{2\pi(y_1^2+y_2^2)} - y_2^2 \frac{\exp(-\frac{1}{2}(y_1^2+y_2^2))}{2\pi(y_1^2+y_2^2)} \right|$$
$$= \left| \frac{\exp(-\frac{1}{2}(y_1^2+y_2^2))}{2\pi(y_1^2+y_2^2)}(-y_1^2 - y_2^2) \right|$$
$$= \left| -\frac{\exp(-\frac{1}{2}(y_1^2+y_2^2))}{2\pi(y_1^2+y_2^2)}(y_1^2 + y_2^2) \right|$$
$$= \left| -\frac{\exp(-\frac{1}{2}(y_1^2+y_2^2))}{2\pi} \right|$$
$$= \frac{\exp(-\frac{1}{2}(y_1^2+y_2^2))}{2\pi}$$

$= \frac{\exp(-\frac{1}{2}y_1^2)\exp(-\frac{1}{2}y_2^2)}{\sqrt{2\pi}\sqrt{2\pi}}$
$= \frac{\exp(-\frac{1}{2}y_1^2)}{\sqrt{2\pi}}\frac{\exp(-\frac{1}{2}y_2^2)}{\sqrt{2\pi}}$
$= N(y_1|0,1)N(y_2|0,1) \quad \square$

The algorithm to sample from a univariate normal distribution is rather straightforward. In the above proof, we showed y_1 and y_2 were both normally distributed variables. We also know x_1 and x_2 are drawn from uniform distributions. Finally, through properties of statistics, since y_1 and y_2 are normally distributed, then $z_1 = y_1\sigma + \mu$ and $z_2 = y_2\sigma + \mu$ are both normally distributed with mean μ and standard deviation σ.

So, the general steps are as follows

1. Sample x_1 from a uniform distribution across all the reals
2. Sample x_2 from a uniform distribution across all the reals
3. Calculate $y = \cos(2\pi x_2)\sqrt{-2\log(x_1)}$ or $y = \sin(2\pi x_2)\sqrt{-2\log(x_1)}$
4. Calculate $z = y\sigma + \mu$
5. Done. z is your random sample from a univariate Normal distribution with mean μ and std σ.

Problem 3

The Ising model can be seen below, both the lattice and "checkerboard" are drawn. The checkerboard, as used as an example in the question, displays white tiles with label w_i and black tiles with label b_j.

We want to show that $p(b_1, b_2, \dots | w_1, w_2, \dots) = p(b_1|w_1, w_2, \dots)p(b_2|w_1, w_2, \dots)\dots$ given that $p(x) \propto \exp(\beta \sum_{i \sim j} \mathbb{1}[x_i = x_j])$ and the model is working on the basis of nearest-neighbor interactions.

Assuming $p(x)$ represents the probability of a node X being in state x with neighbors Y, then because of nearest-neighbour interactions, we have $p(x|Y) = \exp(\beta \sum_{y_j \in Y} \mathbb{1}[x_i = y_j])$.

As such, this general idea can be carried over to the joint probability of the black tiles conditioned on the white tiles. Since each black tile only interacts with its neighbors (i.e. $Y = w_1, w_2, \dots$), which are white, we can easily see none of the black tiles are neighbours of each other. Since none of the black tiles are neighbors of each other, they cannot interact each other (i.e. black tiles do not interact with any other black tiles). Since there are no black-black interactions, we can assume they are independent. Thus, in statistics terms, we know two variables are independent if $P(w, z) = P(w)P(z)$, or in other words, the two variables can be written as pdfs of themselves. A similar situation occurs here. Since for each black tile X, we have Y that consists solely of white tiles, so each black tile can be written as a function of that one black tile along with all the white tiles. Basically, $P(b_1, b_2|w_1, w_2, \dots) = P(b_1|w_1, w_2, \dots)P(b_2|w_1, w_2)$, similar to the example with w and z shown above. This can be similarly expanded to the general case: $p(b_1, b_2, \dots | w_1, w_2, \dots) = p(b_1|w_1, w_2, \dots)p(b_2|w_1, w_2, \dots)\dots$ $\square$

This above idea can be exploited by the Gibbs sampling procedure. In Gibbs sampling, we want to estimate some values $(x_1, x_2, \dots, x_n)$. For each iteration, we select some random value x_i, and we want to make a guess/estimate for that variable's value, given the values of all the other values, or in other words we want to estimate x_i from the distribution $p(x_i|x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = p(x_i|x_{-i})$. Now, want to relate this ideology back to the original nearest-neighbour checkerboard example in this question. We are trying to estimate the distribution $p(b_1, b_2, \dots | w_1, w_2, \dots)$. Since we know the black tiles are independent of each other given the white variables, then assuming we are given values of the white variables, we can easily estimate each black tile's distribution $p(b_i|w_1, w_2, \dots)$ given the remaining black tile variables. In other words, we can estimate $p(b_i|b_1, \dots b_{i-1}, bi + 1, \dots, b_m)$ since we know the distribution of all the other black tiles ($p(b_j) \propto \exp(\beta \sum_{j \sim p} \mathbb{1}[x_j = x_p])$). So, since we know the distribution of all the other black tiles, if we fix them (and assume they are givens), we can estimate the only non-fixed black tile's (b_j) value for that given iteration. So, clearly, since we can see the black variables are independent of each other given the white variables, we can abuse this knowledge to do Gibbs sampling.

Problem 4

Problem 5

$$
\begin{aligned}
& p(d|s, D) \\
&= \tfrac{p(s,D|d)p(d)}{p(s,D)} \\
&= \tfrac{p(s|d)p(D|d)p(d)}{p(s)p(D)} \\
&= \textstyle\int_{W,b,p} \tfrac{p(s,W,b,p|d)p(d)p(W,b,p)p(D|d,W,b,p)}{p(s,W,b,p)} \\
&= \textstyle\int_{W,b,p} p(d|s, W, b, p)p(W, b, p)p(D|d, W, b, p) \\
&= \textstyle\int_{W,b,p} p(d|s, W, b, p)p(W, b, p) \prod_{n=1}^{N} p(s^n|d^n, W, b)p(d^n|p) \\
&= \textstyle\int_{W,b,p} p(d|s, W, b, p)p(W, b, p|D) \qquad \square
\end{aligned}
$$

One way to estimate $p(d_i = 1|s, D)$ using sampling is to just use Gibbs sampling. Since we're given the network, and corresponding likelihood equation, we can proceed as follows:

1. Fix all diseases at 0 or 1 to produce disease vector $\mathbf{d}$
2. At each iteration i, fix all d_{-i} and calculate $p(d_i = 0|s, D)$ and $p(d_i = 1|s, D)$ using the above network.
3. Then, change $\mathbf{d_i}$ to the value with the highest probability (0 or 1).
4. Repeat (go back to step 1) for a certain number of iterations, which is just the burn-in. At this point, the disease vector $\mathbf{d}$ should have converged already.
5. Once finished with the burn-in, iterate over the diseases again and carry out the same calculations as step (2) to find $p(d_i = 1|s, D)$. These are the final disease probabilities.

Problem 6

NOTE: All referenced figures and tables can be found on the next 4 pages.

a) Run "python3 q6.py" to run one simulation of the program. You must install the numpy and matplotlib libraries, which can be done through the following commands: "pip3 install numpy" and "pip3 install matplotlib".

b) Results can be seen in Tables 1, 2, as well as Figures 1, 2. While both values of σ produce the correct estimated means (-5 and 5), their acceptance rates are vastly different. With $\sigma = 0.5$, we get an acceptance rate of 0.1303. Meanwhile, with $\sigma = 5$, we get acceptance rate 0.0025. This makes sense since with a smaller value of σ, the random walk of the MH algorithm (which is produced by the proposal distribution) leads us to making smaller steps. Meanwhile, when the value of σ is larger, those steps are going to be larger at each iteration, which means there is a higher chance of making a step in the wrong direction, thus leading to a larger rejection rate by the algorithm. Basically, smaller σ means smaller, more precise, and subsequently more confident steps.

c) Results can be seen in Table 3 and Figure 3. The estimated mean is roughly where it should be (-5 and 5).

Simulation #	μ_1	μ_2	**Acceptance Rate**
1	-5.0212	5.0119	0.1225
2	-5.1669	5.1004	0.1296
3	-4.8128	4.9267	0.1310
4	-5.0322	5.0308	0.1335
5	-5.2505	4.8600	0.1309
6	-5.0585	4.9253	0.1340
Average	-5.0570	4.9759	0.1303

Table 1: Metropolis-Hastings with $\sigma = 0.5$

Simulation #	μ_1	μ_2	**Acceptance Rate**
1	-4.9132	4.8871	0.0025
2	-4.9984	4.9574	0.0019
3	-5.3520	4.8601	0.0030
4	-5.1066	4.8501	0.0029
5	-5.2658	4.8183	0.0023
6	-5.1072	5.0082	0.0023
Average	-5.1239	4.8969	0.0025

Table 2: Metropolis-Hastings with $\sigma = 5$

Simulation #	μ_1	μ_2
1	-5.0954	5.0613
2	-4.7688	4.8494
3	-4.9848	4.9554
4	-5.2155	4.8305
5	-5.0172	4.8701
6	-4.9543	4.9489
Average	-5.0060	4.9193

Table 3: Gibbs Sampling

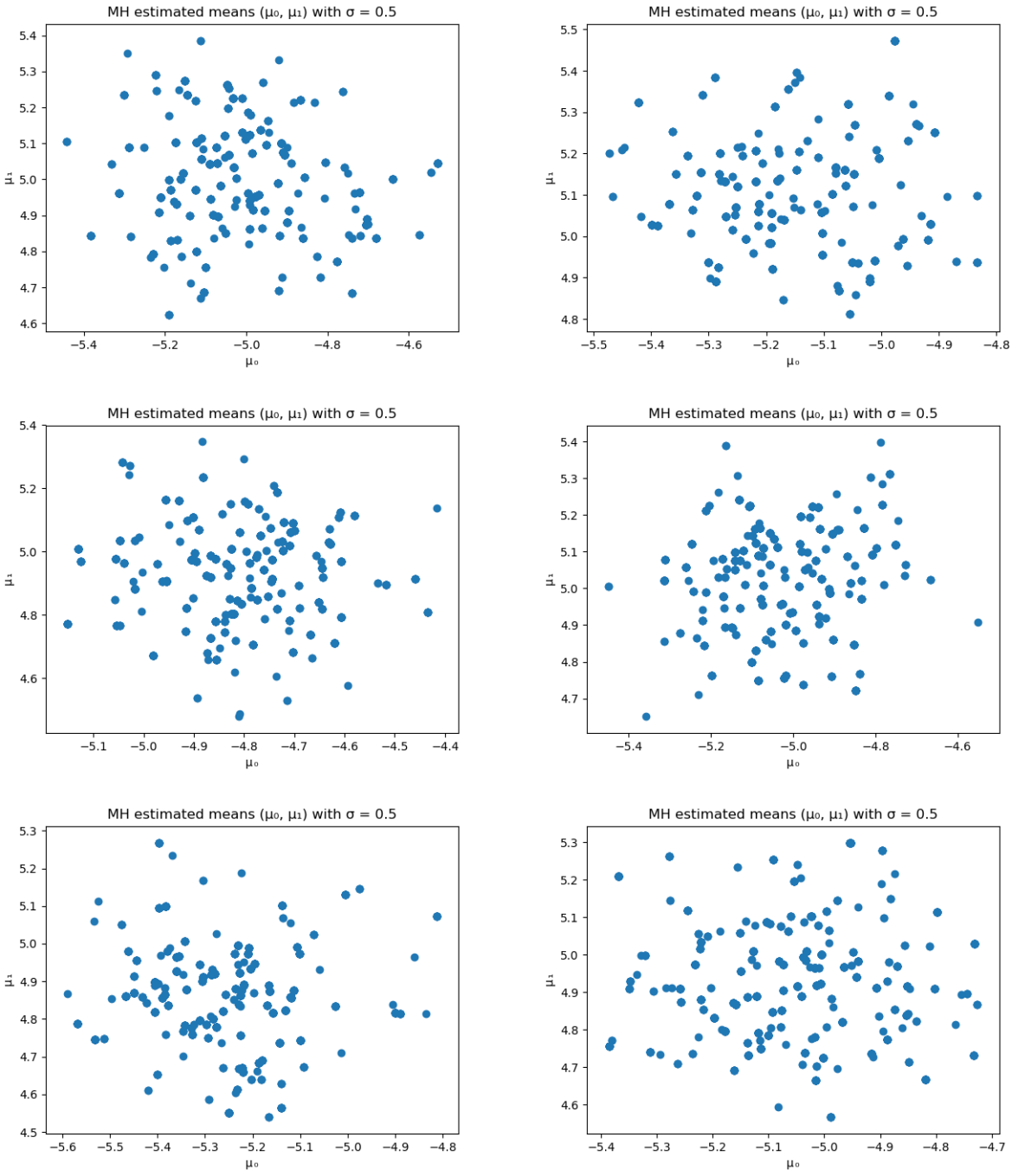

Figure 1: Metropolis-Hastings with $\sigma = 0.5$

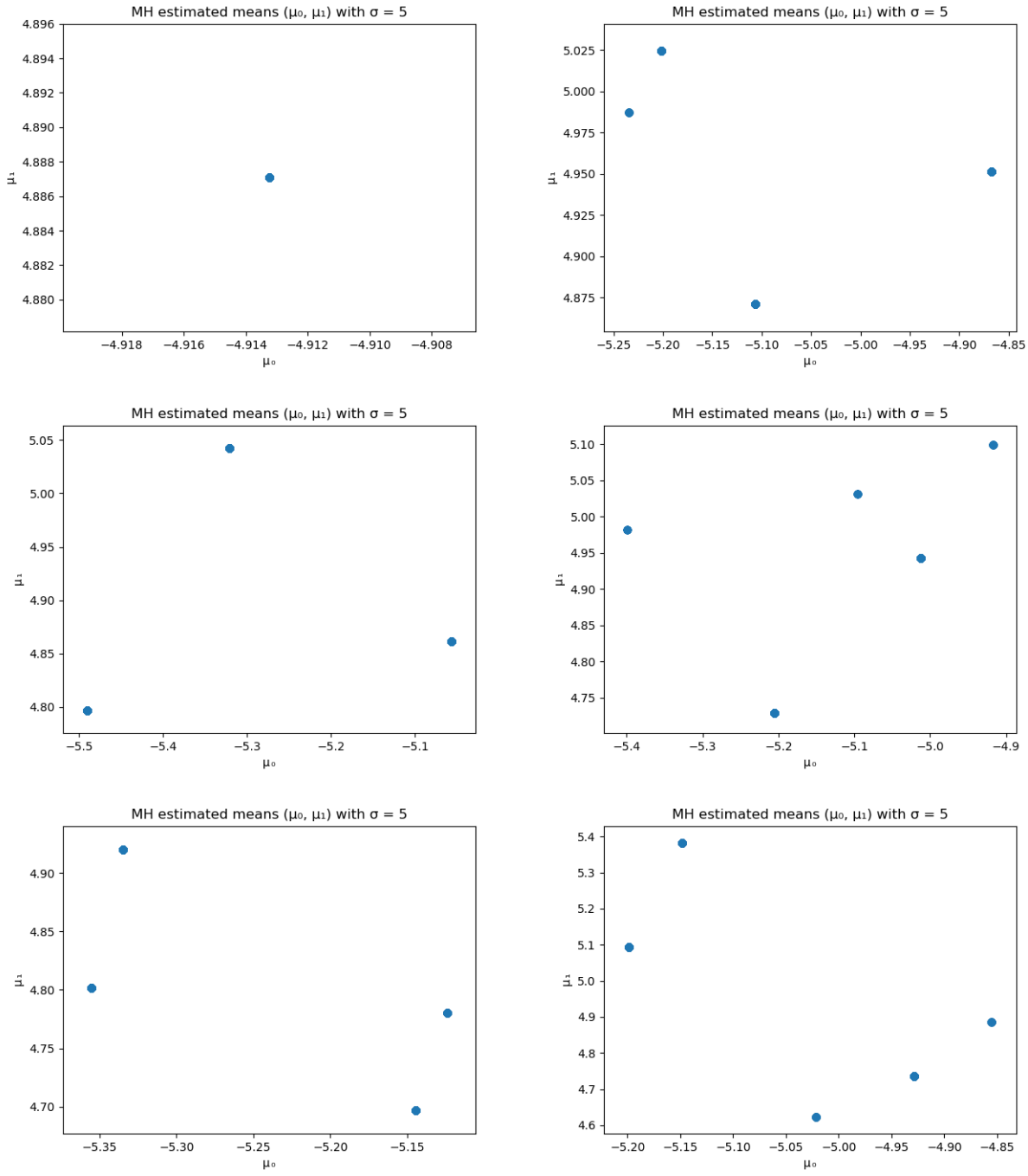

Figure 2: Metropolis-Hastings with $\sigma = 5$

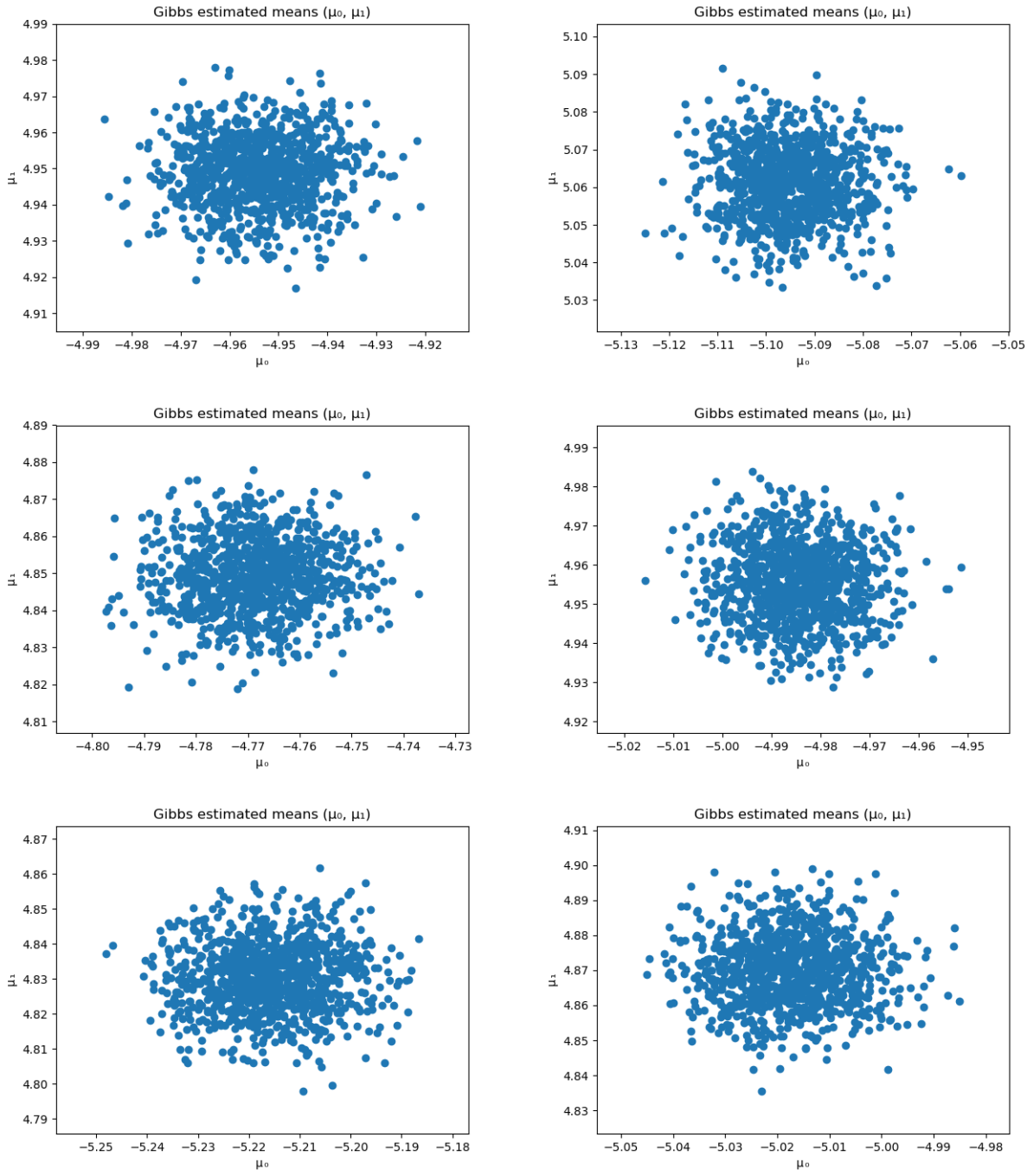

Figure 3: Gibbs Sampling

Problem 7

Problem 8

Problem 9