Student Name: Irene Chang

Collaboration Statement:

Total hours spent: 8 hours

I consulted the following resources:

- Kelsey, Pani
- textbook, lecture slides

Contents

HW-1a Detailed Balance for MH 2
HW-1b Detailed Balance for MH 3
CP-2a: Figure: Metropolis Samples vs. Proposal Standard Deviation 4
CP-2b: Short answer: Which hyperparameter would you recommend and why? 4
CP-2c: Short answer: Can you use a single chain's accept rates to assess convergence? 4
CP-3a: Figure: Gibbs Samples 5
CP-3b: What are the advantages of the Gibbs sampler over the random walk? 5
CP-3c: What are the disadvantages of the Gibbs sampler over the random walk? 6
CP-4a: Figure: MCMC samples vs. Training Set Size 7
CP-4b: Short Answer: Does $N = 1$ plot make sense? 8
CP-4c: Short Answer: What happens as $N \to \infty$? 8

HW-1a Detailed Balance for MH

Show the following for any $x > 0, y > 0$:

$$\frac{\min\left(1, \frac{x}{y}\right)}{\min\left(1, \frac{y}{x}\right)} = \frac{x}{y} \tag{1}$$

Case 1: $0 < x \leq y$: Then $\frac{x}{y} \leq 1, \frac{y}{x} \geq 1$.

Thus,

$$\min\left(1, \frac{x}{y}\right) = \frac{x}{y}, \min\left(1, \frac{y}{x}\right) = 1 \Rightarrow \frac{\min\left(1, \frac{x}{y}\right)}{\min\left(1, \frac{y}{x}\right)} = \frac{\frac{x}{y}}{1} = \frac{x}{y}$$

Case 2: $0 < y \leq x$: Then $\frac{y}{x} \leq 1, \frac{x}{y} \geq 1$.

Thus,

$$\min\left(1, \frac{y}{x}\right) = \frac{y}{x}, \min\left(1, \frac{x}{y}\right) = 1 \Rightarrow \frac{\min\left(1, \frac{x}{y}\right)}{\min\left(1, \frac{y}{x}\right)} = \frac{1}{\frac{y}{x}} = \frac{x}{y}$$

HW-1b Detailed Balance for MH

If $\mathcal{T}$ is the Metropolis-Hastings transition probability distribution, show that it means the detailed balance condition with respect to p^* for any a, b such that $a \neq b$:

$$p^*(a)\mathcal{T}(b|a) = p^*(b)\mathcal{T}(a|b) \tag{2}$$

We have:

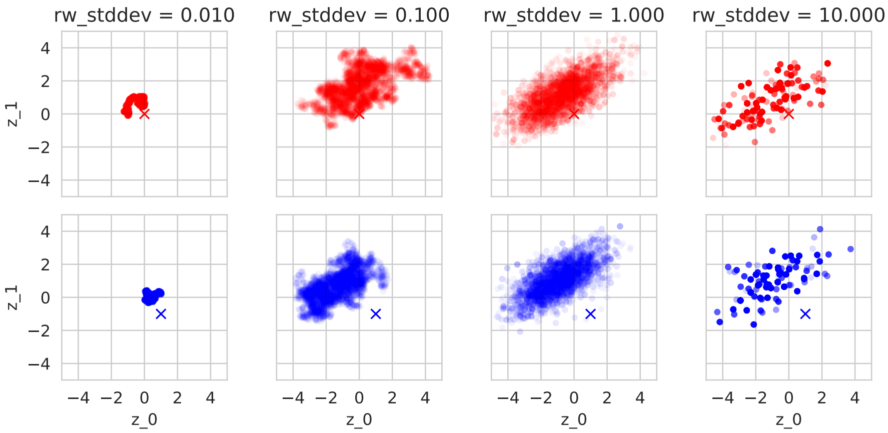
$$\begin{aligned}
\frac{\mathcal{T}(a|b)}{\mathcal{T}(b|a)} &= \frac{\min\left(1, \frac{\tilde{p}(a)Q(b|a)}{\tilde{p}(b)Q(a|b)}\right) Q(a|b)}{\min\left(1, \frac{\tilde{p}(b)Q(a|b)}{\tilde{p}(a)Q(b|a)}\right) Q(b|a)} \text{ (given MH transition)} \\
&= \frac{\tilde{p}(a)Q(b|a)}{\tilde{p}(b)Q(a|b)} \cdot \frac{Q(a|b)}{Q(b|a)} \text{ (based on equation (1) above)} \\
&= \frac{\tilde{p}(a)}{\tilde{p}(b)}
\end{aligned}$$

Since $p^*(z) = c \cdot \tilde{p}(z)$ (we only know how to evaluate the distribution of $p^*(z)$ up to some constant c, and $\tilde{p}(z)$ distribution can be evaluated). Thus:

$$\frac{\mathcal{T}(a|b)}{\mathcal{T}(b|a)} = \frac{\tilde{p}(a)}{\tilde{p}(b)} = \frac{c \cdot p^*(a)}{c \cdot p^*(b)} = \frac{p^*(a)}{p^*(b)}$$

$$\Rightarrow \mathcal{T}(a|b)p^*(b) = p^*(a)\mathcal{T}(b|a)$$

CP-2a: Figure: Metropolis Samples vs. Proposal Standard Deviation



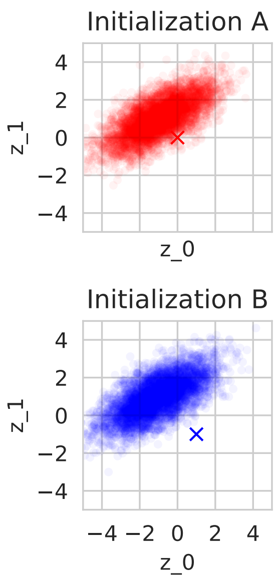
CP-2b: Short answer: Which hyperparameter would you recommend and why?

For both of the cases, I would use standard deviation of 1, since the samples generated concentrate around the true mean of [-1, 1], and they resemble the true variances of the original multivariate distribution: the two variables z0 and z1 has a strong correlation with one another (cov z0, z1 is 0.95); the variance of z1 is 1.0 so most of the data is within 2 standard deviations from the mean, which is from -1 to 3 in z1 direction; similarly, most of the data should be from -3.8 to 1.8 in z0 direction. The graph for std of 10 also shows this variance, but most of the data is spread pretty evenly across the whole range, instead of concentrating near the true mean as much as the case when std is 1.0.

CP-2c: Short answer: Can you use a single chain's accept rates to assess convergence?

The high acceptance rate of 0.8 doesn't imply that the MCMC chain converges. It can be the case that your chain jumps very small steps at a time, which will make the algorithm slow to ”discover” the range of the distribution and hence slow in converging to the stationary distribution.

CP-3a: Figure: Gibbs Samples



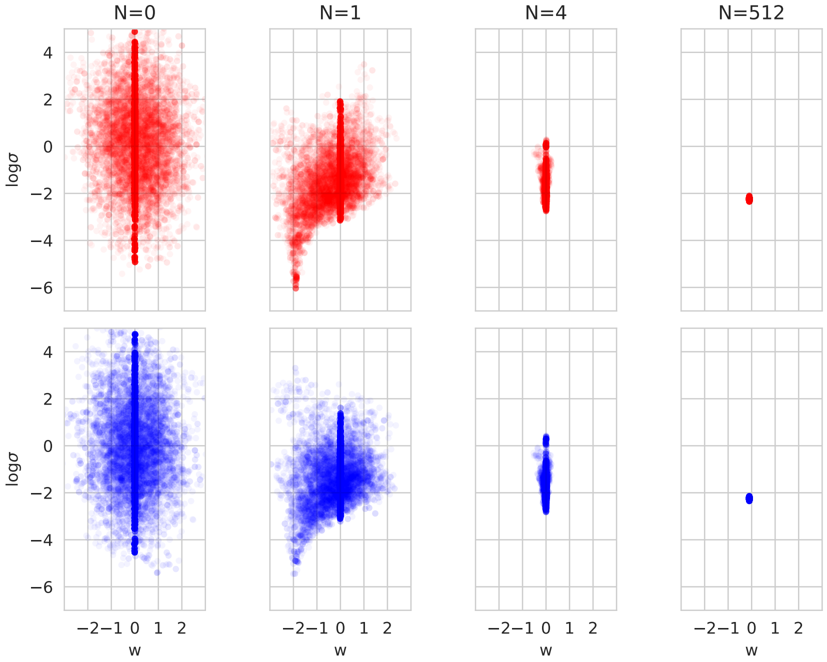
CP-3b: What are the advantages of the Gibbs sampler over the random walk?

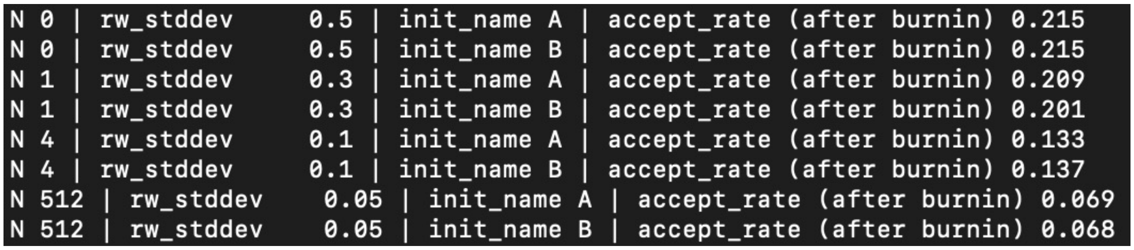
Random walk algorithm is design to give a Markov chain that utilizes proposal-acceptance method to correct the wrong density and arrive at the optimal distribution, but we will still need to sample from a multivariate distribution, which causes the algorithm to suffer from the curse of dimensionality. An advantage of Gibbs sampling is that it breaks the high-dimensionality problem since you've broken down the parameter space into several lower dimensional steps (so that we update one variables at a time which is much easier). In Gibbs sampling, proposals are always accepted, so we are not concerned with choosing a proposal distribution. And it is easier to construct the conditional distributions so that they have closed forms (each step is conjugate), which makes it easier to evaluate, and it's faster to sample from known distributions.

CP-3c: What are the disadvantages of the Gibbs sampler over the random walk?

One disadvantage of Gibbs sampling is that we need to be able to derive the above conditional probability distributions. If there are strong dependencies (correlations) between the variables, it's harder to move around (you can only walk in one direction and can't branch out to other directions to ”explore” the regions beyond what's already known) so it can be slower to converge to an optimal distribution.

CP-4a: Figure: MCMC samples vs. Training Set Size





```
N 0 | rw_stddev     0.5 | init_name A | accept_rate (after burnin) 0.215
N 0 | rw_stddev     0.5 | init_name B | accept_rate (after burnin) 0.215
N 1 | rw_stddev     0.3 | init_name A | accept_rate (after burnin) 0.209
N 1 | rw_stddev     0.3 | init_name B | accept_rate (after burnin) 0.201
N 4 | rw_stddev     0.1 | init_name A | accept_rate (after burnin) 0.133
N 4 | rw_stddev     0.1 | init_name B | accept_rate (after burnin) 0.137
N 512 | rw_stddev    0.05 | init_name A | accept_rate (after burnin) 0.069
N 512 | rw_stddev    0.05 | init_name B | accept_rate (after burnin) 0.068
```

CP-4b: Short Answer: Does $N = 1$ plot make sense?

Yes, $N = 1, x = -2t(x = 0.06, t = -0.12)$, so $w = -2$, we know that the chance of w = -2 is less likely than w=0, yet we observed a $w = -2$ in this case, so from the formula, we can see that $p(w)$ will get very small. Since there's only 1 data point, which makes $t = -2w$ exactly, the standard deviation will be small and so $\log(\sigma)$ will be small. This is reflected on the graph in that there is a trailing tail around $w = -2$, compared to the case when $N = 0$ (when there's no data the parameters just follow the prior distributions). So the distribution is still centered at w = 0 and log σ = 0 as in the priors, but they have more weights in the area of w = -2 and small $\log(\sigma)$. In conclusion, since we only know this one data point, the joint distribution graph is skewed away from the prior distributions and towards the area that will likely give this evidence (this data point) and hence it's shaped like in the graph.

CP-4c: Short Answer: What happens as $N \to \infty$?

The variance of the target distributions decreases as N gets larger since the model is subject to less overfitting with more evidence, and the parameters converge to the true value as variance goes to 0. In this case, for N = 512, we can see that $w \to 0$ and $\log(\sigma) \to -2$. As N increases, MLE also gives us the parameter estimates that are close to the true parameters, so it's in this aspect that they are similar.