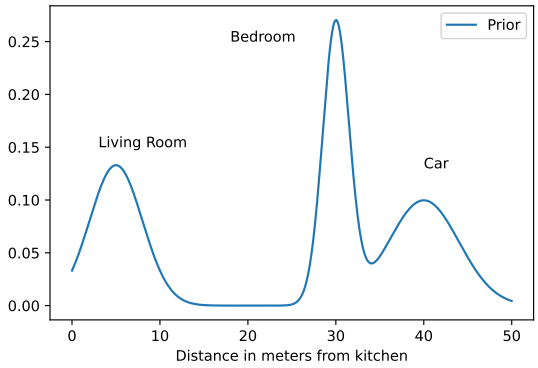
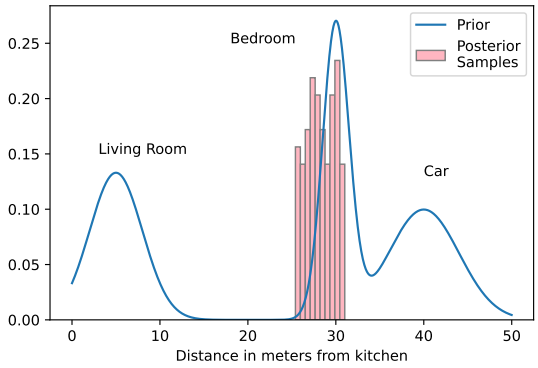
Data 102 (Fall 2022) **Names:**
Discussion Worksheet 5
09/16/2022

---

1. You are searching for your phone and you are not sure where it is. It could be in your bedroom, your living room, or your car. We represent this as a prior over possible locations, which for simplicity we've projected to a 1-dimensional line:

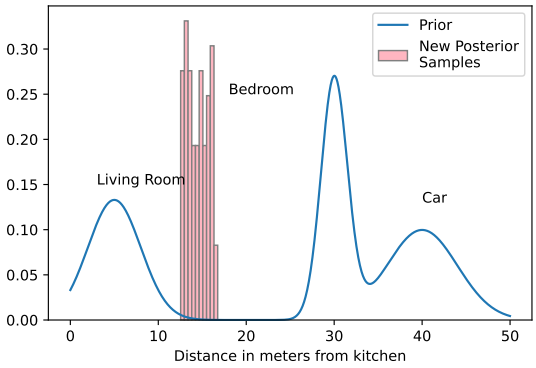


You use the Find Your Phone app, which gives you the GPS location of the phone. The GPS is accurate up to some small error, which we model as a Gaussian with a standard deviation of 2 meters. After observing the GPS signal, you use rejection sampling to sample from your posterior distribution, and observe the following samples:



(a) Approximately where was the GPS signal? Draw it on the x-axis.
Solution: To the left of center of the posterior samples (e.g. 25 meters).

(b) Suppose you had instead seen the following posterior samples:



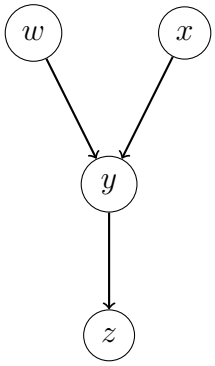
How is this result possible? What does it say about your prior?

Solution: The phone is actually somewhere between the living room and bedroom, so your prior is wrong.

(c) For the posterior from part (b): If you used rejection sampling with the prior as your proposal distribution, would the acceptance rate be high or low? Justify your answer.

Solution: Low. Your prior (proposal) will be very different from the unnormalized posterior (target distribution), so the likelihood of a proposed sample falling under the target curve is low – hence high rejection rate.

2. Which of the following statements are true about this graphical model?



(a) $x \perp\!\!\!\perp w$

(b) $x \perp\!\!\!\perp w|y$

(c) $w \perp\!\!\!\perp z|y$

Solution:
a) True. We can see this algebraically by writing out the joint distribution:

$$p(w, x, y, z) = p(w)p(x)p(y|w, x)p(z|y)$$

Since we're only interested in the relationship between w and x, we need to marginalize over z and y.

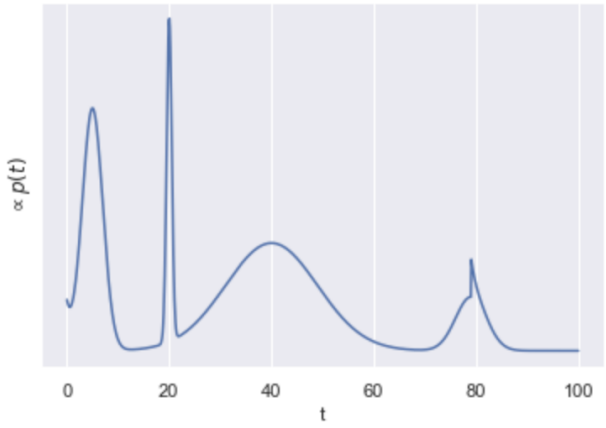
$$\begin{aligned}
p(w, x) &= \sum_y \sum_z [p(w)p(x)p(y|w, x)p(z|y)] && (1)\\
&= p(w)p(x) \sum_y \sum_z [p(y|w, x)p(z|y)] && (2)\\
&= p(w)p(x) \sum_y p(y|w, x) \underbrace{\left[\sum_z p(z|y)\right]}_{=1} && (3)\\
&= p(w)p(x) \underbrace{\sum_y p(y|w, x)}_{=1} && (4)\\
&= p(w)p(x) && (5)
\end{aligned}$$

This is precisely the definition of independence, so we're done. Intuitively, when we marginalize a node in a graphical model, parents of that node usually aren't affected.
b) False. We can follow a similar computation above. From the definition of conditional probability, $p(w, x|y) = \frac{p(w,x,y)}{p(y)}$. Since this is a distribution over w and x, the denominator is a constant.

$$\begin{aligned}
p(w, x|y) &\propto p(w, x, y) && (6)\\
&= \sum_z [p(w)p(x)p(y|w, x)p(z|y)] && (7)\\
&= p(w)p(x)p(y|w, x) \underbrace{\sum_z p(z|y)}_{=1} && (8)
\end{aligned}$$

At this point, we can see that the conditional distribution does *not* factor because of the term $p(y|w, x)$, so we can conclude that conditioned on y, w and x are *not* independent.
c) True. For Bayesian hierarchical models, we know that given a particular node was observed, the node's parents are independent of its children. Therefore, since y was observed, parent node w is independent of child node z.

3. Given the following distribution, rank the Metropolis-Hastings proposal distributions from best to worst. Explain your answers.



(a) A uniform distribution of width 1 centered at the current t (i.e. Uniform$[t-0.5,t+0.5]$)

(b) A shifted exponential distribution starting at the current t with $\lambda = 1$

(c) A normal distribution with mean at the current t and standard deviation 200

(d) A normal distribution with mean at the current t and standard deviation 40

Solution: $d > c > a > b$

- d is the best. It guarantees that we are reasonably likely to generate samples across the support of $p(t)$. The main weakness is that it will take a while for the proposal to find the thin spiky mode at $t = 20$, but other proposal distributions have larger weaknesses.
- b is the worst because it can only move to larger values of t, meaning that we won't ever propose a value smaller than the one we currently have.
- c and a are poor choices. Most of the proposals from 3 will be rejected, and the sequence of samples from using 1 will take a long time to converge, because the proposer will have trouble jumping between different modes of the distribution (in other words, the mixing time of our chain will be very slow if we use 1).

4. In this problem, we'll use a Bayesian hierarchical model to model the number of failures, X_i, for each of n power plant pumps[1]. Consider the following Gamma-Poisson model,

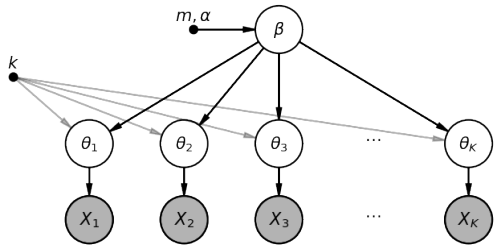
$$\begin{aligned} \beta &\sim \text{Gamma}(m, \alpha) \\ \theta_i \mid \beta &\sim \text{Gamma}(k, \beta), \quad i = 1, \ldots, n \\ X_i \mid \theta_i &\sim \text{Poisson}(\theta_i), \quad i = 1, \ldots, n, \end{aligned}$$

where the θ_i are independent of each other and represent the rate of failures for each power plant pump. The parameters β and θ_i are unknown, and m, α, and k are fixed and known.

We'd like to infer the parameters β and θ_i from the data X. That is, we'd like to sample from the posterior distribution $\mathbb{P}(\beta, \theta \mid X)$. We will do so using Gibbs sampling.

(a) Draw a graphical model that represents the specified Gamma-Poisson model.

Solution: We have the following graphical model:



To run Gibbs sampling, we need to sample each parameter conditional on the current values of the other parameters. We'll do this in the next two steps.

(b) *Conditional Distribution of β.* What is the distribution $\mathbb{P}(\beta \mid \theta_{1:n}, X_{1:n})$? (*Hint:* The answer lies within a common distribution family.)

[1] E I George, U E Makov, and A F M Smith. Conjugate likelihood distributions. Scandinavian Journal of Statistics, 20:147–156, 1993.

Solution:

$$\begin{aligned}
\mathbb{P}(\beta \mid \theta_1, \ldots, \theta_n, X_1, \ldots, X_n) &= \mathbb{P}(\beta \mid \theta_1, \ldots, \theta_n) \\
&= \frac{\mathbb{P}(\beta) \prod_{i=1}^n \mathbb{P}(\theta_i \mid \beta)}{\int_0^\infty \mathbb{P}(b) \prod_{i=1}^n \mathbb{P}(\theta_i \mid b)\, db} \\
&\propto_\beta \beta^{m-1} e^{-\alpha\beta} \prod_{i=1}^n \beta^k e^{-\beta\theta_i} \\
&\propto_\beta \beta^{nk+m-1} e^{-(\alpha+\sum_{i=1}^n \theta_i)\beta} \\
&\propto_\beta \text{Gamma}(nk + m, \alpha + \sum_{i=1}^n \theta_i).
\end{aligned}$$

(c) *Conditional Distribution of θ_i.* What is the conditional distribution $\mathbb{P}(\theta_i \mid \beta, \theta_{1:i-1}, \theta_{i+1:n}, X_{1:n})$? (*Hint:* This also lies within a common distribution family.)

Solution:

$$\begin{aligned}
\mathbb{P}(\theta_i \mid \beta, \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n, X_1, \ldots, X_n) &= \mathbb{P}(\theta_i \mid \beta, X_i) \\
&= \frac{\mathbb{P}(\theta_i, \beta, X_i)}{\mathbb{P}(\beta, X_i)} \\
&= \frac{\mathbb{P}(X_i \mid \theta_i, \beta)\mathbb{P}(\theta_i \mid \beta)\mathbb{P}(\beta)}{\mathbb{P}(\beta, X_i)} \\
&= \frac{\mathbb{P}(\beta)\mathbb{P}(\theta_i \mid \beta)\mathbb{P}(X_i|\beta, \theta_i)}{\mathbb{P}(\beta) \int_0^\infty \mathbb{P}(u|\beta)\mathbb{P}(X_i \mid \beta, u)\, du} \\
&= \frac{\mathbb{P}(\theta_i \mid \beta)\mathbb{P}(X_i|\beta, \theta_i)}{\int_0^\infty \mathbb{P}(u|\beta)\mathbb{P}(X_i \mid \beta, u)\, du} \\
&\propto_{\theta_i} \theta_i^{k-1} e^{-\beta\theta_i} \theta_i^{X_i} e^{-\theta_i} \\
&\propto_{\theta_i} \theta_i^{X_i+k-1} e^{-(\beta+1)\theta_i} \\
&\propto_{\theta_i} \text{Gamma}(X_i + k, \beta + 1).
\end{aligned}$$

(d) Using the results from the last two parts, write out pseudocode for Gibbs sampling from the posterior distribution.

Solution: Initialize $\beta^{(0)} \sim \text{Gamma}(m, \alpha)$ and $\theta_i^{(0)} \mid \beta = \beta^{(0)} \sim \text{Gamma}(k, \beta^{(0)})$ for all i. For $t = 1, \ldots, T$ for some large stopping time T:

(a) Start with $(\beta^{(t-1)}, \theta_1^{(t-1)}, \ldots, \theta_n^{(t-1)})$ from the previous iteration.

(b) Sample $\beta^{(t)}$ according to

$$\beta^{(t)} \sim \mathbb{P}(\beta \mid \theta_1 = \theta_1^{(t-1)}, \ldots, \theta_n = \theta_n^{(t-1)}) = \text{Gamma}(nk + m, \alpha + \sum_{i=1}^n \theta_i^{(t-1)})$$

(c) Sample the $\theta_i^{(t)}$ in parallel according to

$$\theta_i^{(t)} \sim \mathbb{P}(\theta_i \mid \beta = \beta^{(t)}, X_i = x_i) = \text{Gamma}(x_i + k, \beta^{(t)} + 1)$$

Note that we are allowed to sample $\theta_i^{(t)}$ in parallel because each θ_i's updates do not depend on each other.

5. (Challenge Question)

(a) Consider a Markov chain over the integers $\{1, \ldots, n\}$ such that i transitions to $\max(1, i-1)$ or $\min(n, i+1)$ with equal probability. How fast does the mixing time grow as a function of n? (I.e., is it $O(n)$, $O(n^2)$, ...?)

(b) Construct a Markov chain over $\{1, \ldots, n\}$ whose mixing time grows exponentially in n.

Feedback Form

On a scale of 1-5, where 1 = much too slow and 5 = much too fast, how was the pace of the discussion section?

1 2 3 4 5

Which problem(s) did you find most useful?

Which were least useful?

Any other feedback?