Assignment 3: Approximation via Monte Carlo (worth 6 points)
BMI/COMPSCI/PSYCH 841: Computational Cognitive Science
Prof. Austerweil
Due April 9 at 8:59pm

In this problem set, you will explore how to use different Monte Carlo methods.

1. (Monte Carlo). In this problem, you will compare two Monte Carlo methods for estimating the probability that a standard normal distribution has a value greater than two (i.e., $P(Y > 2|Y \sim N(0,1)))$. For all parts of problem 1, we will use $T = 1000$.

   (a) The first method we will use is naïve Monte Carlo. Monte Carlo estimates the expected value of a function of a random variable. The probability a random variable is greater than two can be written as an expectation of a function of a random variable — it is the expected number of times that the random variable is greater than two if we drew a lot of samples from the distribution. So, $P(Y > 2) = E[I(Y > 2)]$, where $I(\cdot)$ is an indicator function, which returns 1 when its argument is true and 0 otherwise.[1]

   Write code that samples T standard normally distributed variables $x_{MC}^{(1)}, \ldots, x_{MC}^{(T)} \sim N(0,1)$ Create a vector which cumulatively estimates the Monte Carlo estimate of the probability of seeing a value greater than two after observing each subsequent data point. In other words, create a vector $\mathbf{f}_{MC}$ such that the value of element t in $\mathbf{f}_{MC}$ is the Markov Carlo estimate of the probability that a standard normal randomly generated value is greater than 2 from t samples (the number of samples in $x_{MC}^{(1)}, \ldots, x_{MC}^{(t)}$ that are greater than 2). Plot $\mathbf{f}_{MC}$ vector. Is there anything strange about it? (Does it look like the example we did in class of the Monte Carlo estimate of the average number of dots seen when you roll a die?) Write 1-2 sentences explaining why the plot of the updated Monte Carlo estimate looks the way it does. Also report the final estimate according to the Monte Carlo estimate ($f_{MC}(T)$).

   (b) The second method we will use is called *importance sampling*. This is another method for estimating the expectation of a function of a random variable. Rather than using the distribution of the random variable itself to generate samples, it uses a different distribution (q – called the *proposal* distribution)[2] and then re-weights each sample by how probable they would have been under the true distribution. So, if $x_{IS}^{(1)}, \ldots, x_{IS}^{(T)}$ are my T samples generated from distribution q rather than p, my estimate would be

   $$\frac{1}{T}\sum_{t=1}^{T}\frac{p\left(x_{IS}^{(t)}\right)}{q\left(x_{IS}^{(t)}\right)} I\left(x_{IS}^{(t)} > 2\right)$$

   Lets apply this to calculating the probability that $X > 2$ given $X \sim N(0,1)$. Because we are interested in the probability $X > 2$, lets use a distribution that is closer to 2: $q = N(2,1)$. So, samples from the proposal distribution come from $N(2,1)$ (a normal distribution with a mean of 2 and variance of 1). This can be drawn from a standard

[1] In your favorite programming language, it is easy to calculate the probability that a standard normal distribution is greater than two. However, I want you to implement these estimators to get a feel for how they behave.

[2] If you are curious why we use q instead of p for the probability under the proposal distribution, at least one reason is to remind us that q is not the actual probability., Rather, it is a surrogate that we are using as part of our approximation method.

normal distribution (N(0,1)) from by adding two to a random variable generated from standard normal distribution.

One way to solve the problem then is to sample each data point from q and then use a function that returns the probability density of some value under a Normal distribution with a specified mean and variance. Alternatively, you could divide p by q explicitly. The explicit solution is

$$\frac{1}{T}\sum_{t=1}^{T} e^{-2x_{IS}^{(t)}+2} I(x_{IS}^{(t)} > 2)$$

Now, make the same plot as in 1a, but now using the importance sampling estimate (i.e., plot a vector $\mathbf{f}^{IS}$ such that f_t^{IS} is the importance sampling estimate of the probability that a standard normal randomly generated variable is greater than 2 using $q \sim N(2,1)$ after t samples). Write 1-2 sentences explaining why the plot of the updated importance sampling estimate looks the way it does. Also, please report the importance sampling estimate given all of the samples.

(c) Write a few sentences comparing Figures 1a and 1b: How and why do they differ? What is a broad lesson about Monte Carlo approximation that we can learn from this comparison?

2. (Markov chain Monte Carlo). For this problem, you will be implementing a hierarchical Bayesian model and using a Markov chain Monte Carlo technique (Gibbs sampling) to approximate its predictions. In particular, you will implement the Beta-binomial model with an exponential prior used in **?**[3] to understand how people learn *overhypotheses* — beliefs about how properties tend to be distributed over different situations. So, the Beta distribution is prominent in their model, which is a probability distribution over values between 0 and 1. The generative model for Beta-Binomial model with an exponential prior is

$$\alpha \sim \text{Exponential}(1) \qquad \beta \sim \text{Beta}(1,1)$$
$$\theta_i|\alpha,\beta \sim \text{Beta}\left(\alpha\beta, \alpha(1-\beta)\right) \qquad y_i|n_i,\theta_i \sim \text{Binomial}(\theta_i; n_i)$$

Remember their cover story. Imagine that you come across a box filled with bags of marbles. Each bag has a lot of marbles. Each marble is either black or white. You get to observe n_i marbles from bag i and write down that y_i of them are white. The proportion of white marbles in bag i is θ_i (and because there are lots and lots of marbles in each bag, we can safely assume that y_i is Binomial distributed with parameters θ_i and n_i).

(a) First, you will estimate the posterior distribution of a simplified version of the model: The posterior probability of the proportion of white marbles after seeing some marbles from a single bag without the hierarchical priors. So, the generative process for this part is: $\theta_i|\alpha,\beta \sim \text{Beta}\left(\alpha\beta, \alpha(1-\beta)\right)$ and $y_i|n_i,\theta_i \sim \text{Binomial}(\theta_i; n_i)$. As you can look up on the Conjugate Prior page of Wikipedia[4], the posterior distribution for a Beta-Binomial model is $\text{Beta}\left(\alpha\beta + y_i, \alpha(1-\beta) + (n_i - y_i)\right)$. This is the equation you need to complete this task. For each α and β value, please graph the prior distribution, (i.e. , no marbles observed) the posterior after observing one white marble (and zero black marbles), the

[3] Note that **?** used the more general Dirichlet-multinomial model, which is essentially the same as a Beta-binomial model, but generalized to more than two choices.

[4] Note that the notation here follows the notation used by **?**. Wikipedia uses a different notation, which is actually the more standard one in statistics.

posterior distribution after observing 5 white and 5 black marbles, and the posterior distribution after observing 9 white and 1 black marble all on the same plot.

i. $\alpha = 0.1$ and $\beta = 0.5$ ($a = 0.05, b = 0.05$)
ii. $\alpha = 0.1$ and $\beta = 0.9$ ($a = 0.09, b = 0.01$)
iii. $\alpha = 1$ and $\beta = 0.5$ ($a = 0.5, b = 0.5$)
iv. $\alpha = 1$ and $\beta = 0.9$ ($a = 0.9, b = 0.1$)
v. $\alpha = 5$ and $\beta = 0.5$ ($a = 2.5, b = 2.5$)
vi. $\alpha = 5$ and $\beta = 0.9$ ($a = 4.5, b = 0.5$)

Although I expect you to execute and think about each of these, you are only required to turn in one of the graphs. Please turn in make sure that the graph is visible and each distribution is distinct and clearly labelled. Please write a paragraph describing how changing α and β affect the posterior update.

(b) Now you will implement samplers to estimate the posterior distribution on α, β, and $\theta_1, \ldots \theta_M$ after observing some marbles from M bags (i.e., given $y_1, \ldots, y_M, n_1, \ldots, n_M$) and the predictive probability for the probability of white marbles in new bags after observing some marbles from some bags (i.e., given $y_1, \ldots, y_M, n_1, \ldots, n_M$). We will be approximating the posterior distribution of $\theta_1 \ldots, \theta_M$ using Gibbs sampling steps and interleaving these steps with Metropolis-Hastings Markov chain Monte Carlo (MH MCMC) sampling for α and β.

Here is an overview of the process: First, you will initialize α and β by sampling them from their priors (Exponential(1) and Beta(1,1), respectively). Then you will use α and β to initialize θ for each of the M bags. Then, you will sample new proposal values for α and β. Then, (via an equation provided in the detailed procedure below) you will use the M θ values you just sampled and your proposal α and β values to determine whether to accept the proposed α and β, or to keep the previous values. Once that is done, the process repeats.

To implement the Gibbs sampling portion, we iterate through the bags and for each bag, we resample the current value for the proportion of white marbles in bag i ($\theta_i^{(t)}$) given the value of all other variables (so, given $\theta_1^{(t-1)}, \ldots, \theta_{i-1}^{(t-1)}, \ldots, \theta_{i+1}^{(t-1)}, \ldots, \theta_M^{(t-1)}, \alpha^{(t)}, \beta^{(t)}, \mathbf{y}, \mathbf{n}$). Given α and β all of the bags are independent of each other (if you're not sure, use the conditional independence rules for graphical models we discussed in class), and so, we sample $\theta_i^{(t)} | \alpha^{(t)}, \beta^{(t)}, y_i, n_i$. Fortunately, we already know this from the last subproblem! It's a Beta-Binomial model and so

$$\theta_i^{(t)} | \alpha^{(t)}, \beta^{(t)}, y_i, n_i \sim \text{Beta}\left(\alpha^{(t)}\beta^{(t)} + y_i, \alpha^{(t)}(1 - \beta^{(t)}) + (n_i - y_i)\right)$$

The samplers for α is a bit more tricky because an explicit solution for the posterior distribution of $\alpha | \theta_1, \ldots, \theta_M$ is unknown. To implement the MH MCMC, we successively generate new proposed values for α and β given its previous value. For α, we will use the Student's t distribution with one-degree of freedom and a mean centered at the current sample for α, $\alpha^{(t)}$ (under the constraint that $\alpha^{(t+1)} > 0$).[5] We do so that we normally

[5] Student was the pseudonym for Gosset, a statistician who developed statistical tests while for the purpose of ensuring the quality of Guiness beer. Check out the Wikipedia page (https://en.wikipedia.org/wiki/Student%27s_t-distribution) on the Student's t distribution for lore on why Gosset wrote under this pseudonym. Also, Student's t distribution and its generalization (the Cauchy distribution) have extremely bizarre properties: e.g., it has infinite mean and variance when the degrees of freedom is 1 — meaning that a single sample is just as good of an estimator of the mean as the empirical average!

propose to change α to values near the current one, but also occasionally propose very different values, which can help the Markov chain "mix" better. Student's t probability distribution with one degree of freedom is

$$f(x) = \frac{1}{\pi(1+x^2)}$$

So, our proposal distribution $q(\alpha'|\alpha^{(t)}) = (\alpha^{(t)} + X^{(t)})I(\alpha^{(t)} + X^{(t)} > 0)$, where $X \sim$ Student$'$s $t(1)$ (Note the $'$ is used to denote the proposed new value of the corresponding variable – in this case α). Choosing a proposal distribution for β is simpler. It will be a Beta distribution whose mean is β_t and has a bit of variance. So, $q(\beta'|\beta^{(t)})$ is Beta$(1+\beta^{(t)}, 2-\beta^{(t)})$.

To implement MH MCMC, we first initialize the Markov chain by a random sample from the prior ($\alpha^{(1)} \sim$ Exponential(1) and $\beta^{(1)} \sim$ Beta(1, 1)) and then for $t = 1, \ldots, T$ follow the following steps:

I. Do a "sweep" (resample each $\theta_i^{(t)}$ once) of Gibbs sampling[6]. So, resample each θ_i: $\theta_i^{(t)}|\alpha^{(t)}, \beta^{(t)}, y_i, n_i \sim \text{Beta}\left(\alpha^{(t)}\beta^{(t)} + y_i, \alpha^{(t)}(1-\beta^{(t)}) + (n_i - y_i)\right)$

II. Generate a proposed next state for the sampler using the current state. To do so, sample α' by adding $X \sim$ Student$'$s $t(1)$ to $\alpha^{(t)}$ and sample $\beta'|\beta^{(t)} \sim$ Beta$(1+\beta^{(t)}, 2-\beta^{(t)})$.

III. Calculate the probability c that we accept α' and β' as the next state. Because the proposal distribution for α is symmetric, we do not need to include it in our acceptance term. However, the proposal distribution for β is not symmetric, and so the MH requires a correction term to ensure that the probability that we move back and forth between states. So, the acceptance probability is the product of two probabilities: (1) the ratio of the probability of transitioning from the proposal back to the current state to the probability of the reverse transition and (2) the ratio of the posterior probability of the proposal to the posterior probability of the previous sample. Don't worry. I've done the derivation for you (it is straightforward, so it's a good exercise to do if you are curious!).It is

$$\begin{aligned} c &= \frac{q(\beta^{(t)}|\beta')p(\alpha', \beta'|\theta_1^{(t)}, \ldots, \theta_M^{(t)})}{q(\beta'|\beta^{(t)})p(\alpha^{(t)}, \beta^{(t)}|\theta_1^{(t)}, \ldots, \theta_M^{(t)})} \\ &= \frac{\texttt{betapdf}(\beta^{(t)}; 1+\beta', 2-\beta')p(\alpha')p(\beta')\prod_i p\left(\theta_i^{(t)}|\alpha', \beta'\right)}{\texttt{betapdf}(\beta'; 1+\beta^{(t)}, 2-\beta^{(t)})p(\alpha^{(t)})p(\beta^{(t)})\prod_i p\left(\theta_i^{(t)}|\alpha^{(t)}, \beta^{(t)}\right)} \end{aligned}$$

where `betapdf(x;y,z)` is the amount of density at x given by a Beta function with parameters y and z.

IV. Sample $Y \sim U(0,1)$. If $Y \leq c$, $\left(\alpha^{(t+1)}, \beta^{(t+1)}\right) = (\alpha', \beta')$. Otherwise, reject and the next sample is the same as t: $\left(\alpha^{(t+1)}, \beta^{(t+1)}\right) = \left(\alpha^{(t)}, \beta^{(t)}\right)$

Now we use the sampler to explore the model's behavior given different sets of marble bags. Please use the following sets of marble bags:

I. 10 bags where each bag has 9 white and 11 black marbles.

[6] You should do this in a random order each sweep. Please generate a random permutation of M elements and use the order provided by it for sampling your variables.

II. 5 bags with 1 white and 19 black marbles, and 5 bags with 19 white and 1 black marble.

For each of these, please run the sampler for 3000 sweeps ($T = 3000$). Please include (two separate) histograms of the α samples and β samples for both I. and II. Also, calculate the average α and β value, and the expected value for a new bag of marbles (θ_{M+1}) after observing the bags of marbles for each (note that this is $\frac{\bar{\alpha}\bar{\beta}}{\bar{\alpha}\bar{\beta}+\bar{\alpha}(1-\bar{\beta})}$, where $\bar{\cdot}$ is the average of that variable).

Please write two paragraphs. In the first paragraph, describe the histograms for α and β given each set of marble bags and explain what their values mean. In the second paragraph, compare the histograms and the values you were asked to calculate. What is equal and unequal? Explain why.

Variable/Equation	Meaning	Matlab command
M	Number of bags of marbles	NA
y_i	Number of white marbles in bag i	NA
n_i	Total number of marbles in bag i	NA
θ_i	Percentage of white marbles in bag i	NA
$\mathbf{y}$	A vector containing the observed number of white marbles in each bag. $\mathbf{y} = (y_1, \ldots, y_M$	NA
$\mathbf{n}$	A vector containing the number of observed marbles from each bag $\mathbf{n} = (n_1, \ldots, n_M)$	NA
$y_i \mid n_i, \theta_i \sim \text{Binomial}(\theta_i; n_i)$	Draw n_i marbles from bag i. This says the probability of observing y_i white ones is Binomial distributed with parameters θ_i and n_i	`binopdf`(y_i, n_i, θ_i)
α	Parameter in the Kemp parameterization of the Beta distribution. It controls the distribution's shape. Roughly, $\alpha < 1$ will put most of its probability mass at 0 or 1 and $\alpha > 1$ will have a more Gaussian-like shape.	NA
Beta distribution	This is a probability distribution over numbers between 0 and 1. As probabilities are between 0 and 1 and it is conjugate to the Binomial distribution, it is commonly used as a prior distribution over probabilities.	NA
β	Parameter in the Kemp parameterization of the Beta distribution. It encodes the prior belief of the proportion of white to black balls in a bag of marbles. $0 \leq \beta \leq 1$.	NA
$\theta_i \mid \alpha, \beta \sim \text{Beta}\left(\alpha\beta, \alpha(1-\beta)\right)$	θ_i (the percentage of white marbles in bag i) is Beta distributed with parameters $a = \alpha\beta$ and $b = \alpha(1-\beta)$.	`betapdf`(θ_i, a, b); `betarnd(a,b)`
$\text{Beta}\left(\alpha\beta + y_i, \alpha(1-\beta) + (n_i - y_i)\right)$	The posterior distribution in a Beta-Binomial model. It is Beta distributed with $a = \alpha\beta + y_i$ and $b = \alpha(1-\beta) + (n_i - y_i)$.	`betapdf`(θ_i, a, b)
$\theta_i^{(t)}$	The t-th sample of the percentage of white marbles in bag i	NA
$X \sim \text{Student}'\text{s}t(1)$	X is the change to α that we propose in each MH step. The Student t distribution is like the Normal distribution, but has heavier tails, meaning that it is more likely to produce a large deviation than a Normal distribution would be.	`trnd`(1)
$q(\alpha' \mid \alpha^{(t)})$	q is the proposal distribution for α. It is q instead of p to remind you that it is not the real distribution on α.	NA
$q(\beta' \mid \beta^{(t)})$	q is the proposal distribution for β. It is q instead of p to remind you that it is not the real distribution on β.	NA

Table 1: The different symbols and variables in Problem 2, their meaning, and Matlab commands for calculating them (when they are distributions).