### Overview

Submit your writeup, including any code, as a PDF via gradescope.[1] We recommend reading through the entire homework beforehand and carefully using functions for testing procedures, plotting, and running experiments. Taking the time to reuse code will help in the long run!

Data science is a collaborative activity. While you may talk with others about the homework, please write up your solutions individually. If you discuss the homework with your peers, please include their names on your submission. Please make sure any handwritten answers are legible, as we may deduct points otherwise.

## The One with all the Beetles

1. (9 points) Cindy has an inordinate fondness for beetles and for statistical modeling. She observes one beetle everyday and keeps track of their lengths. From her studies she feels that the beetle lengths she sees are uniformly distributed. So she chooses a model that the lengths of the beetles come from a uniform distribution on $[0, w]$: here $w$ is an unknown parameter corresponding to the size of the largest possible beetle. Since the maximum size $w$ is unknown to her, she would like to estimate it from the data. She observes lengths of $n$ beetles, and calls them $x_1, \ldots, x_n$.

   (a) (1 point) What is the likelihood function of the observations $x_1, \ldots, x_n$? Express your answer as a function of the largest size parameter $w$.

   **Solution:** The liklihood function for each sample is that of the uniform distribution on $[0, w]$. This is given by

   $$f_1(x; w) = \frac{1}{w} \mathbb{1}\left[x \leq w\right] \tag{1}$$

   for $x \geq 0$. Here the subscript indicates the number of samples. Since all the samples are independent, we get

   $$f_n(x_1, \ldots, x_n; w) = \frac{1}{w^n} \prod_i \mathbb{1}\left[x_i \leq w\right] \tag{2}$$

   $$= \frac{1}{w^n} \mathbb{1}\left[\max_i x_i \leq w\right]. \tag{3}$$

   The last equality follows by noting that the each of the $x_i$ is less than $w$ if and only if the maximum is less than $w$. This makes intuitive sense since the uniform distribution puts no probability on points outside its domain.

   *Hint: Your answer should include the indicator function* $\mathbb{1}(\max_i x_i \leq w)$. *To see why, consider what happens if* $w = 3$ *cm and* $x_1 = 5$ *cm.*

   (b) (2 points) Use your answer from part (a) to explain why the maximum likelihood estimate for $w$ is $\max_i x_i$.

[1] In Jupyter, you can download as PDF or print to save as PDF

**Solution:** We know from our earlier calculation that the liklihood is given by

$$f_n(x_1, \ldots, x_n; w) = \frac{1}{w^n} \mathbb{1}\left[\max_i x_i \leq w\right] \tag{4}$$

Now we consider this as a function of $w$ for fixed $x_1 \ldots x_n$. First note that this is zero for $w \leq \max_i x_i$. Next for $w \geq \max_i x_i$ this is a strictly decreasing function in $w$. Thus, the maximum is achieved at $w = \max_i x_i$.

(c) (2 points) Cindy decides to instead use a Bayesian approach. She has a prior belief that $w$ follows a *Pareto distribution* with parameters $\alpha, \beta > 0$. We can write:

$$w \sim \text{Pareto}(\alpha, \beta)$$
$$p(w) = \frac{\alpha \beta^\alpha}{w^{\alpha+1}} \mathbb{1}(w > \beta)$$

If she observed values $x_1, ...., x_n$, show that the posterior distribution for $w$ is also a Pareto distribution, and compute the parameters as a function of $\alpha$, $\beta$, and the observations $x_1, \ldots, x_n$.

**Solution:** Let us denote the prior distribution as $p$ and the posterior as $p_1$. Recall from Bayes rule, we have the the posterior is given by

$$p_1\left(w | x_1, \ldots, x_n\right) = \frac{f_n(x_1, \ldots, x_n; w) p(w)}{\int f_n(x_1, \ldots, x_n; t) p(t) dt}$$

Here the demoninator uses the law of total probability. Let us look at the numerator and denominator separately.

$$\begin{aligned}
f_n\left(x_1, \ldots x; w\right) p\left(w\right) &= \alpha \beta^\alpha w^{-\alpha-1} w^{-n} \mathbb{1}\left[\max_i x_i \leq w\right] \mathbb{1}\left[w > \beta\right] \\
&= \alpha \beta^\alpha w^{-\alpha-n-1} \mathbb{1}\left[\max_i x_i \leq w\right] \mathbb{1}\left[w > \beta\right] \\
&= \alpha \beta^\alpha w^{-\alpha-n-1} \mathbb{1}\left[\max_i x_i \leq w \wedge w > \beta\right] \\
&= \alpha \beta^\alpha w^{-\alpha-n-1} \mathbb{1}\left[w > \max\{\beta, \max_i x_i\}\right].
\end{aligned}$$

Let us denote $\max\{\beta, \max_i x_i\}$ as $\beta'$. Now looking at the denominator, we get

$$\begin{aligned}
\int_0^\infty f_n(x_1, \ldots, x_n; t) p(t) dt &= \int_0^\infty \alpha \beta^\alpha t^{-\alpha-n-1} \mathbb{1}\left[t > \beta'\right] dt \\
&= \alpha \beta^\alpha \int_{\beta'}^\infty t^{-\alpha-n-1} dt \\
&= \alpha \beta^\alpha \left[\frac{t^{-\alpha-n}}{-\alpha-n}\right]_{\beta'}^\infty \\
&= \alpha \beta^\alpha \frac{\beta'^{-\alpha-n}}{\alpha+n}
\end{aligned}$$

Let $\alpha' = \alpha + n$. Then, taking the ratio, we get

$$p_1(w|x_1, \ldots, x_n) = \frac{\alpha' \beta'^{\alpha'}}{w^{\alpha'+1}} \mathbb{1}\left[w > \beta'\right]$$

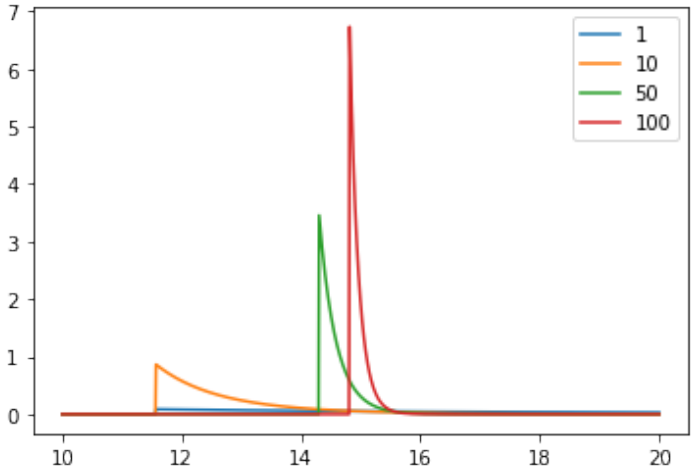
(d) (2 points) Provide a short description in plain English that explains what the parameters of the Pareto distribution mean, in the context of the Pareto-uniform conjugate pair.

*Hint: For the Beta-Binomial conjugate pair that we explored in class, the answer would be that the Beta parameters acted as pseudo-counts of observed positive and negative examples.*

**Solution:** From the previous part, we saw that $\beta' = \max\{\beta, \max_i x_i\}$ and $\alpha' = \alpha + n$. This indicates that the parameters $\alpha$ and $\beta$ keep track of the number of samples seen and maximum of the samples seen so far respectively.

(e) (2 points) Let us say Cindy started with the initial prior values $\alpha = 1$ and $\beta = 10$ on day 0. Use the code in beetledata.py to generate the data for the lengths of the beetles she sees, starting with day one to day one hundred. Use the data to make a graph of one curve for each of the days $1, 10, 50$ and $100$ (so four curves total), where each curve is the probability density function of Cindy's posterior for the respective day. As with other familiar distributions, you can find the Pareto distribution in scipy.

**Solution:**




Note that as the days progress the plot shift to the right as we see larger and larger maximum values. Also our confidence our estimate increases which can be seen by the strength of the peaks.

(f) (0 points) (Optional) Use pymc3 to sample from the posterior for days $1, 10, 50$ and $100$ and plot a density function for each of the cases. Compare the results from the

analytic and simulation based computation of the densities.

## Bayesian Fidget Spinners

2. (10 points) Nat's company manufactures fidget spinners. The company uses two factories, which we'll call factory 0 and factory 1. Each fidget spinner from factory $k$ is defective with probability $q_k$ ($k \in \{0, 1\}$). Nat knows that factory 0 produces fewer defective fidget spinners than factory 1 (in other words, $q_0 < q_1$).

   She receives $n$ boxes full of fidget spinners, but the boxes aren't labeled (in other words, she doesn't know which box is from which factory). For each box, she starts randomly pulling out fidget spinners until she finds a defective one, and records how many fidget spinners she pulled out (including the defective one). She calls this number $x_i$ for box $i$, for $i = 1, \ldots, n$.

   She wants to estimate the following pieces of information:

   - Which boxes came from factory 0, and which came from factory 1? She defines a binary random variable for each box $z_i$ with the factory label (i.e., $z_i = 0$ if box $i$ is from factory 0, and $z_i = 1$ if box $i$ is from factory 1).
   - How reliable is each factory? In other words, what are $q_0$ and $q_1$?

   Inspired by what she learned about Gaussian mixture models, she sets up the following probability model:

$$\begin{aligned} z_i &\sim \text{Bernoulli}(\pi) \\ q_k &\sim \text{Beta}(\alpha_k, \beta_k) \\ x_i | z_i, q_0, q_1 &\sim \text{Geometric}(q_{z_i}) \end{aligned}$$

   (a) (1 point) Draw a graphical model for the probability model described above if $n = 2$ (i.e., there are only two boxes of fidget spinners).

   Nat decides to implement the model above setting the following hyperparameters:

$$\pi = 0.45, \quad q_0 \sim \text{Beta}(1, 5), \quad q_1 \sim \text{Beta}(5, 1)$$

   (b) (2 points) The choices of the parameters in Nat's model represents her beliefs about the factories.

      i. (1 point) From her choice for $\pi$, what can we infer about her beliefs about the two factories?

      **Solution:** Note that Nat's choice of $\pi < 1/2$. So we can infer that she thinks that factory 0 produces more boxes of fidget spinners.

ii. (1 point) Similarly, from her choices for $\alpha$ and $\beta$, what can we infer about her beliefs about the factories?

> **Solution:** Since Beta$(1, 5)$ is more concentrated on smaller values than Beta$(5, 1)$ one can infer that Nat believes that factory 0 has lower rate of defects than factory 1.

(c) (5 points) Use `fidget_data.py` to generate the data that Nat observes, then, using PyMC3, fit the model outlined above, setting the hyperparameters to the values that Nat chose. Obtain 1000 samples from the posterior distribution $p(q_0, q_1|x_1, \ldots, x_n)$, and generate a scatterplot (one point per sample).

You can use the code below (also provided to you in `fidget_model.py`) to help you get started.

```
import pymc3 as pm


alphas = ...
betas = ...
pi = ...

with pm.Model() as model:
    z = pm.Bernoulli(
        # Define the Bernoulli Model Here
    )

    # Hint: you should use the shape= parameter here so that
    # q is a PyMC3 array with both q0 and q1.
    q = ...

    # Hint: it may be useful to use "fancy indexing" like we did in
    class.
    # See below for an example
    X = pm.Geometric(
        # DEFINE THE GEOMETRIC MODEL HERE
    )

    trace = ....

# FANCY INDEXING

my_binary_array = np.array([0, 0, 1, 1, 0, 1])
my_real_array = np.array([0.27, 0.34])
print(my_real_array[my_binary_array])
```

i. Under the posterior, what is the probability that factory 0 produces more boxes than factory 1?

> **Solution:** Under the posterior, the value for $\pi$ is approximately 0.45 (ie the model should tend toward estimating the ground truth of the simulated data). Thus, a probability that a box is from factory 0 is approximately 0.55 and that from factory 1 is approximately 0.45. Thus that probability that

factory 0 produces more boxes is

$$\Pr\left[\text{Binom}(50, 0.45) \leq 50\right].$$

One can explitly compute this as

$$\sum_{i=0}^{50} \binom{100}{i} (0.45)^i (0.55)^{100-i} = 0.8654$$

An acceptable way for this problem set to compute this is take samples from the posterior. Count the number of number of instances in which $\frac{1}{m}\sum_i \mathbb{1}[\sum_i z_i \leq 50]$ where $m$ is the sample size and output this as the empirical estimate for the probability. Note that the model-fitting isn't super stable over various runs and this empirical estimate can deviate substantially from the analytical estimate.

ii. What is your median estimate of factory 0's defect rate, based on the samples from the posterior?

**Solution:** This would be a number close to 0.05.

(d) (2 points) Nat's friend Yaro suggests using Gibbs sampling. What is the Gibbs sampling update for $q_k$? Your answer should be in the form of a well-known distribution, along with values for the parameter(s) of that distribution. Justify your answer.

*Hint*: you can derive the update analytically, or you can use the fact that the Beta distribution is a conjugate prior for a Geometric likelihood.

**Solution:**

$$q_0 \sim \text{Beta}(\alpha_0 + n_0, \beta_0 - n_0 + \sum_{i:z_i=0} x_i)$$

$$q_1 \sim \text{Beta}(\alpha_1 + n_1, \beta_1 - n_1 + \sum_{i:z_i=1} x_i)$$

where $n_0 = \sum_{i=1}^n 1_{z_i=0}$, and $n_1 = \sum_{i=1}^n 1_{z_i=1}$.

This follows by assuming that the likelihood for a generic $x$, coming from firm $k \in \{0, 1\}$, is $(1 - q_k)^{x-1} q_k$ (indeed you get $x_i = x$ if you sample $k - 1$ non-defective fidget spinners and a defective one).

*Note*: when grading, we will also accept the less precise solution where you assume that the likelihood is $(1-q_k)^x q_k$, and you get the following Gibbs sampling updates

$$q_0 \sim \text{Beta}(\alpha_0 + n_0, \beta_0 + \sum_{i:z_i=0} x_i)$$

$$q_1 \sim \text{Beta}(\alpha_1 + n_1, \beta_1 + \sum_{i:z_i=1} x_i).$$

## Rejection Sampling

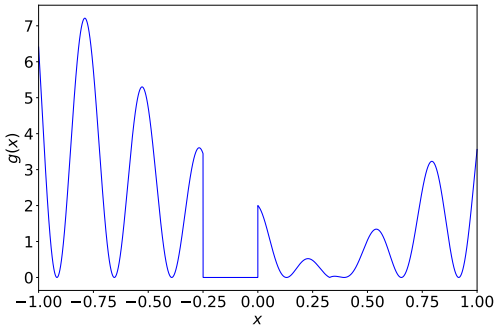
3. (6 points) Consider the function

$$g(x) = \cos^2(12x) \times |x^3 + 6x - 2| \times \mathbb{1}_{x \in (-1,-.25) \cup (0,1)}.$$

In this problem, we use rejection sampling to generate random variables with pdf $f(x) = cg(x)$.

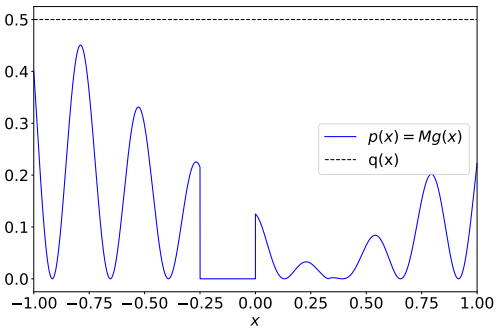
(a) (2 points) Plot $g$ over its domain. What is a uniform proposal distribution $q$ that covers the support of $f$? What is the largest possible constant $M$ such that the scaled target distribution $p(x) = Mg(x)$ satisfies $p(x) \leq q(x)$ for all $x$?

**Solution:** One proposal is $q(x) = \frac{1}{2}\mathbb{1}_{x \in (-1,1)}$. Since $g(x) \leq 8$ on the range $-1 \leq x \leq 1$ we will take $M = \frac{1}{16}$ so that $p(x) \leq q(x)$. Since $\max_x g(x) \approx 7.21$ we can really take any $M \leq \frac{1}{14.42}$ for this proposal distribution.




(b) (2 points) Suppose you run rejection sampling with target $p$ and proposal $q$ from part (a) until you generate $n$ samples and your sampler runs a total of $N \geq n$ times, including $n$ acceptances and $N - n$ rejections. Explain how you can use $n, N$ and $M$ to estimate $c$.
*Hint*: the ratio of acceptances $n$ to total runs $N$ is an approximation of the ratio between the area under the curve $p(x)$ and the area under $q(x)$.
*Hint*: remember what happens if you integrate a pdf over its entire support.

**Solution:** Below we plot the target $p$ and proposal $q$.



Rejection sampling is set up so that we accept a sample if it lies below the blue target curve. In particular

$$\frac{n}{N} \approx \frac{\int_{-1}^{1} p(x)dx}{\int_{-1}^{1} q(x)dx} = M \int_{-1}^{1} g(x)dx = \frac{M}{c} \int f(x)dx = \frac{M}{c}.$$

Thus, we shall use $\hat{c} = \frac{NM}{n}$ to estimate $c$.

(c) (2 points) Use rejection sampling to generate a sample of size $10^3$ from $p(x)$. Since $f(x)$ is a pdf and it's proportional to $p(x)$, we can display its estimate easily: plot a normalized histogram of your sample, and overlay a smooth kernel density estimate, that will provide more information on the shape of the estimated distribution. Repeat the previous steps increasing the number of samples to $10^6$.

**Solution:**