6.7900: Pratice Problems for Exam 1, Fall 2024

Solutions

These are not the **only** acceptable answers. Some other answers also received credit.

Answer the questions in the spaces provided. Show your work neatly. **We will only grade answers that appear in the answer boxes or on answer lines.**

If a question seems vague or under-specified to you, make an assumption, write it down, and solve the problem given your assumption.

You may prepare and use both sides of one 8.5 inch x 11 inch sheet of paper upon which you may write/print anything you like. You may not use any electronic device or any other resource other than your two-sided sheet of paper.

Write your name on every page.

Come to the front if you need to ask a question.

Name: _____ MIT Email: _____

Name: _____

Classification

1. (13 points) Consider two-class classification problems, where each sample $x_i \in \mathbb{R}^d$ is either labeled as 1 or -1. Recall that logistic regression attempts to find a *linear classifier* (i.e., a hyperplane), and it does so by modeling the conditional probability as:

$$\hat{P}(y = 1|x; w) = \sigma(w^\top x), \tag{1}$$

where $\sigma(x) = 1/(1 + e^{-x})$ is the sigmoid function, $w \in \mathbb{R}^d$ is the parameter vector, and $x \in \mathbb{R}^d$ represents an input.

The parameters of a logistic regression model may be fit by minimizing the expression

$$L(w) = \frac{1}{N} \sum_{i=1}^{N} \ell((x_i, y_i); w)$$

where $\ell((x, y); w) = \log(1 + e^{-y w^\top x}) = -\log \hat{P}(y|x; w)$ and N is the number of data points. Let $\hat{w}$ be the argument that minimizes L(w), i.e., the result of ERM. For this problem, we assume that $\hat{w}$ exists and is unique.

For each of the following questions, give your answer and provide a 1-sentence justification for each bullet point.

(a) (4 points) Suppose you are working with a real-life dataset. You then discover there was a feature in the dataset that was deleted by accident. You discover the mistake and include this new feature in addition to the existing features and then run ERM. Let $\hat{w}^{(\text{expanded})}$ be the result of ERM with the extra feature now included in the model.

- How does the training loss compare between the model without the new feature (i.e., $\hat{w}$) and the model with the new feature (i.e., $\hat{w}^{(\text{expanded})}$)?

> **Solution:** The training loss with the new feature is smaller than (or equal to) the previous training loss.

- Suppose the $N = 10000$ and $d = 10$. What can you say about the relationship between the test loss of the models? (Note: your answer may be a rough statement, but answering "the test loss of the new model can be larger or smaller than the old model depending on the dataset" is not sufficient for credit.)

> **Solution:** The test loss with the new feature can be anywhere from much smaller (if the feature is useful) to very slightly larger (if the feature is spurious). In particular, the test loss with the new feature cannot be much worse than the test loss without the new feature.
> Give full credit if the student indicates anything about how the test loss cannot be too much worse. Do not assign credit if they say "it depends on the data" with demonstrating any further insight.

(b) (2 points) Suppose we add ridge regularization and instead minimize the penalized loss function:

$$\hat{w}_\lambda = \operatorname*{argmin}_w L(w) + \lambda \|w\|^2. \tag{2}$$

- How does $L(\hat{w})$ compare to $L(\hat{w}_\lambda)$? (Greater than, less than, or about the same, it depends, etc.) Explain briefly.

> **Solution:** $L(\hat{w}) \leqslant L(\hat{w}_\lambda)$ since $\hat{w}$ minimizes L.

(c) (4 points) You encounter a Caltech student who tells you it's better to minimize the hinge loss:

$$L_{\text{hinge}}(w) = \frac{1}{N} \sum_{i=1}^{N} \ell_{\text{hinge}}((x_i, y_i); w) \tag{3}$$

where $\ell_{\text{hinge}}((x, y), w) = \max(0, 1 - y w^\top x)$. Call the minimizer $\hat{w}_{\text{hinge}}$.

- How does $L(\hat{w})$ compare to $L(\hat{w}_{\text{hinge}})$? (Greater than, less than, or about the same, it depends, etc.) Explain briefly.

> **Solution:** $L(\hat{w}) \leqslant L(\hat{w}_{\text{hinge}})$ since $\hat{w}$ minimizes L.

- Suppose that as $N \to \infty$, the models converge to unique minimizers, so that $\hat{w} \to w^*$ and $\hat{w}_{\text{hinge}} \to w^*_{\text{hinge}}$ with probability 1. What can you say about the test logistic loss using w^* compared to the test logistic loss using w^*_{hinge}? (Note: your answer may be a rough statement, but answering "the test loss of the new model can be larger or smaller than the old model depending on the dataset" is not sufficient for credit.)

> **Solution:** The loss with $\hat{w} \to w^*$ will be at least as small, since it is the minimizer of the population risk.
> It will be slightly smaller in most cases, but students need not explicitly say this for full credit.

(d) (3 points) You fit your model and find the logistic loss on an infinite test sample is 0.01. You are ultimately interested in classification accuracy, however. Suppose you turn your model into a binary classifier in the natural way:

$$h(x) = \begin{cases} -1 & \text{if } \hat{P}(y = 1 \mid x; \hat{w}) \leqslant .5 \\ 1 & \text{if } \hat{P}(y = 1 \mid x; \hat{w}) > .5 \end{cases} \tag{4}$$

Which of the following misclassification rates of $h(x)$ are possible? Check all that are possible.

( ) Misclassification rate of 0

( ) Misclassification rate of $1/2$

( ) Misclassification rate of 1

Solution: It cannot be $1/2$ or 1, because logistic loss is at least $\log 2$ for any misclassification.

It turns out it cannot be zero either, since if there was zero test loss on an infinite data set, then the data would be linearly separable and the model fitting would not converge on the training set. This was not the point of the question, however, so no points will be deducted for either answer to the first checkbox.

Name: _____

Regression

2. (8 points) Consider a data set with N data points, where the inputs are $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, and the corresponding targets (labels) are $t_1, \ldots, t_N$. N is finite; and each data point is independently drawn.

   The targets t is given by a noisy linear model:

   $$t = \mathbf{w}_{\text{unknown}}{}^T \boldsymbol{\phi}(\mathbf{x}) + \epsilon$$

   where $\boldsymbol{\phi}$ is a known featurization; $\mathbf{w}_{\text{unknown}}$ is an unknown model parameter; ϵ is a uni-variate Gaussian noise with zero mean and a fixed variance σ^2.

   We try to fit a linear regression model $\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)$ on the data set, but with the n^{th} data point weighted by a factor $r_n > 0$. In other words, we look for $\mathbf{w}$ to minimize the error:

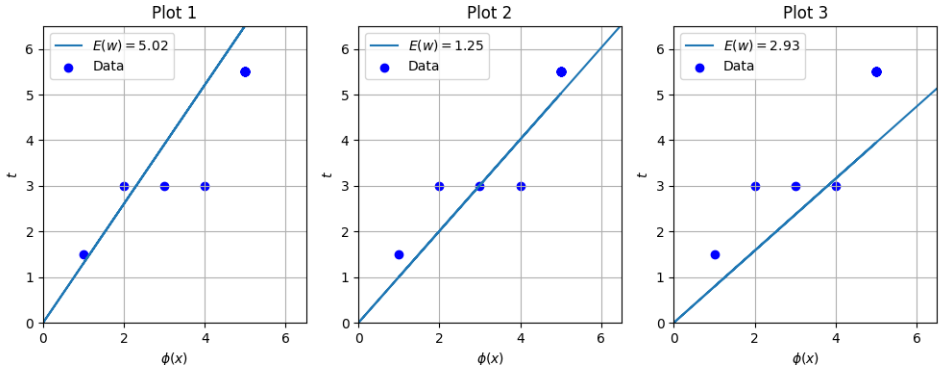
   $$E(\mathbf{w}) := \frac{1}{2}\sum_{n=1}^{N} r_n \left\{t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)\right\}^2.$$

   (a) (3 points) Let's first look at a concrete example with one-dimensional input. Suppose the data set has 5 data points, where the features are $(1, 2, 3, 4, 5)$ and the targets are $(1.5, 3, 3, 3, 5.5)$.

   Consider three possible weighting schemes:

   - Scheme A: $r_2 = 20$, and $r_n = 1, \forall n \neq 2$
   - Scheme B: $r_3 = 20$, and $r_n = 1, \forall n \neq 3$
   - Scheme C: $r_4 = 20$, and $r_n = 1, \forall n \neq 4$

   Plotted blow are the resulting optimal models $\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x})$ and the optimal error. Which plot corresponds to Scheme C?



   ( ) Plot 1

   ( ) Plot 2

   ( ) Plot 3

   Solution: Plot 3.

(b) (3 points) Continuing with the part (a) setup. Given the experiment results in part (a), among the 2^{nd}, 3^{rd}, and 4^{th} data points, which one has the smallest noise $\|\epsilon\|$?

( ) The 2^{nd} data point.
( ) The 3^{rd} data point.
( ) The 4^{th} data point.
( ) We do not have enough info from part (a) results to determine which data point has the smallest noise $\|\epsilon\|$.

Solution: We do not have enough info from part (a) results to determine which data point has the smallest noise $\|\epsilon\|$.

(c) (2 point) Consider now the general setup, described before part (a).
If you are free to set your own weighting scheme r_n, $n = 1, 2, \ldots N$, in $E(\mathbf{w})$, is it true that there exists an algorithm whereby you can figure out which one(s) of the finitely-many N data points have the smallest noise $\|\epsilon\|$?

( ) True
( ) False

Solution: False.
In order to compare the noise of the data points, for each data point, we need to subtract from the target the term $\mathbf{w}_{\text{unknown}}{}^{T}\boldsymbol{\phi}(\mathbf{x})$.
With finitely-many training samples, we can not figure out the unknown true $\mathbf{w}_{\text{unknown}}$ and therefore we can not figure out the true noises.

Uncertainty

3. (6 points) A probability-valued predictor is a function h that assigns to every domain point x a probability value, $h(x) \in [0, 1]$. Formally, we define a probability-vallued predictor as a function, $h : X \to [0, 1]$. We assume a binary response: $y \in \{0, 1\}$.

We will consider the loss of such h on a data point (x, y) to be

$$\ell(y, h(x)) = -y\log(h(x)) - (1 - y)\log(1 - h(x)),$$

which is minus the log of the probability assigned to the observed outcome. This is sometimes called "log loss" or "cross-entropy loss".

(a) (4 points) Suppose the true distribution is $P(Y = 1 \mid X = x) = f(x)$ for a fixed function f. What is the Bayes-optimal probability-valued predictor for this loss function? Show your work.

Solution: The Bayes-optimal predictor is $h(x) = f(x)$. That is, the true probability of $Y = 1$ given $X = x$ minimizes the loss.

To see this, for each x, we wish to minimize the expected loss

$$-f(x)\log(h(x)) - (1 - f(x))\log(1 - h(x)) \tag{5}$$

over our output $h(x)$ with $f(x)$ fixed. This is convex in the value of $h(x)$, so the minimum is where the derivative is zero. Taking the derivative and setting it to zero gives:

$$\frac{f(x)}{1 - f(x)} = \frac{h(x)}{1 - h(x)}. \tag{6}$$

This implies $h(x) = f(x)$ minimizes the loss.

(b) (2 points) *Based on the previous part*, why might you prefer to use this loss function rather than hinge loss when training a classifier? (1 sentence is sufficient.)

Solution: This loss function encourages the model to output values close to the probability of observing $y = 1$ for a given x. (Hinge loss does not have this property.)

Any answer remotely close to "This tells you more about uncertainty because it targets the probability" is worth full credit.

"This gives you a number between 0 and 1" is incorrect because it does not build on the previous part.

Name: _____

Uniformly naive

4. (18 points) Consider a generative approach to classification, in which we estimate $P(Y)$ and $P(X|Y)$ from data. There are two classes, 0 and 1. We will make the same independence assumption as in Naive Bayes, that the features X_j are independent of each other given the class Y, but the features are d real-valued random variables, with independent uniform distributions. So:

$$Y \sim \text{Bernoulli}(q_1) \tag{7}$$

$$X_j \mid Y = c \sim \text{Uniform}(a_{cj}, b_{cj}) \quad \text{for } 1 \leqslant j \leqslant d \tag{8}$$

where $c \in \{0, 1\}$ and $q_0 = 1 - q_1$.
So, the parameter vector $\theta = q_1, a_{01}, b_{01}, a_{11}, b_{11}, \ldots, a_{0d}, b_{0d}, a_{1d}, b_{1d}$.

(a) For a data set $D = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$, write a formula for the log likelihood $P(D; \theta)$ in terms of x and y values in the data set and parameter values in θ.

Solution:

$$\begin{aligned}\log P(D;\theta) \quad = \quad & \sum_{i=1}^{n} y^{(i)} \log q_1 \left(\begin{cases} \frac{1}{\prod_{j=1}^{d}(b_{1j}-a_{1j})} & \text{if } a_{1j} \leqslant x_j \leqslant b_{1j} \text{ for all } 1 \leqslant j \leqslant d \\ 0 & \text{otherwise} \end{cases} \right) \\ & + (1 - y^{(i)}) \log q_0 \left(\begin{cases} \frac{1}{\prod_{j=1}^{d}(b_{0j}-a_{0j})} & \text{if } a_{0j} \leqslant x_j \leqslant b_{0j} \text{ for all } 1 \leqslant j \leqslant d \\ 0 & \text{otherwise} \end{cases} \right)\end{aligned}$$

(b) Given parameters θ and a new example x, such that for all feature indices j, $a_{1j} \leqslant x_j \leqslant b_{1j}$ and $a_{0j} \leqslant x_j \leqslant b_{0j}$, under what conditions would you predict that it belongs to class 1? Express your answer in terms of elements of x and θ.

Solution:

$$\begin{aligned} P(Y = 1 \mid X) &> P(Y = 0 \mid X) \\ P(X \mid Y = 1)P(Y = 1) &> P(X \mid Y = 0)P(Y = 0) \\ P(X \mid Y = 1)P(Y = 1) &> P(X \mid Y = 0)P(Y = 0) \\ \frac{q_1}{\prod_{j=1}^{d}(b_{1j} - a_{1j})} &> \frac{q_0}{\prod_{j=1}^{d}(b_{0j} - a_{0j})} \end{aligned}$$

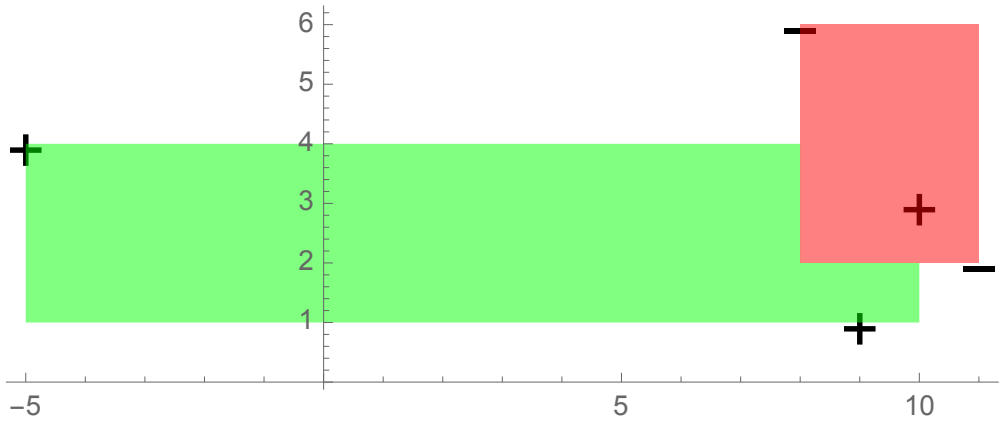
(c) Given training data

x	y
$(10, 3)$	1
$(9, 1)$	1
$(-5, 4)$	1
$(11, 2)$	0
$(8, 6)$	0

What are the maximum-likelihood parameter estimates?

Solution: $q_1 = 3/5; q_0 = 2/5; a_{11} = -5; b_{11} = 10; a_{01} = 8; b_{01} = 11; a_{12} = 1; b_{12} = 4; a_{02} = 2; b_{02} = 6$

(d) Given the same training data (plotted below), and using the maximum-likelihood parameter estimates, label very clearly all regions of the space that would be classified as positive and those that would be classified as negative.



Solution: The positive box has dimensions 15 x 3 and probability 3/5. The negative box has dimensions 3 x 4 and probability 2/5.

Compare (1/75) with (1/30). That means that, inside the negative box, a negative label is most likely.

Name: _____

Off to the races

5. (10 points) You are trying to decide what fraction, g, of your wealth to bet on the next horse race. You can observe a vector x of features of the horse. This particular horse will either win ($y = 1$) or lose ($y = 0$) the race. Your loss function is, for some fixed positive constant $c > 0$,

$$L(g, y) = \begin{cases} -cg & \text{if } y = 1 \\ g & \text{if } y = 0 \end{cases}$$

That is, if the horse wins and you bet fraction g of your money, then you win cg in profit (that is, your loss is $-cg$); if the horse loses, then you lose your bet g.

You have a data set $D = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$, representing previous examples of horses labeled with whether or not they have won their races.

We would like to use the principle of empirical risk minimization to estimate parameters w to fit a model of the form $g = \sigma(w \cdot x)$ to the data, where σ is the sigmoid function.

(a) Write an expression for the empirical risk as a function of w, in terms of σ, w, c, and elements of D.

Solution:

$$ER(w) = \sum_{i=1}^{n} \sigma(w \cdot x^{(i)})(-c)^{y^{(i)}}$$

(b) What would the update rule for a stochastic gradient optimizer be? Please write it in terms of σ, w, c, $x^{(j)}$, and $y^{(j)}$, where $(x^{(j)}, y^{(j)})$ is a new training example.

Solution:

$$w = w - \eta\sigma(w \cdot x^{(j)})(1 - \sigma(w \cdot x^{(j)})) \cdot x^{(j)} \cdot (-c)^{y^{(i)}}$$

Name: _____

Regression with variances

6. (18 points) Your friend Dana is an astronomer who is trying to predict the degree of sunspot activity, Y, as a function of a vector of observable parameters X. Dana believes the observations of X are very reliable, but the Y observations are corrupted by Gaussian noise. Furthermore, the noise depends on the atmospheric conditions and may be different on every observation. Luckily, last year, the astronomers developed a good way of predicting the level of noise, and so Dana has a data set consisting of triples $D = \{(x^{(i)}, y^{(i)}, v^{(i)})\}_{i=1}^{n}$. We make the modeling assumption that, for some weight vector w, and for all i,

$$Y^{(i)} \mid X^{(i)} = x^{(i)} \sim \text{Normal}(w \cdot x^{(i)}, v^{(i)})$$

Note: $v^{(i)}$ is a variance.

(a) Write an expression for the log likelihood of the data in terms of parameters w and elements of D.

Solution:

$$\log P(D; w) = \sum_{i=1}^{n} -\frac{1}{2}\log(2v^{(i)}) - \frac{(w \cdot x^{(i)} - y^{(i)})^2}{2v^{(i)}}$$

(b) Derive a stochastic gradient descent update rule for w. Please write it in terms of w, $x^{(j)}$, $y^{(j)}$, and $v^{(j)}$ where $(x^{(j)}, y^{(j)}, v^{(j)})$ is a new training example.

Solution:

$$w := w - \eta \frac{x^{(i)}}{v^{(i)}}(y^{(i)} - w \cdot x^{(i)})$$

(c) If all the $v^{(i)}$ are equal is this the same as ordinary least squares? Explain briefly.

Solution: Yes. All samples have the same variance, which is the standard OLS assumption.

(d) Is there a value of $v^{(i)}$ that would cause the maximum likelihood weight estimates to be independent of $x^{(i)}$ and $y^{(i)}$? Explain briefly.

Solution: Infinity. We're dividing by the variance in the update, so the bigger the variance, the less effect it has on the result.

(e) Is there a value of $v^{(i)}$ that would cause the maximum likelihood weight estimates to be independent of $x^{(j)}$ and $y^{(j)}$, for $j \neq i$, irrelevant? Explain briefly.

Solution: 0. We're dividing by the variance in the update, so the smaller the variance, the more effect it has on the result (and the less the other observations have, in comparison).

Name: _____

Piecewise Linear Regression

7. (20 points) Suppose you were trying to do regression on a one-dimensional input space using a piecewise linear (but not necessarily continuous) function. A predictor with m pieces is parameterized with $m-1$ breakpoints $c_1, \ldots, c_{m-1}$ and m pairs $\beta_0^{(j)}, \beta^{(j)}$, so the regression function is

$$h(x) = \begin{cases} \beta^{(1)} \cdot x + \beta_0^{(1)} & \text{if } x \leqslant c_1 \\ \beta^{(2)} \cdot x + \beta_0^{(2)} & \text{if } c_1 < x \leqslant c_2 \\ \ldots & \\ \beta^{(m)} \cdot x + \beta_0^{(m)} & \text{if } c_{m-1} < x \end{cases} \tag{9}$$

The decision of how many pieces to use is part of the model-fitting process.

(a) If you were given 4 training points $\{(x^{(i)}, y^{(i)})\}_{i=1}^4$, give a set of parameters that would minimize sum squared error on the data. If it is useful, assume that $x^{(i)} < x^{(j)}$ for $i < j$.

> **Solution:** Choose three breakpoints, between each successive pair of points. Set $\beta^{(i)} = 0$ for all i. Set $\beta_0^{(i)} = y^{(i)}$ for all i.

(b) Is that set of parameters unique? Briefly explain why or why not.

> **Solution:** No. There are many choices of breakpoints and linear segments that will get 0 error on the training data.

(c) If you were required to limit yourself to predictors with $m = 2$, sketch an algorithm for finding the model parameters to minimize sum squared error on the data set with 4 training points.

> **Solution:** Put a breakpoint between points 2 and 3; then fit a line to points 1 and 2 and fit another line to points 3 and 4.
>
> In fact, it was a mistake on our part to specialize this question to 4 data points. In the general case (with $m = 2$ but arbitrary amounts of data), you'd have to consider putting the breakpoint between each adjacent pair of points, doing OLS on each half to get the predictors, and then picking the breakpoint that gives the smallest SSE overall.

(d) You are given 100 training examples, and you'd like to find a predictor (including m and parameter values) that you think will minimize expected squared loss on unseen data drawn from that same distribution. Sketch a procedure for doing this.

> **Solution:** For $m = 1, ..., 100$ fit the minimum MSE hypothesis, then use a validation set to pick which one generalizes the best. Or, use cross-validation in a similar way.

Name: _____

Skee Ball

8. (24 points) Skee Ball is a carnival game, where a player tries to roll a ball up a ramp and get it to fall into a hole. Different holes win the player different numbers of points.

   Your skee ball game has three holes: a, b, and c.

   You can throw the ball soft (0) or hard (1).

   Initially, you don't know very much about how the game works; in particular, you don't know how your choice of throwing hard or soft affects which hole the ball falls into.

   So, you do some experiments!

   - You throw the ball soft 3 times, and it lands in a, a, and b.
   - You throw the ball hard 3 times and it lands in c, c, and c.

   Let H be the random variable indicating which hole the ball falls into (a, b, or c) and F be the random variable indicating how forcefully you throw the ball (0 or 1).

   For simplicity, we'll define $\theta_{hf} = P(H = h \mid F = f)$.

   In the all parts of this question, feel free to write out an expression with numbers plugged into it; you don't have to evaluate the expression numerically.

   (a) Having collected your experimental data, what are the maximum likelihood estimates of θ_{hf} for all values of $h \in \{a, b, c\}$ and $f \in \{0, 1\}$?

   Solution:

   $$P(H = a \mid F = 0) = 2/3$$
   $$P(H = b \mid F = 0) = 1/3$$
   $$P(H = c \mid F = 0) = 0$$
   $$P(H = a \mid F = 1) = 0$$
   $$P(H = b \mid F = 1) = 0$$
   $$P(H = c \mid F = 1) = 1$$

(b) We can think about the three parameters associated with a single conditional distribution as a point in a higher dimensional space: $\theta_f = (\theta_{af}, \theta_{bf}, \theta_{cf})$. Describe the set of valid values of θ_f?

> **Solution:** The 3-dimensional simplex. (Or, the set of 3-D vectors of positive numbers that sum to 1.)

(c) You want to be Bayesian and start with a uniform prior on θ_0 and a uniform prior on θ_1. What family of distributions, with what parameters, would you use for this purpose?

> **Solution:** Two Dirichlet distributions with parameters $(1, 1, 1)$.

(d) What would the Bayesian posteriors on θ_0 and on θ_1 be, after conditioning on your experimental data? Provide distribution family (e.g. Gaussian) and numerical values (or detailed expressions) of the parameters.

> **Solution:** Dirichlet$(3, 2, 1)$ and Dirichlet$(1, 1, 4)$

Now assume that getting a ball into hole a is worth 1 point, into b is worth 5 points and into c is worth 4 points. We want to earn as many points as possible, and the loss relative to putting the ball into hole b (worth 5 points) is therefore $4, 0, 1$ for holes a, b, c.

(e) Let $\hat{\theta}_{hf}$ be the maximum likelihood estimate of getting a ball into hole h given how forcefully it was thrown. If we approximate θ_{hf} by using the MLE, write an expression for the approximate risk of each choice of how to throw the ball. What is the action that minimizes this approximate risk for the MLE calculated above?

Solution:

$$\begin{aligned}
\text{Risk}(F = f) &= 4 * P(a \mid f) + 0 * P(b \mid f) + 1 * P(c \mid f) \\
&\approx 4 * \hat{P}(a \mid f) + 0 * \hat{P}(b \mid f) + 1 * \hat{P}(c \mid f) \\
\text{Risk}(F = 0) &\approx 4 * 2/3 + 0 * 1/3 + 1 * 0 = 8/3 \\
\text{Risk}(F = 1) &\approx 4 * 0 + 0 * 0 + 1 * 1 = 1 \\
f^* &= 1
\end{aligned}$$

(f) Assuming the Bayesian posterior is $p(\theta_f \mid \mathcal{D})$, write an expression for the posterior risk of each choice of how to throw the ball (i.e., write the risk where $p(\theta_f)$ is approximated by the posterior $p(\theta_f \mid \mathcal{D})$). What is the action that minimizes this approximate risk for the MLE calculated above?

Solution: Use posterior predictive distribution of Dirichlet

$$\begin{aligned}
\text{Risk}(F = f) &= 4 * P(a \mid f, \mathcal{D}) + 0 * P(b \mid f, \mathcal{D}) + 1 * P(c \mid f, \mathcal{D}) \\
\text{Risk}(F = 0) &= 4 * 3/6 + 0 * 2/6 + 1 * 1/6 = 13/6 \\
\text{Risk}(F = 1) &= 4 * 1/6 + 0 * 1/6 + 1 * 4/6 = 8/6 \\
f^* &= 1
\end{aligned}$$

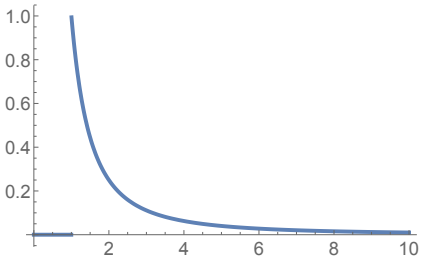
Name: _____

Pareto Optimal?

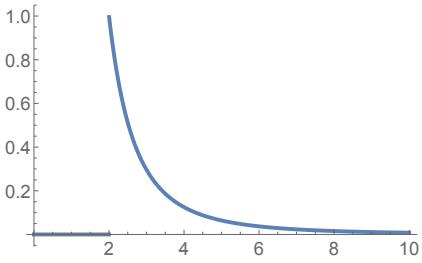
9. (24 points) You can get samples of a random variable X which is drawn uniformly at random from the interval $[0, M]$, but you don't know M. You model your prior belief on M using a Pareto distribution with parameters 1, 1, which is shown in graph A below.

   A *Pareto* distribution has two parameters α and β both of which are real values greater than 0. Its pdf is

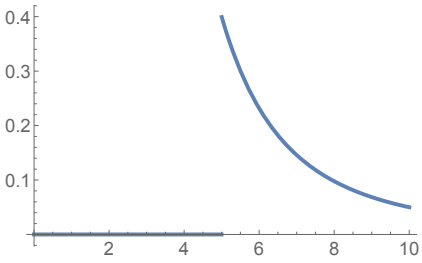
$$f_M(m) = \begin{cases} \frac{\alpha \beta^\alpha}{m^{\alpha+1}} & \text{if } m > \beta \\ 0 & \text{otherwise} \end{cases} \tag{10}$$



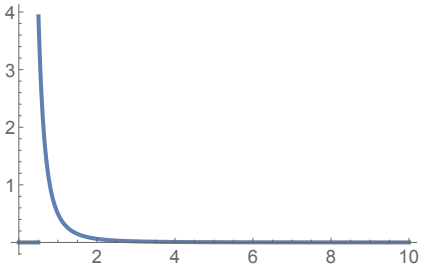
(a) A



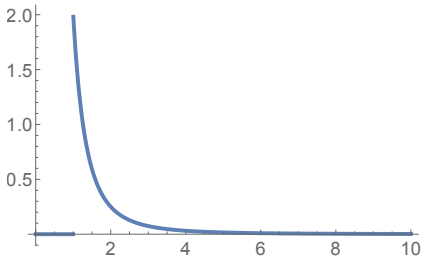
(b) B



(c) C



(d) D



(e) E

   (a) What is the pdf of the conditional distribution $p(M \mid X)$?
   (Hint: the Pareto distribution is a conjugate prior for uniform observations.)

Solution:

$$\begin{aligned} p_{M|X}(m \mid x) &\propto p_{X|M}(x \mid m) f_M(m) \\ &\propto \frac{1}{m} \cdot \begin{cases} \frac{\alpha \beta^\alpha}{m^{\alpha+1}} & \text{if } m > \beta \text{ and } m > x \\ 0 & \text{otherwise} \end{cases} \\ &\propto \begin{cases} \frac{\alpha \beta^\alpha}{m^{\alpha+2}} & \text{if } m > \max(\beta, x) \\ 0 & \text{otherwise} \end{cases} \\ &\propto \begin{cases} \frac{(\alpha+1)\max(\beta,x)^{(\alpha+1)}}{m^{\alpha+2}} & \text{if } m > \max(\beta, x) \\ 0 & \text{otherwise} \end{cases} \\ &= \text{Pareto}(\alpha + 1, \max(\beta, x)) \end{aligned}$$

(b) If you start with a prior distribution Pareto$(1, 1)$ and observe $x^{(1)} = 0.5$, what is the family and parameters of the posterior distribution?

Solution:

Pareto$(2, 1)$

(c) Which of the graphs above does it correspond to?

◯ A ◯ B ◯ C ◯ D √ **E**

(d) If you start with a prior distribution Pareto$(1, 1)$ and observe $x^{(1)} = 5$, what is the family and parameters of the posterior distribution?

Solution:

Pareto$(2, 5)$

(e) Which of the graphs above does it correspond to?

◯ A ◯ B √ **C** ◯ D ◯ E