

## 6.7900: Practice Problems for Exam 1, Fall 2024

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: \_\_\_\_\_

## 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 hyper-plane), and it does so by modeling the conditional probability as:

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

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.,  $\hat{w}^{(\text{expanded})}$ )?

- 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.)

Name: \_\_\_\_\_

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

$$\hat{w}_\lambda = \underset{w}{\operatorname{argmin}} L(w) + \lambda \|w\|^2. \quad (2)$$

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

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

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

where  $\ell_{\text{hinge}}((x, y), w) = \max(0, 1 - yw^\top x)$ . 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.

- Suppose that as  $N \rightarrow \infty$ , the models converge to unique minimizers, so that  $\hat{w} \rightarrow w^*$  and  $\hat{w}_{\text{hinge}} \rightarrow w_{\text{hinge}}^*$  with probability 1. What can you say about the test logistic loss using  $w^*$  compared to the test logistic loss using  $w_{\text{hinge}}^*$ ? (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.)

Name: \_\_\_\_\_

- (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:

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

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

## Regression

2. (8 points) Consider a data set with  $N$  data points, where the inputs are  $\{x_1, \dots, x_N\}$ , and the corresponding targets (labels) are  $t_1, \dots, t_N$ .  $N$  is finite; and each data point is independently drawn.

The targets  $t$  is given by a noisy linear model:

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

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

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

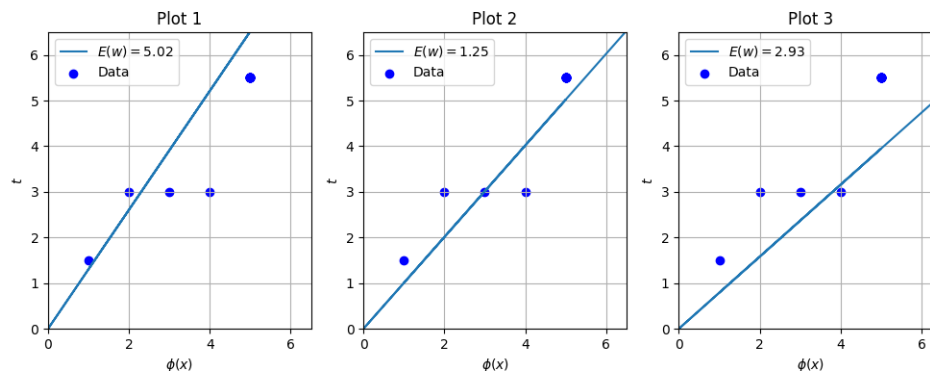
$$E(\mathbf{w}) := \frac{1}{2} \sum_{n=1}^N r_n \{t_n - \mathbf{w}^T \phi(x_n)\}^2.$$

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



- ( ) Plot 1  
 ( ) Plot 2  
 ( ) Plot 3

**Name:** \_\_\_\_\_

Name: \_\_\_\_\_

- (b) (3 points) Continuing with the part (a) setup. Given the experiment results in part (a), among the 2<sup>nd</sup>, 3<sup>rd</sup>, and 4<sup>th</sup> data points, which one has the smallest noise  $\|\epsilon\|$ ?
- ( ) The 2<sup>nd</sup> data point.
  - ( ) The 3<sup>rd</sup> data point.
  - ( ) The 4<sup>th</sup> data point.
  - ( ) 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  $r_n$ ,  $n = 1, 2, \dots, N$ , 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

## 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-valued predictor as a function,  $h : X \rightarrow [0, 1]$ . 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”.

Name: \_\_\_\_\_

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

- (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.)



**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:

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

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

where  $c \in \{0, 1\}$  and  $q_0 = 1 - q_1$ .

So, the parameter vector  $\theta = q_1, a_{01}, b_{01}, a_{11}, b_{11}, \dots, a_{0d}, b_{0d}, a_{1d}, b_{1d}$ .

- (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  $\theta$ .

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

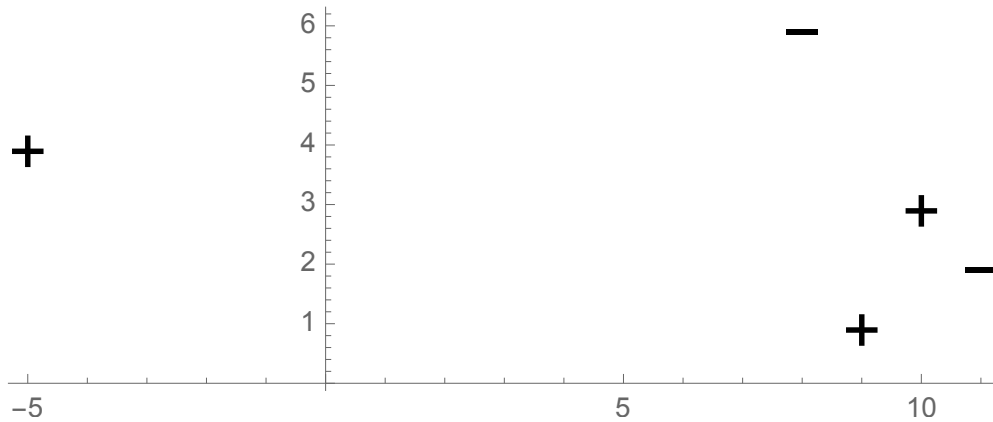
Name: \_\_\_\_\_

(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?

(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.



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  $\sigma$  is the sigmoid function.

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

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

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  $i$ ,

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

**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  $D$ .

- (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.

Name: \_\_\_\_\_

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

(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.

(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.

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  $c_1, \dots, c_{m-1}$  and  $m$  pairs  $\beta_0^{(j)}, \beta^{(j)}$ , so the regression function is

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

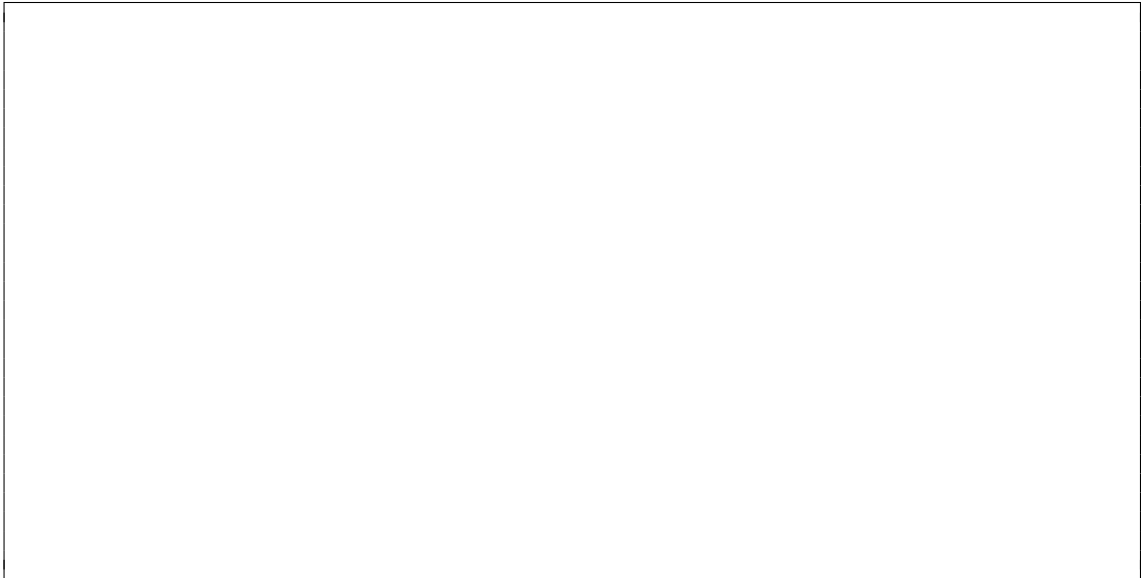
The decision of how many pieces to use is part of the model-fitting process.

- (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  $i < j$ .

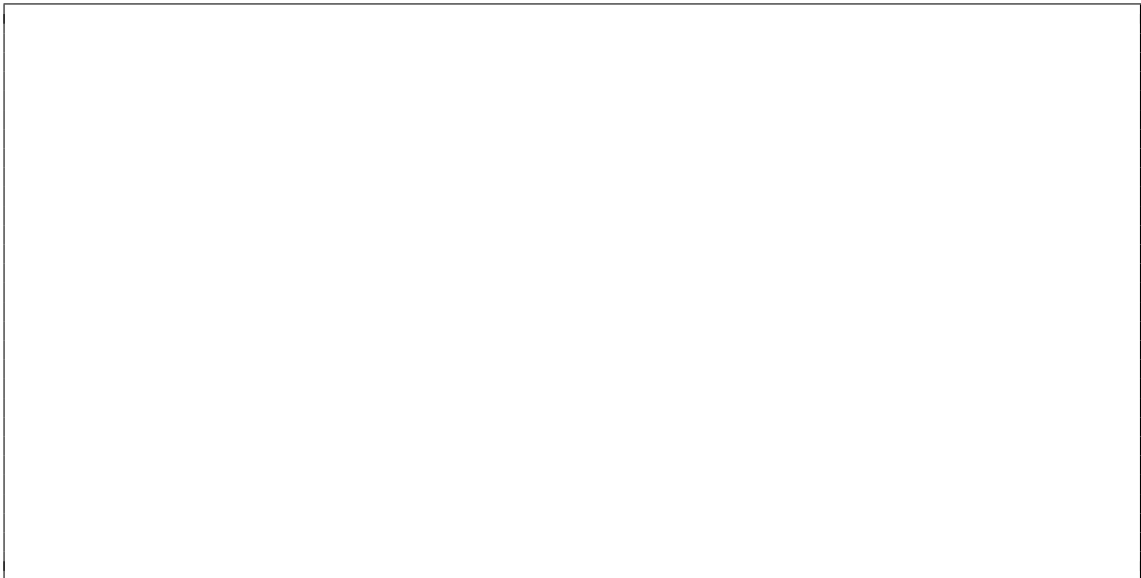
- (b) Is that set of parameters unique? Briefly explain why or why not.

Name: \_\_\_\_\_

- (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.



- (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.



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  $\theta_{hf}$  for all values of  $h \in \{a, b, c\}$  and  $f \in \{0, 1\}$ ?



Name: \_\_\_\_\_

- (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  $\theta_f$ ?

- (c) You want to be Bayesian and start with a uniform prior on  $\theta_0$  and a uniform prior on  $\theta_1$ . What family of distributions, with what parameters, would you use for this purpose?

- (d) What would the Bayesian posteriors on  $\theta_0$  and on  $\theta_1$  be, after conditioning on your experimental data? Provide distribution family (e.g. Gaussian) and numerical values (or detailed expressions) of the parameters.

Name: \_\_\_\_\_

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  $\theta_{hf}$  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?

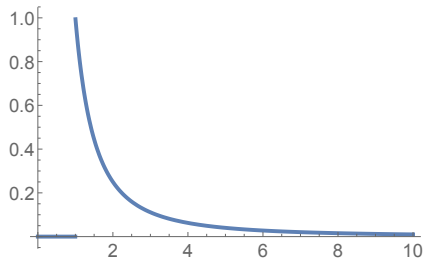
- (f) Assuming the Bayesian posterior is  $p(\theta_f | \mathcal{D})$ , write an expression for the posterior risk of each choice of how to throw the ball (i.e., write the risk where  $p(\theta_f)$  is approximated by the posterior  $p(\theta_f | \mathcal{D})$ ). What is the action that minimizes this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

### Pareto Optimal?

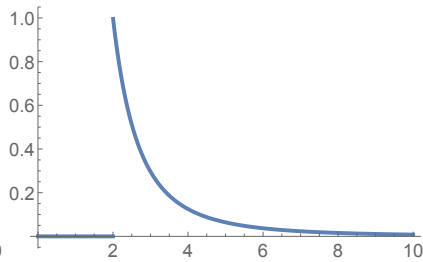
9. (24 points) You can get samples of a random variable  $X$  which is drawn uniformly at random from the interval  $[0, M]$ , but you don't know  $M$ . You model your prior belief on  $M$  using a Pareto distribution with parameters  $1, 1$ , which is shown in graph A below.

A *Pareto* distribution has two parameters  $\alpha$  and  $\beta$  both of which are real values greater than 0. Its pdf is

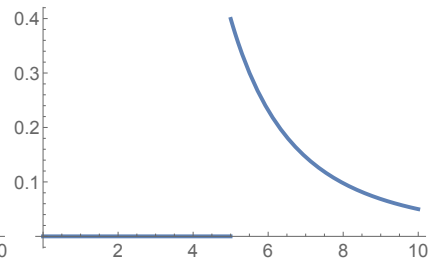
$$f_M(m) = \begin{cases} \frac{\alpha\beta^\alpha}{m^{\alpha+1}} & \text{if } m > \beta \\ 0 & \text{otherwise} \end{cases} \quad (10)$$



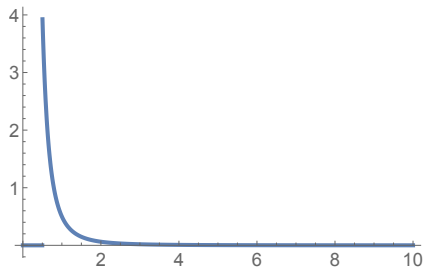
(a) A



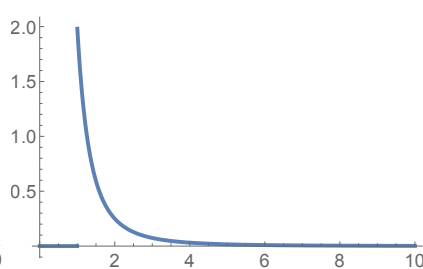
(b) B



(c) C



(d) D

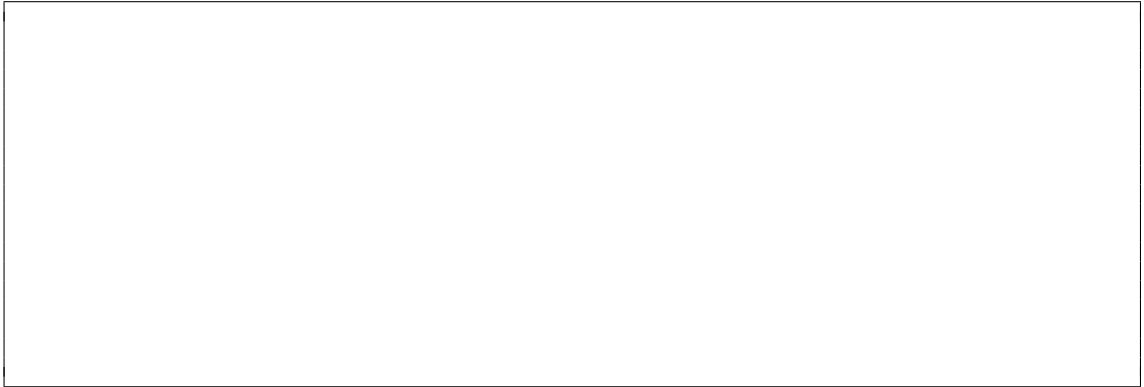


(e) E

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

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

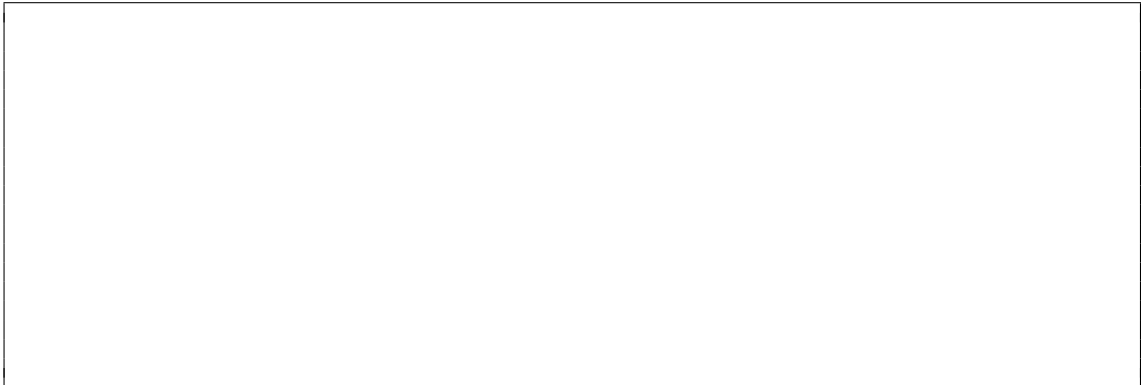
Name: \_\_\_\_\_



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

- A    B    C    D    E

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



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

- A    B    C    D    E