6.867: Exam 1, Fall 2015

Solutions

Answer the questions in the spaces provided on the question sheets. If you run out of room for an answer, continue on the back of the page. Show your work neatly.

You may use any printed or written material. You may not use any electronic device (calculator, phone, tablet, laptop, etc).

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

Write your name on every page.

Come to the front to ask questions.

Name: _____ Athena ID: _____

Question	Points	Score
1	24	
2	20	
3	18	
4	20	
5	24	
6	24	
Total:	130	

Name: _____

Regularization

1. (24 points) You've collected data of the form $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$ with $x^{(i)}, y^{(i)} \in \mathbb{R}$. You decide to use a conditional Gaussian model $Y \mid X = x \sim \mathcal{N}(w_0 + w_1 x, \sigma^2)$. For each of the expressions below, indicate whether it corresponds to: ordinary least-squares regression (OLS), MAP for Bayesian regression (BR), ridge regression (RR), LASSO, LAD, something else that is sensible (Good), or something else that is not sensible (Bad).

   Only if it is one of these last two cases, explain why it is or is not a reasonable thing to do.

   (a) $\text{argmax}_w \, p(\mathcal{D} \mid w)$
   √ **OLS** ◯ BR ◯ RR ◯ LASSO ◯ LAD ◯ Good ◯ Bad

   (b) $\text{argmax}_w \, p(w \mid \mathcal{D})$
   ◯ OLS √ **BR** ◯ RR ◯ LASSO ◯ LAD ◯ Good ◯ Bad

   (c) $\text{argmax}_w \, p(\mathcal{D} \mid w) p(w)$
   ◯ OLS √ **BR** ◯ RR ◯ LASSO ◯ LAD ◯ Good ◯ Bad

   (d) $\text{argmax}_w \prod_i p(x^{(i)} \mid y^{(i)}, w)$
   ◯ OLS ◯ BR ◯ RR ◯ LASSO ◯ LAD ◯ Good √ **Bad**

   > **Solution:** Our application is to predict y given x, so we should pick weights to make the conditional model $p(y \mid x, w)$ as accurate as possible.

   (e) $\text{argmax}_w \prod_i p(y^{(i)} \mid x^{(i)}, w)$
   √ **OLS** ◯ BR ◯ RR ◯ LASSO ◯ LAD ◯ Good ◯ Bad

   (f) $\text{argmax}_w \prod_i p((x^{(i)}, y^{(i)}) \mid w)$
   ◯ OLS ◯ BR ◯ RR ◯ LASSO ◯ LAD ◯ Good √ **Bad**

   > **Solution:** We are not optimizing the conditional model, which is what we will use for predictions.

   (g) $\text{argmax}_w \sum_i p(y^{(i)} \mid x^{(i)}, w)$
   ◯ OLS ◯ BR ◯ RR ◯ LASSO ◯ LAD ◯ Good √ **Bad**

   > **Solution:** Summing probabilities doesn't have an interpretation is optimizing any kind of likelihood; this would allow a model to, for example, assign near-zero probability to one data point without necessarily incurring a huge penalty.

   (h) $\text{argmax}_w \sum_i \log p(y^{(i)} \mid x^{(i)}, w)$
   √ **OLS** ◯ BR ◯ RR ◯ LASSO ◯ LAD ◯ Good ◯ Bad

   (i) $\text{argmin}_w \sum_i (w_0 + w_1 \cdot x^{(i)} - y^{(i)})^2$
   √ **OLS** ◯ BR ◯ RR ◯ LASSO ◯ LAD ◯ Good ◯ Bad

(j) $\text{argmin}_w \sum_i |w_0 + w_1 \cdot x^{(i)} - y^{(i)}|$

◯ OLS ◯ BR ◯ RR ◯ LASSO √ **LAD** ◯ Good ◯ Bad

(k) $\text{argmin}_w \sum_i (w_0 + w_1 \cdot x^{(i)} - y^{(i)})^2 - \lambda w \cdot w$

◯ OLS ◯ BR ◯ RR ◯ LASSO ◯ LAD ◯ Good √ **Bad**

Solution: This is good if you let λ be negative, but if λ is positive, then it encourages large weight magnitudes.

(l) $\text{argmin}_w \sum_i (w_0 + w_1 \cdot x^{(i)} - y^{(i)})^2 + \lambda w \cdot w$

◯ OLS ◯ BR √ **RR** ◯ LASSO ◯ LAD ◯ Good ◯ Bad

(m) $\text{argmin}_w \sum_i (w_0 + w_1 \cdot x^{(i)} - y^{(i)})^2 + \lambda |w_0| + \lambda |w_1|$

◯ OLS ◯ BR ◯ RR √ **LASSO** ◯ LAD ◯ Good ◯ Bad

Name: _____

Piecewise Linear Regression

2. (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

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

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: _____

XOR

3. (18 points) Consider the following data set: $\{((0,0),-1),((1,0),+1),((0,1),+1),((1,1),-1)\}$.

   For each of the following feature representations, write down the corresponding kernel and specify whether or not it makes the data linearly separable. Draw the data in the new feature space and if it is linearly separable, draw in the separator that would be found by a hard-margin SVM.

   (a) $\phi(x) = (x_1 \cdot x_2, 1)$

   Kernel

   Solution:

   $$K(x,z) = x_1 x_2 z_1 z_2 + 1$$

   Separable? ◯ True √ **False**

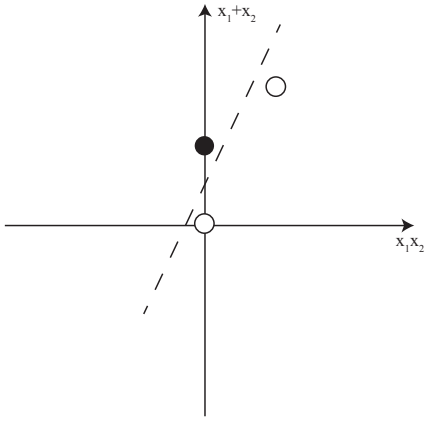
   The data points are not linearly separable. Because both $(0,0)$ and $(1,0)$ maps to the point $(0,1)$ in the feature space, but they have different labels.

(b) $\phi(x) = (x_1 \cdot x_2, x_1 + x_2)$
Kernel

Solution:
$$K(x, z) = x_1x_2z_1z_2 + x_1z_1 + x_1z_2 + x_2z_1 + x_2z_2$$

Separable? $\surd$ **True** ◯ False

Data in feature space, separator if it exists (Black dots are positive, white dots are negative)

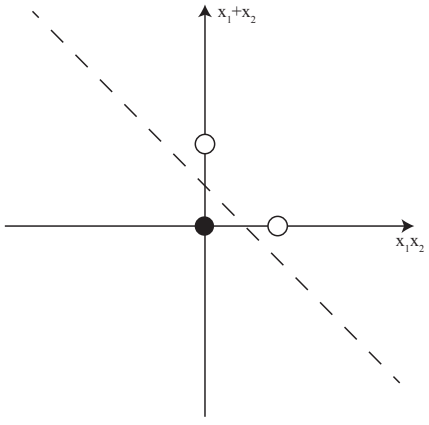


(c) $\phi(x) = (x_1 \cdot x_2, (1 - x_1) \cdot (1 - x_2))$
Kernel

Solution:
$$K(x, z) = x_1x_2z_1z_2 + (1 - x_1)(1 - x_2)(1 - z_1)(1 - z_2)$$

Separable? $\surd$ **True** ◯ False

Data in feature space, separator if it exists (Black dots are positive, white dots are negative)



Name: _____

SVM

4. (20 points) *Credit to Alex Smola*

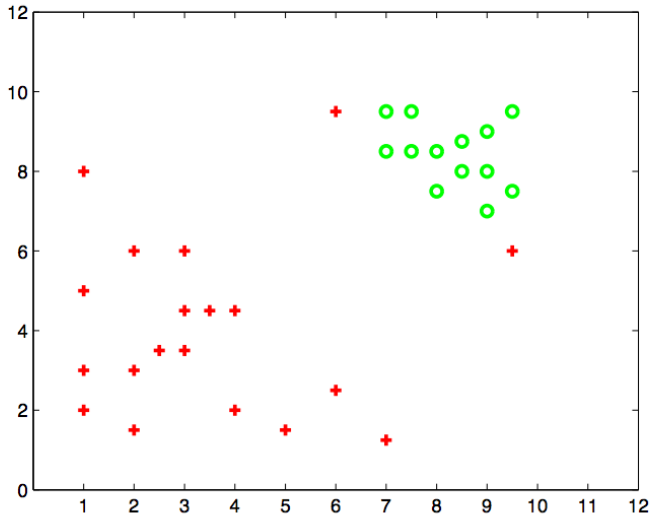
Recall that the primal form of the soft SVM optimization problem is

$$\min_{\theta,\theta_0} \left[\frac{1}{2}\|\theta\|^2 + C\sum_{i=1}^{n} \xi_i \right] \tag{2}$$

$$\text{s.t. } y^{(i)}(\theta^\top x^{(i)} + \theta_0) \geqslant 1 - \xi_i \tag{3}$$

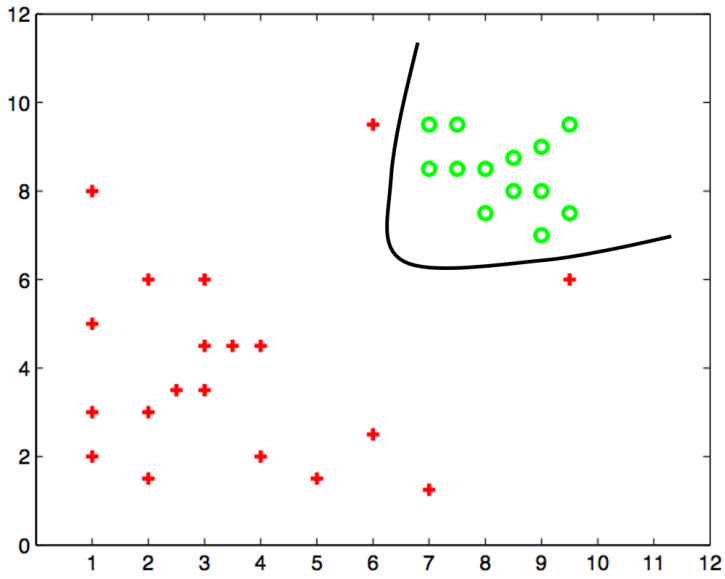
$$\xi_i \geqslant 0 \tag{4}$$

Assume we are training an SVM with a **polynomial kernel** of degree 2. You are given the data set shown below. Please answer the following questions *qualitatively*, by giving a one sentence answer for each and drawing your solution, if one exists, on the given figure.



(a) Where would the decision boundary be if C is extremely large (i.e., as C goes to infinity)?

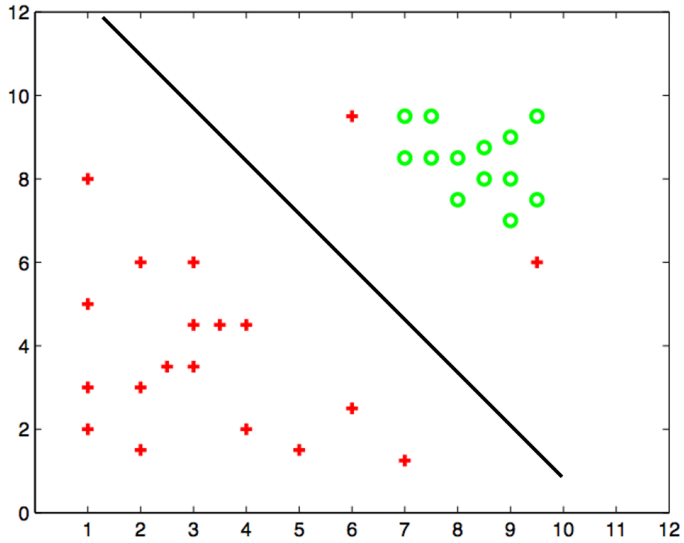
Name: _____



Solution:

(b) For C near 0, where would the decision boundary be?

Solution: One possible solution, with a small but non-zero C, will be:



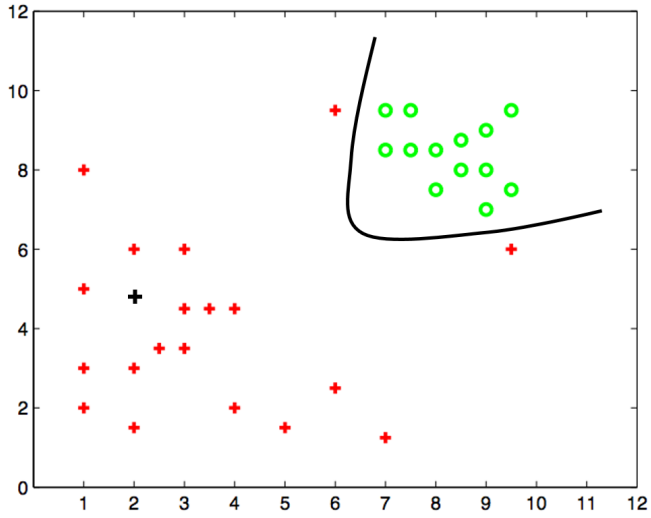
But, really, the problem was too ambiguous (one could argue that the weights would be 0, or that, as they tend toward 0 their ratio is undefined, etc.) so we gave everyone full credit.

(c) In a classification task, which solution would work better: using a very large C or setting C to be near 0? Explain briefly.

Solution: There is a trade-off here, and we accepted a wide variety of answers as long as they were well argued. Using a very large C could lead to overfitting and, hence, bad predictions on unseen data. Using a C near zero asks the weights to be small magnitude with very little misclassification penalty: this can lead to bad performance on the training data and even worse performance on test data.

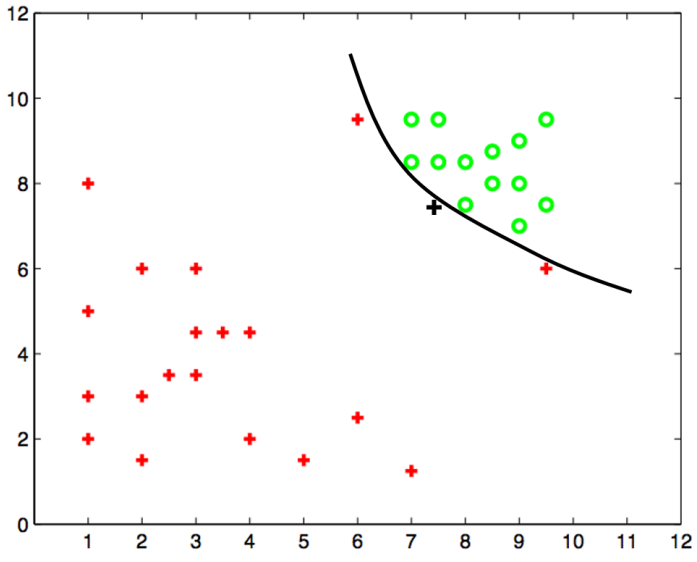
(d) Draw a data point that *would not* change the decision boundary for very large C.

Solution: There are many possible solutions to this. One possible data point that will not change the decision boundary is:



(e) Draw a data point that *would significantly* change the decision boundary for very large C.

Solution: There are many possible solutions to this. One possible data point that will significantly change the decision boundary is:



Name: _____

Skee Ball

5. (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

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

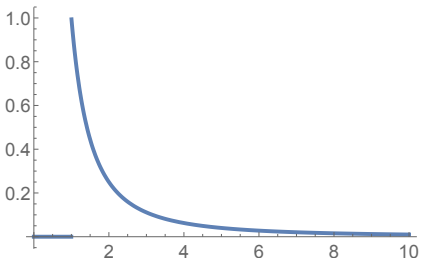
Name: _____

Pareto Optimal?

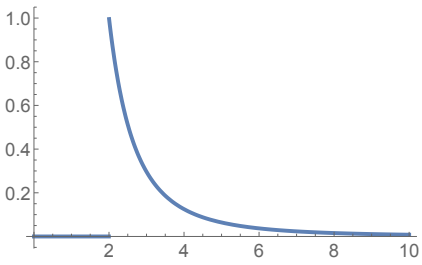
6. (24 points) You can get samples of a random variable X which is drawn uniformly at random from the interval $[0, M]$, but you don't know M. You model your prior belief on M using a Pareto distribution with parameters $1, 1$, which is shown in graph A below.

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

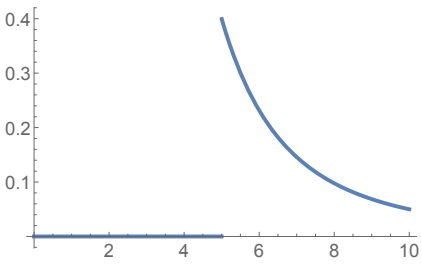
$$f_M(m) = \begin{cases} \frac{\alpha \beta^\alpha}{m^{\alpha+1}} & \text{if } m > \beta \\ 0 & \text{otherwise} \end{cases} \tag{5}$$



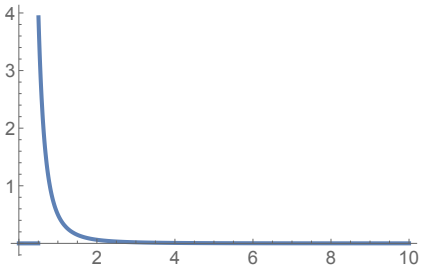
(a) A



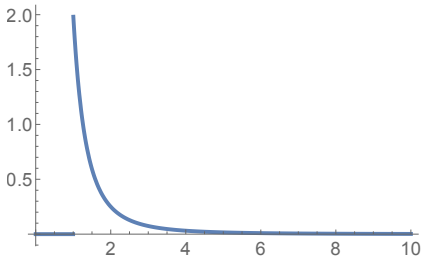
(b) B



(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:

$$\text{Pareto}(2, 1)$$

(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:

$$\text{Pareto}(2, 5)$$

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

◯ A ◯ B √ **C** ◯ D ◯ E