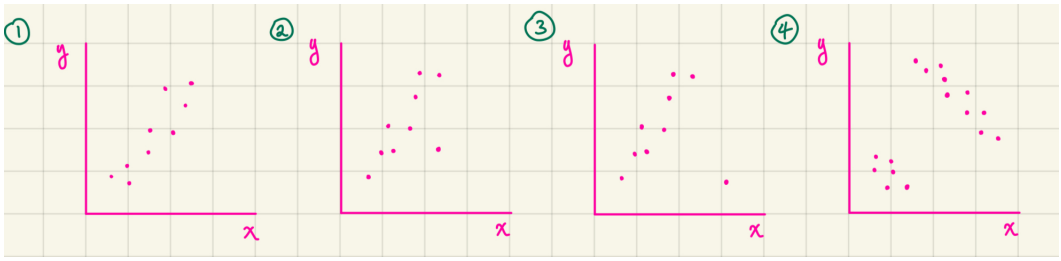
CS 1810 Concept Checks

1 Linear Regression

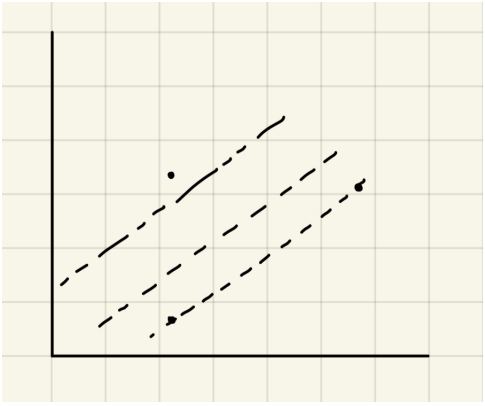
1. What would the linear regression model look like for each of the following datasets?



2. Consider the following distributions for the noise in a probablistic linear regression model:

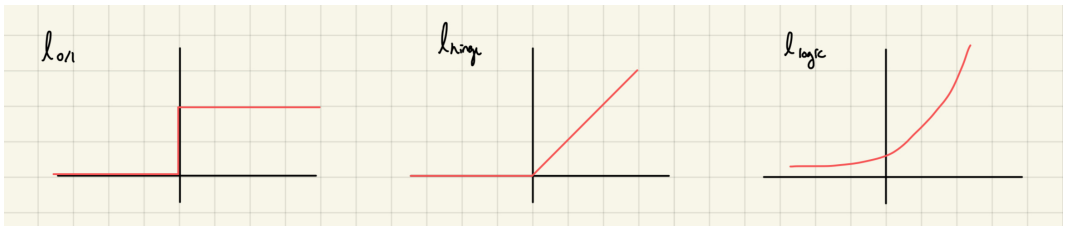
   - **Laplace** ($\lambda = 1$): $p(\epsilon) \propto \exp(-|\epsilon|)$
   - **Gaussian** ($\sigma^2 = 1$): $p(\epsilon) \propto \exp(-\frac{1}{2}\epsilon^2)$
   - **Student-t** ($\nu = 1$): $p(\epsilon) \propto (1 + \epsilon^2)^{-1}$

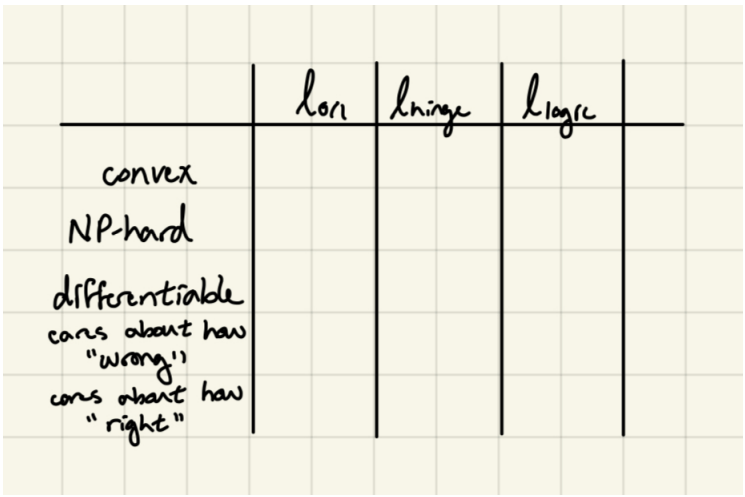
   For each of the dotted lines below, explain which distribution would correspond with that particular model.



2 Classification

3. Fill out the following table to describe the different types of classification loss functions.

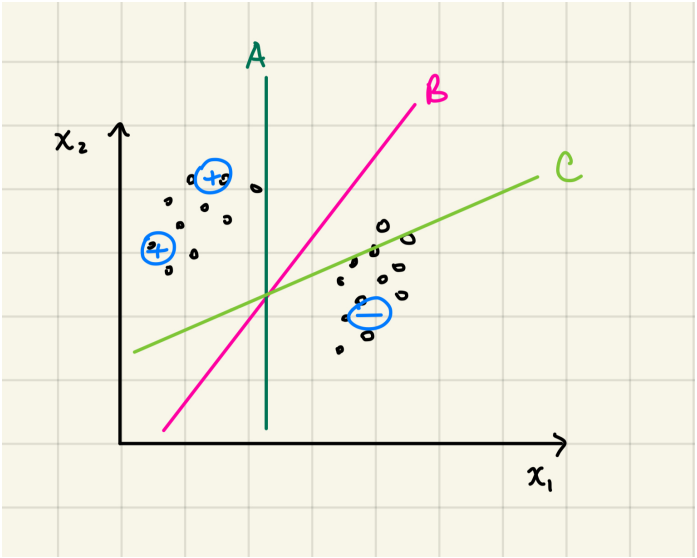




4. Consider the following semi-supervised data set, meaning some of the points are labeled and some of them are not. We have three possible decision boundaries, lines A, B, and C. All of the boundaries are correct given labeled data.

   (a) Does seeing unlabeled data effect your preference?

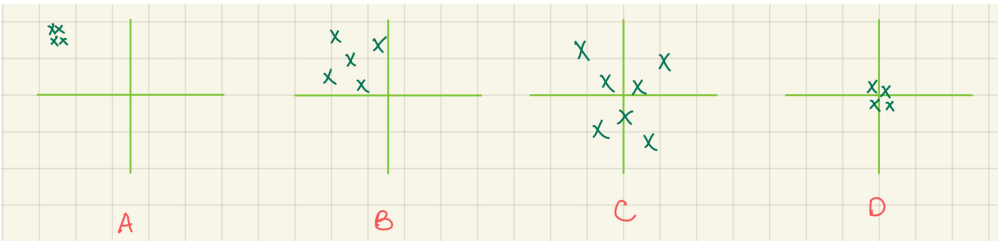
   (b) Does it matter/justify with respect to discriminative story vs. generative story?



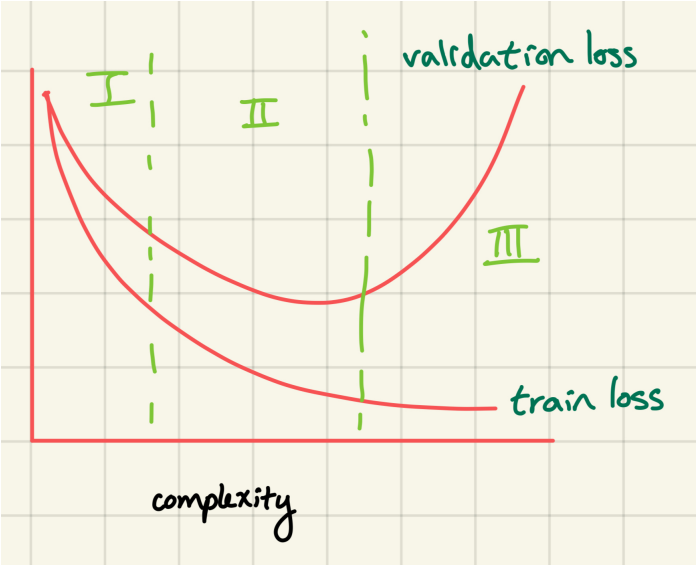
3 Model Selection and Neural Networks

5. Two questions:

   (a) For each of the four graphs below, our goal is the center of the graph and the green X's represent our prediction. Describe whether each graph has 1) high or low bias and 2) high or low variance.



(b) Label each out of I, II, III as overfitting, underfitting, or good fit.



6. Suppose we have two data points $(0, 0), (1, 1)$. Now consider the following 3 models:

   - Model 1: $\hat{y} = a_0$
   - Model 2: $\hat{y} = a_0 + a_1 x$
   - Model 3: $\hat{y} = a_0 + a_1 x + a_2 x^2$

   We also have the priors

   $$p(a_0) = \text{Unif}(-100, 100)$$
   $$p(a_1) = \text{Unif}(-1, 1)$$
   $$p(a_2) = \text{Unif}(-100, 100)$$

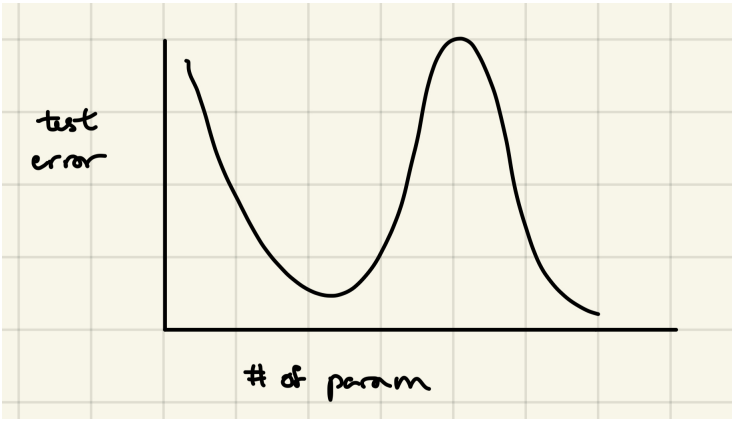
   What is the posterior predictive at $x = 1/2$?

7. A couple of questions on bias-variance for neural networks:

   (a) As the number of parameters increases, can bias increase?

   (b) If the number of parameters is way less than the number of data points, can we fit the data perfectly? If the number of parameters is equal to the number of data points, can we fit the data perfectly? If the number of parameters is way greater than the number of data points, can we fit the data perfectly?

(c) Examine the following graph:



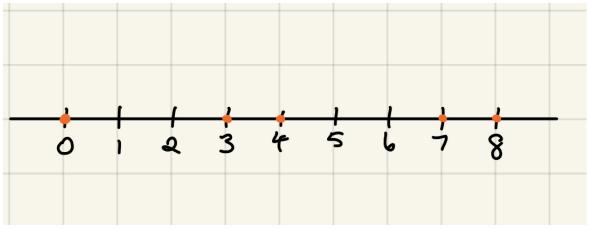
Why might we see this second descent in the loss?

4 Support Vector Machines

8. We are considering the regularization parameter C in the soft margin SVM formulation. For each of the following questions, answer with small C or big C:
   (a) Which would you choose if the goal is to maximize the margin?
   (b) Which would you choose if the goal is to minimize error on the training set?
   (c) Which would you choose if the goal is to overfit?
   (d) Which would you choose if the goal is to underfit?
   (e) Which would you choose if the goal is to have a less "flat" boundary?
   (f) Which would you choose if the goal is to have a more "flat" boundary?
9. A couple of questions on overfitting in SVMs:
   (a) What would happen if we use the RBF kernel with a small σ^2? (in terms of over and underfitting)
   (b) What would happen if we use the RBF kernel with a large σ^2? (in terms of over and underfitting)
   (c) How do we know if an SVM is overfitting?

5 Clustering and Mixture Models

10. Examine the following data:

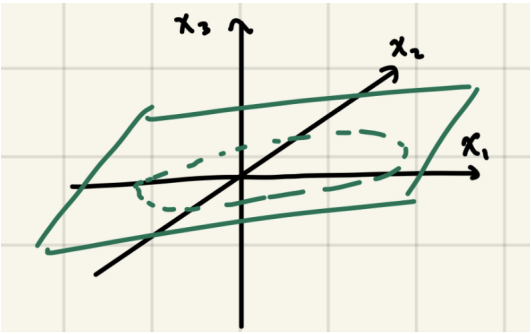


(a) HAC with average linkage, what does my cluster look like?

(b) Initialize K-means with $\{0\}, \{3, 4, 7, 8\}$. Does the solution change?

6 Principal Component Analysis

11. Examine the following image of data living on the dotted circle, and suppose $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$ are the PCs.



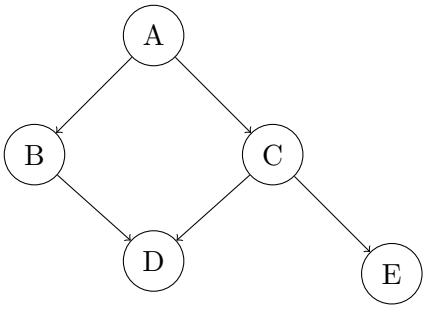
(a) Order the corresponding eigenvalues $\lambda_1, \lambda_2, \lambda_3$ from largest to smallest.

(b) Suppose we project the PCs $\boldsymbol{V}_{1:3}$ using orthogonal matrix $\boldsymbol{Q}$, i.e. our principal component matrix becomes $\boldsymbol{V}_{1:3}\boldsymbol{Q}$. Does the loss change compared to what it was before?

(c) Does $\boldsymbol{V}_{1:3}\boldsymbol{Q}$ still contain the directions of greatest variance in the data?

7 Topic Models and Graphical Models

12. Suppose we have the following graph:



How many parameters are needed to describe the model if:

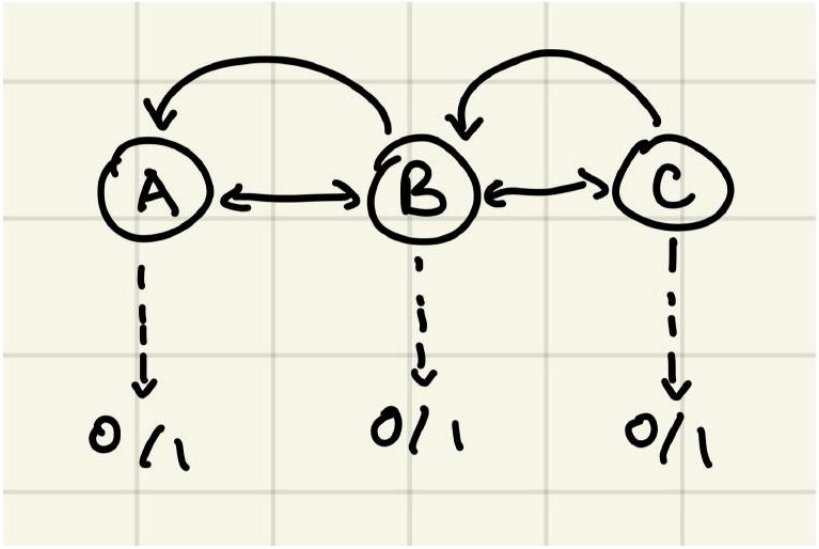
(a) All the variables are discrete over 4 categories.

(b) The variables are 1D Gaussian, i.e. $x_A \sim N(\mu_A, \sigma^2), x_B \sim N(w_{AB}x_A, \sigma^2), \ldots$

(c) The variables are 4D Gaussian, i.e. $\boldsymbol{x_A} \sim N(\boldsymbol{\mu_A}, \boldsymbol{\Sigma}), \boldsymbol{x_B} \sim N(\boldsymbol{W_{AB}x_A}, \boldsymbol{\Sigma}), \ldots$

8 Hidden Markov Models

13. We have the following HMM:



Let us say initially, the probability is concentrated on the A state. We have initial probabilities:

$$\boldsymbol{\theta} = p(\boldsymbol{s}_0) = \begin{array}{c} \\ \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \end{array}$$

Then, we have a transition matrix that tells us the probability from moving from one state to another:

$$\boldsymbol{T} = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/4 & 1/2 & 1/4 \\ 0 & 1/2 & 1/2 \end{pmatrix} \end{array}$$

Finally, we have the emission matrix which gives the probability of observing 0 or 1 given the latent state:

$$\boldsymbol{\pi} = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \\ 1 & 0 \end{pmatrix} \end{array}$$

(a) Assume at time $t = 1$, you observe 0 . What is your $p(\boldsymbol{s}_1 | \boldsymbol{x}_1 = 0)$?

(b) Assume at time $t = 2$, you observe 0 . What is your $p(\boldsymbol{s}_2 | \boldsymbol{x}_1 = 0, \boldsymbol{x}_2 = 0)$?

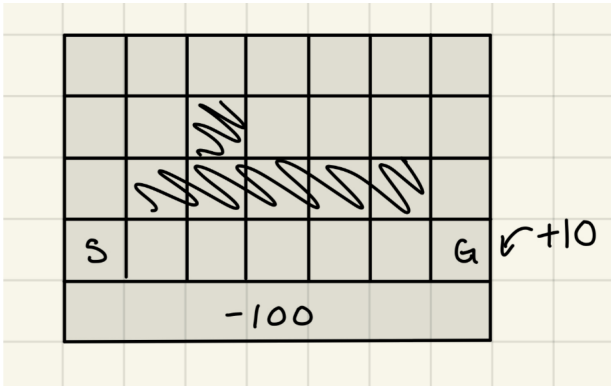
(c) Write the Viterbi path for the observed data $\boldsymbol{x}_1 = 0, \boldsymbol{x}_2 = 0$.

9 Markov Decision Processes

14. Suppose we replace the reward $r(s, a)$ with $r(s, a) + \beta$. Would the optimal policy change?

10 Reinforcement Learning

15. Consider the following grid world:



There is a cliff and you do not want to fall off the bottom of the cliff. Uncertainty is coming from two places: (1) is the world and (2) is our choice to use the ϵ-greedy method. Say that our actions err with probability δ, meaning that the action taken is different from what was intended by the policy.

(a) If $\delta \approx 0$, what is π^*?

(b) If δ is moderate, what is π^*?

(c) Suppose $\delta \approx 0$ and ϵ is moderate. What does SARSA learn?

(d) Suppose $\delta \approx 0$ and ϵ is moderate. What does Q-learning learn?