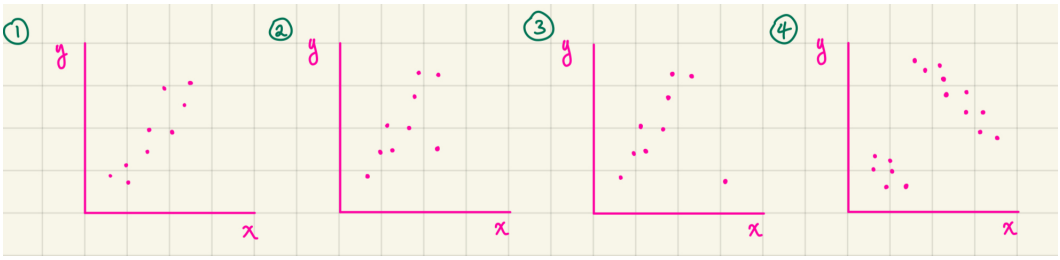
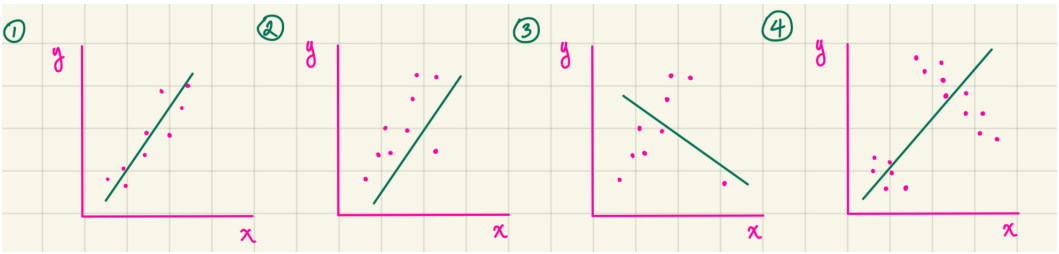
CS 1810 Concept Checks

1 Linear Regression

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



Solution:



(1) The line of best fit has a positive slope.

(2) The line of best fit still has a positive slope but the outlier pulls the line to the right.

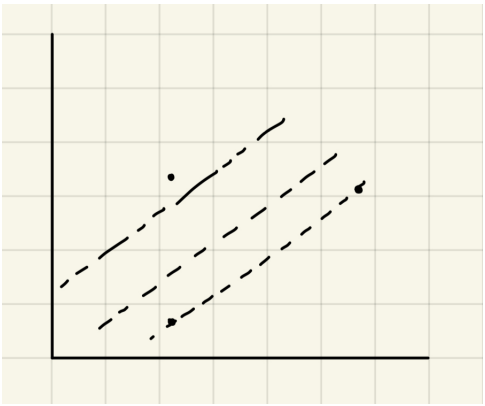
(3) When the outlier is very far out, it can mess with the line of best fit, causing the slope to flip.

(4) We can see that the linear regression model does not do a good job expressing the dataset. But the best that a linear regression model can do is capture the two clusters by having a positive slope.

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.

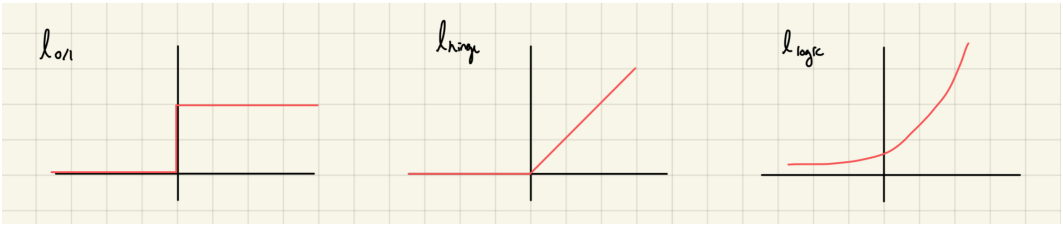


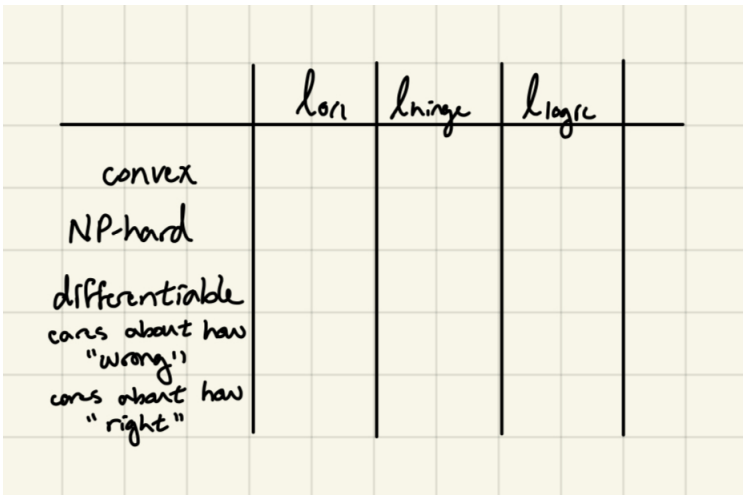
Solution:

To determine the order, we must consider how these distributions penalize the outlier. We can do this by examining how the distributions grow with ϵ. In particular, let's take the natural log of all 3. We get that the Laplace becomes $-|\epsilon|$, the Gaussian becomes $-\frac{1}{2}\epsilon^2$, and the Student-t becomes $-\log(1+\epsilon^2)$. We see that the order of the magnitude of these terms (from small to large) when $\epsilon \to \infty$ is Student-t, Laplace, and then Gaussian. Hence, we expect the lines from bottom to top to correspond to this order, since being further from the outlier corresponds to a larger ϵ.

2 Classification

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





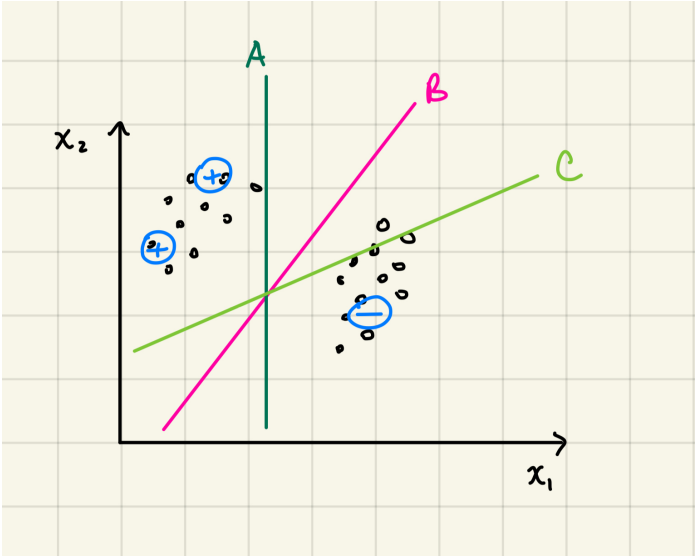
Solution:

Note that the l_{hinge} could also be considered differentiable if we exclude the singular point that is not differentiable. The 0/1 loss will not care how "right" or "wrong" a model prediction is, it only cares about whether or not the prediction is right or wrong. The l_{hinge} and l_{logic} graphs both give a higher penalty the more wrong an output is, so the functions "care" about how wrong an output is. Only the l_{logic} graphs "cares" about how right an output is.

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?

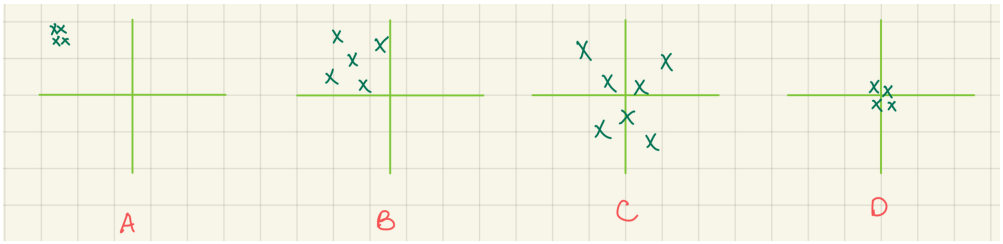


Solution: B is the best boundary. With boundary B, there is an implicit assumption that all the data points in the upper left corner are in the positive class and the data points in the bottom right corner are in the negative class. With the generative story, the geometry of the labels determine how the $\boldsymbol{x}$'s are classified. In the discriminative story, $\boldsymbol{x}$ makes y, and it is hard for us to tell the story when we only have a few data points labeled. This illustrates why the generative model might be beneficial. In defense of the discriminative model, the discriminative model is simple. So both models have their pros and cons.

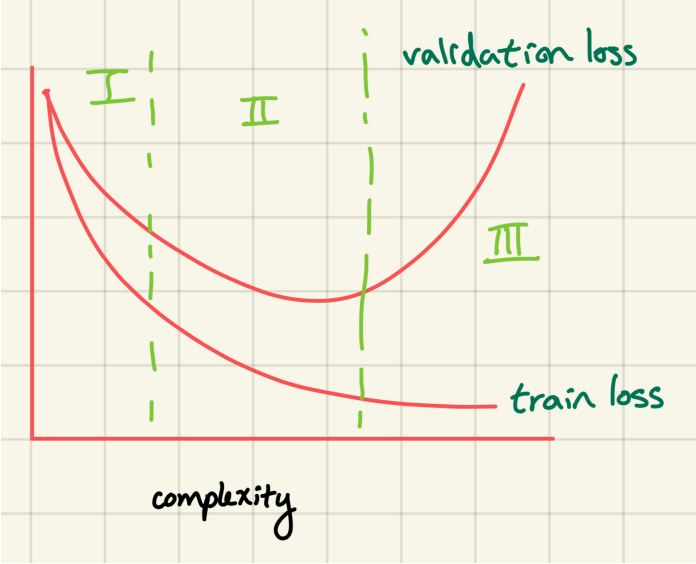
3 Model Selection and Neural Networks

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.



Solution:

(a) Graph A is high bias, because the X's are far from the center. Graph A is low variance because the X's are all close to each other. Graph B is high bias and high variance. Graph C is low bias and high variance. Graph D is low bias and low variance.

(b) For the second question: Graphs that lie in section I. are underfitting. When the loss from the training data and the loss from the validation set are both high, then the graph is underfitting. Graphs in III. are overfitting. A good sign that your model is overfitting is when the train loss is low but the validation set loss is high. This means that the model is overfitting to the training dataset but performs poorly on the validation set. Graphs in II. are therefore a good fit ; we want the training and validation losses to be relatively low.

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 $x = 1/2$?

Solution: We first note that Model 1 is incompatible with the data since it states that the output is a constant, yet this was not reflected in the data. For Model 2, we see that we can simply fit a_0, a_1 via the system of equations defined by the two datapoints. This yields

$a_0 = 0, a_1 = 1$. Hence, the posterior predictive at $x = 1/2$ yields $\hat{y} = 1/2$. For Model 3, we again see that $a_0 = 0$ is fit. However, we now have that $1 = a_1 + a_2$. Hence,

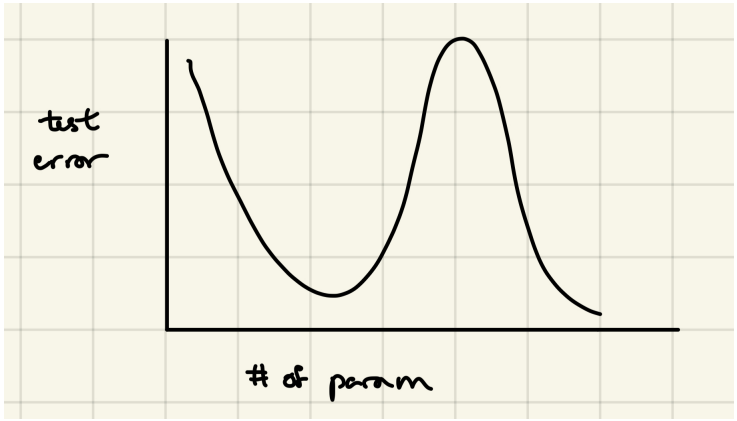
$$\hat{y} = a_0 + a_1/2 + a_2/4 = (a_1 + a_2)/4 + a_1/4 = 1/4 + \text{Unif}(-1, 1)/4 = \text{Unif}(0, 0.5)$$

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?

Solution:

(a) No, the model becomes more flexible.

(b) If the number of parameters is way less than the number of data points, we cannot fit the data perfectly. If the number of parameters is equal to the number of data points, we can fit the model perfectly (system of n unknowns and n equations). There is one model that does this. If the number of parameters is way greater than the number of data points, we can fit the model perfectly in multiple ways!

(c) The idea is that there is some implicit regularization. In particular, when you have lots of functions to chose from, you tend to choose one that is more smooth.

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?

Solution:

(a) Small C because it is less sensitive to outliers.

(b) Large C to avoid misclassifications.

(c) Large C.

(d) Small C.

(e) Large C.

(f) Small C.

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?

Solution:

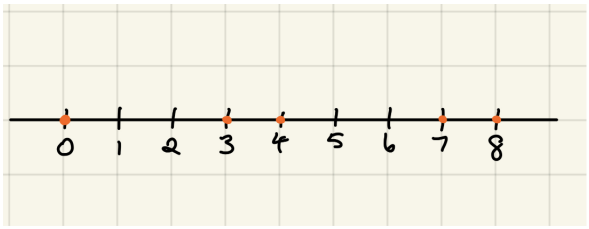
(a) Under this kernel, we can choose large, nonzero $\alpha^{(n)}$'s without incurring a large penalty due to the fact that the kernel will be very small except for points that are very similar to each other. Nearly every point will thus be a support vector, and we end up with a very jagged line in our model. This is overfitting.

(b) Now, we have fewer support points and larger margin and a flatter decision boundary. So, here we have underfitting.

(c) Signs of overfitting: lots of support points, lots of points n for which $\alpha^{(n)} > 0$, smaller margins. Cross validation can be used to check if we have overfitting.

5 Clustering and Mixture Models

10. Examine the following data:

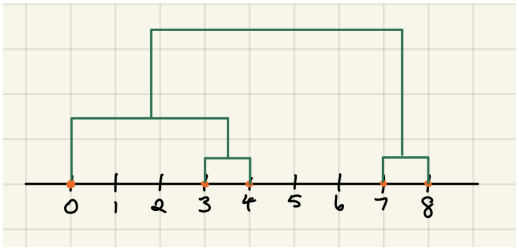


(a) HAC with average linkage, what does my cluster look like?

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

Solution:

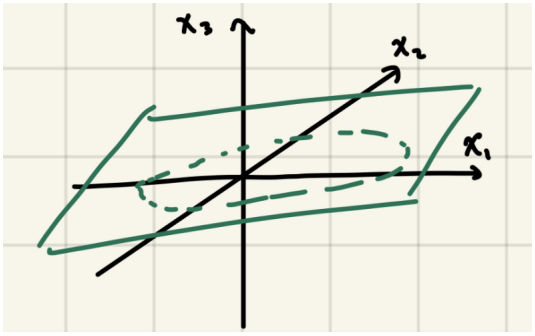
(a) See the following image:



(b) The solution will not change! The only thing that can change is 3 moving to cluster with 0.

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?

Solution:

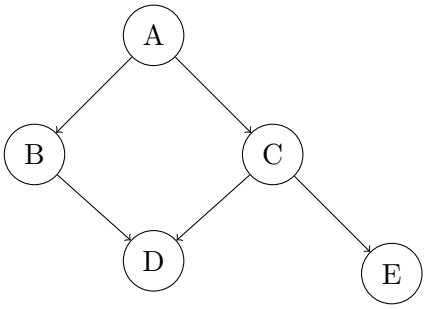
(a) $\lambda_1, \lambda_2, \lambda_3$.

(b) No because the reconstruction loss uses $\boldsymbol{V}_{1:3}\boldsymbol{Q}\left(\boldsymbol{V}_{1:3}\boldsymbol{Q}\right)^\top = \boldsymbol{V}_{1:3}\boldsymbol{V}_{1:3}^\top$.

(c) Not, unless $\boldsymbol{Q}$ is the identity matrix.

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. $\boldsymbol{x_A} \sim N(\boldsymbol{\mu_A}, \boldsymbol{\Sigma}), \boldsymbol{x_B} \sim N(\boldsymbol{W_{AB}x_A}, \boldsymbol{\Sigma}), \ldots$

Solution:

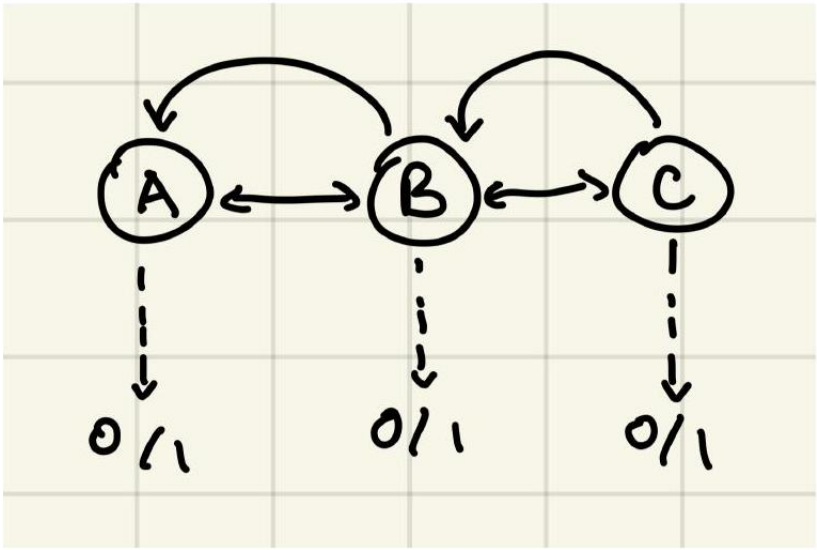
(a) We need 3 parameters for A. B, C each need $4 \times 3 = 12$. D needs $4 \times 4 \times 3 = 48$. Finally, E needs $4 \times 3 = 12$. This is a total of 87 parameters!

(b) We only need $\mu_A, w_{AB}, w_{AC}, w_{BD}, w_{CD}, w_{CE}$. This is a total of 6 parameters.

(c) We need $\boldsymbol{\mu_A}, \boldsymbol{W_{AB}}, \boldsymbol{W_{AC}}, \boldsymbol{W_{BD}}, \boldsymbol{W_{CD}}, \boldsymbol{W_{CE}}$. The first quantity contains 4 entries, and the latter 5 quantities each contain $4 \times 4 = 16$ parameters. So we have a total of 84 parameters!

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:

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

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

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

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

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

(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 $\boldsymbol{x}_1 = 0, \boldsymbol{x}_2 = 0$.

Solution:

(a) The starting matrix is: We have

$$\boldsymbol{\theta T} = \begin{pmatrix} 1/2 & 1/2 & 0 \end{pmatrix}$$

which gives us $p(\boldsymbol{s}_1)$. Now to get $p(\boldsymbol{s}_1|\boldsymbol{x}_1 = 0)$, we find the element-wise product of the above matrix with the first column of $\boldsymbol{\pi}$. This gives us

$$\begin{pmatrix} 0 & 1/4 & 0 \end{pmatrix}$$

which we can normalize to get

$$\begin{pmatrix} 0 & 1 & 0 \end{pmatrix}$$

as the final answer.

(b) We take a very similar approach to the previous part: First,

$$\begin{pmatrix} 0 & 1 & 0 \end{pmatrix} \boldsymbol{T} = \begin{pmatrix} 1/4 & 1/2 & 1/4 \end{pmatrix}$$

which gives us $p(\boldsymbol{s}_2|\boldsymbol{x}_1 = 0)$. Now to get $p(\boldsymbol{s}_2|\boldsymbol{x}_1 = 0, \boldsymbol{x}_2 = 0)$, we again find the element-wise product of the above matrix with the first column of $\boldsymbol{\pi}$. This gives us

$$\begin{pmatrix} 0 & 1/4 & 1/4 \end{pmatrix}$$

which we can normalize to get

$$\begin{pmatrix} 0 & 1/2 & 1/2 \end{pmatrix}$$

as the final answer.

(c) Using the matrices from part 1 and part 2, we see that B, B or B, C both work.

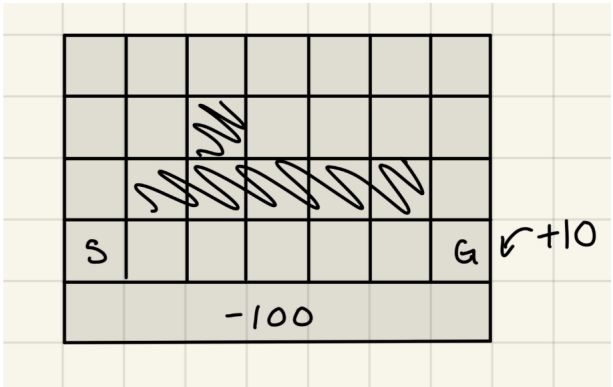
9 Markov Decision Processes

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

Solution: No, since if we previously had the objective $V(s_0)$, we would now just have $V(s_0) + \beta \sum_t \gamma^t$, where the right term is not affected by the choice of policy π.

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?

Solution:

(a) With no risk of falling off, the agent should just walk straight across.

(b) Given the risk of falling off, the agent should walk up, right, down.

(c) SARSA's update step will bake in the danger of falling off the cliff. Hence, it will learn to walk up, right, down.

(d) Q-learning's update step leads it to learn the policy of walking straight across, since the best case future action at each step is to continue along the path (rather than walk off).