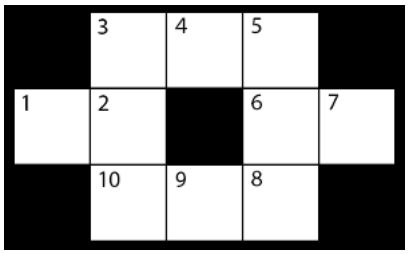
Name: _____ Student ID: _____

1. (29 points) Particle Filtering

   In this question, we will use a particle filter to track the state of a robot that is lost in the small map below:



   The robot's state is represented by an integer $1 \leq X_t \leq 10$ corresponding to its location in the map at time t. We will approximate our belief over this state with $N = 8$ particles.

   You have no control over the robot's actions. At each timestep, the robot either stays in place, or moves to any one of its neighboring locations, all with equal probability. For example, if the robot starts in state $X_t = 7$, it will move to state $X_{t+1} = 6$ with probability $\frac{1}{2}$ or $X_{t+1} = 7$ with probability $\frac{1}{2}$. Similarly, if the robot starts in state $X_t = 2$, the next state X_{t+1} can be any element of $\{1, 2, 3, 10\}$, and each occurs with probability $\frac{1}{4}$.

   At each time step, a sensor on the robot gives a reading $E_t \in \{H, C, T, D\}$ corresponding to the *type* of state the robot is in. The possible types are:

   * Hallway (H) for states bordered by two parallel walls (4,9). * Corner (C) for states bordered by two orthogonal walls (3,5,8,10). * Tee (T) for states bordered by one wall (2,6). * Dead End (D) for states bordered by three walls (1,7).

   The sensor is not very reliable: it reports the correct type with probability $\frac{1}{2}$, but gives erroneous readings the rest of the time, with probability $\frac{1}{6}$ for each of the three other possible readings.

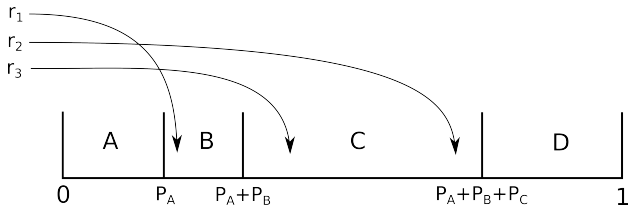
   (a) (4 points) Sensor Model

   Fill in the sensor model below:

   P(Sensor Reading = H | State Type = H) = _____
   P(Sensor Reading = C | State Type = H) = _____
   P(Sensor Reading = T | State Type = H) = _____
   P(Sensor Reading = D | State Type = H) = _____
   P(Sensor Reading = H | State Type = C) = _____
   P(Sensor Reading = C | State Type = C) = _____
   P(Sensor Reading = T | State Type = C) = _____
   P(Sensor Reading = D | State Type = C) = _____
   P(Sensor Reading = H | State Type = T) = _____
   P(Sensor Reading = C | State Type = T) = _____
   P(Sensor Reading = T | State Type = T) = _____
   P(Sensor Reading = D | State Type = T) = _____
   P(Sensor Reading = H | State Type = D) = _____
   P(Sensor Reading = C | State Type = D) = _____
   P(Sensor Reading = T | State Type = D) = _____
   P(Sensor Reading = D | State Type = D) = _____

(b) (4 points) Sampling Review and Initial Belief State

Suppose that we want to sample from a set of 4 events, $\{A, B, C, D\}$, which occur with corresponding probabilities P_A, P_B, P_C, P_D. First, we form the set of cumulative weights, given by $\{0, P_A, P_A + P_B, P_A + P_B + P_C, 1\}$. These weights partition the $[0, 1)$ interval into bins, as shown below. We then draw a number r uniformly at random from $[0, 1)$ and pick A, B, C, or D based on which bin r lands in. The process is illustrated in the diagram below. If r_1, uniformly chosen from $[0, 1)$, lands in the interval $[P_A, P_A + P_B]$, then the resulting sample would be B. Similarly, if r_2 lands in $[P_A + P_B, P_A + P_B + P_C]$, the sample would be C, and r_3 landing in $[P_A + P_B, P_A + P_B + P_C]$ would also be C.



Now we will sample the starting positions for our particles at time $t = 0$. For each particle p_i, we have generated a random number r_i sampled uniformly from $[0, 1)$. Your job is to use these numbers to sample a starting location for each particle. As a reminder, locations are integers from the range $[1, 10]$, as shown in the map. You should assume that the locations go in ascending order and that each location has equal probability. The random number generated for particle i, denoted by r_i, is provided. Please fill in the locations of the eight particles.

$r_1 = 0.914$ $\quad p_1 =$ _____ $\quad r_2 = 0.473$ $\quad p_2 =$ _____

$r_3 = 0.679$ $\quad p_3 =$ _____ $\quad r_4 = 0.879$ $\quad p_4 =$ _____

$r_5 = 0.212$ $\quad p_5 =$ _____ $\quad r_6 = 0.024$ $\quad p_6 =$ _____

$r_7 = 0.458$ $\quad p_7 =$ _____ $\quad r_8 = 0.154$ $\quad p_8 =$ _____

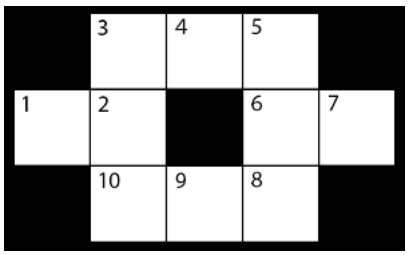
At this point, it is *highly recommended* that you copy down the starting locations for each particle as you will need them to answer Part 1(c).

(c) (4 points) Time Update

Now we'll perform a time update from $t = 0$ to $t = 1$ using the transition model. Stated again, the transition model is as follows: At each timestep, the robot either stays in place, or moves to any one of its neighboring locations, all with equal probability.

For each particle, take the starting position you found in Part 1(b), and perform the time update for that particle. You should again sample from the range $[0, 1)$, where the bins are the possible locations **sorted in ascending numerical order**. As an example, if $X_t = 2$, the next state can be one of $\{1, 2, 3, 10\}$, each with equal probability, so the $[0, 0.25)$ bin would be for $X_{t+1} = 1$, the $[0.25, 0.5)$ bin would be for $X_{t+1} = 2$, the $[0.5, 0.75)$ bin would be for $X_{t+1} = 3$, and the $[0.75, 1)$ bin would be for $X_{t+1} = 10$.

The map is shown again below:



$r_1 = 0.674$	$p_1 =$ _____	$r_2 = 0.119$	$p_2 =$ _____
$r_3 = 0.748$	$p_3 =$ _____	$r_4 = 0.802$	$p_4 =$ _____
$r_5 = 0.357$	$p_5 =$ _____	$r_6 = 0.736$	$p_6 =$ _____
$r_7 = 0.425$	$p_7 =$ _____	$r_8 = 0.058$	$p_8 =$ _____

At this point, it is **highly recommended** that you copy down the new locations for each particle as you will need them to answer Parts 1(d)–(f).

(d) (5 points) Probability Distribution Induced by the Particles

Recall that a particle filter just keeps track of a list of particles, but at any given time, we can compute a probability distribution from these particles. Using the current newly updated set of particles (that you found in Part 3) , give the estimated probability that the robot is in each location.

$P(X_1 = 1) =$ _____	$P(X_1 = 6) =$ _____
$P(X_1 = 2) =$ _____	$P(X_1 = 7) =$ _____
$P(X_1 = 3) =$ _____	$P(X_1 = 8) =$ _____
$P(X_1 = 4) =$ _____	$P(X_1 = 9) =$ _____
$P(X_1 = 5) =$ _____	$P(X_1 = 10) =$ _____

(e) (5 points) Incorporating Evidence

The sensor reading at $t = 1$ is: $E_1 = D$

Using the sensor model you specified in Part 1, incorporate the evidence by reweighting the particles. Also enter the normalized and cumulative weights for each particle. The normalized weight for a specific particle can be calculated by taking that particle's weight and dividing by the sum of all the particle weights. The cumulative weight keeps track of a running sum of all the weights of the particles seen so far (meaning, particle i will have a cumulative weight equal to the sum of the weights of all particles j such that $j \leq i$).

Refer back to Part 3 to get the positions of your particles.

The map is shown again below:

Particle p_1 weight: _____	Particle p_5 weight: _____
p_1 normalized weight: _____	p_5 normalized weight: _____
p_1 cumulative weight: _____	p_5 cumulative weight: _____
Particle p_2 weight: _____	Particle p_6 weight: _____
p_2 normalized weight: _____	p_6 normalized weight: _____
p_2 cumulative weight: _____	p_6 cumulative weight: _____
Particle p_3 weight: _____	Particle p_7 weight: _____
p_3 normalized weight: _____	p_7 normalized weight: _____
p_3 cumulative weight: _____	p_7 cumulative weight: _____
Particle p_4 weight: _____	Particle p_8 weight: _____
p_4 normalized weight: _____	p_8 normalized weight: _____
p_4 cumulative weight: _____	p_8 cumulative weight: _____

(f) (4 points) Resampling

Finally, we'll resample the particles. This reallocates resources to the most relevant parts of the state space in the next time update step.

Notice that your cumulative weights effectively tell you where the bins used in resampling the particles lie. For example, for particle 1, you calculated the cumulative weight to be some value, p. Then, on a random value draw, if a value between 0 and p was chosen, you would generate a new particle where particle 1 is. Use these bounds to resample the eight particles. In the "New Particle" row, enter the particle corresponding to the bin that the random value chose. In the "New Location" row, enter the location corresponding to this new particle. You may need to look back at Part 3 to get the locations of the particles.

$r_1 = 0.403$
New particle for p_1: _____
New location for p_1: _____

$r_2 = 0.218$
New particle for p_2: _____
New location for p_2: _____

$r_3 = 0.217$
New particle for p_3: _____
New location for p_3: _____

$r_4 = 0.826$
New particle for p_4: _____
New location for p_4: _____

$r_5 = 0.717$
New particle for p_5: _____
New location for p_5: _____

$r_6 = 0.460$
New particle for p_6: _____
New location for p_6: _____

$r_7 = 0.794$
New particle for p_7: _____
New location for p_7: _____

$r_8 = 0.016$
New particle for p_8: _____
New location for p_8: _____

(g) (3 points) Analysis

The sensor provided a reading $E_1 = D$.

What fraction of the particles are now on a dead end? *Answer.* _____

This completes everything for the first time step, $t = 0 \rightarrow t = 1$. Of course, we would now continue by repeating the time update, evidence incorporation by reweighting, and resampling. We'll leave that to the computers, though.

2. (10 points) Particle Filtering Implementation

Consider the following particle filter implementations.

Default Implementation: Resample after Evidence Incorporation

1. Initialize particles by sampling from initial state distribution.
2. Repeat:
    1. Perform time update
    2. Weight according to evidence
    3. Resample according to weights

Alternative Implementation: Resample after Time Update

1. Initialize particles by sampling from initial state distribution and assigning uniform weights.
2. Repeat:
    1. Perform time update, retaining weights
    2. Resample according to weights
    3. Weight according to evidence

For each of the following statements about the two implementations, select whether they are true or false.

The default implementation will typically provide a better estimate of the distribution than the alternate implementation.

☐ True ☐ False

If the observation model is deterministic then the default implementation will typically provide a better estimate of the distribution than the alternate implementation.

☐ True ☐ False

If the transition model is deterministic then the default implementation will typically provide a better estimate of the distribution than the alternate implementation.

☐ True ☐ False

3. (12 points) Maximum Likelihood Estimation

We will begin with a short derivation. Consider a probability distribution with a domain that consists of $|\ X\ |$ different values. We get to observe N total samples from this distribution. We use n_i to represent the number of the N samples for which outcome i occurs. Our goal is to estimate the probabilities θ_i for each of the events $i = 1, 2, \ldots, |\ X\ | - 1$. The probability of the last outcome, $|\ X\ |$, equals $1 - \sum_{i=1}^{|X|-1} \theta_i$.

In *maximum likelihood estimation* (mle), we choose the θ_i that maximize the likelihood of the observed samples,

$$L(\text{samples}, \theta) \propto (1 - \theta_1 - \theta_2 - \cdots - \theta_{|X|-1})^{n_{|X|}} \prod_{i=1}^{|X|-1} \theta_i^{n_i}.$$

For this derivation, it is easiest to work with the log of the likelihood. Maximizing log-likelihood also maximizes likelihood, since the quantities are related by a monotonic transformation. Taking logs we obtain

$$\log\big[L(\text{samples}, \theta)\big] \propto n_{|X|} \log(1 - \theta_1 - \theta_2 - \cdots - \theta_{|X|-1}) + \Sigma_{i=1}^{|X|-1} n_i \log \theta_i.$$

We denote the maximum-likelihood estimate of θ by θ^{ML}; that is,

$$\theta^{\text{ML}} = \operatorname*{argmax}_{\theta} \{n_{|X|} \log(1 - \theta_1 - \theta_2 - \cdots - \theta_{|X|-1}) + \Sigma_{i=1}^{|X|-1} n_i \log \theta_i\}.$$

To find θ^{ML}, we take derivatives with respect to each θ_i, setting these derivatives equal to zero, and solving for θ_i in each case. Setting derivatives with respect to θ_i equal to zero, we obtain $|\ X\ | - 1$ equations in the $|\ X\ | - 1$ unknowns, $\theta_1, \theta_2, \ldots, \theta_{|X|-1}$:

$$\frac{-n_{|X|}}{1 - \theta_1 - \theta_2 - \cdots - \theta_{|X|-1}} + \frac{n_i}{\theta_i} = 0.$$

Multiplying by $\theta_i(1 - \theta_1 - \theta_2 - \cdots - \theta_{|X|-1})$ makes the original $|\ X\ | - 1$ nonlinear equations into $|\ X\ | - 1$ linear equations:

$$-n_{|X|}\theta_i + n_i \left(1 - \theta_1 - \theta_2 - \cdots - \theta_{|X|-1}\right) = 0.$$

That is, the maximum likelihood estimates, $\theta^{\text{ML}} = (\theta_1^{\text{ML}}, \theta_2^{\text{ML}}, \ldots, \theta_{|X|-1}^{\text{ML}})$, are found by solving a linear system of $|\ X\ | - 1$ equations in $|\ X\ | - 1$ unknowns. Doing so shows that the maximum likelihood estimate corresponds to simply the count for each outcome divided by the total number of samples. I.e., we have that $\theta_i^{\text{ML}} = \frac{n_i}{N}$.

Now, consider a sampling process with 3 possible outcomes: R, G, and B. We observe the following sample counts:

outcome	R	G	B
count	0	3	10

(a) What is the total sample count N? *Answer.* _____

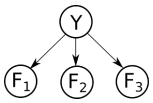
(b) What are the maximum likelihood estimates for the probabilities of each outcome?
$\theta_R^{\text{ML}} =$ _____ $\theta_G^{\text{ML}} \approx$ _____ $\theta_B^{\text{ML}} \approx$ _____

(c) Now, use Laplace smoothing with strength $k = 4$ to estimate the probabilities of each outcome.
$\theta_R^{\text{LAP},4} =$ _____ $\theta_G^{\text{LAP},4} =$ _____ $\theta_B^{\text{LAP},4} =$ _____

(d) Now, consider Laplace smoothing in the limit $k \to \infty$. Fill in the corresponding probability estimates.
$\theta_R^{\text{LAP},\infty} =$ _____ $\theta_G^{\text{LAP},\infty} =$ _____ $\theta_B^{\text{LAP},\infty} =$ _____

4. (24 points) NAÏVE BAYES

In this question, we will train a Naive Bayes classifier to predict class labels Y as a function of input features F_i.



We are given the following 15 training points:

F_1	0	0	1	1	0	0	1	0	0	0	0	1	0	1	0
F_2	0	1	0	1	0	0	1	0	1	1	0	0	0	0	0
F_3	1	1	1	0	1	0	0	0	0	1	1	1	0	1	0
Y	A	A	A	A	A	A	A	A	A	A	A	B	C	C	C

(a) (6 points) What is the maximum likelihood estimate of the prior $P(Y)$?

$P(Y = A)$: _____ $P(Y = B)$: _____ $P(Y = C)$: _____

What are the maximum likelihood estimates of the conditional probability distributions? Fill in the probability values below (tables for the 2nd and 3rd features are done for you).

$P(F_1 = 0 \mid Y = A)$: _____ $P(F_1 = 1 \mid Y = A)$: _____

$P(F_1 = 0 \mid Y = B)$: _____ $P(F_1 = 1 \mid Y = B)$: _____

$P(F_1 = 0 \mid Y = C)$: _____ $P(F_1 = 1 \mid Y = C)$: _____

F_2	Y	$P(F_2\|Y)$
0	A	0.545
1	A	0.455
0	B	1.000
1	B	0.000
0	C	1.000
1	C	0.000

F_3	Y	$P(F_3\|Y)$
0	A	0.455
1	A	0.545
0	B	0.000
1	B	1.000
0	C	0.667
1	C	0.333

(b) (6 points) Now consider a new data point (F_1=1, F_2=1, F_3=1). Use your classifier to determine the joint probability of causes Y and this new data point, along with the posterior probability of Y given the new data:

$P(Y = A, F_1 = 1, F_2 = 1, F_3 = 1)$: _____

$P(Y = B, F_1 = 1, F_2 = 1, F_3 = 1)$: _____

$P(Y = C, F_1 = 1, F_2 = 1, F_3 = 1)$: _____

$P(Y = A \mid F_1 = 1, F_2 = 1, F_3 = 1)$: _____

$P(Y = B \mid F_1 = 1, F_2 = 1, F_3 = 1)$: _____

$P(Y = C \mid F_1 = 1, F_2 = 1, F_3 = 1)$: _____

What label does your classifier give to the new data point? (Break ties alphabetically)

☐ A ☐ B ☐ C

(c) (6 points) The training data is repeated here for your convenience:

F_1	0	0	1	1	0	0	1	0	0	0	0	1	0	1	0
F_2	0	1	0	1	0	0	1	0	1	1	0	0	0	0	0
F_3	1	1	1	0	1	0	0	0	0	1	1	1	0	1	0
Y	A	A	A	A	A	A	A	A	A	A	A	B	C	C	C

Use Laplace Smoothing with strength k = 2 to estimate the prior $P(Y)$ for the same data.

$P(Y = A)$: _____ $P(Y = B)$: _____ $P(Y = C)$: _____

Use Laplace Smoothing with strength k = 2 to estimate the conditional probability distributions below (again, the second two are done for you).

$P(F_1 = 0 \mid Y = A)$: _____ $P(F_1 = 1 \mid Y = A)$: _____

$P(F_1 = 0 \mid Y = B)$: _____ $P(F_1 = 1 \mid Y = B)$: _____

$P(F_1 = 0 \mid Y = C)$: _____ $P(F_1 = 1 \mid Y = C)$: _____

F_2	Y	$P(F_2 \mid Y)$
0	A	0.533
1	A	0.467
0	B	0.600
1	B	0.400
0	C	0.714
1	C	0.286

F_3	Y	$P(F_3 \mid Y)$
0	A	0.467
1	A	0.533
0	B	0.400
1	B	0.600
0	C	0.571
1	C	0.429

(d) (6 points) Now consider again the new data point ($F_1 = 1$, $F_2 = 1$, $F_3 = 1$). Use the Laplace-Smoothed version of your classifier to determine the joint probability of causes Y and this new data point, along with the posterior probability of Y given the new data:

$P(Y = A, F_1 = 1, F_2 = 1, F_3 = 1) =$ _____

$P(Y = B, F_1 = 1, F_2 = 1, F_3 = 1) =$ _____

$P(Y = C, F_1 = 1, F_2 = 1, F_3 = 1) =$ _____

$P(Y = A \mid F_1 = 1, F_2 = 1, F_3 = 1) =$ _____

$P(Y = B \mid F_1 = 1, F_2 = 1, F_3 = 1) =$ _____

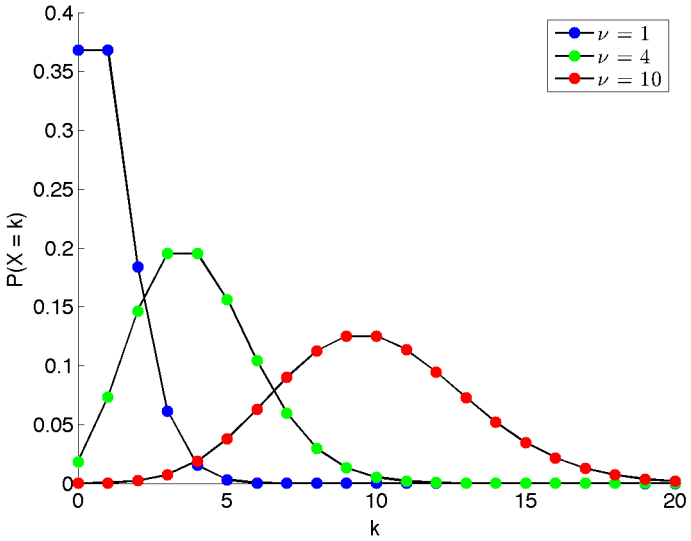
$P(Y = C \mid F_1 = 1, F_2 = 1, F_3 = 1) =$ _____

What label does your (Laplace-Smoothed) classifier give to the new data point? (Break ties alphabetically)

☐ A ☐ B ☐ C

5. (13 points) Poisson Parameter Evaluation

We will now consider maximum likelihood estimation in the context of a different probability distribution. Under the Poisson distribution, the probability of an event occuring $X = k$ times is: $P(X = k) = \frac{\nu^k e^{-\nu}}{k!}$. Here ν is the *parameter* we wish to estimate. The distribution is plotted for several values of ν below.



On a sheet of scratch paper, work out the maximum likelihood estimate for ν, given observations of several k_i. Hints: start by taking the product of the equation above over all the k_i, and then taking the log. Then, differentiate with respect to ν, set the result equal to 0, and solve for ν in terms of the k_i.

You observe the samples $k_1 = 5$, $k_2 = 6$, $k_3 = 2$, $k_4 = 2$, $k_5 = 5$. What is your maximum likelihood estimate of ν?

Answer. _____

6. (12 points) Datasets

When training a classifier, it is common to split the available data into a training set, a hold-out set, and a test set, each of which has a different role.

Which data set is used to learn the conditional probabilities?

☐ Training Data ☐ Hold-Out Data ☐ Test Data

Which data set is used to tune the Laplace Smoothing hyperparameters?

☐ Training Data ☐ Hold-Out Data ☐ Test Data

Which data set is used for quantifying performance results?

☐ Training Data ☐ Hold-Out Data ☐ Test Data