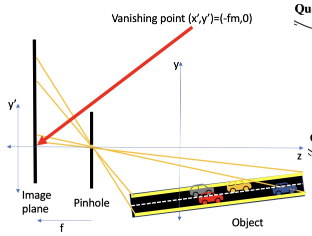
Vanishing point

- Plug equations for the lines into the pinhole camera equations:

$$\frac{x_1'}{f} = -\frac{az + c_1}{z}, \quad \frac{y_1'}{f} = -\frac{bz + d_1}{z}$$
$$\frac{x_2'}{f} = -\frac{az + c_2}{z}, \quad \frac{y_2'}{f} = -\frac{bz + d_2}{z}$$

- As $z \to \infty$, the two lines converge to the vanishing point, which depends only on the ***slope*** of the lines, not on their shift:

$$(x', y') = (-fa, -fb)$$



Question 7 *(7 points)*

You are standing on a downward-sloping hillside, with your camera pointed straight ahead of you. Parallel to your line of sight, on your left-hand side (at position $x = -2$ meters), there is a low fence (height 1 meter). The fence descends the hill in front of you, vanishing into a point far in the distance. Let (x', y') denote the position of the fence's vanishing point on your photograph, where x' is horizontal position, y' is vertical position, and $(0,0)$ is the point directly corresponding to your line of sight.

- Is $x' < 0$, $x' = 0$, or $x' > 0$? Explain.
- Is $y' < 0$, $y' = 0$, or $y' > 0$? Explain.

Solution:

- $x' = 0$. The fence is parallel to your line of sight, thus $x = -b$ for some offset b. If we divide by z, and let z go to infinity, we find that $x' = 0$.
- $y' > 0$. The fence is descending, so $y = az + c$ for some negative value of a. Dividing by z, substituting $y/z = -y'/f$, and letting z go to infinity, we find that $y' = -af$, which is positive.

(c) (3 points) Write the transition probability table $P(s'|s,a)$.

Solution:

s'

(s,a)	0	1	2	3
(0,1)	0	1	0	0
(0,2)	1/2	0	1/2	0
(1,1)	0	0	1	0
(1,2)	0	1/2	0	1/2
(2,1)	0	0	0	1
(2,2)	0	0	1/2	1/2

(d) (2 points) Use value iteration to find $U(s)$, the utility of each state, assuming a discount factor of $\gamma = 1$.

Solution: The utility estimate at each iteration is given as follows; after $t = 5$, the utility no longer changes.

t	0	1	2	3
1	0	0	0	0
2	-1	-1	-1	10
3	-2	3.5	9	10
4	2.5	8	9	10
5	7	8	9	10

$R(0) = R(1) = R(2) = -1 \quad R(3) = 10$

$-1 + \gamma\max\left(1\cdot(-1), \frac{1}{2}\cdot(-1) + \frac{1}{2}(-1)\right) = -2$

$-1 + \gamma\max\left(1\cdot(-1), \frac{1}{2}(-1) + \frac{1}{2}(10)\right) = 3.5$

Question 17 *(0 points)*

After t iterations of the "Value Iteration" algorithm, the estimated utility $U(s)$ is a summation including terms $R(s')$ for the set of states s' that can be reached from state s in at most $t-1$ steps.

√ **True**

Question 19 *(10 points)*

A cat lives in a two-room apartment. It has two possible actions: purr, or walk. It starts in room $s_0 = 1$, where it receives the reward $r_0 = 2$ (petting). It then implements the following sequence of actions: a_0 =walk, a_1 =purr. In response, it observes the following sequence of states and rewards: $s_1 = 2$, $r_1 = 5$ (food), $s_2 = 2$.

(a) (3 points) The cat starts out with a Q-table whose entries are all $Q(s,a) = 0$, then performs one iteration of TD-learning using each of the two SARS sequences described above (one iteration/time step, for two time steps). Because the cat doesn't like to worry about the distant future, it uses a relatively high learning rate ($\alpha = 0.05$) and a relatively low discount factor ($\gamma = \frac{3}{4}$). Which entries in the Q-table have changed, after this learning, and what are their new values?

Solution:

- $t = 0$:

$$Q_{local} = r_0 + \gamma \max_a Q(s_1, a) = 2 + 0 = 2$$
$$Q(1,\text{walk}) \leftarrow Q(1,\text{walk}) + \alpha(Q_{local} - Q(1,\text{walk}))$$
$$= 0 + 0.05(2 - 0) = 0.1$$

- $t = 1$:

$$Q_{local} = r_1 + \gamma \max_a Q(s_2, a) = 5 + 0 = 5$$
$$Q(2,\text{purr}) \leftarrow Q(2,\text{purr}) + \alpha(Q_{local} - Q(2,\text{purr}))$$
$$= 0 + 0.05(5 - 0) = 0.25$$

So the changed values are $Q(1,\text{walk}) \leftarrow 0.1$ and $Q(2,\text{purr}) \leftarrow 0.25$.

(b) (2 points) Instead of model-free learning, the cat decides to implement model-based learning. It estimates $P(s'|s,a)$ using Laplace smoothing, with a smoothing parameter of $k = 1$, using the two SARS observations listed at the start of this problem. What are the new values of $P(s'|s = 2, a = \text{purr})$ for $s' \in \{1,2\}$?

Solution:

$$P(s' = 1|s = 2, a = \text{purr}) = \frac{1 + \text{Count}(s_t = 2, a_t = \text{purr}, s_{t+1} = 1)}{2 + \sum_{s'} \text{Count}(s_t = 2, a_t = \text{purr}, s_{t+1} = s')} = \frac{1}{3}$$
$$P(s' = 2|s = 2, a = \text{purr}) = \frac{1 + \text{Count}(s_t = 2, a_t = \text{purr}, s_{t+1} = 2)}{2 + \sum_{s'} \text{Count}(s_t = 2, a_t = \text{purr}, s_{t+1} = s')} = \frac{2}{3}$$

(c) (3 points) After many rounds of model-based learning, the cat has deduced that $R(1) = 2$, $R(2) = 5$, and $P(s'|s,a)$ has the following table:

a:	purr		walk	
s:	1	2	1	2
$P(s' = 1\|s,a)$	2/3	1/3	1/3	2/3
$P(s' = 2\|s,a)$	1/3	2/3	2/3	1/3

Question 20 *(0 points)*

What is the optimal policy defined by the Bellman equation?

Solution:

$$\pi^*(s) = \arg\max_a \sum_{s'} P(s'|s,a)U(s')$$

Question 21 *(0 points)*

When we apply the Q-learning algorithm to learn the state-action value function, one big problem in practice may be that the state space of the problem is continuous and high-dimensional. Discuss at least two possible methods to address this.

Solution:

1. Discretize the state space.
2. Design a lower-dimensional set of discrete features to represent the states.
3. Use a parametric approximator (e.g., a neural network) to estimate the Q function values and learn the parameters instead of directly learning the state-action value functions.

Question 22 *(0 points)*

In a Markov Decision Process with finite state and action sets, model-based reinforcement learning needs to learn a larger number of trainable parameters than model-free reinforcement learning.

√ **True**

○ False

Explain:

Solution: Model-based learning needs to learn $P(s'|s,a)$, a set of $N_S^2 N_a$ parameters, where N_s is the number of states, N_a the number of actions. Model-free learning needs to learn $Q(s,a)$, a set of only $N_s N_a$ trainable parameters.

The robot performs the following action:

- Starting state s: the state with $R(s) = 0.00$).
- Action a: robot tries to move to the horizontally neighboring state.
- Ending state s': the move is successful.

Given this one training observation, use Laplace smoothing, with a smoothing parameter of $k = 1$, to estimate the value of $P(s'|s,a)$ for this particular combination of (s,a,s').

Solution: The starting state is $s = (1,2)$. According to the problem description, the possible outcomes are known to be $s \in \{(1,1),(1,2),(2,2)\}$, so Laplace smoothing gives:

$$P(s' = (1,1)|s = (1,2), a = L) = \frac{1+k}{1+3k}$$
$$= \frac{2}{4}$$

Question 24 *(0 points)*

A cat lives in a two-room apartment; its current state is given by the room number it currently occupies ($s \in \{1,2\}$). It has two possible actions: walk, or purr. The cat attempts to determine the optimum policy using Q-learning. It starts out with an empty Q-table ($Q(s,a) = 0$ for all s and a). Starting in state $s_1 = 1$, it receives the following rewards, performs the following actions, and observes the following resulting states:

t	s	R	a	s
1	1	2	purr	1
2	1	2	purr	1

The cat performs one iteration of time-difference Q-learning with each of these two observations, using a learning rate of $\alpha = 0.1$ and a discount factor of $\gamma = 1$.

(a) After these two iterations of Q-learning, what values in the Q-table have changed?

Solution: The cat has only observed the (s,a) combination (1,purr), so the only entry in the Q table that has changed is $Q(1,\text{purr})$.

(b) After these two iterations of Q-learning, what is $Q(1,\text{purr})$?

Solution: Using the formulas

$$Q_{\text{local}}(s,a) = R(s) + \gamma Q_t(s',a)$$
$$Q_t(s,a) = Q_{t-1}(s,a) + \alpha\left(Q_{\text{local}}(s,a) - Q_{t-1}(s,a)\right)$$

We get the following:

$$Q_{\text{local}}(1,\text{purr}) = 2 + 0$$
$$Q_1(1,\text{purr}) = 0 + 0.1\,(2 - 0) = 0.2$$
$$Q_{\text{local}}(1,\text{purr}) = 2 + 0.2$$
$$Q_2(1,\text{purr}) = 0.2 + 0.1\,(2.2 - 0.2) = 0.4$$

The cat decides to use policy iteration to find a new optimal policy under this model. It starts with the following policy: $\pi(1) = \text{purr}$, $\pi(2) = \text{walk}$. Now it needs to find the policy-dependent utility, $U^\pi(s)$. Again, because the cat doesn't care about the distant future, it uses a relatively low discount factor ($\gamma = 3/4$). Write two linear equations that can be solved to find the two unknowns $U^\pi(1)$ and $U^\pi(2)$; your equations should have no variables in them other than $U^\pi(1)$ and $U^\pi(2)$.

Solution: The two equations are

$$U^\pi(1) = R(1) + \frac{3}{4}\sum_{s'} P(s'|1,\pi(1))U^\pi(s')$$
$$U^\pi(2) = R(2) + \frac{3}{4}\sum_{s'} P(s'|2,\pi(2))U^\pi(s')$$

Plugging in the given values of all variables, we have

$$U^\pi(1) = 2 + \frac{3}{4}\left(\frac{2}{3}U^\pi(1) + \frac{1}{3}U^\pi(2)\right)$$
$$U^\pi(2) = 5 + \frac{3}{4}\left(\frac{2}{3}U^\pi(1) + \frac{1}{3}U^\pi(2)\right)$$

(2 points) Since it has some extra time, and excellent python programming skills, the cat decides to implement deep reinforcement learning, using an actor-critic algorithm. Inputs are one-hot encodings of state and action. What are the input and output dimensions of the actor network, and of the critic network?

Solution: The actor network takes a state as input, thus its input dimension is 2 (if the input is a one-hot encoding of two states). It computes the probability that any given action is the best action, so its output dimension is 2 (if there are two possible actions). The critic takes, as input, an encoding of the state (two dimensions), and an encoding of the action (two dimensions, if the action is a one-hot encoding of two possible actions), for a total of 4 input dimensions. It computes, as output, a real-valued score $Q(s,a)$, which is a 1-dimensional (scalar) output.

estion 24 *(0 points)*

A cat lives in a two-room apartment; its current state is given by the room number it currently occupies ($s \in \{1,2\}$). It has two possible actions: walk, or purr. The cat attempts to determine the optimum policy using Q-learning. It starts out with an empty Q-table ($Q(s,a) = 0$ for all s and a). Starting in state $s_1 = 1$, it receives the following rewards, performs the following actions, and observes the following resulting states:

t	s	R	a	s
1	1	2	purr	1
2	1	2	purr	1

The cat performs one iteration of time-difference Q-learning with each of these two observations, using a learning rate of $\alpha = 0.1$ and a discount factor of $\gamma = 1$.

(a) After these two iterations of Q-learning, what values in the Q-table have changed?

Solution: The cat has only observed the (s,a) combination (1,purr), so the only entry in the Q table that has changed is $Q(1,\text{purr})$.

(b) After these two iterations of Q-learning, what is $Q(1,\text{purr})$?

Solution: Using the formulas

$$Q_{\text{local}}(s,a) = R(s) + \gamma Q_t(s',a)$$
$$Q_t(s,a) = Q_{t-1}(s,a) + \alpha\left(Q_{\text{local}}(s,a) - Q_{t-1}(s,a)\right)$$

We get the following:

$$Q_{\text{local}}(1,\text{purr}) = 2 + 0$$
$$Q_1(1,\text{purr}) = 0 + 0.1\,(2 - 0) = 0.2$$
$$Q_{\text{local}}(1,\text{purr}) = 2 + 0.2$$
$$Q_2(1,\text{purr}) = 0.2 + 0.1\,(2.2 - 0.2) = 0.4$$

A robot fire truck is able to manipulate its own horizontal location (D), the angle of its ladder (θ), and the length of its ladder (L). The ladder has a length of L, and an angle (relative to the x axis) of θ ($0 \le \theta \le \frac{\pi}{2}$ radians), so that the position of the tip of the ladder is

$$(x,z) = (D + L\cos\theta, L\sin\theta)$$

(a) What is the dimension of the configuration space of this robot?

Solution: There are three dimensions: D, L, and θ.

(b) The robot must operate between two buildings, positioned at $x = 0$ and at $x = 10$ meters. No part of the robot (neither its base, nor the tip of the ladder) may ever come closer than 1 meter to either building. What portion of configuration space is permitted? Express your answer as a set of inequalities involving only the variables D, L, and θ; the variables x and z should not appear in your answer.

Solution:

$$1 \le D + L\cos\theta \le 9, \quad 1 \le D \le 9$$

(c) The robot's objective is to save a cat from a tree. The cat is at position $(x,z) = (5,5)$. The robot begins at position $(D = 5, L = 3, \theta = 0)$. The final position of the robot depends on how much it costs to raise the ladder by one radian, as compared to the relative cost of extending the ladder by one meter, and the relative cost of moving the truck by one meter. Why?

Solution: Minimum-cost search for the solution will explore steps, in configuration space, that are of equal cost in each direction. The resulting shortest path will depend on the number of steps required to raise the ladder by $\pi/4$ radians, versus shifting the truck by 4m and extending the ladder. Equivalently: if raising the ladder is more expensive, then the truck will move 4m away, and raise the ladder only a little, but if raising the ladder is cheap, then the truck will raise the ladder up to vertical.

In order to lift the iron bar, the robot must reach an OBJECTIVE where one of its bodies is at position $(1,0)$ and the other is at position $(11,0)$. In terms of your notation from part (a), specify the OBJECTIVE as a set of points in configuration space. You may specify the OBJECTIVE as a set of discrete points, or as a set of equalities and inequalities.

Solution:

$$\text{OBJECTIVE} = \{(1,0,11,0),(11,0,1,0)\}$$

If the TBR touches the bar (with either of its bodies) at any location other than the endpoints ($(1,0)$ and $(11,0)$), then the bar falls off its tripods. This constitutes a FAILURE. Characterize FAILURE as a set of points in configuration space. You may specify FAILURE as a set of discrete points, or as a set of equalities and inequalities.

Solution: FAILURE is the set of all vectors (x_1, y_1, x_2, y_2) such that

$$1 < x_1 < 11 \text{ and } y_1 = 0$$

or

$$1 < x_2 < 11 \text{ and } y_2 = 0$$

The Actor-Critic Algorithm

$$\pi_a(s) = \text{Probability that } a \text{ is the best action in state } s$$
$$Q(s,a) = \text{Expected sum of future rewards if } (s,a)$$

- The critic is trained as a normal deep Q-learner:

$$\mathcal{L}_{critic} = \frac{1}{2}E[(f(\vec{s}_t, \vec{a}_t) - Q_{local}(\vec{s}_t, \vec{a}_t))^2]$$

- The actor is trained as an imitation learner, trying to compute a policy that will maximize the expected value of future rewards:

$$\mathcal{L}_{actor} = -\sum_a \pi_a(s)Q(s,a)$$

(10 points) Captain Marble is flying over a thick cloud bank. Her estimated velocity is v m/s northward, with a standard deviation of p m/s. Once per second, she uses her x-ray vision to scan the ground below, and recognizes landmarks that specity her lattitude, x, with a standard deviation of e meters. Based on all observations up through time t, her estimated lattitude at time t is y meters north of the equator, with a standard deviation of s meters. At time $t+1$ she is about to make another observatino, but just then she wonders: what are the expected value and standard deviation of her position at time $t+1$, on the basis of only the measurements up through and including time t?

Solution: The expected value is $y+v$, with a standard deviation of $\sqrt{s^2+p^2}$.

(10 points) Leira the merman is swimming eastward from Honolulu at a somewhat variable velocity: his velocity averages a m/s, but with a standard deviation of b m/s. Once every second, he checks his location usign sonar. Based on all of his observations up through time $t-1$, he believes that his longitude at time t is c meters east of Honolulu, with an uncertainty (standard deviation) of d meters. Since he's not sure, he takes another sonar measurement at time t; his sonar reading says that he is e meters east of Honolulu, but with a measurement uncertainty (standard deviation) of f meters. Based on all of his observations up through and including the observation at time t, what are the expected value and standard deviation of his longitude relative to Honolulu?

Solution: The Kalman gain is $k=d^2/(d^2+f^2)$. The expected value of his position at time t, given all observations up through and including time t, is $(1-k)c+ke$, with a standard deviation of $d\sqrt{1-k}$.

Question 10 *(7 points)*

Once every ten years, Briar Rose awakens, and checks the value of her stock portfolio. With probability 60%, she finds that the Dow Jones Industrial Average (DJIA) has doubled while she slept; with probability 40%, she finds that it has halved. She then sells whatever assets she had during the past ten years, and uses all of the money for one of two things: either she purchases stocks (a portfolio whose value equals the DJIA), or she purchases gold (suppose that, in her world, the value of gold never changes). After taking one of these two actions, she goes back to sleep for ten more years.

Define the state of her finances in decade d to be a tuple specifying the value of her wealth ten years ago, w_{d-1}, and the value of her wealth now, w_d, thus $s_d=(w_{d-1},w_d)$. Because the stock market either doubles or halves, while gold never changes, this tuple is always $s\in\{(x,0.5x),(x,x),(x,2x)\}$ for some real number x. Define the reward to be the amount she has earned this decade: $R(s_d)=w_d-w_{d-1}$. Consider using value iteration to find $U_t(s)$, the t-step approximation of the utility of state s. Starting with $U_0(s)=0$ for all states, find $U_1(s)$ and $U_2(s)$ for each of the states $s\in\{(x,0.5x),(x,x),(x,2x)\}$. Your answer should be six equations, showing $U_1((x,0.5x)),\ldots,U_2((x,2x))$, each as a function of x and γ, where γ is the discount factor.

Solution:

$$U_1((x,0.5x))=-0.5x$$
$$U_1((x,x))=0$$
$$U_1((x,2x))=x$$

For $U_2(s)$, the correct solution takes into account that s' is defined in terms of x', which may be $2x$ or $0.5x$. That yields the following solution:

$$\begin{aligned}
U_2((x,0.5x)) &= -0.5x+\gamma\max(0,0.6x'+0.4(-0.5x'))\\
&= -0.5x+\gamma\max(0,0.3x+0.4(-0.25x))\\
U_2((x,x)) &= \gamma\max(0,0.6x+0.4(-0.5x))\\
U_2((x,2x)) &= x+\gamma\max(0,0.6x'+0.4(-0.5x'))\\
&= x+\gamma\max(0,1.2x+0.4(-x))
\end{aligned}$$

(b) Consider the set of all rooms that can be reached by starting in room 431, and taking t tunnels in sequence (i.e., choose a tunnel, take it to a new room, then repeat this $t-1$ more times). Call the set of rooms that can be reached in this way $\mathbb{S}_t$. As a function of t and ε, how many times must you play the game in order to estimate $P(s'|s,a)$ for every room $s\in\mathbb{S}_t$, and for every action $a\in\{L,R,F\}$, with a precision of ε?

Solution: There are 3^t rooms in the set, thus there are 3^{t+1} possible combinations of (s,a). For each one, we need to play the game $1/\varepsilon$ times to estimate the probability, thus in total we need to play the game $3^{t+1}/\varepsilon$ times.

Question 12 *(7 points)*

One reason to use a replay buffer, in Q-learning, is that the Q-learning algorithm does not immediately learn about the long-term consequences of an action. For example, suppose that your replay buffer contains the following $(s_t,a_t,R(s_t),s_{t+1},a_{t+1})$ tuples:

s_t	a_t	$R(s_t)$	s_{t+1}	a_{t+1}
431	L	-0.04	1024	R
1024	R	0.5	516	R

Consider using SARSA Q-learning to estimate $Q_t(s,a)$, starting with $Q_0(s,a)=0$ for all s and a. When you estimate $Q_t(s,a)$, use every row of the replay buffer, not just the t^{th} row. Let α be the learning rate, and let γ be the discount factor; as a function of α and γ, find $Q_1(431,L)$, $Q_1(1024,R)$, and $Q_2(431,L)$.

Solution:

$$\begin{aligned}
Q_1(431,L) &= Q_0(431,L)+\alpha(R(431)+\gamma Q_0(1024,R)-Q_0(431,L))\\
&= -0.04\alpha\\
Q_1(1024,R) &= Q_0(1024,R)+\alpha(R(54)+\gamma Q_0(516,R)-Q_0(1024,R))\\
&= 0.5\alpha\\
Q_2(431,L) &= Q_1(431,L)+\alpha(R(431)+\gamma Q_1(1024,R)-Q_1(431,L))\\
&= -0.04\alpha-0.04\alpha+0.5\alpha^2\gamma+0.04\alpha^2
\end{aligned}$$

Solution: The reachable area of the room is $(D-2R)^2-(L+2R)(W+2R)$. Dividing that area into $R\times R$ squares gives approximately

$$\frac{(D-2R)^2-(L+2R)(W+2R)}{R^2}$$

squares.

If one considers that only the *center* of each square needs to be within the reachable area, then the total area covered by the entire *periphery* of each square could be as large as $(D-R)^2-(L+R)(W+R)$. If one makes this assumption, then the number of squares is

$$\frac{(D-R)^2-(L+R)(W+R)}{R^2}$$

Question 13 *(7 points)*

Consider a deep Q-learning algorithm, in which $Q(\vec{s},a)$ is approximated, after t iterations of training, by

$$Q_t(\vec{s},a)=\vec{w}_t^T\vec{h}_t(\vec{s},a),$$

where $\vec{w}_t^T=[w_{t,1},\ldots,w_{t,5}]$ is a 5-dimensional weight vector, and $\vec{h}_t^T(\vec{s},a)=[h_{t,1}(\vec{s},a),\ldots,h_{t,5}(\vec{s},a)]$ is a 5-dimensional hidden node activation vector. Consider two particular state vectors, $\vec{s}=[0.3,0.6]^T$ and $\vec{s}=[-0.2,0.5]^T$, and two particular actions, $a\in\{L,R\}$. Suppose that, in response to these state vectors and these actions, the network currently computes the following hidden node activation vectors:

$\vec{s}^T$	a	$\vec{h}_t^T(\vec{s},a)$
$[0.3,0.6]$	L	$[a,b,c,d,e]$
$[0.3,0.6]$	R	$[f,g,h,i,j]$
$[-0.2,0.5]$	L	$[k,l,m,n,o]$
$[-0.2,0.5]$	R	$[p,q,r,s,t]$

Suppose that, in the t^{th} iteration of deep Q-learning, using only the t^{th} entry from the replay buffer, the output weight vector is $\vec{w}_t=[1,0,0,0,0]^T$, the input state is $\vec{s}_t=[0.3,0.6]^T$, the action is $a_t=R$, the reward is $R(\vec{s}_t)=-0.04$, the resulting state vector is $\vec{s}_{t+1}=[-0.2,0.5]^T$, and the loss function is

$$\mathscr{L}=\frac{1}{2}(Q_t(\vec{s}_t,a_t)-Q_{local}(\vec{s}_t,a_t))^2$$

Assume a discount factor of γ, and assume TD-learning, not SARSA. In terms of any or all o variables $a,b,c,\ldots,r,s,t$, and/or γ, what is the value of $\nabla_{\vec{w}_t}\mathscr{L}$?

Solution: First, find Q_{local}.

$$\begin{aligned}
Q_{local}(\vec{s}_t,a_t) &= R(\vec{s}_t)+\gamma\max_{a'}Q_t(\vec{s}_{t+1},a')\\
&= -0.04+\gamma\max(k,p)
\end{aligned}$$

Now in terms of Q_{local}, we can write:

$$\begin{aligned}
\nabla_{\vec{w}_t}\mathscr{L} &= (Q_t(\vec{s}_t,a_t)-Q_{local}(\vec{s}_t,a_t))\nabla_{\vec{w}_t}Q_t(\vec{s}_t,a_t)\\
&= (Q_t(\vec{s}_t,a_t)-Q_{local}(\vec{s}_t,a_t))\vec{h}_t(\vec{s}_t,a_t)\\
&= (f+0.04-\gamma\max(k,p))\begin{bmatrix} f\\ g\\ h\\ i\\ j\end{bmatrix}
\end{aligned}$$

Question 17 *(0 points)*

In a pinhole camera, a light source at (x,y,z) is projected onto a pixel at $(x',y',-f)$ through a pinhole at $(0,0,0)$. Write $\sqrt{(x')^2+(y')^2}$ in terms of x, y, z, and f.

Solution: From the idea of similar triangles, we have

$$\frac{x'}{f}=-\frac{x}{z},\quad\frac{y'}{f}=-\frac{y}{z}$$

from which we derive

$$\sqrt{(x')^2+(y')^2}=\frac{f}{z}\sqrt{x^2+y^2}$$

Question 18 *(0 points)*

The real world contains two parallel infinite-length lines, whose equations, in terms of the coordinates (x,y,z), are parameterized as $ax+by+cz=d$ and $ax+by+cz=e$; in addition, both of these lines are on the ground plane, $y=g$, for some constants (a,b,c,d,e,g). Show that the images of these two lines, as imaged by a pinhole camera, converge to a vanishing point, and give the coordinates (x',y') of the vanishing point.

Solution: From the idea of similar triangles, we have

$$\frac{x'}{f}=-\frac{x}{z},\quad\frac{y'}{f}=-\frac{y}{z}$$

From which we derive

$$x=\frac{-zx'}{f},\quad y=\frac{-zy'}{f}$$

So the equations of the two lines are

$$\begin{aligned}
-\frac{ax'}{f}-\frac{by'}{f}+c &= \frac{d}{z}\\
-\frac{ax'}{f}-\frac{by'}{f}+c &= \frac{e}{z}
\end{aligned}$$

As $z\rightarrow\infty$, the right-hand-sides of these two equations both go to zero, and the equations of both lines converge to

$$ax'+by'=cf$$

In addition, we have $y=g$, so $y'=-fg/z\rightarrow 0$, and therefore $x'=cf/a$. The coordinates are $(x',y')=(cf/a,0)$.

Question 19 *(0 points)*

Consider the convolution equation

$$Z(x',y')=\sum_m\sum_n h(m,n)Y(x'-m,y'-n)$$

Where $Y(x',y')$ is the original image, $Z(x',y')$ is the filtered image, and the filter $h(m,n)$ is given by

$$h(m,n)=\begin{cases}\frac{1}{21} & 1\le m\le 3,\ -3\le n\le 3\\ -\frac{1}{21} & -3\le m\le -1,\ -3\le n\le 3\end{cases}$$

Would this filter be more useful for smoothing, or for edge detection? Why?

Solution: The sum of $h(m,n)$, over all m and n, is 0. So if it is filtering a constant-color region, the output would always be zero, regardless of the input color. So it's not very useful for smoothing.

Any given pixel of $Z(x',y')$ is the difference between the pixels $Y(x',y')$ to its left, minus those to its right. Since it's computing a difference, it would be useful for edge detection.

Question 20 *(0 points)*

Under what circumstances is a difference-of-Gaussians filter more useful for edge detection than a simple pixel difference?

Solution: A difference-of-Gaussians filter first smooths the input image (using a Gaussian smoother), then computes a pixel difference. The smoothing step can reduce random noise. Therefore, this procedure is more useful if the input image has some random noise in it.

Robotics: Inverse kinematics

- Obstacles are things in the workspace, $\mathcal{W}$, that we don't want to run into.
- We want to plan a path through configuration space, C, such that we don't run into any obstacle.
- In order to do that, we need **inverse kinematics**: a function that converts obstacles in the workspace, $\mathcal{W}_{\text{obs}}$, into equivalent obstacles in configuration space, C_{obs}.
$$C_{\text{obs}}=\{q:\exists b:\varphi_b(q)\in\mathcal{W}_{\text{obs}}\}$$
- For example: we usually do this by just exhaustively testing every point in configuration space, to see if it runs into an obstacle.

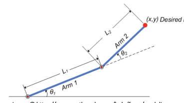


Image © https://www.mathworks.com/help/fuzzy/modeling-inverse-kinematics-in-a-robotic-arm.html

Question 6 *(7 points)*

Slarti is in Paris, attempting to walk home from a pizza restaurant. He is on a sidewalk whose west edge is a steep drop into the river; he wants to make sure he does not fall into the river. Let Y_t be the true distance between Slarti and the cliff edge at time t, measured in meters (m). Suppose you know that $Y_0=3$m, for sure. Since Slarti is too full to walk straight, he wobbles as he walks. Since it's foggy, he does not always see clearly: X_t is how far away the cliff edge looks, at time t, which may or may not be equal to the true distance Y_t. The transition probabilities and observation probabilities are

$$P(Y_t=k|Y_{t-1}=j)=\begin{cases}\frac{1}{4} & k=j-1\\ \frac{1}{2} & k=j\\ \frac{1}{4} & k=j+1\end{cases},\quad P(X_t=k|Y_t=j)=\begin{cases}\frac{1}{3} & k=j-1\\ \frac{1}{3} & k=j\\ \frac{1}{3} & k=j+1\end{cases}$$

What is $P(Y_2=2,X_2=2)$?

Solution:

$$\begin{aligned}
P(Y_2=1,X_2=2) &= \sum_{y_1}P(Y_0=3,Y_1=y_1,Y_2=2,X_2=2)\\
&= P(Y_0=3,Y_1=2,Y_2=2,X_2=2)+P(Y_0=3,Y_1=3,Y_2=2,X_2=2)\\
&= \left(\frac{1}{2}\right)\left(\frac{1}{4}\right)\left(\frac{1}{3}\right)+\left(\frac{1}{4}\right)\left(\frac{1}{2}\right)\left(\frac{1}{3}\right)
\end{aligned}$$