Mini-Course 1, Module 2
Markov Decision Processes
Worksheet Class

CMPUT 365
Fall 2021

Reminders: Sept 15, 2021

- There is no programming assignment this week. There is an assignment—just not code
    - So today (Wednesday Lab Session) will not be about programming this one week
- We will do some worksheet questions
- I will go over some questions from Discord
- Friday we will have a background review lecture on Stats

Late Policy

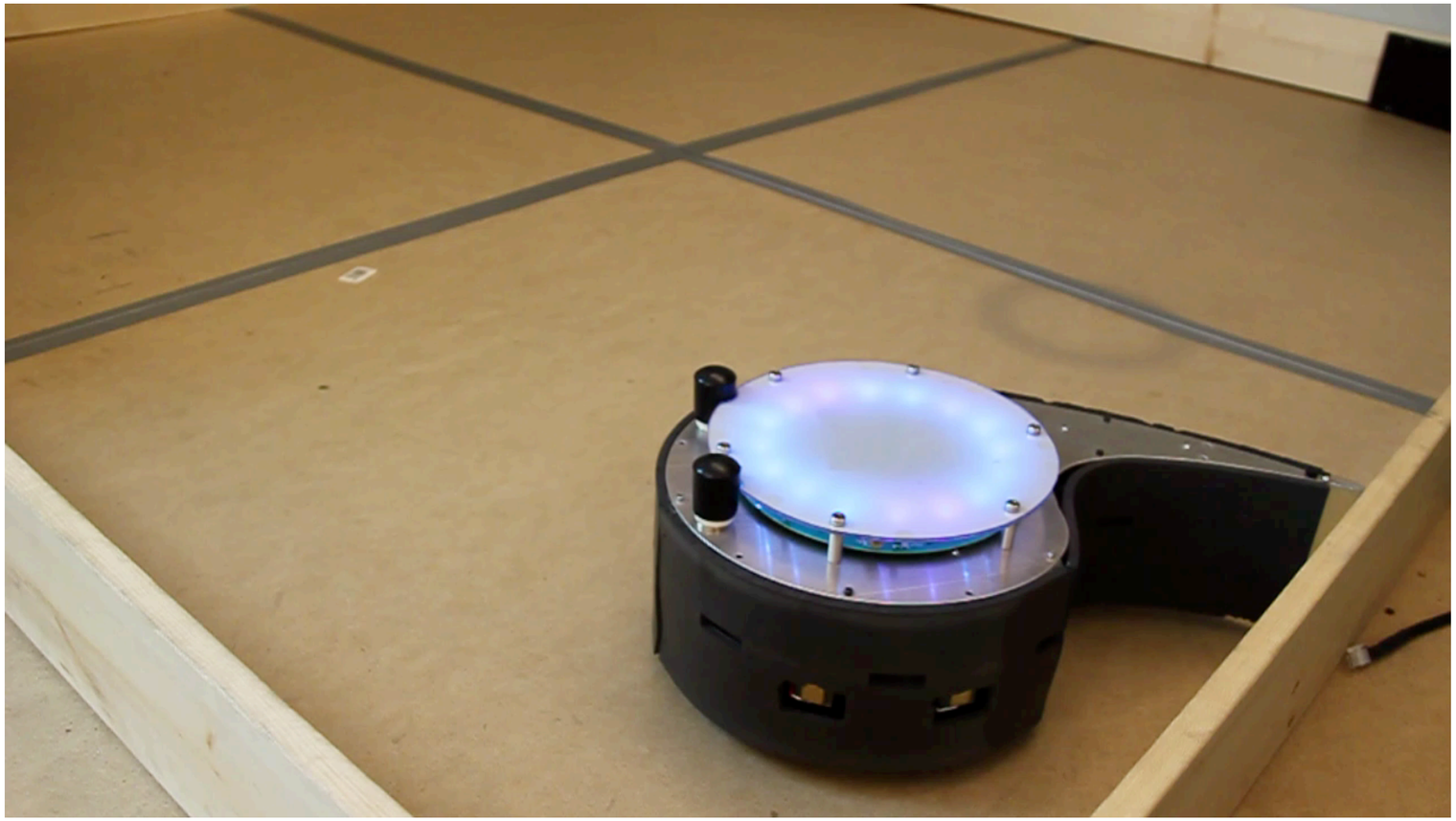
- No late submissions! Late = 0
- No transferring marks to the final!
- You are allow to miss **one assignment and two practice quizzes**. Choose wisely!
- If some extreme situation (exceptional circumstance) arises email me personally and we ill discuss accommodations.

Weekly work flow

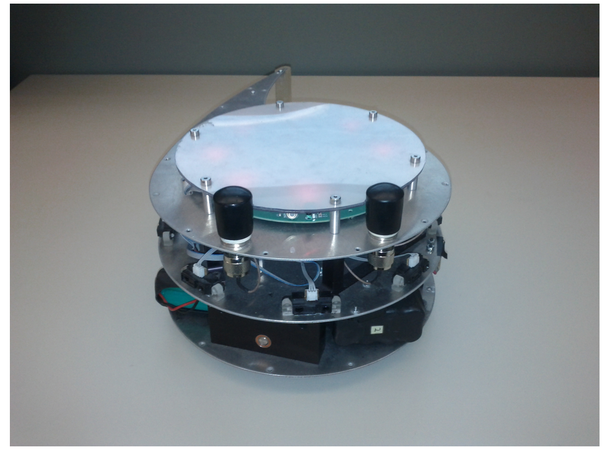
- **Friday / Saturday / Sunday:**
  - watch the videos [time required 30 mins, max 1hr]
  - Skim textbook and read about things that were not clear to you; ask questions on Discord
  - Do practice quiz
- **Monday**: we review of videos / concepts / quiz + Q&A
- **Wednesday**: lab day with TAs, help with your notebook assignment
- **Friday**: bonus lectures and practice exercises
- *Throughout the week: **read** book, work on **programming** assignment [use Discord & office hours]*

Example MDPs

- Consider this robot called the critterbot

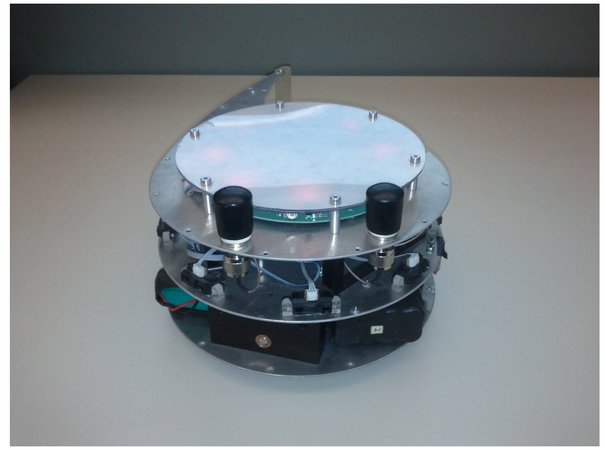


Robot sensors

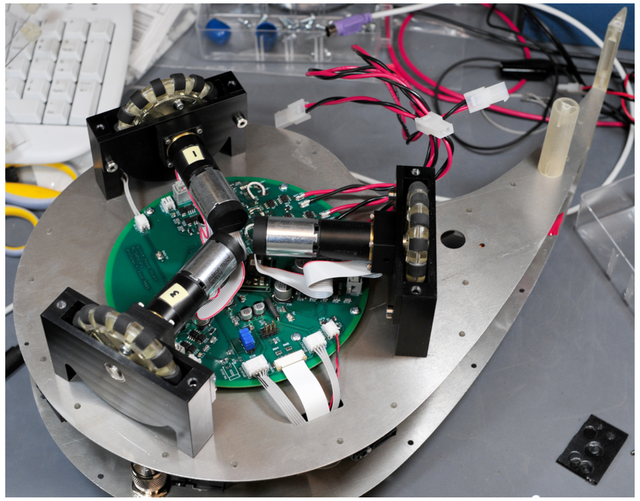


- This robot has **sensors** all over it:
  - Distance sensors
  - Light sensors
  - Thermal sensors
  - Motor information like speed, current, velocity, and temp
  - Accel XYZ
  - Rotational vel
  - etc

Robot actuation



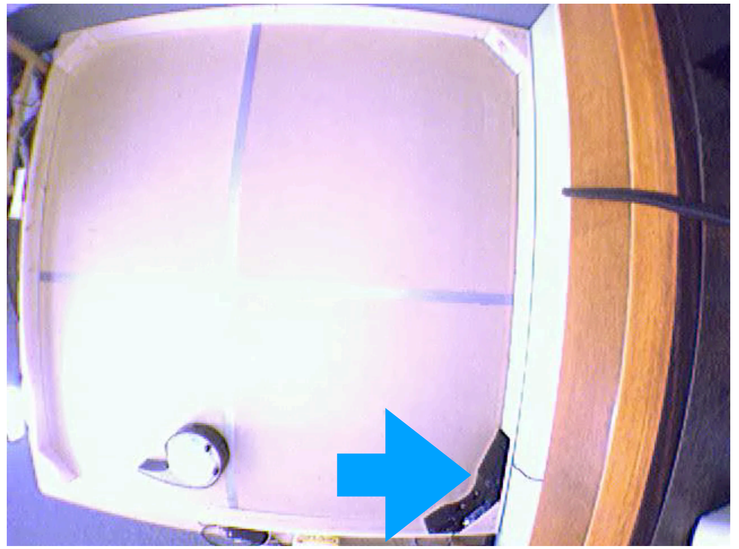
- This robot has an omni directional drive system:
  - Can move in any direction on a surface



An RL robot MDP

- Imagine we wanted this robot to **goto the light quickly**
  - What is the **state**?
  - What are the **actions**?
  - What is the **reward**?
- In other words how to formulate this as an **RL problem**?

Robot state



- Perhaps the sensors are so good that every situation in the pen looks **unique** to the agent:
  - The agent would not do better by **remembering** a history of the sensors
  - This is true because of the **IR beacon on the charger**, and **magnetic** sensors + all the other sensors
- We could also use an overhead camera + a localization algorithm to extract X,Y, theta position —Markov state
- It would be easy to imagine that if the robot only had distance sensors and the pen was square, that the state would **not be fully observable**

Robot actions

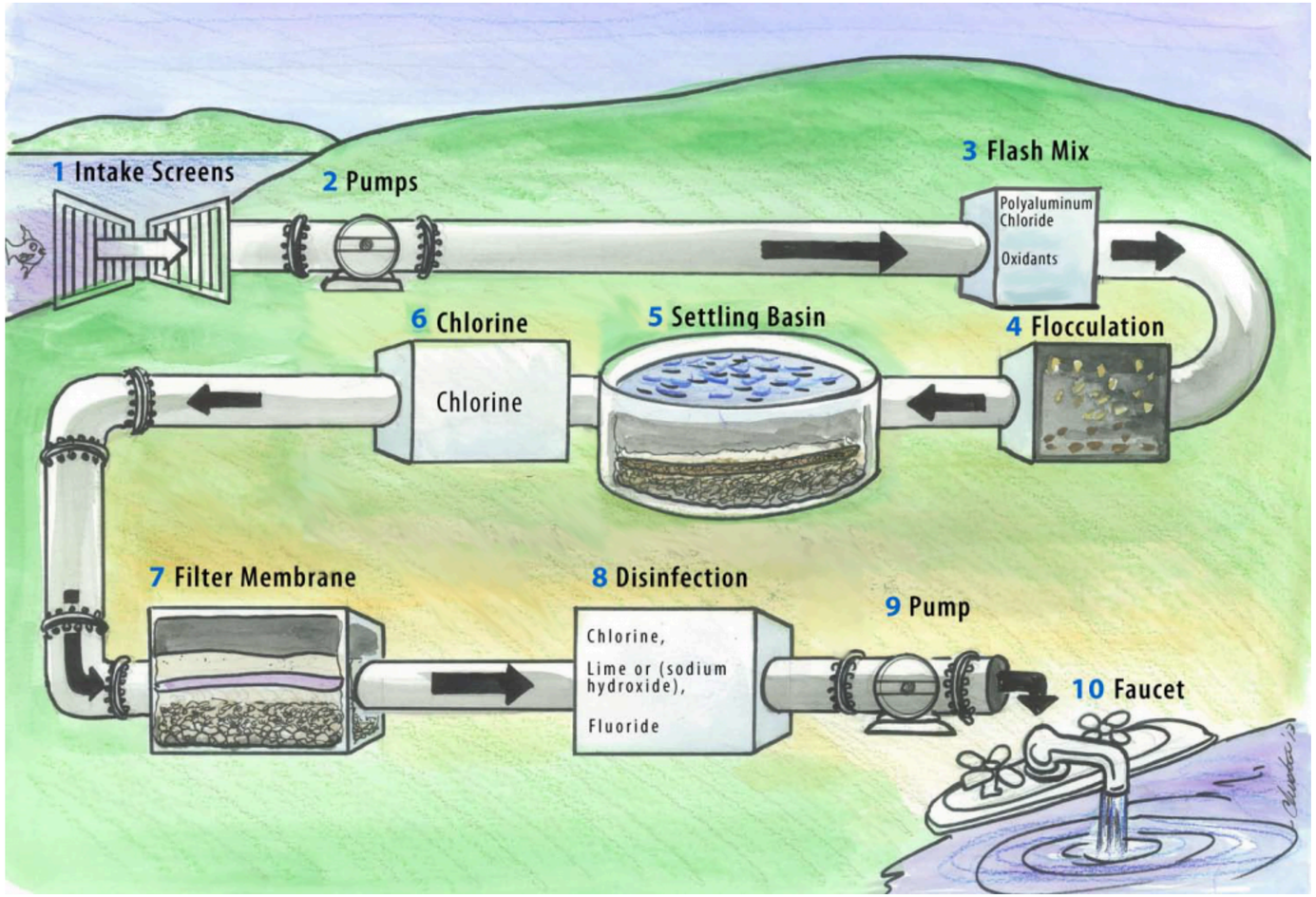
- This robot can be controlled in two basic ways:
  - X,Y,theta mode -> specify amount of translation in X and Y, and rotation
  - Voltage mode -> how much voltage to send to each of the three motors

Robot Reward

- To encourage light seeking
  - -1 per step, until front light sensor > threshold
  - OR
  - Reward = front_light_sensor_reading
- In both cases terminate when light sensor > threshold; **Episodic problem**
  - What is gamma?
- How do we reset to the state state? ME :(

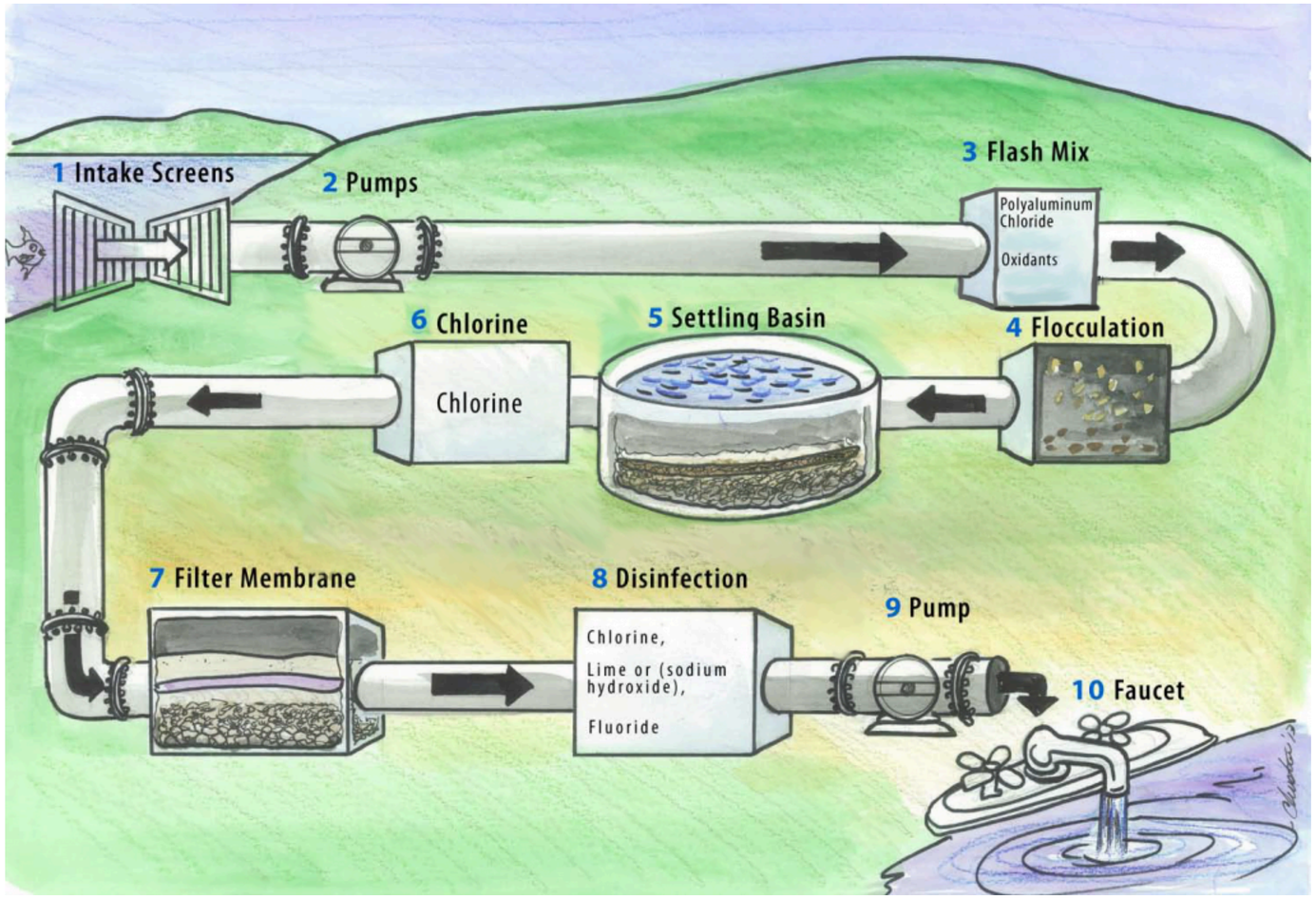
Fresh Water Treatment

- Water from the river
- Passes through many stages:
  - Chemical treatment
  - Settling
  - Mixing
  - Filters
  - Then to you to drink



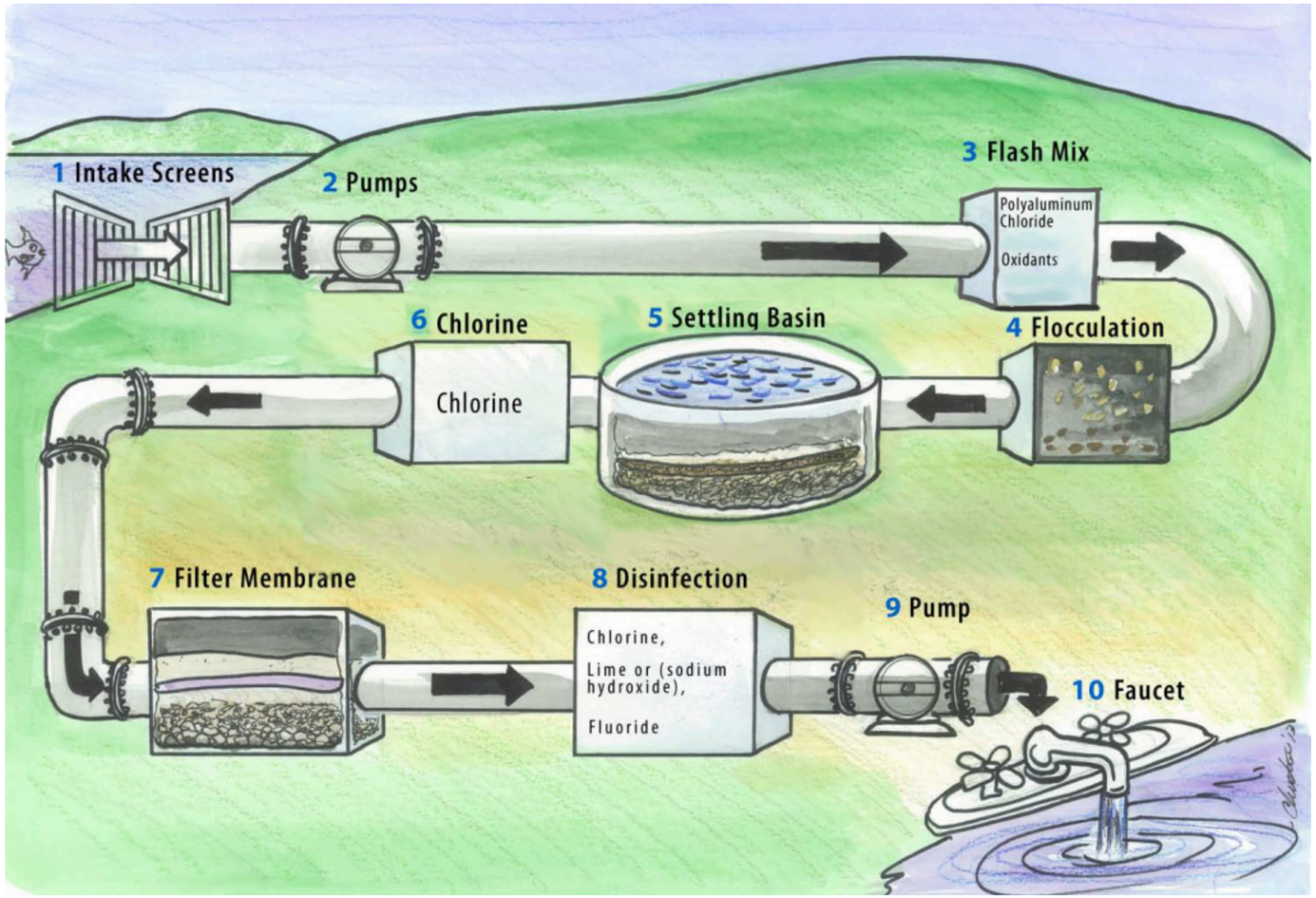
Fresh Water Treatment

- The reward is easy
  - R_t = Water flow - energy
- The system is **safe** by design:
  - Hard to break the components
  - Plant cannot produce unclean water
- Good for trial and error learning



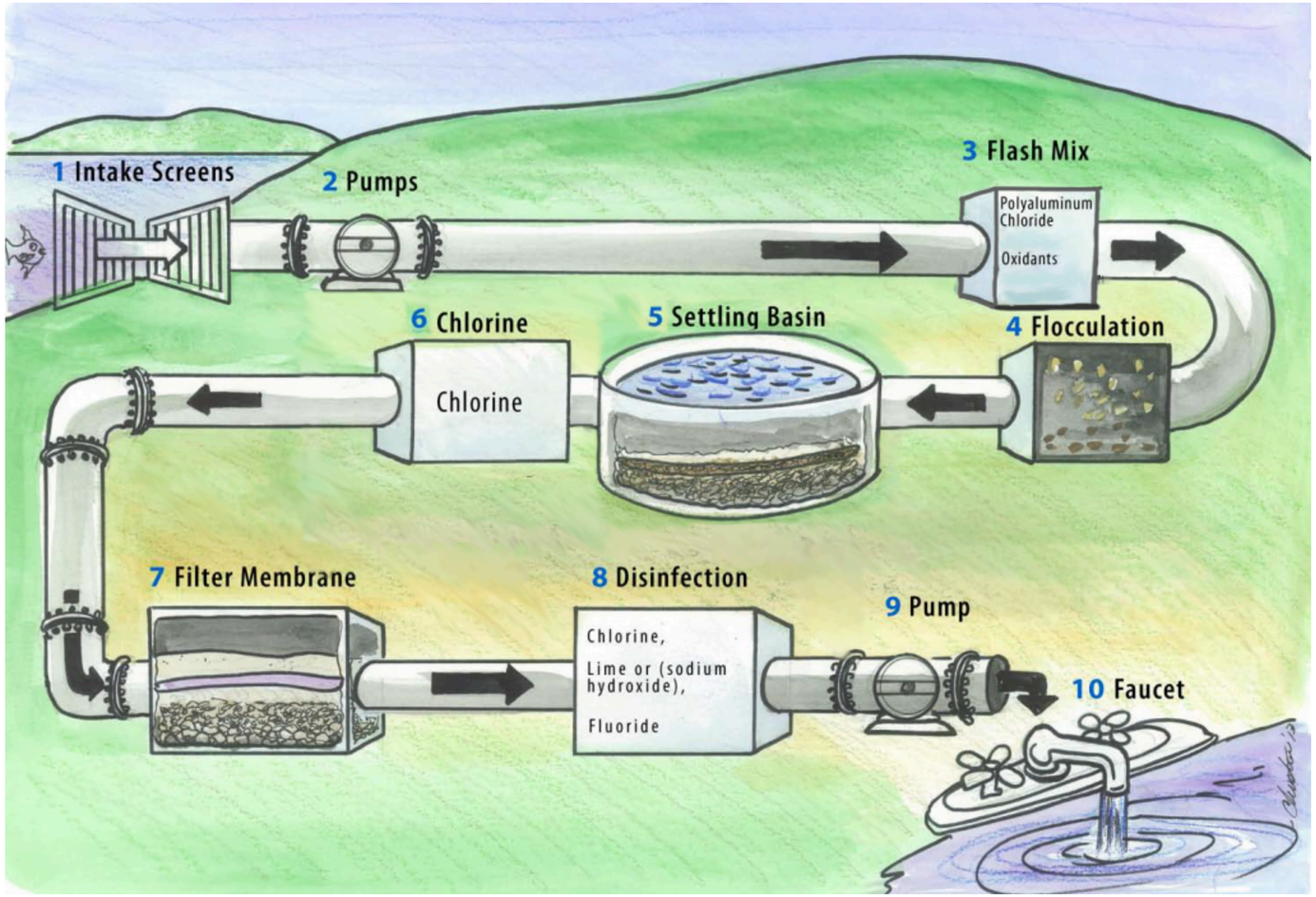
Fresh Water Treatment

- **Many sensors:**
  - Pump speeds and temp
  - TMP & other water data
  - Turbidity
  - Weather & future weather
  - River conditions
  - State of the filter
- Could be very close to **Markov state**



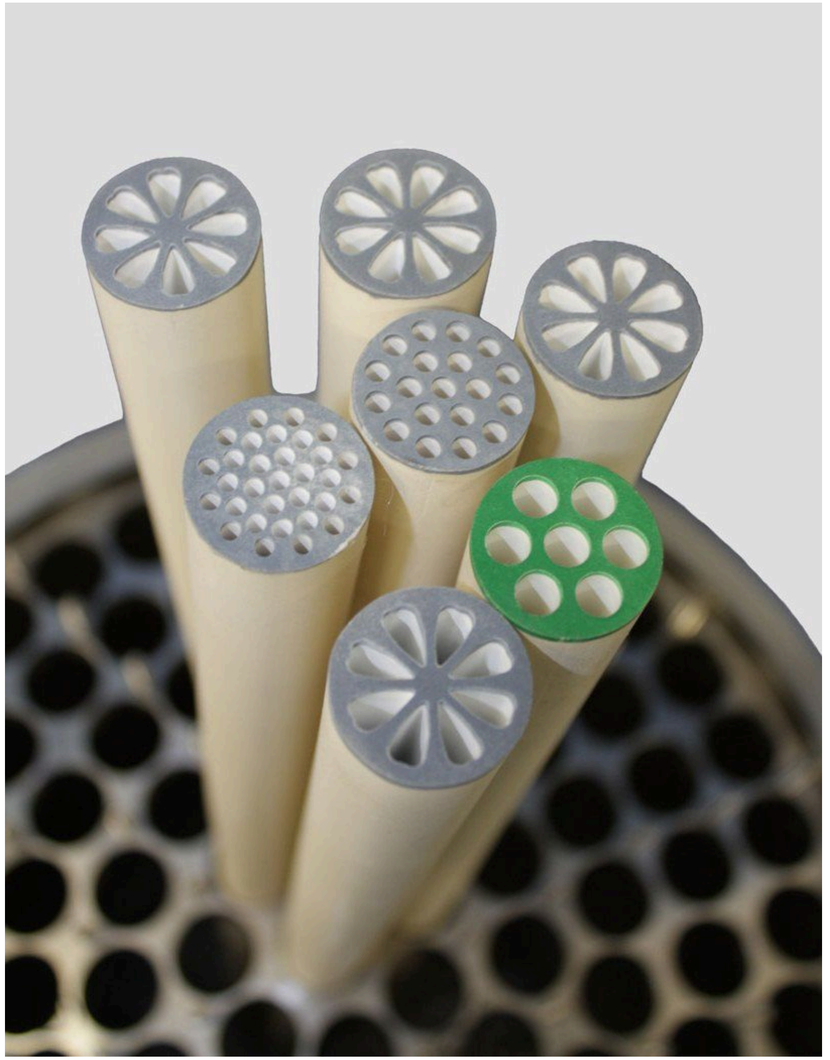
Fresh Water Treatment

- Many possible **actions**:
  - Pump speeds
  - Chemical treatments
  - **Backwashing the filter …**



Fresh Water Treatment

- The filter is comprised of hundreds of filter tubes
- They get full of dirt and other stuff
- To clean them we just run the system backwards (**backwashing**)
    - But backwashing uses a **lot of energy** and **stops water production**
- **Action**: backwash or not on every step



More example MDPs

- White in 1969 (not me!!)
  - Salmon Harvesting
  - Agriculture: how much to plant based on weather and soil state.
  - Water resources: keep the correct water level at reservoirs.
  - Inspection, maintenance and repair: when to replace/inspect based on age, condition, etc.
  - Purchase and production: how much to produce based on demand.
  - Queues

More example MDPs

- Flying REAL (small) Helicopters ~Andrew Ng et al

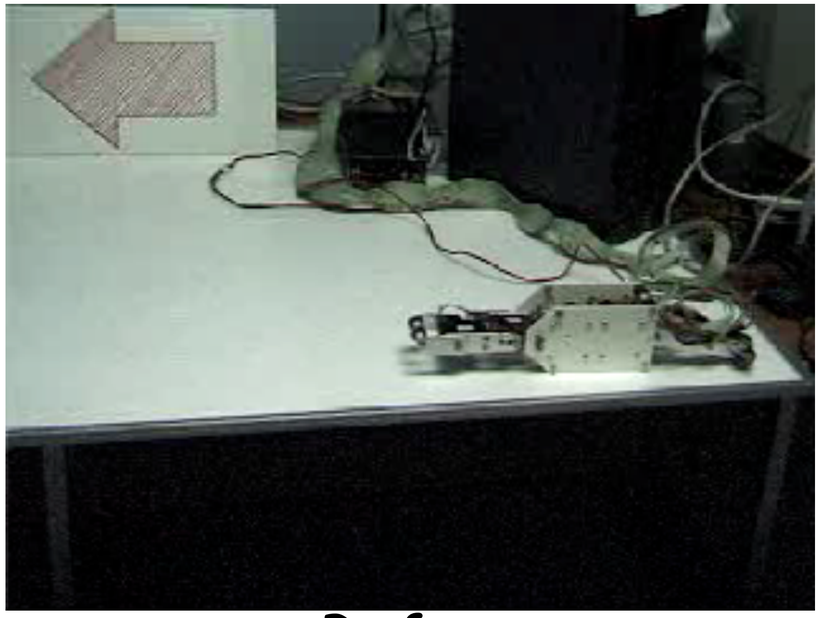
Heli MDP

- **Actions (multi-dimensional & continuous!!):**
  - Longitudinal and latitudinal Pitch controls
  - Man rotor pitch
  - Tail rotor pitch
  - Throttle

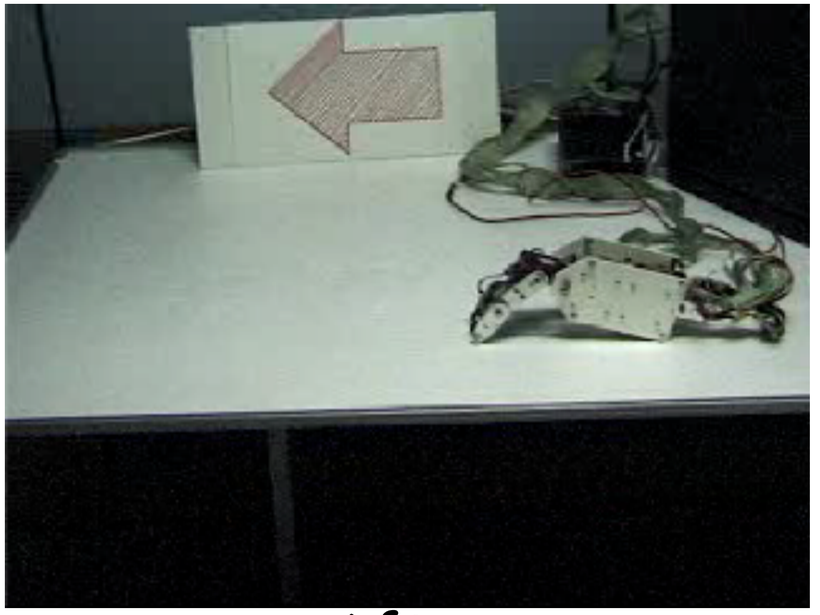
$$R(s^s) = -(\alpha_x(x - x^*)^2 + \alpha_y(y - y^*)^2 + \alpha_z(z - z^*)^2 + \alpha_{\dot{x}}\dot{x}^2 + \alpha_{\dot{y}}\dot{y}^2 + \alpha_{\dot{z}}\dot{z}^2 + \alpha_\omega(\omega - \omega^*)^2),$$

- **State (8 dim):**
  - helicopter's position (x, y, z), orientation (roll ϕ, pitch θ, yaw ω), velocity ($\dot{x}$,$\dot{y}$,$\dot{z}$) and angular velocities ($\dot{\phi}$,$\dot{\theta}$,$\dot{\omega}$

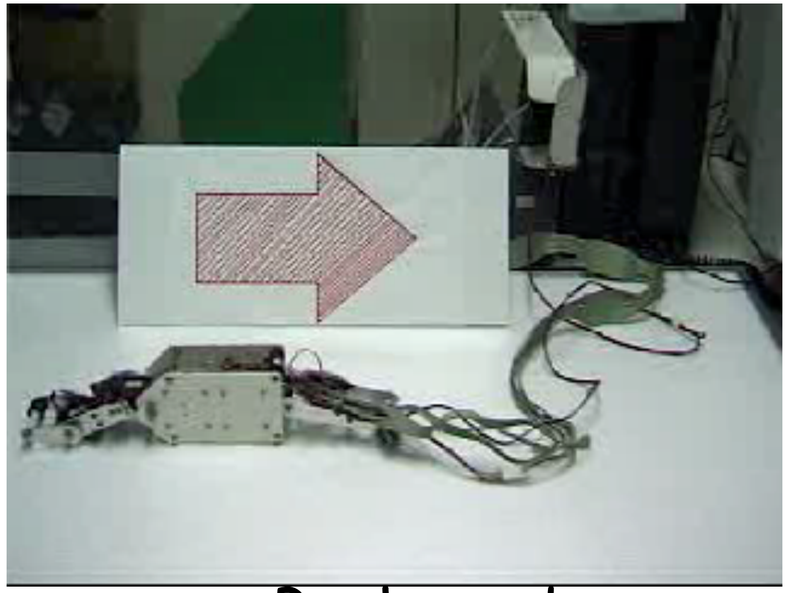
Hajime Kimura's RL Robots



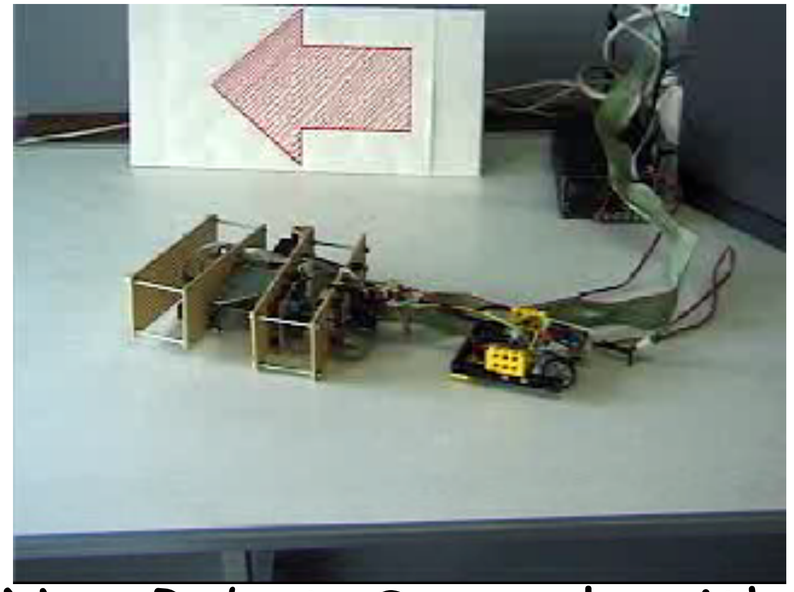
Before



After

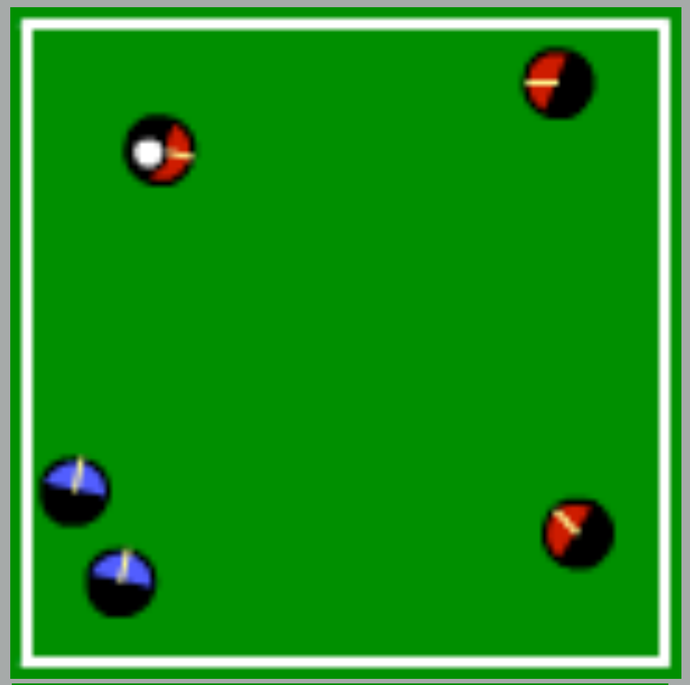


Backward



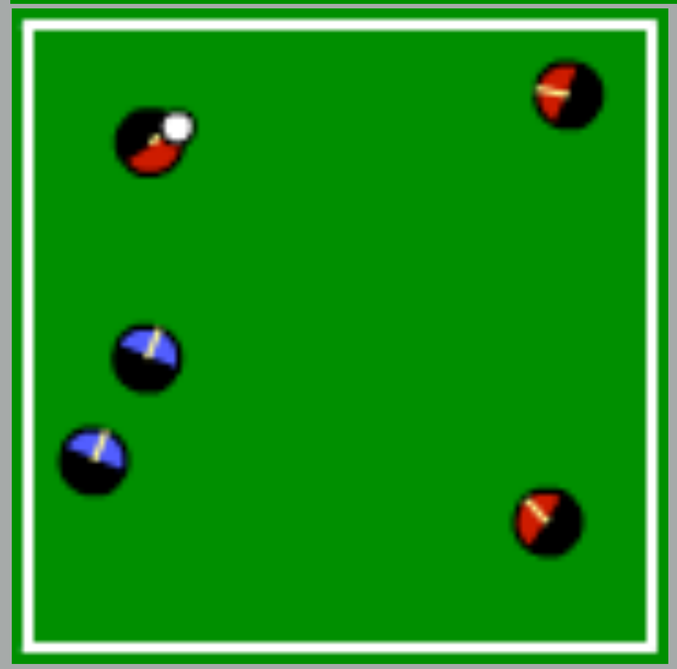
New Robot, Same algorithm

- Stone & Sutton



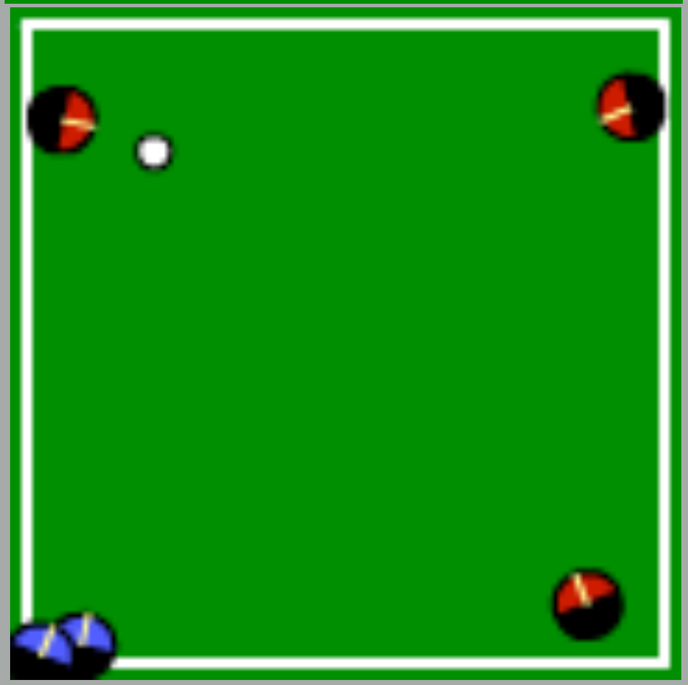
Random

Learned

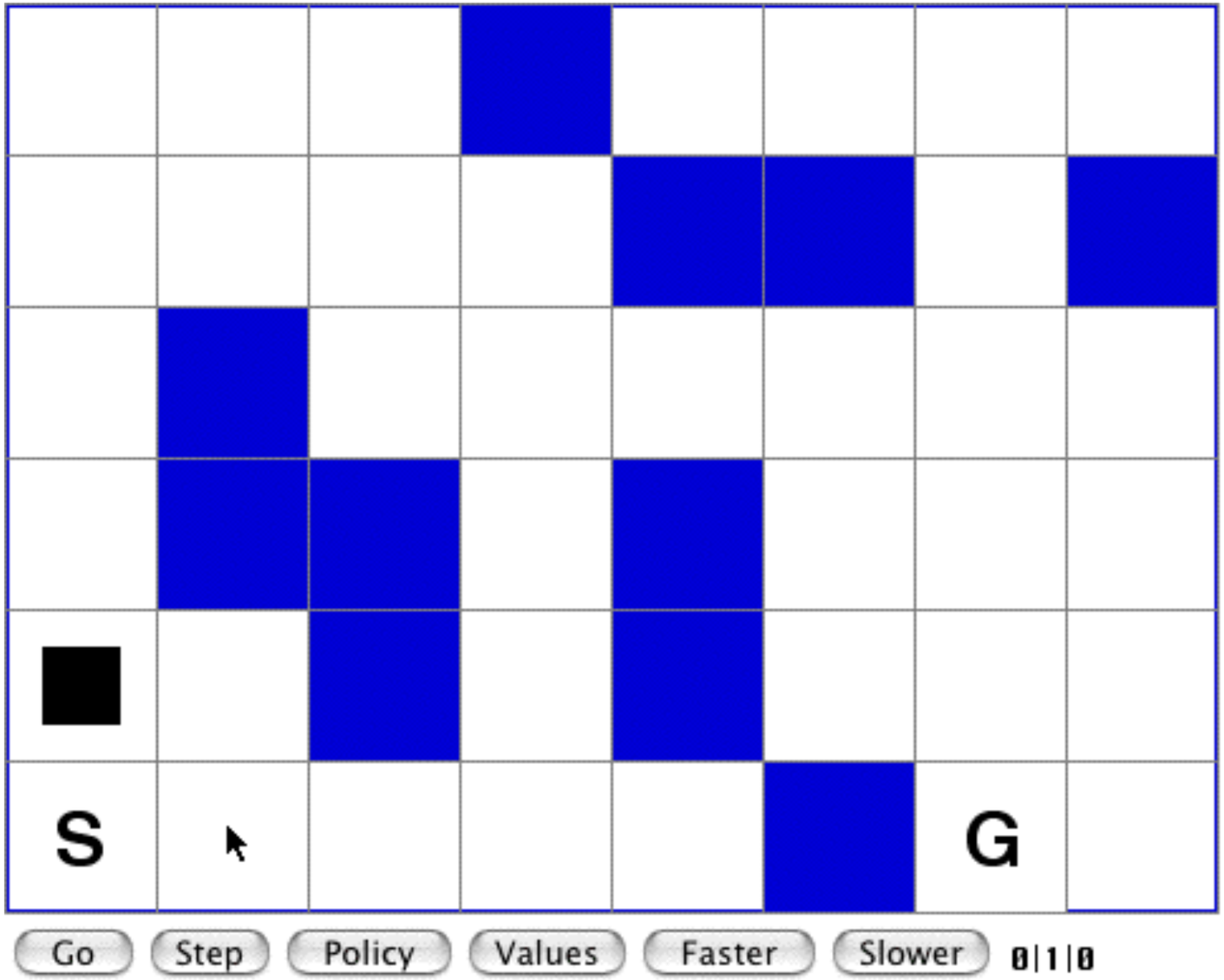


Hand-coded

Hold



GridWorld Example



Worksheet Review

3. (Exercise 2.2 from S&B 2nd edition) Consider a k-armed bandit problem with $k = 4$ actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using ϵ-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$, for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1, R_1 = -1, A_2 = 2, R_2 = 1, A_3 = 2, R_3 = -2, A_4 = 2, R_4 = 2, A_5 = 3, R_5 = 0$. On some of these time steps the ϵ case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?

T	Q1	Q2	Q3	Q4	{A*_t}	A_t	Explore?	R_1
1	0	0	0	0	{1,2,3,4}	1	Maybe	-1
2	-1	0	0	0	{2,3,4}	2	Maybe	1
3	-1	1	0	0		2		-2
4	-1	-0.5	0	0		2		2
5	-1	0.3333	0	0		3		0

Key learnings

T	Q1	Q2	Q3	Q4	{A*_t}	A_t	Explor	R_1
1	0	0	0	0	{1,2,3,	1	Maybe	-1
2	-1	0	0	0	{2,3,4}	2	Maybe	1
3	-1	1	0	0		2		-2
4	-1	-0.5	0	0		2		2
5	-1	0.3333	0	0		3		0

- Initial Q-values (all zeros) do not impact the computation of the sample average
- We don’t change the values of actions no taken
- The explore step or \epsilon step might choose the greedy action: can only know for sure when agent explores
- Try to imagine the agents life:
  - Look at the Q-values; pick an action; observe the reward; update one of the Q-values
  - Look at the Q-values; pick an action; observe the reward; update one of the Q-values
  - …. forever

Worksheet Review

Q3

Suppose that in a lottery you have 0.01% chance of winning and the prize is \$1000. The ticket to enter the lottery costs you \$10. What is the expected amount you would earn, when buying a ticket for this lottery?

$$\mathbb{E}[X] \doteq \sum_{i=1}^{k} x_i p_i = x_1 p_1 + x_2 p_2 + \ldots + x_k p_k$$

Two outcomes: win or not

Worksheet Review

Q3

Suppose that in a lottery you have 0.01% chance of winning and the prize is \$1000. The ticket to enter the lottery costs you \$10. What is the expected amount you would earn, when buying a ticket for this lottery?

$$\mathbb{E}[X] \doteq \sum_{i=1}^{k} x_i p_i = x_1 p_1 + x_2 p_2 + \ldots + x_k p_k$$

Two outcomes: win or not

$$\mathbb{E}[X] =$$

Worksheet Review

Q3

Suppose that in a lottery you have 0.01% chance of winning and the prize is \$1000. The ticket to enter the lottery costs you \$10. What is the expected amount you would earn, when buying a ticket for this lottery?

$$\mathbb{E}[X] \doteq \sum_{i=1}^{k} x_i p_i = x_1 p_1 + x_2 p_2 + \ldots + x_k p_k$$

Two outcomes: win or not

$$\mathbb{E}[X] = \left(\frac{0.01}{100}\right)(1000 - 10)$$

Prob of win

Outcome on win

Worksheet Review

Q3

Suppose that in a lottery you have 0.01% chance of winning and the prize is \$1000. The ticket to enter the lottery costs you \$10. What is the expected amount you would earn, when buying a ticket for this lottery?

$$\mathbb{E}[X] \doteq \sum_{i=1}^{k} x_i p_i = x_1 p_1 + x_2 p_2 + \ldots + x_k p_k$$

Two outcomes: win or not

$$\mathbb{E}[X] = \left(\frac{0.01}{100}\right)(1000 - 10) +$$

Prob of win

Outcome on win

Worksheet Review

Q3

Suppose that in a lottery you have 0.01% chance of winning and the prize is \$1000. The ticket to enter the lottery costs you \$10. What is the expected amount you would earn, when buying a ticket for this lottery?

$$\mathbb{E}[X] \doteq \sum_{i=1}^{k} x_i p_i = x_1 p_1 + x_2 p_2 + \ldots + x_k p_k$$

Two outcomes: win or not

$$\mathbb{E}[X] = \left(\frac{0.01}{100}\right)(1000 - 10) + \left(1 - \frac{0.01}{100}\right)(0 - 10)$$

Prob of win

Outcome on win

Prob of loss

Outcome on loss

Worksheet Review

1. Suppose $\gamma = 0.9$ and the reward sequence is $R_1 = 2, R_2 = -2, R_3 = 0$ followed by an infinite sequence of 7s. What are G_1 and G_0?

- Work backwards
- Sequence of rewards is: R1 = 2, R2 = -2, R3 = 0, R4 = 7, R5 = 7, …..
- Let's start with R4=7, the first of the unending sequence of 7's
- Using our formula G_t = R_{t+1} + gamma R_{t+2} + gamma^2 R_{t+3} + … we write it as:
  - G_3 = R_4 + gamma R_5 + gamma^2 R_6 + gamma^3 R_7 + …
    = 7 + 0.9*7 + (0.9)^2*7 + + (0.9)^3*7…
  - Use our special formula to work out G_3: $\sum_{k=0}^{\infty} \gamma^k R = \frac{R}{1-\gamma}$

Worksheet Review

1. Suppose $\gamma = 0.9$ and the reward sequence is $R_1 = 2, R_2 = -2, R_3 = 0$ followed by an infinite sequence of 7s. What are G_1 and G_0?

- Continue working backwards from G_3 using our other special formula

  - $G_t = R_{t+1} + \gamma G_{t+1}$

Answer:

$G_3 = 7 + 0.9 \times 7 + 0.9^2 \times 7 + ... = 7 \times \frac{1}{1-0.9} = 70$
$G_2 = R_3 + 0.9 \times G_3 = 0 + 0.9 \times 70 = 63$
$G_1 = R_2 + 0.9 \times G_2 = -2 + 0.9 \times 63 = 54.7$
$G_0 = R_1 + 0.9 \times G_1 = 2 + 0.9 \times 54.7 = 51.23$

Worksheet Question

4. Prove that the discounted sum of rewards is always finite, if the rewards are bounded: $|R_{t+1}| \leq R_{\max}$ for all t for some finite $R_{\max} > 0$.

$$\left| \sum_{i=0}^{\infty} \gamma^i R_{t+1+i} \right| < \infty \qquad \text{for } \gamma \in [0, 1)$$

Recall $\sum_{i=0}^{\infty} |a_i| < \infty$ then $\left| \sum_{i=0}^{\infty} a_i \right| < \infty$

4. Prove that the discounted sum of rewards is always finite, if the rewards are bounded: $|R_{t+1}| \leq R_{\max}$ for all t for some finite $R_{\max} > 0$.

$$\left| \sum_{i=0}^{\infty} \gamma^i R_{t+1+i} \right| < \infty \qquad \text{for } \gamma \in [0, 1)$$

If $\sum_{i=0}^{\infty} |\gamma^i R_{t+1+i}| < \infty$ then $\left| \sum_{i=0}^{\infty} \gamma^i R_{t+1+i} \right| < \infty$

4. Prove that the discounted sum of rewards is always finite, if the rewards are bounded: $|R_{t+1}| \leq R_{\max}$ for all t for some finite $R_{\max} > 0$.

$$\left| \sum_{i=0}^{\infty} \gamma^i R_{t+1+i} \right| < \infty \qquad \text{for } \gamma \in [0, 1)$$

If $\sum_{i=0}^{\infty} |\gamma^i R_{t+1+i}| < \infty$ then $\left| \sum_{i=0}^{\infty} \gamma^i R_{t+1+i} \right| < \infty$

$$= \sum_{i=0}^{\infty} \gamma^i |R_{t+1+i}|$$

4. Prove that the discounted sum of rewards is always finite, if the rewards are bounded: $|R_{t+1}| \leq R_{\max}$ for all t for some finite $R_{\max} > 0$.

$$\left| \sum_{i=0}^{\infty} \gamma^i R_{t+1+i} \right| < \infty \qquad \text{for } \gamma \in [0, 1)$$

If $\sum_{i=0}^{\infty} |\gamma^i R_{t+1+i}| < \infty$ then $\left| \sum_{i=0}^{\infty} \gamma^i R_{t+1+i} \right| < \infty$

$$= \sum_{i=0}^{\infty} \gamma^i |R_{t+1+i}|$$

$$\leq \sum_{i=0}^{\infty} \gamma^i R_{\max}$$

4. Prove that the discounted sum of rewards is always finite, if the rewards are bounded: $|R_{t+1}| \leq R_{\max}$ for all t for some finite $R_{\max} > 0$.

$$\left| \sum_{i=0}^{\infty} \gamma^i R_{t+1+i} \right| < \infty \qquad \text{for } \gamma \in [0, 1)$$

If $\sum_{i=0}^{\infty} |\gamma^i R_{t+1+i}| < \infty$ then $\left| \sum_{i=0}^{\infty} \gamma^i R_{t+1+i} \right| < \infty$

$$\begin{aligned} &= \sum_{i=0}^{\infty} \gamma^i |R_{t+1+i}| \\ &\leq \sum_{i=0}^{\infty} \gamma^i R_{\max} \\ &= R_{\max} \sum_{i=0}^{\infty} \gamma^i \end{aligned}$$

4. Prove that the discounted sum of rewards is always finite, if the rewards are bounded: $|R_{t+1}| \leq R_{\max}$ for all t for some finite $R_{\max} > 0$.

$$\left| \sum_{i=0}^{\infty} \gamma^i R_{t+1+i} \right| < \infty \qquad \text{for } \gamma \in [0, 1)$$

If $\sum_{i=0}^{\infty} |\gamma^i R_{t+1+i}| < \infty$ then $\left| \sum_{i=0}^{\infty} \gamma^i R_{t+1+i} \right| < \infty$

$$= \sum_{i=0}^{\infty} \gamma^i |R_{t+1+i}|$$

$$\leq \sum_{i=0}^{\infty} \gamma^i R_{\max}$$

$$= R_{\max} \sum_{i=0}^{\infty} \gamma^i = \frac{R_{\max}}{1-\gamma}$$

$R_{\max}$ and $\frac{1}{1-\gamma}$ are finite.