“The test of a man isn’t what you think he’ll do. It’s what he actually does.”

Frank Herbert, *Dune*

CMPUT 365
Introduction to RL

Marlos C. Machado

https://openart.ai/discovery/sd-1006656316309258270

Class 34/35

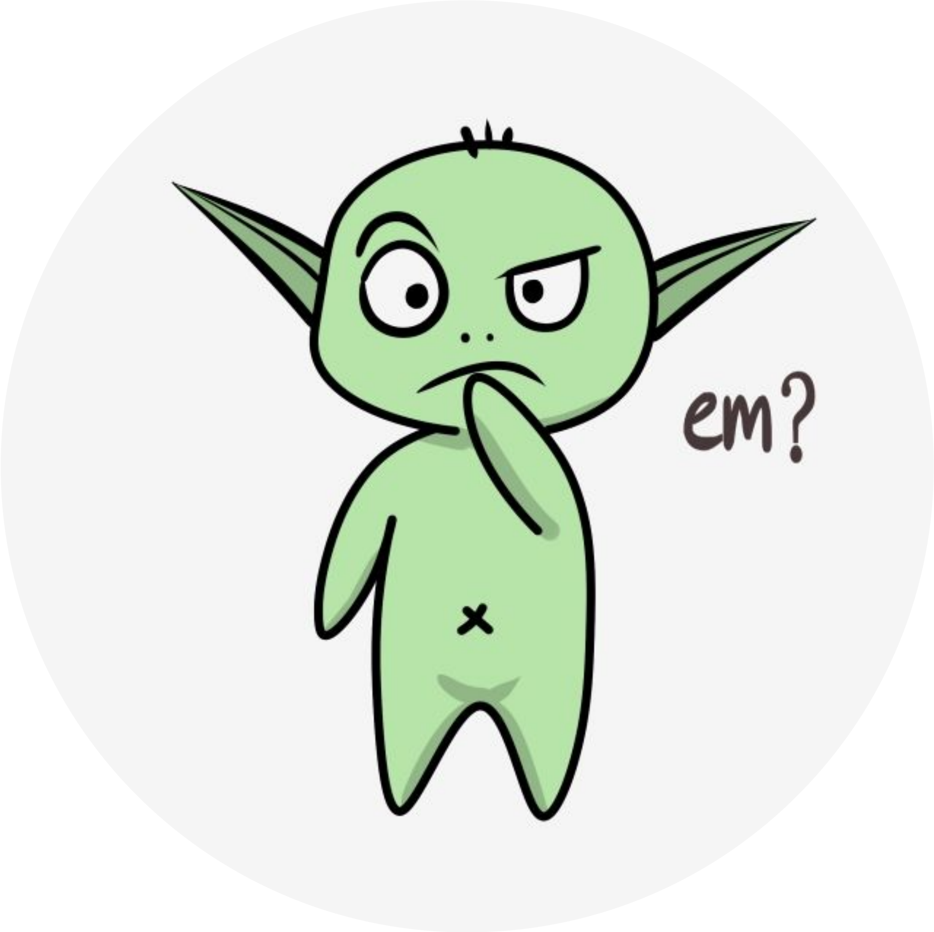
Reminders and Notes

- Next week, my office hours will be shifted to 12:30pm to 2:30pm.
- There are no more quizzes nor programming assignments.
  Your grades are already up on eClass!
- Rich Sutton will give a guest lecture next Monday. Spread the word.
- A note on the final exam:
  - The required reading from the syllabus does not mean that's what will be covered in the final exam. There are some mismatches. Anything we discussed in class is fair game, including Maximization Bias and Double Learning (Sec. 6.7), and Nonlinear Function Approximation: Artificial Neural Networks (Sec. 9.7).
  - Final will be *2 hours long*, and questions will cover the whole term.
  - The final covers everything, but there's a bigger focus (~50%) on the last third of the course.
- SPOT Survey is still available for you.

Last Class: The Policy Gradient Theorem

- The Policy Gradient Theorem [Marbach and Tsitsiklis, 1998, 2001; Sutton et al. 2000] provides an analytic expression for the gradient of performance w.r.t. the policy parameter that does not involve the derivative of the state distribution.

$$\nabla J(\boldsymbol{\theta}) \propto \sum_s \mu(s) \sum_a q_\pi(s, a) \nabla \pi(a|s, \boldsymbol{\theta})$$



https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

A First Policy Gradient Method

$$\nabla J(\boldsymbol{\theta}) \propto \sum_s \mu(s) \sum_a q_\pi(s,a) \nabla \pi(a|s,\boldsymbol{\theta})$$

$$= \mathbb{E}_\pi\left[\sum_a q_\pi(S_t,a) \nabla \pi(a|S_t,\boldsymbol{\theta})\right].$$

Any constant of proportionality can be absorbed into the step size α

It weighs the sum by how often the states occur under the target policy π

$$\boldsymbol{\theta}_{t+1} \doteq \boldsymbol{\theta}_t + \alpha \sum_a \hat{q}(S_t,a,\mathbf{w}) \nabla \pi(a|S_t,\boldsymbol{\theta})$$

REINFORCE: Monte Carlo Policy Gradient [Willliams, 1992]

$$\boldsymbol{\theta}_{t+1} \doteq \boldsymbol{\theta}_t + \alpha \sum_a \hat{q}(S_t, a, \mathbf{w}) \nabla \pi(a|S_t, \boldsymbol{\theta})$$

We want an update that at time t involves just A_t. We need to replace a sum over the RV's possible values by an expectation under π, and then sampling the expectation.

$$\begin{aligned}
\nabla J(\boldsymbol{\theta}) &\propto \mathbb{E}_\pi\left[\sum_a q_\pi(S_t, a) \nabla \pi(a|S_t, \boldsymbol{\theta})\right] \\
&= \mathbb{E}_\pi\left[\sum_a \pi(a|S_t, \boldsymbol{\theta}) q_\pi(S_t, a) \frac{\nabla \pi(a|S_t, \boldsymbol{\theta})}{\pi(a|S_t, \boldsymbol{\theta})}\right] \\
&= \mathbb{E}_\pi\left[q_\pi(S_t, A_t) \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta})}{\pi(A_t|S_t, \boldsymbol{\theta})}\right] \quad = \mathbb{E}_\pi\left[G_t \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta})}{\pi(A_t|S_t, \boldsymbol{\theta})}\right]
\end{aligned}$$

REINFORCE: Monte Carlo Policy Gradient [Willliams, 1992]

$$\nabla J(\boldsymbol{\theta}) \propto \mathbb{E}_\pi\left[G_t \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta})}{\pi(A_t|S_t, \boldsymbol{\theta})}\right]$$

$$\boldsymbol{\theta}_{t+1} \doteq \boldsymbol{\theta}_t + \alpha G_t \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta}_t)}{\pi(A_t|S_t, \boldsymbol{\theta}_t)}$$

REINFORCE uses the full return,
thus it is a Monte Carlo method.

REINFORCE: Monte Carlo Policy Gradient [Willliams, 1992]

$$\boldsymbol{\theta}_{t+1} \doteq \boldsymbol{\theta}_t + \alpha G_t \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta}_t)}{\pi(A_t|S_t, \boldsymbol{\theta}_t)}$$

REINFORCE: Monte-Carlo Policy-Gradient Control (episodic) for π_*

Input: a differentiable policy parameterization $\pi(a|s, \boldsymbol{\theta})$
Algorithm parameter: step size $\alpha > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ (e.g., to $\mathbf{0}$)

Loop forever (for each episode):
  Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot|\cdot, \boldsymbol{\theta})$
  Loop for each step of the episode $t = 0, 1, \ldots, T-1$:
    $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$ (G_t)
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha \gamma^t G \nabla \boxed{\ln \pi(A_t|S_t, \boldsymbol{\theta})}$

Recall:

$$\nabla \ln x = \frac{\nabla x}{x}$$

Policy Gradient with Baseline

- As before, in the gradient bandits algorithm, we can generalize the policy gradient theorem to include a comparison of q_π to an arbitrary baseline *b(s)*:

$$\nabla J(\boldsymbol{\theta}) \propto \sum_s \mu(s) \sum_a \Big(q_\pi(s,a) - b(s)\Big) \nabla \pi(a|s,\boldsymbol{\theta}).$$

- The baseline can be any function, even a random variable, as long as it does not vary with a; because the subtracted quantity is zero (as before):

$$\sum_a b(s) \nabla \pi(a|s,\boldsymbol{\theta}) \; = \; b(s) \nabla \sum_a \pi(a|s,\boldsymbol{\theta}) \; = \; b(s) \nabla 1 \; = \; 0$$

REINFORCE with Baseline [Willliams, 1992]

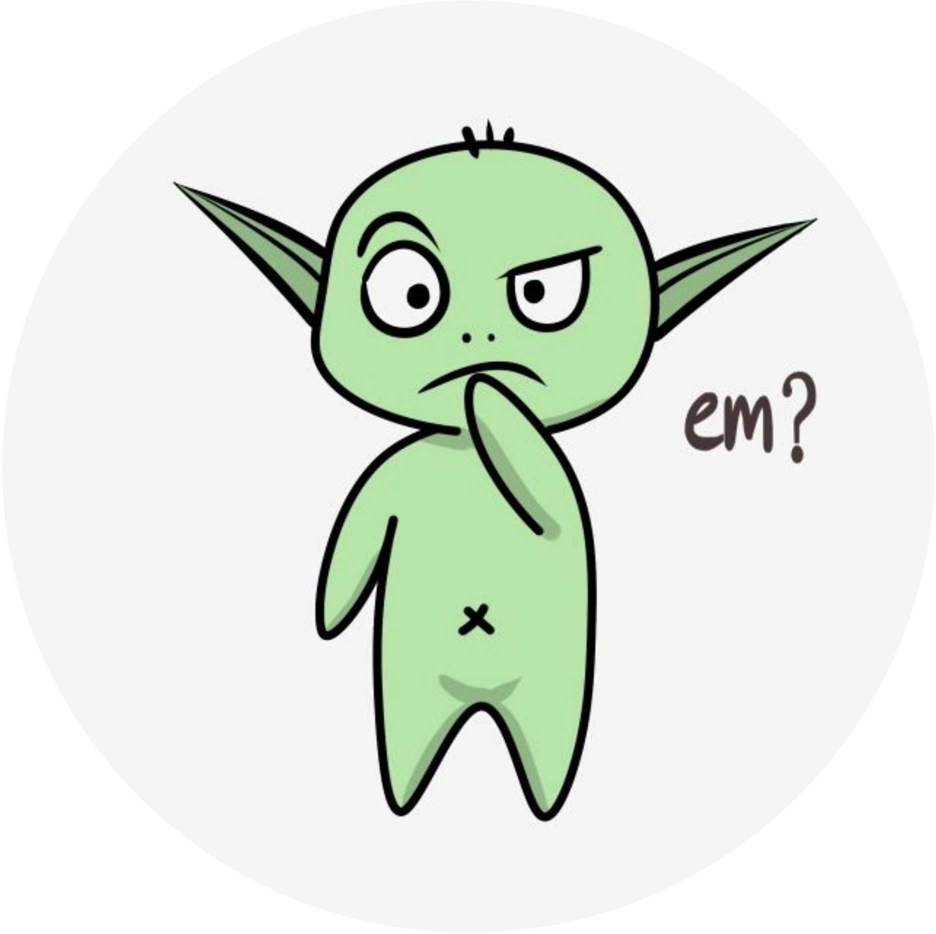
$$\boldsymbol{\theta}_{t+1} \doteq \boldsymbol{\theta}_t + \alpha\Big(G_t - b(S_t)\Big)\frac{\nabla\pi(A_t|S_t,\boldsymbol{\theta}_t)}{\pi(A_t|S_t,\boldsymbol{\theta}_t)}$$

REINFORCE with Baseline [Willliams, 1992]

REINFORCE with Baseline (episodic), for estimating $\pi_{\theta} \approx \pi_*$

Input: a differentiable policy parameterization $\pi(a|s,\boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s,\mathbf{w})$
Algorithm parameters: step sizes $\alpha^{\boldsymbol{\theta}} > 0$, $\alpha^{\mathbf{w}} > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^{d}$ (e.g., to $\mathbf{0}$)

Loop forever (for each episode):
    Generate an episode $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$, following $\pi(\cdot|\cdot,\boldsymbol{\theta})$
    Loop for each step of the episode $t = 0, 1, \ldots, T-1$:
        $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$ $\qquad (G_t)$
        $\delta \leftarrow G - \hat{v}(S_t,\mathbf{w})$
        $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} \delta \nabla \hat{v}(S_t,\mathbf{w})$
        $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} \gamma^t \delta \nabla \ln \pi(A_t|S_t,\boldsymbol{\theta})$



https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Actor-Critic Methods

- In REINFORCE with baseline, the learned state-value function estimates the value of the first state of each state transition.
- In actor-critic methods, the state-value function is applied also to the second state of the transition.
- When the state-value function is used to assess actions in this way it is called a critic, and the overall policy-gradient method is termed an actor–critic method.

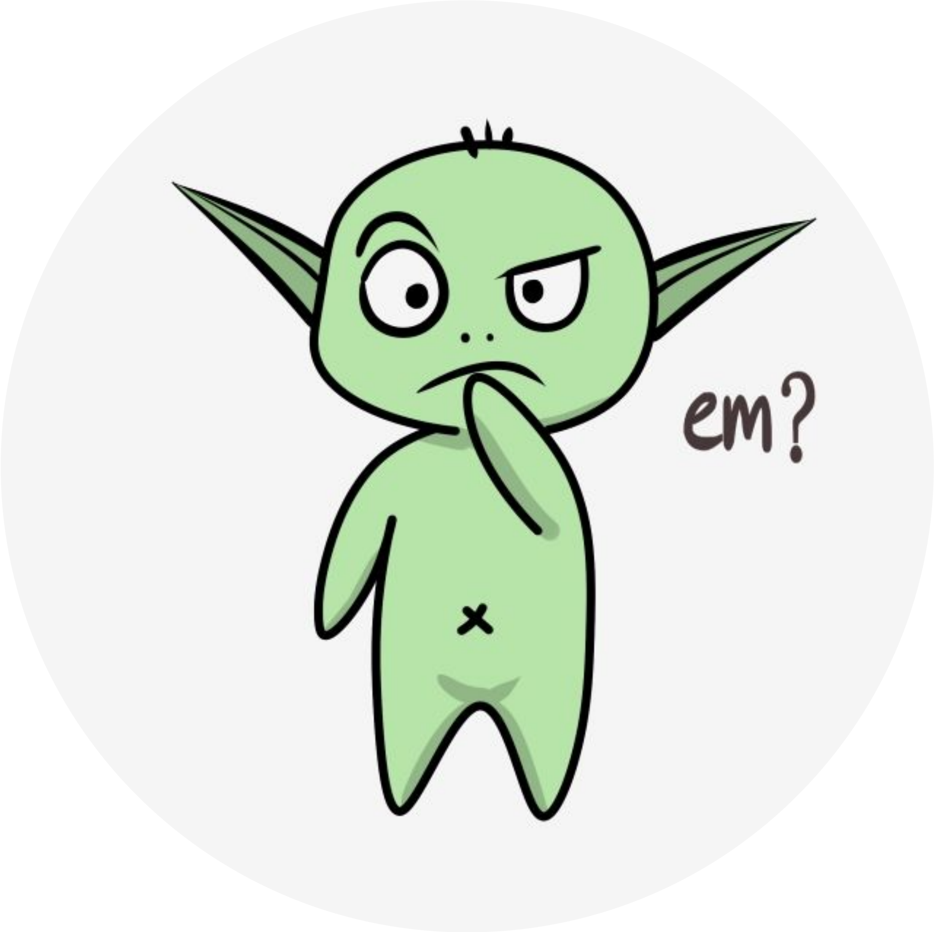
One-Step Actor-Critic

$$
\begin{aligned}
\boldsymbol{\theta}_{t+1} &\doteq \boldsymbol{\theta}_t + \alpha\Big(G_{t:t+1} - \hat{v}(S_t,\mathbf{w})\Big)\frac{\nabla\pi(A_t|S_t,\boldsymbol{\theta}_t)}{\pi(A_t|S_t,\boldsymbol{\theta}_t)} \\
&= \boldsymbol{\theta}_t + \alpha\Big(R_{t+1} + \gamma\hat{v}(S_{t+1},\mathbf{w}) - \hat{v}(S_t,\mathbf{w})\Big)\frac{\nabla\pi(A_t|S_t,\boldsymbol{\theta}_t)}{\pi(A_t|S_t,\boldsymbol{\theta}_t)} \\
&= \boldsymbol{\theta}_t + \alpha\delta_t\frac{\nabla\pi(A_t|S_t,\boldsymbol{\theta}_t)}{\pi(A_t|S_t,\boldsymbol{\theta}_t)}.
\end{aligned}
$$

One-Step Actor-Critic

One-step Actor–Critic (episodic), for estimating $\pi_{\boldsymbol{\theta}} \approx \pi_*$

Input: a differentiable policy parameterization $\pi(a|s, \boldsymbol{\theta})$
Input: a differentiable state-value function parameterization $\hat{v}(s,\mathbf{w})$
Parameters: step sizes $\alpha^{\boldsymbol{\theta}} > 0$, $\alpha^{\mathbf{w}} > 0$
Initialize policy parameter $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ and state-value weights $\mathbf{w} \in \mathbb{R}^{d}$ (e.g., to $\mathbf{0}$)
Loop forever (for each episode):
    Initialize S (first state of episode)
    $I \leftarrow 1$
    Loop while S is not terminal (for each time step):
        $A \sim \pi(\cdot|S, \boldsymbol{\theta})$
        Take action A, observe S', R
        $\delta \leftarrow R + \gamma \hat{v}(S',\mathbf{w}) - \hat{v}(S,\mathbf{w})$ (if S' is terminal, then $\hat{v}(S',\mathbf{w}) \doteq 0$)
        $\mathbf{w} \leftarrow \mathbf{w} + \alpha^{\mathbf{w}} \delta \nabla \hat{v}(S,\mathbf{w})$
        $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha^{\boldsymbol{\theta}} I \delta \nabla \ln \pi(A|S, \boldsymbol{\theta})$
        $I \leftarrow \gamma I$
        $S \leftarrow S'$



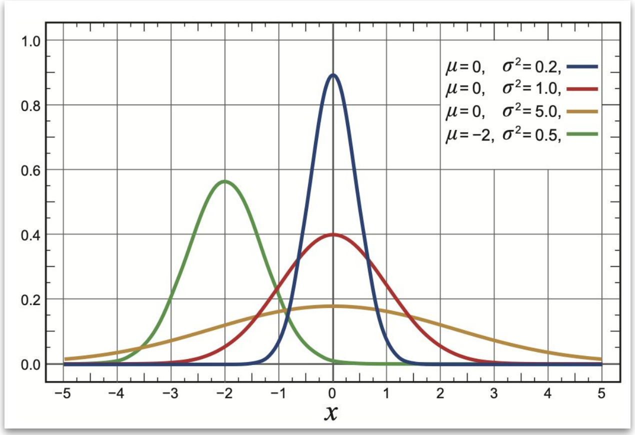
https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

Policy Parameterization for Continuous Actions

- But how do we deal with large action spaces?

- We learn statistics of the probability distribution over actions to take.
  - E.g., the action set might be the real numbers, with actions chosen from a normal distribution.

$$p(x) \doteq \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

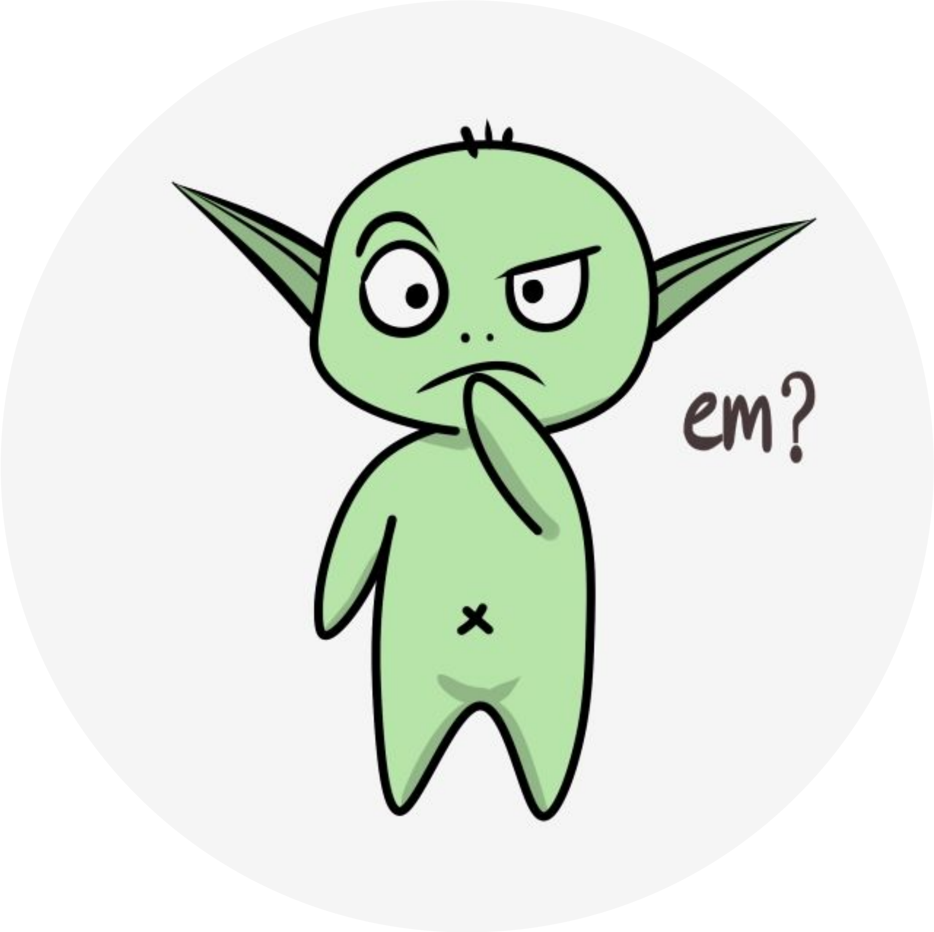


Policy Parameterization for Continuous Actions

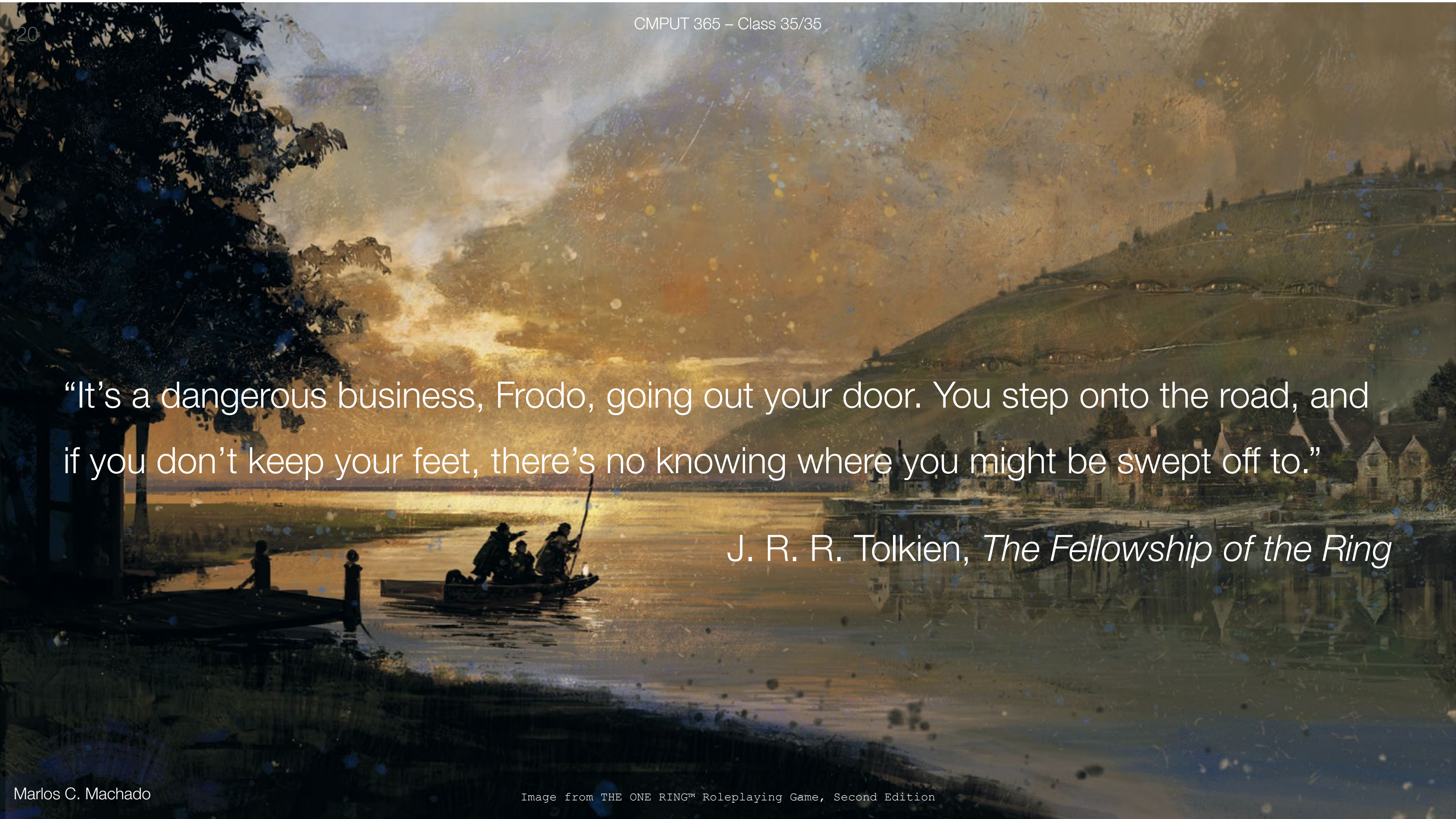
- We need a policy parameterization, as always ¯_(ツ)_/¯

$$\pi(a|s,\boldsymbol{\theta}) \doteq \frac{1}{\sigma(s,\boldsymbol{\theta})\sqrt{2\pi}} \exp\left(-\frac{(a-\mu(s,\boldsymbol{\theta}))^2}{2\sigma(s,\boldsymbol{\theta})^2}\right)$$

$$\mu(s,\boldsymbol{\theta}) \doteq \boldsymbol{\theta}_\mu{}^\top \mathbf{x}_\mu(s) \quad \text{and} \quad \sigma(s,\boldsymbol{\theta}) \doteq \exp\Big(\boldsymbol{\theta}_\sigma{}^\top \mathbf{x}_\sigma(s)\Big)$$



https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht



"It's a dangerous business, Frodo, going out your door. You step onto the road, and if you don't keep your feet, there's no knowing where you might be swept off to."

J. R. R. Tolkien, *The Fellowship of the Ring*

Image from THE ONE RING™ Roleplaying Game, Second Edition

What's next?

Undergraduate courses:

- CMPUT 366 Search and Planning
- CMPUT 455 Search, Knowledge and Simulation
- CMPUT 466 Machine Learning Essentials
- CMPUT 467 Machine Learning II

Grad school 😀

- CMPUT 656 Human-in-the-Loop Reinforcement Learning by M. E. Taylor
- CMPUT 628 Deep Reinforcement Learning by M. C. Machado
- CMPUT 609 Reinforcement Learning II by R. Sutton
- CMPUT 653 Theoretical Foundations of Reinforcement Learning by C. Szepesvari
- CMPUT 653 Real-Time Policy Learning by A. R. Mahmood

“I am glad you are here with me. Here at the end of all things, Sam.”

J. R. R. Tolkien, *The Return of the King*

Image from THE ONE RING™ Roleplaying Game, Second Edition