Final Review

CMPUT 397
Fall 2020

Comments: Dec 4

- The Final will be of the same format as the Midterm
    - on eClass, with long answer questions
    - 2 hours to complete the exam
- Practice Finals have been released

Reviewing the Entire Course

- We recently reviewed Course 1 and Course 2
- To review Course 3, we primarily ask: how are thing different under function approximation?

How are things the same?

- We were always try to estimate a function: v_π or q_π or $v*$ or $q*$ or π itself
  - $v_\pi : \mathcal{S} \to \mathbb{R}, q_\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and so on
- The main difference is in the type of function we learn to try to get an accurate estimate of these true functions

How are things different?

- Before, we were able to represent this function exactly, with a table (tabular)
  - V is a vector (or array) of size number of states, with entry $V(s)$ estimating $v_\pi(s)$
- More generally, we can only approximate these functions
  - e.g., $\hat{v}(s, w) \approx v_\pi(s)$ for each s, for parameterized function $\hat{v}(\cdot, w)$
- Tabular is a special case: $V(s) = \hat{v}(s, w)$ under tabular features (x(s) one-hot encoding)
- Can also see it as a generalization, rather than a difference

What is the source of inaccuracy?

- For tabular estimate V, the inaccuracy is due to insufficient samples and/or updates
  - Question: when and why is V inaccurate in DP? Do we eventually converge to v_π?
  - Question: when and why is V inaccurate in TD? Do we eventually converge to v_π?
- For more general parameterized functions, $\hat{v}(\cdot, w)$, the inaccuracy is due also to the inability to represent v_π

Inaccuracy under FA

- All algorithms (tabular and more general parameterized functions) initialize the weights w to some value

- $\hat{v}(\cdot, w)$ is inaccurate after 10 updates because it has not seen enough samples

    - just like in the tabular setting

- In the limit of samples and updates, $\hat{v}(\cdot, w)$ might still be inaccurate

    - even for w^*, the best weights we could obtain under infinite data and samples, we might have that $\hat{v}(s, w^*) \neq v_\pi(s)$

Estimation error and Approximation error

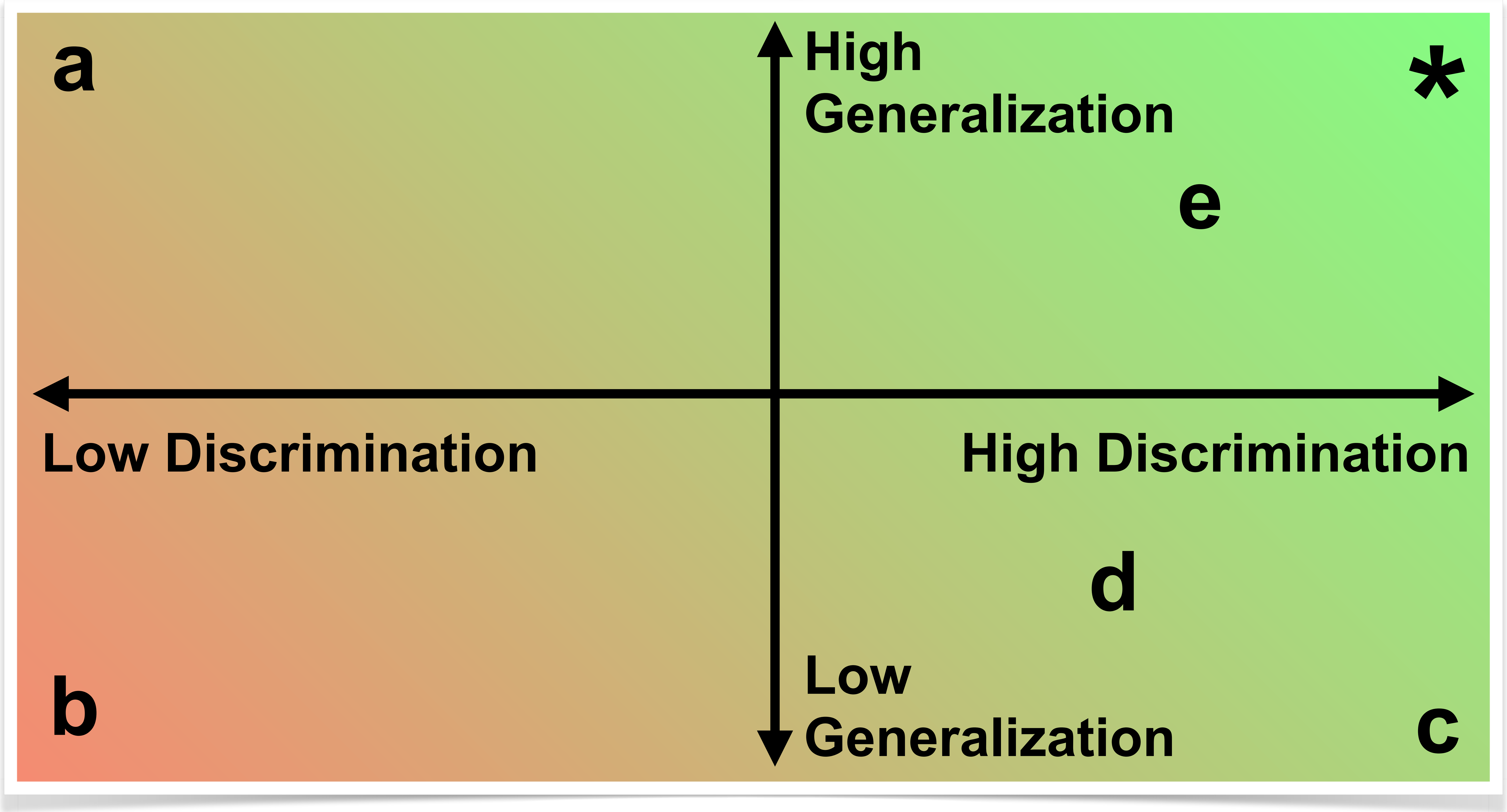
- $\hat{v}(\cdot, w)$ is inaccurate after 10 updates because it has not seen enough samples
  - this is called **estimation error** (due to insufficient data)
- In the limit of samples and updates, $\hat{v}(\cdot, w)$ might still be inaccurate
  - this is called **approximation error** (also called the bias of the estimator)


Can we represent the value function with a table?
No
Yes
Are we using average reward?
Do we already have access to a model?
No
Yes
Yes
No
Will we learn on each timestep?
Are the actions continuous?
Is this a control problem?
Will we learn a model?
Yes
No
Yes
No
Yes
No
Yes
No
Is this a control problem?
Gradient Monte Carlo
Gaussian Actor-Critic
Softmax Actor-Critic
Differential Semi-Gradient SARSA
Q-Planning
Dyna-Q+
Dyna-Q
Will we learn on each timestep?
No
Yes
Semi-Gradient TD
Expected SARSA
Q-Learning
SARSA
Value Iteration
Policy Iteration
Iterative Policy Evaluation
Yes
No
Is this a control problem?
Is this a control problem?
Yes
No
No
Yes
Q-Learning
Expected SARSA
SARSA
TD
Off-Policy MC
MC Prediction
e-soft MC
Exploring Starts MC

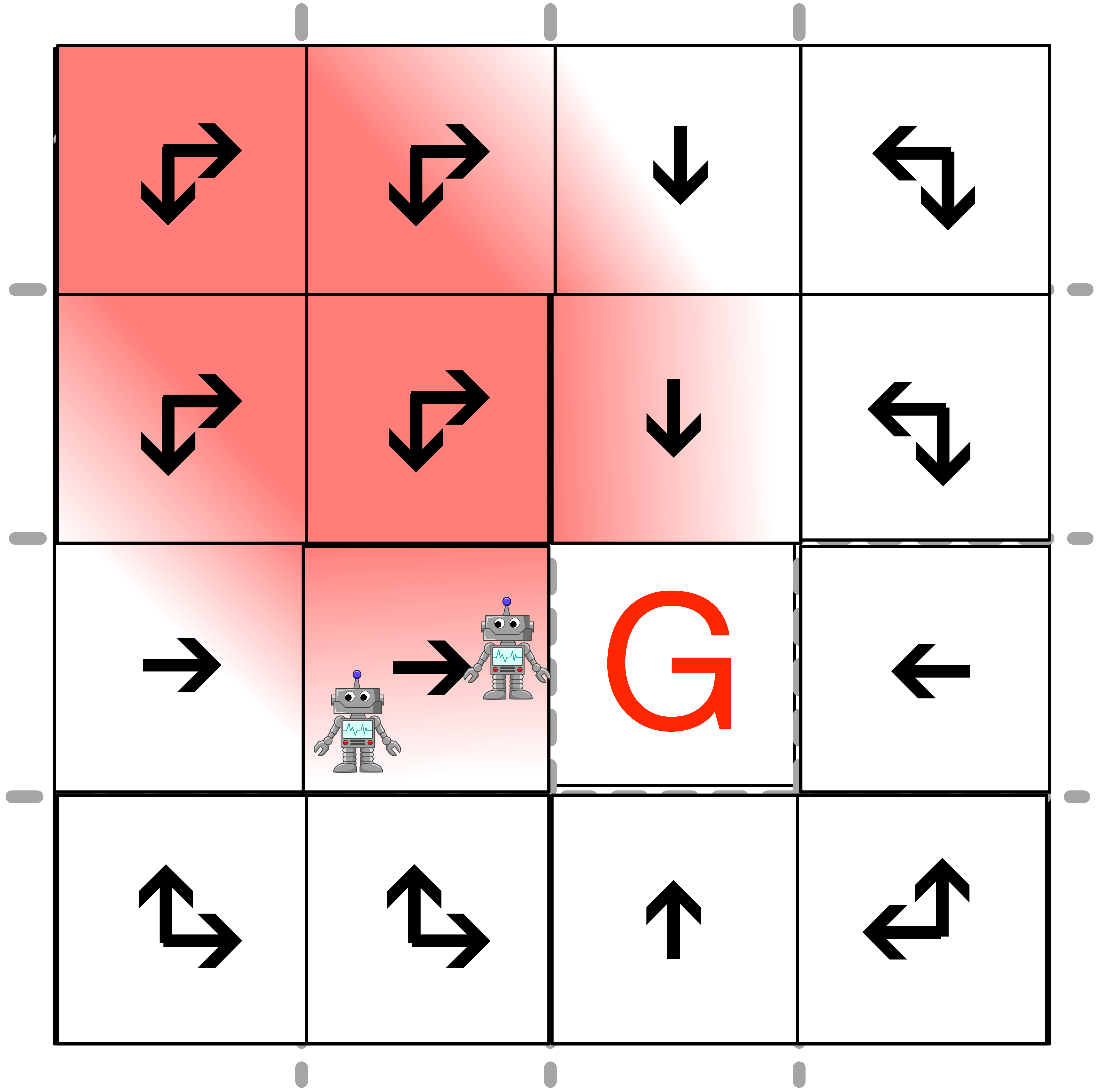
Self-test: How would we use FA for bandits?

- When we did bandits, we estimated $Q(a)$
- What might you use for $\hat{q}(a, w)$? First, what is the tabular form, to represent $Q(a)$?
- Now imagine I tell you there are 1000 actions, and actions 1 - 20 are similar, 21-25 are similar, 25-40 are similar and 40-1000 are similar. What features might you use?
- Is this FA better than the tabular one? Why or why not?

Generalization and Discrimination



Due to approximation error from FA we need a way to decide trade-offs in error across states

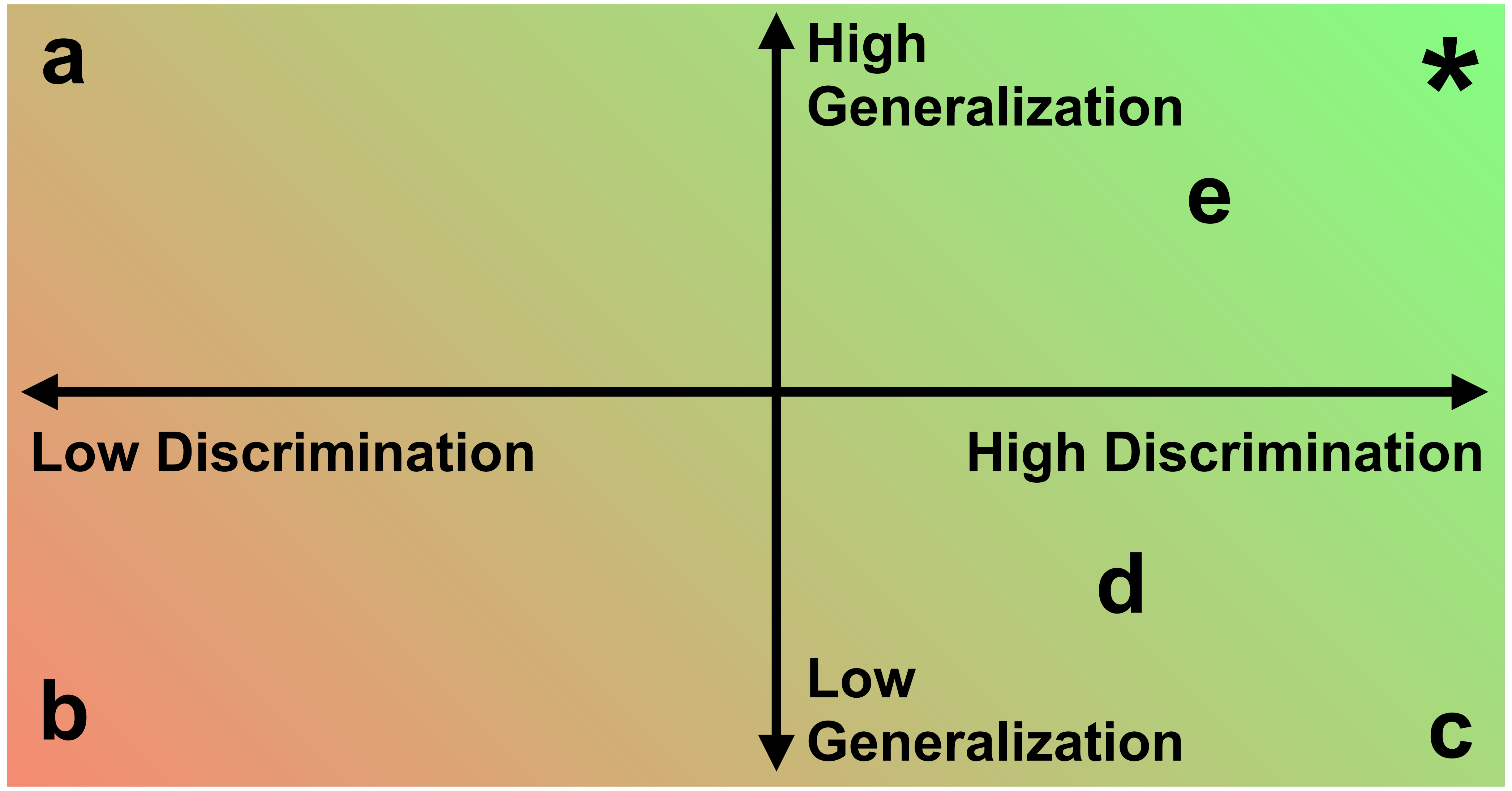


Mean Squared Value Error

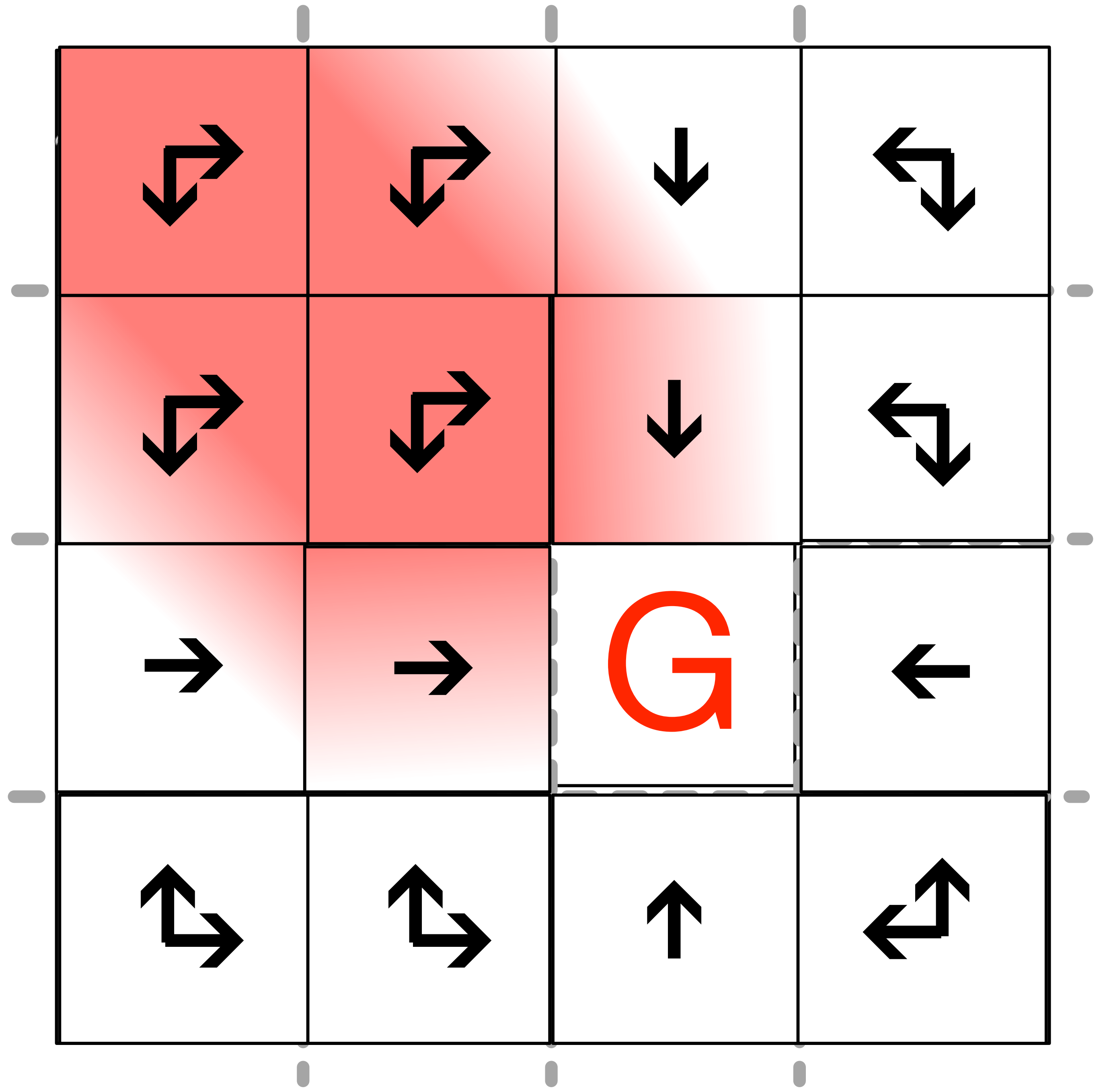
$$\sum_{s} \mu(s)[v_{\pi}(s) - \hat{v}(s, \mathbf{w})]^2$$

The fraction of time we spend in S when following policy π

Self-test: How does the FA and the Weighting mu in the VE interact?



$$\sum_{s} \mu(s)[v_{\pi}(s) - \hat{v}(s, \mathbf{w})]^2$$



Question 1: Imagine we have features that correspond to a. Will vhat be accurate? If no, where will it be more inaccurate and where will it be more accurate?

Question 2: How about e? And how about at the star?

Let’s go through some quiz questions

- C3M1: Q5, Q8, Q9, Q10, Q11
- C3M2: Q3, Q5, Q9, Q11
- C3M3: Q1, Q2, Q8
- Final comment: The final will only include concepts from Course 1, 2 and 3, and will not include anything about average reward and importance sampling.
- You will not be tested on guest lectures, nor on content in Course 4