

Lecture 2: Immediate RL and Bandits

B. Ravindran

Reinforcement Learning

- ❑ Familiar models of machine learning
 - ❑ Learning from data
- ❑ How did you learn to cycle?
 - ❑ Trial and error!
 - ❑ Falling down hurts!
 - ❑ Evaluation, not instruction
 - ❑ Reinforcement Learning
- ❑ Walk, Talk, etc.



Immediate Reinforcement

- ❑ The payoff accrues immediately after an action is chosen
- ❑ One key question - the dilemma between exploration and exploitation
- ❑ *Bandit problems* encapsulate ‘Explore vs Exploit’

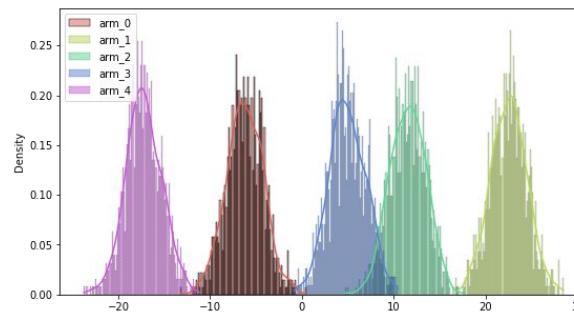


The Explore-Exploit Dilemma

- ❑ Explore to find profitable actions
- ❑ Exploit to act according to the best observations already made
- ❑ Always exploiting might not be optimal
- ❑ Always exploring might not be optimal either
- ❑ Hence, there is an explore-exploit dilemma

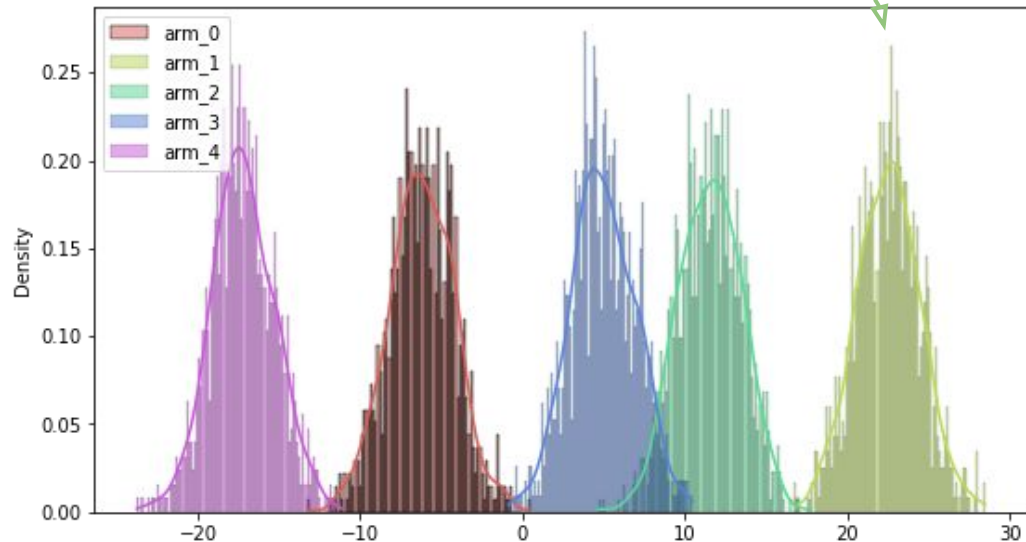
Multi-arm Bandits

- ❑ n-arm bandit problem is to learn to preferentially select a particular action (arm) from a set of n actions $(1, 2, 3, \dots, n)$
- ❑ Each selection results in a reward derived from the respective probability distribution
- ❑ Arm i has a reward distribution with mean μ_i and $\mu^* = \max\{\mu_i\}$



Objective

- ❑ Identify the **correct** arm eventually



Traditional Approaches

□ Let $r_{i,k}$ be the reward sample acquired when i^{th} arm is selected for the k^{th} time

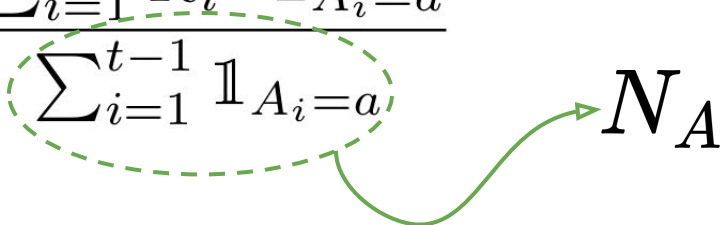
□ Define:

$$Q_t(a) \doteq \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t}$$

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

$$A_t \doteq \operatorname{argmax}_a Q_t(a) \quad (\text{greedy action})$$

Traditional Approaches

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$


N_A Number of times arm A is sampled

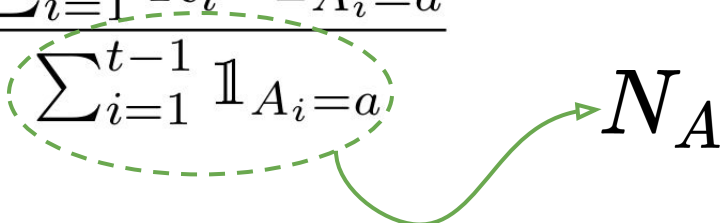
$$NewEstimate \leftarrow OldEstimate + StepSize \left[Target - OldEstimate \right]$$

$$Q_{n+1} = Q_n + \alpha \left[R_n - Q_n \right]$$

□ Setting $\alpha = \frac{1}{N_A}$ yields the same average

Traditional Approaches

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

 N_A Number of times arm A is sampled

$$NewEstimate \leftarrow OldEstimate + StepSize [Target - OldEstimate]$$

$$Q_{n+1} = Q_n + \alpha [R_{n+1} - Q_n]$$

□ Setting $\alpha = \frac{1}{N_A}$ yields the same average

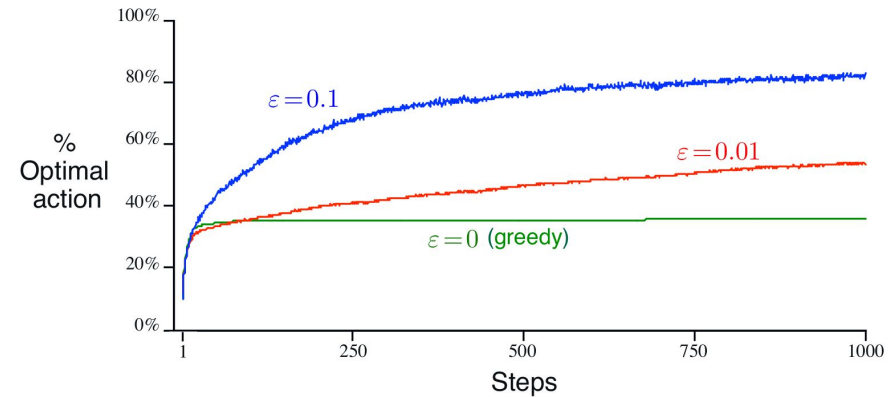
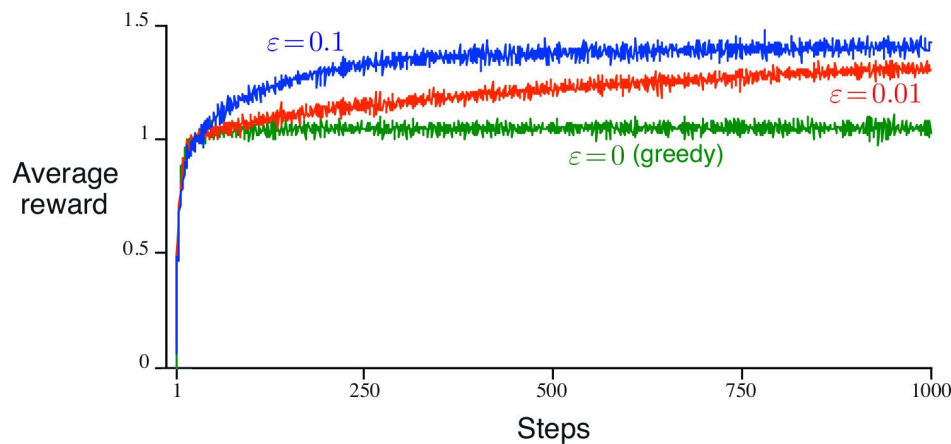
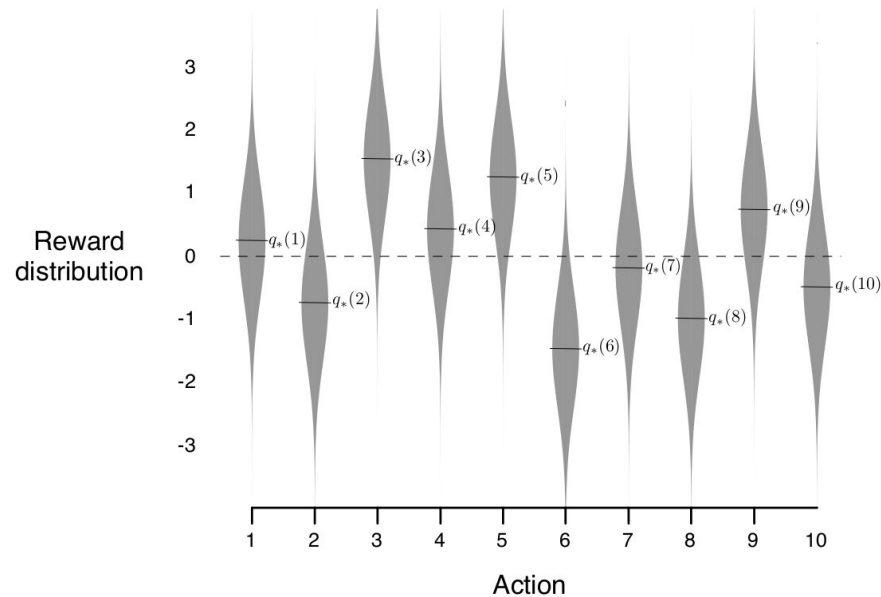
Traditional Approaches

- ❑ **Epsilon Greedy** : Select arm $a^* = \arg \max_a Q_t(a)$ with probability $1 - \epsilon$ and select any arbitrary arm with probability ϵ
- ❑ **Softmax** : Select arms with probability proportional to the current value estimates


$$\Pr\{A_t = a\} \doteq \frac{e^{(Q_t(a) / \tau)}}{\sum_{b=1}^k e^{(Q_t(b) / \tau)}}$$

- ❑ **Asymptotic Convergence guarantees**

ϵ -Greedy Example



Customization



YAHOO!
NEWS




[Search](#)

[Trending News](#) [Taft Point Yosemite](#) [Marine Hawaii](#) [South Korea](#) [Biker brawl](#) [BMW X5](#) [Southwest Airlines](#)

News Home

- U.S.
- World
- Politics
- Tech
- Science
- Health
- Odd News
- Local
- Dear Abby
- Comics
- ABC News
- Yahoo Originals
- Photos


Recommended Games



[More games »](#)


Thank you for helping us improve your Yahoo experience

[Learn more about your feedback.](#)




Can anything stop ISIS in Iraq?
Yahoo's Bianna Golodryga talks with experts about the fall of the major Iraqi city of Ramadi. ... [Read More »](#)


White House: Ramadi capture by Islamic State a 'setback'



Heightened security in Waco after deadly biker gang shootout

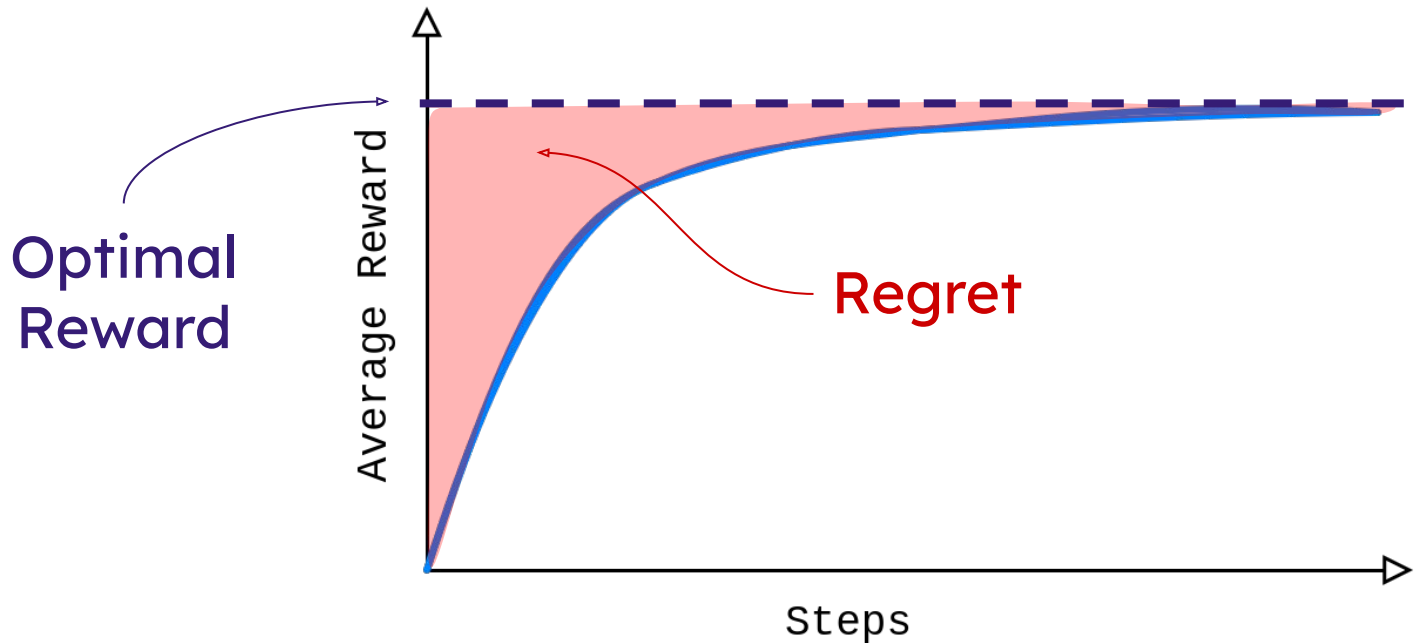


Lindsey Graham: 'I am running because the world is falling apart'

[All News](#) [Yahoo Originals](#)  [AP](#) [Reuters](#)

Objective

- ❑ Identify the **correct** arm eventually
- ❑ Maximize the total rewards obtained
- ❑ Minimize **regret (= loss)** while learning



Objective

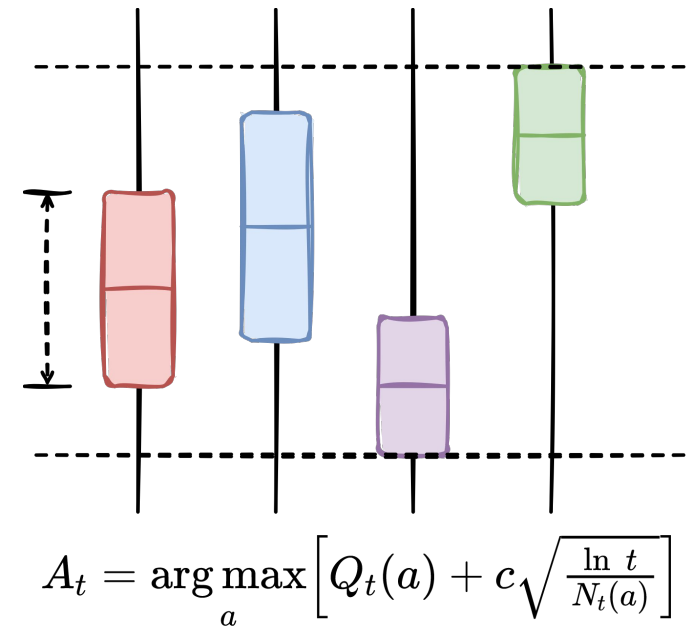
- ❑ Identify the correct arm eventually
- ❑ Maximize the total rewards obtained
- ❑ Minimize regret (= loss) while learning
- ❑ Probably Approximately Correct (PAC) frameworks
 - ❑ Identification of an ϵ -optimal arm with probability $1 - \delta$
 - ❑ ϵ -Optimal: Mean of the selected arm satisfies
 - ❑ Minimize sample complexity: Order of samples required for such an arm identification

Other Approaches

- ❏ Median Elimination (Even-Dar et al., 2006)
- ❏ Upper Confidence Bounds (UCB) (Auer et al., 1998, 2010)
- ❏ Thompson Sampling (Chappelle & Li, 2001, Agrawal & Goyal, 2012)

UCB

- ❑ ϵ -greedy action selection forces the non-greedy actions to be tried
- ❑ With no preference for arms that are nearly greedy or particularly uncertain
- ❑ Opt for non-greedy actions based on their potential for optimality & consider estimate uncertainties



UCB


$$A_t = \arg \max_a \left[Q_t(a) + c \sqrt{\frac{\ln t}{N_t(a)}} \right]$$

Upper bound on
true $Q_*(a)$

Uncertainty in the
estimate of $Q_*(a)$

- ❑ $c > 0$ - controls the degree of exploration
- ❑ Sub-optimal arm j played fewer than $\frac{8 \ln t}{\Delta_j^2}$ times
- ❑ Further improvements focus on reducing the constants

Customization






YAHOO!
NEWS

Trending News Taft Point Yosemite Marine Hawaii South Korea Biker brawl BMW X5 Southwest Airlines

News Home

U.S.
World
Politics
Tech
Science
Health
Odd News
Local
Dear Abby
Comics
ABC News
Yahoo Originals
Photos


Recommended Games



More games »

Thank you for helping us improve your Yahoo experience


[Learn more about your feedback.](#)




Can anything stop ISIS in Iraq?

Yahoo's Bianna Golodryga talks with experts about the fall of the major Iraqi city of Ramadi. ... [Read More »](#)


White House: Ramadi capture by Islamic State a 'setback'



Heightened security in Waco after deadly biker gang shootout



Lindsey Graham: 'I am running because the world is falling apart'

All News Yahoo Originals  News AP Reuters

Ad Selection

Shop for Florists in Chennai on Google

Sponsored ⓘ



Online Flowers
Delivery
₹1,749
Ferns N Petals



Message In A
Bottle with teg
₹349
Ferns N Petals



Classic Bunch -
online flower ...
₹499
FlowerAura
Special offer



Online Flower
Delivery
₹599
Ferns N Petals



Relish Of
Heavenly Treat
₹1,399
Ferns N Petals



Florists In Chennai - Same Day Delivery Within 4 Hrs - floweraura.com

(Ad) www.floweraura.com/Online-Florist/Chennai ▼

Online **Flowers** & Gifts Delivery @ Rs 399. Best Price, 100% Smile Guaranteed.

Delivery in 4 Hrs · Mid-Night Delivery · No Hidden Cost · Free Shipping · Flowers Starting @ Rs 399

Types: Cakes, Flowers, Gifts, Chocolate

Flowers Delivery in Chennai - Express Delivery in 2-3 Hrs

(Ad) www.flowersnfruits.com/Flower_Delivery/Chennai ▼ 099300 06747

Order **Flowers** Now For Express Delivery within 2-3 hrs Anywhere in **Chennai**.

Contextual Bandits

- ❑ Different ads for different users
 - ❑ One bandit for each user!
- ❑ Hard to train - Need several rounds of experience with same user
- ❑ Assume that the parameters of the reward distributions themselves are determined by a set of hyperparameters
 - ❑ Typical assumption is a linear parameterization of the expectation

Contextual Bandits

- ❑ Assume that each user is represented by a set of features
 - ❑ Can be joint features of user and arm
- ❑ The “statistic” used for choosing arms is now dependent on these features
- ❑ Could correspond to the presence or absence of different signals

LinUCB

- ❑ One of the more popular contextual bandit algorithms
- ❑ *Predicted expected reward* assumed to be a linear function of the features
- ❑ Use ridge regression to fit parameters
- ❑ Can derive upper confidence bounds for the regression fit
- ❑ Use UCB like action selection
- ❑ Gives better performance with lesser “training” data