

# Regret Bound for UCB (Sketch)

Let  $\Delta_i = \mu^* - \mu_i$ , where  $\mu_i = \mathbb{E}[a_i]$  and  $\mu^*$  is the mean reward of the optimal arm. Assume arm  $a_1$  is optimal. Let  $N_i(n)$  denote the number of times arm  $i$  is pulled up to time  $n$ .

$$R_n = \sum_{i:\Delta_i>0} \Delta_i \mathbb{E}[N_i(n)].$$

Define the “good event” where the true mean of the optimal arm is always below its UCB and after  $U_i$  samples of action  $i$ , its UCB is below  $\mu_1$ :

$$G_i := \left\{ \mu_1 \leq \min_t \text{UCB}_1(t, \delta) \right\}, \& \left\{ \hat{\mu}_{i, U_i} + \sqrt{2 \log(2/\delta)/U_i} < \mu_1 \right\},$$

**Claim.** If  $G_i$  holds, then arm  $i \neq a^*$  is played at most  $U_i$  times.

**Proof (by contradiction).** Assume  $G_i$  holds. Then there exists a time when arm  $i$ 's UCB must have been highest and it was pulled again for some  $N + t(i) > U_i$ .

$$\text{UCB}_i(t, \delta) \leq \mu_i + \sqrt{2 \log(1/\delta)/U_i} \leq \mu_1 \leq \text{UCB}_1(t, \delta)$$

where the first inequality holds because  $N + t(i) > U_i$  and the second and third hold due to the definition of the good event holding. But this implies arm  $i$  cannot be selected at time  $t$ , a contradiction. Hence  $N_t(i) \leq U_i$  when  $G_i$  holds. Therefore,

$$\mathbb{E}[N_i(n)] \leq U_i + \Pr(G_i^c) n.$$

If we choose

$$U_i := 2 \log(1/\delta) / \Delta_i^2.$$

and set  $\delta = 1/n^2$  and applying a union bound ( $\Pr(G_i^c) \leq \sum_n \delta = n\delta = \frac{1}{n}$ ) yields

$$R_n \lesssim \sum_{i:\Delta_i>0} \frac{4 \log n}{\Delta_i} + \sum_i \Delta_i.$$

The full proof requires more details on how to derive  $U_i$ : see Theorem 7.1 attimore and Szepesvari, Bandit Algorithms for a formal proof.

# Lecture 9: Data Efficient Reinforcement Learning

Emma Brunskill

CS234 Reinforcement Learning

Winter 2026

- Many slides from or derived from David Silver, Examples new and proof new.

*poll everywhere*

## Select all that are true

- F • RLHF and DPO both learn an explicit representation of a reward model from preference data
- F • Both are constrained to be at most as good as the best examples in the pairwise preference data
- F • DPO does not use a reference policy
- None of the above
- Not Sure

## Select all that are true

- RLHF and DPO both learn an explicit representation of a reward model from preference data
- Both are constrained to be at most as good as the best examples in the pairwise preference data
- DPO does not use a reference policy
- None of the above
- Not Sure

# Class Structure

- Last time: Midterm
- **This time: Data Efficient Reinforcement Learning – Bandits**
- Next time: Data Efficient Reinforcement Learning

# Recall RL Involves

- Generalization
- Optimization
- Delayed outcomes
- Exploration

# Evaluation Criteria

- How do we evaluate how "good" an algorithm is?
- If converges?
- If converges to optimal policy?
- How quickly reaches optimal policy?
- Mistakes made along the way?
- Will introduce different measures to evaluate RL algorithms

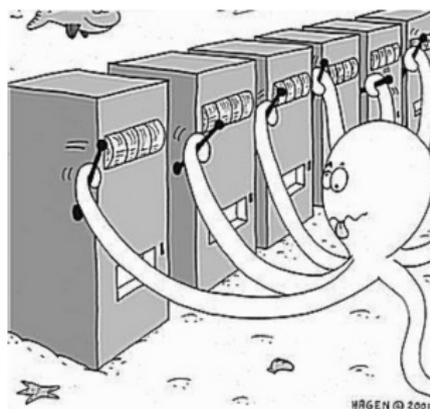
# Settings, Frameworks & Approaches

- Over next couple lectures will consider 2 settings, multiple frameworks, and approaches
- Settings: Bandits (single decisions), MDPs
- Frameworks: evaluation criteria for formally assessing the quality of a RL algorithm
- Approaches: Classes of algorithms for achieving particular evaluation criteria in a certain set
- Note: We will see that some approaches can achieve multiple frameworks in multiple settings

- Setting: Introduction to multi-armed bandits & Approach: greedy methods
- Framework: Regret
- Approach:  $\epsilon$ -greedy methods
- Approach: Optimism under uncertainty
- Framework: Bayesian regret
- Approach: Probability matching / Thompson sampling

# Multiarmed Bandits

- Multi-armed bandit is a tuple of  $(\mathcal{A}, \mathcal{R})$
- $\mathcal{A}$  : known set of  $m$  actions ('arms')
- $\mathcal{R}^a(r) = \mathbb{P}[r \mid a]$  is an unknown probability distribution over rewards
- At each step  $t$  the agent selects an action  $a_t \in \mathcal{A}$
- The environment generates a reward  $r_t \sim \mathcal{R}^{a_t}$
- Goal: Maximize cumulative reward  $\sum_{\tau=1}^t r_{\tau}$



# Toy Example: Ways to Treat Broken Toes

- Consider deciding how to best treat patients with broken toes
- Imagine have 3 possible options: (1) surgery (2) buddy taping the broken toe with another toe, (3) do nothing
- Outcome measure / reward is binary variable: whether the toe has healed (+1) or not healed (0) after 6 weeks, as assessed by x-ray

**Note:** This is a made up example. This is not the actual expected efficacies of the various treatment options for a broken toe

# L11N2 Check Your Understanding: Bandit Toes

- Consider deciding how to best treat patients with broken toes
- Imagine have 3 common options: (1) surgery (2) buddy taping the broken toe with another toe (3) doing nothing
- Outcome measure is binary variable: whether the toe has healed (+1) or not (0) after 6 weeks, as assessed by x-ray
- Model as a multi-armed bandit with 3 arms, where each arm is a Bernoulli variable with an unknown parameter  $\theta_i$
- Select all that are true

F ① Pulling an arm / taking an action corresponds to whether the toe has healed or not

② A multi-armed bandit is a better fit to this problem than a MDP because treating each patient involves multiple decisions *but it involves a sigl dec.*

T ③ After treating a patient, if  $\theta_i \neq 0$  and  $\theta_i \neq 1 \forall i$  sometimes a patient's toe will heal and sometimes it may not

④ Not sure

# L11N2 Check Your Understanding: Bandit Toes Solution

- Consider deciding how to best treat patients with broken toes
- Imagine have 3 common options: (1) surgery (2) buddy taping the broken toe with another toe (3) doing nothing
- Outcome measure is binary variable: whether the toe has healed (+1) or not (0) after 6 weeks, as assessed by x-ray
- Model as a multi-armed bandit with 3 arms, where each arm is a Bernoulli variable with an unknown parameter  $\theta_i$
- Select all that are true
  - 1 Pulling an arm / taking an action corresponds to whether the toe has healed or not
  - 2 A multi-armed bandit is a better fit to this problem than a MDP because treating each patient involves multiple decisions
  - 3 After treating a patient, if  $\theta_i \neq 0$  and  $\theta_i \neq 1 \forall i$  sometimes a patient's toe will heal and sometimes it may not
  - 4 Not sure

# Greedy Algorithm

- We consider algorithms that estimate  $\hat{Q}_t(a) \approx Q(a) = \mathbb{E}[R(a)]$
- Estimate the value of each action by Monte-Carlo evaluation

$$\hat{Q}_t(a) = \frac{1}{N_t(a)} \sum_{i=1}^t r_i \mathbb{1}(a_i = a) \quad \text{empirical avg}$$

- The **greedy** algorithm selects the action with highest value

$$a_t^* = \arg \max_{a \in \mathcal{A}} \hat{Q}_{t-1}(a)$$

# Toy Example: Ways to Treat Broken Toes

- Imagine true (unknown) Bernoulli reward parameters for each arm (action) are
  - surgery:  $Q(a^1) = \theta_1 = .95$
  - buddy taping:  $Q(a^2) = \theta_2 = .9$
  - doing nothing:  $Q(a^3) = \theta_3 = .1$

# Toy Example: Ways to Treat Broken Toes, Greedy

- Imagine true (unknown) Bernoulli reward parameters for each arm (action) are
  - surgery:  $Q(a^1) = \theta_1 = .95$
  - buddy taping:  $Q(a^2) = \theta_2 = .9$
  - doing nothing:  $Q(a^3) = \theta_3 = .1$
- Greedy
  - 1 Sample each arm once
    - Take action  $a^1$  ( $r \sim \text{Bernoulli}(0.95)$ ), get 0,  $\hat{Q}(a^1) = 0$
    - Take action  $a^2$  ( $r \sim \text{Bernoulli}(0.90)$ ), get +1,  $\hat{Q}(a^2) = 1$
    - Take action  $a^3$  ( $r \sim \text{Bernoulli}(0.1)$ ), get 0,  $\hat{Q}(a^3) = 0$
  - 2 What is the probability of greedy selecting each arm next? Assume ties are split uniformly.

100% prob for  $a^2$

# Toy Example: Ways to Treat Broken Toes, Greedy

- Imagine true (unknown) Bernoulli reward parameters for each arm (action) are
  - surgery:  $Q(a^1) = \theta_1 = .95$
  - buddy taping:  $Q(a^2) = \theta_2 = .9$
  - doing nothing:  $Q(a^3) = \theta_3 = .1$
- Greedy
  - 1 Sample each arm once
    - Take action  $a^1$  ( $r \sim \text{Bernoulli}(0.95)$ ), get 0,  $\hat{Q}(a^1) = 0$
    - Take action  $a^2$  ( $r \sim \text{Bernoulli}(0.90)$ ), get +1,  $\hat{Q}(a^2) = 1$
    - Take action  $a^3$  ( $r \sim \text{Bernoulli}(0.1)$ ), get 0,  $\hat{Q}(a^3) = 0$
  - 2 Will the greedy algorithm ever find the best arm in this case? *NO*

# Greedy Algorithm

- We consider algorithms that estimate  $\hat{Q}_t(a) \approx Q(a) = \mathbb{E}[R(a)]$
- Estimate the value of each action by Monte-Carlo evaluation

$$\hat{Q}_t(a) = \frac{1}{N_t(a)} \sum_{t=1}^T r_t \mathbb{1}(a_t = a)$$

- The **greedy** algorithm selects the action with highest value

$$a_t^* = \arg \max_{a \in \mathcal{A}} \hat{Q}_{t-1}(a)$$

- **Greedy can lock onto suboptimal action, forever**

- Setting: Introduction to multi-armed bandits & Approach: greedy methods
- **Framework: Regret**
- Approach:  $\epsilon$ -greedy methods
- Approach: Optimism under uncertainty
- Framework: Bayesian regret
- Approach: Probability matching / Thompson sampling

# Assessing the Performance of Algorithms

- How do we evaluate the quality of a RL (or bandit) algorithm?
- So far: computational complexity, convergence, convergence to a fixed point, & empirical performance performance
- Today: introduce a formal measure of how well a RL/bandit algorithm will do in any environment, compared to optimal

- **Action-value** is the mean reward for action  $a$

$$Q(a) = \mathbb{E}[r \mid a]$$

- **Optimal value**  $V^*$

$$V^* = Q(a^*) = \max_{a \in \mathcal{A}} Q(a)$$

- **Regret** is the opportunity loss for one step, where the expectation is taken over the decision policy used to select  $a_t$

$$l_t = \mathbb{E}[V^* - Q(a_t)]$$

# Regret

- **Action-value** is the mean reward for action  $a$

$$Q(a) = \mathbb{E}[r \mid a]$$

- **Optimal value**  $V^*$

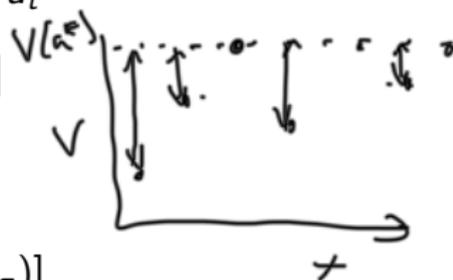
$$V^* = Q(a^*) = \max_{a \in \mathcal{A}} Q(a)$$

- **Regret** is the opportunity loss for one step, where the expectation is taken over the decision policy used to select  $a_t$

$$l_t = \mathbb{E}[V^* - Q(a_t)]$$

- **Total Regret** is the total opportunity loss

$$L_t = \mathbb{E}\left[\sum_{\tau=1}^t V^* - Q(a_\tau)\right]$$



- Maximize cumulative reward  $\iff$  minimize total regret

# Evaluating Regret

- **Count**  $N_t(a)$  is number of times action  $a$  has been selected at time step  $t$
- **Gap**  $\Delta_a$  is the difference in value between action  $a$  and optimal action  $a^*$ ,  $\Delta_a = V^* - Q(a)$
- Regret is a function of gaps and counts

$$\begin{aligned}L_t &= \mathbb{E} \left[ \sum_{\tau=1}^t V^* - Q(a_\tau) \right] \\&= \sum_{a \in \mathcal{A}} \mathbb{E}[N_t(a)] (V^* - Q(a)) \\&= \sum_{a \in \mathcal{A}} \mathbb{E}[N_t(a)] \Delta_a\end{aligned}$$

- A good algorithm ensures small counts for large gaps, but gaps are not known

# Toy Example: Ways to Treat Broken Toes, Optimism, Assessing Regret of Greedy

- True (unknown) Bernoulli reward parameters for each arm (action) are
  - surgery:  $Q(a^1) = \theta_1 = .95$
  - buddy taping:  $Q(a^2) = \theta_2 = .9$
  - doing nothing:  $Q(a^3) = \theta_3 = .1$

$$E[R(a^r)] - E[R(a^i)]$$

- Greedy

Action	Optimal Action	Observed Reward	Regret
$a^1$	$a^1$	0	0
$a^2$	$a^1$	1	0.05
$a^3$	$a^1$	0	0.85
$a^2$	$a^1$	1	0.05
$a^2$	$a^1$	0	0.05

# Toy Example: Ways to Treat Broken Toes, Optimism, Assessing Regret of Greedy

- True (unknown) Bernoulli reward parameters for each arm (action) are
  - surgery:  $Q(a^1) = \theta_1 = .95$
  - buddy taping:  $Q(a^2) = \theta_2 = .9$
  - doing nothing:  $Q(a^3) = \theta_3 = .1$

- Greedy

Action	Optimal Action	Observed Reward	Regret
$a^1$	$a^1$	0	0
$a^2$	$a^1$	1	0.05
$a^3$	$a^1$	0	0.85
$a^2$	$a^1$	1	0.05
$a^2$	$a^1$	0	0.05

- Regret for greedy methods can be linear in the number of decisions made (timestep)

$\Delta t$

# Toy Example: Ways to Treat Broken Toes, Optimism, Assessing Regret of Greedy

- Greedy

Action	Optimal Action	Observed Reward	Regret
$a^1$	$a^1$	0	0
$a^2$	$a^1$	1	0.05
$a^3$	$a^1$	0	0.85
$a^2$	$a^1$	1	0.05
$a^2$	$a^1$	0	0.05

- **Note: in real settings we cannot evaluate the regret because it requires knowledge of the expected reward of the true best action.**
- Instead we can prove an upper bound on the potential regret of an algorithm in **any bandit** problem

- Setting: Introduction to multi-armed bandits & Approach: greedy methods
- Framework: Regret
- **Approach:  $\epsilon$ -greedy methods**
- Approach: Optimism under uncertainty
- Framework: Bayesian regret
- Approach: Probability matching / Thompson sampling

# $\epsilon$ -Greedy Algorithm

*assume unique best action*

- The  $\epsilon$ -**greedy** algorithm proceeds as follows:
  - With probability  $1 - \epsilon$  select  $a_t = \arg \max_{a \in \mathcal{A}} \hat{Q}_t(a)$
  - With probability  $\epsilon$  select a random action
- Always will be making a sub-optimal decision  $\epsilon$  fraction of the time
- Already used this in prior homeworks

$$\epsilon \left( \frac{|\mathcal{A}| - 1}{|\mathcal{A}|} \right)$$

$\uparrow$   
# actions

# Toy Example: Ways to Treat Broken Toes, $\epsilon$ -Greedy

- Imagine true (unknown) Bernoulli reward parameters for each arm (action) are
  - surgery:  $Q(a^1) = \theta_1 = .95$
  - buddy taping:  $Q(a^2) = \theta_2 = .9$
  - doing nothing:  $Q(a^3) = \theta_3 = .1$
- $\epsilon$ -greedy
  - Sample each arm once
    - Take action  $a^1$  ( $r \sim \text{Bernoulli}(0.95)$ ), get +1,  $\hat{Q}(a^1) = 1$
    - Take action  $a^2$  ( $r \sim \text{Bernoulli}(0.90)$ ), get +1,  $\hat{Q}(a^2) = 1$
    - Take action  $a^3$  ( $r \sim \text{Bernoulli}(0.1)$ ), get 0,  $\hat{Q}(a^3) = 0$
  - Let  $\epsilon = 0.1$
  - What is the probability  $\epsilon$ -greedy will pull each arm next? Assume ties are split uniformly.

$$1 - \epsilon = .9 \quad a_1 \quad a_2 \quad .1 \quad a_1 \quad a_2 \quad a_3$$

# Toy Example: Ways to Treat Broken Toes, Optimism, Assessing Regret of Greedy

- True (unknown) Bernoulli reward parameters for each arm (action) are
  - surgery:  $Q(a^1) = \theta_1 = .95$
  - buddy taping:  $Q(a^2) = \theta_2 = .9$
  - doing nothing:  $Q(a^3) = \theta_3 = .1$

- UCB1 (Auer, Cesa-Bianchi, Fischer 2002)

Action	Optimal Action	Regret
$a^1$	$a^1$	0
$a^2$	$a^1$	0.05
$a^3$	$a^1$	0.85
$a^1$	$a^1$	0
$a^2$	$a^1$	0.05

- Will  $\epsilon$ -greedy ever select  $a^3$  again? If  $\epsilon$  is fixed, how many times will each arm be selected?  $\infty$

# Recall: Bandit Regret

- **Count**  $N_t(a)$  is expected number of selections for action  $a$
- **Gap**  $\Delta_a$  is the difference in value between action  $a$  and optimal action  $a^*$ ,  $\Delta_i = V^* - Q(a_i)$
- Regret is a function of gaps and counts

$$\begin{aligned} L_t &= \mathbb{E} \left[ \sum_{\tau=1}^t V^* - Q(a_\tau) \right] \\ &= \sum_{a \in \mathcal{A}} \mathbb{E}[N_t(a)](V^* - Q(a)) \\ &= \sum_{a \in \mathcal{A}} \mathbb{E}[N_t(a)]\Delta_a \end{aligned}$$

- A good algorithm ensures small counts for large gap, but gaps are not known

# L11N3 Check Your Understanding: $\epsilon$ -greedy Bandit Regret

- **Count**  $N_t(a)$  is expected number of selections for action  $a$
- **Gap**  $\Delta_a$  is the difference in value between action  $a$  and optimal action  $a^*$ ,  $\Delta_i = V^* - Q(a_i)$
- Regret is a function of gaps and counts

$$L_t = \sum_{a \in \mathcal{A}} \mathbb{E}[N_t(a)] \Delta_a$$

- Informally an algorithm has linear regret if it takes a non-optimal action a constant fraction of the time
- Assume  $\exists a$  s.t.  $\Delta_a > 0$
- Select all
  - 1  $\epsilon = 0.1$   $\epsilon$ -greedy can have linear regret
  - 2  $\epsilon = 0$   $\epsilon$ -greedy can have linear regret
  - 3 Not sure

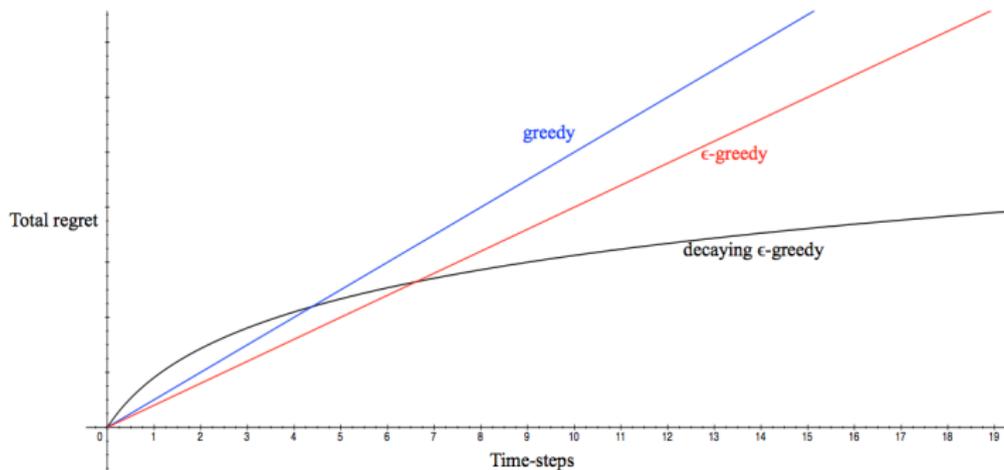
# L11N3 Check Your Understanding: $\epsilon$ -greedy Bandit Regret Answer

- **Count**  $N_t(a)$  is expected number of selections for action  $a$
- **Gap**  $\Delta_a$  is the difference in value between action  $a$  and optimal action  $a^*$ ,  $\Delta_i = V^* - Q(a_i)$
- Regret is a function of gaps and counts

$$L_t = \sum_{a \in \mathcal{A}} \mathbb{E}[N_t(a)] \Delta_a$$

- Informally an algorithm has linear regret if it takes a non-optimal action a constant fraction of the time
- Assume  $\exists a$  s.t.  $\Delta_a > 0$
- Select all
  - 1  $\epsilon = 0.1$   $\epsilon$ -greedy can have linear regret
  - 2  $\epsilon = 0$   $\epsilon$ -greedy can have linear regret
  - 3 Not sure

# "Good": Sublinear or below regret



- **Explore forever:** have linear total regret
- **Explore never:** have linear total regret
- Is it possible to achieve sublinear (in the time steps/number of decisions made) regret?

# Types of Regret bounds

- **Problem independent:** Bound how regret grows as a function of  $T$ , the total number of time steps the algorithm operates for
- **Problem dependent:** Bound regret as a function of the number of times we pull each arm and the gap between the reward for the pulled arm and  $a^*$

6

# Lower Bound

- Use lower bound to determine how hard this problem is
- The performance of any algorithm is determined by similarity between optimal arm and other arms
- Hard problems have similar looking arms with different means
- This is described formally by the gap  $\Delta_a$  and the similarity in distributions  $D_{KL}(\mathcal{R}^a \parallel \mathcal{R}^{a^*})$
- Theorem (Lai and Robbins): Asymptotic total regret is at least logarithmic in number of steps

$$\lim_{t \rightarrow \infty} L_t \geq \log t \sum_{a | \Delta_a > 0} \frac{\Delta_a}{D_{KL}(\mathcal{R}^a \parallel \mathcal{R}^{a^*})}$$

- Promising in that lower bound is sublinear

- Setting: Introduction to multi-armed bandits & Approach: greedy methods
- Framework: Regret
- Approach:  $\epsilon$ -greedy methods
- **Approach: Optimism under uncertainty**
- Framework: Bayesian regret
- Approach: Probability matching / Thompson sampling

# Approach: Optimism in the Face of Uncertainty

- Choose actions that that might have a high value
- Why?
- Two outcomes:

get a high value  
learn something

# Approach: Optimism in the Face of Uncertainty

- Choose actions that that might have a high value
- Why?
- Two outcomes:
  - Getting high reward: if the arm really has a high mean reward
  - Learn something: if the arm really has a lower mean reward, pulling it will (in expectation) reduce its average reward and the uncertainty over its value

# Upper Confidence Bounds

- Estimate an upper confidence  $U_t(a)$  for each action value, such that  $Q(a) \leq U_t(a)$  with high probability *↙ at time step  $t$  how many times did we*
- This depends on the number of times  $N_t(a)$  action  $a$  has been selected *for*
- Select action maximizing Upper Confidence Bound (UCB) *action*  
*or*

$$a_t = \arg \max_{a \in \mathcal{A}} [U_t(a)]$$

# Confidence Bounds for SubGaussian Variables

Corollary 5.5 [Lattimore and Szepesvari, Bandit Algorithms]. Assume that  $X_i - \mu$  are independent,  $\sigma$ -sub-Gaussian random variables. Let  $\hat{\mu} = \frac{1}{n} \sum_{t=1}^n X_t$ . Then for any  $\varepsilon \geq 0$ ,

$$\mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right), \quad \mathbb{P}(\hat{\mu} \leq \mu - \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right),$$

# Confidence Bounds for SubGaussian Variables to UCB

Corollary 5.5 [Lattimore and Szepesvari, Bandit Algorithms]. Assume that  $X_i - \mu$  are independent,  $\sigma$ -sub-Gaussian random variables. Let  $\hat{\mu} = \frac{1}{n} \sum_{t=1}^n X_t$ . Then for any  $\varepsilon \geq 0$ ,

$$\mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right), \quad \mathbb{P}(\hat{\mu} \leq \mu - \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right),$$

# Confidence Bounds for SubGaussian Variables to UCB

Corollary 5.5 [Lattimore and Szepesvari, Bandit Algorithms]. Assume that  $X_i - \mu$  are independent,  $\sigma$ -sub-Gaussian random variables. Let  $\hat{\mu} = \frac{1}{n} \sum_{t=1}^n X_t$ . Then for any  $\varepsilon \geq 0$ ,

$$\mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right), \quad \mathbb{P}(\hat{\mu} \leq \mu - \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right),$$

Therefore: For any  $\delta \in (0, 1]$ , with probability at least  $1 - \delta$ ,

$$\mu \leq \hat{\mu} + \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}.$$

# Confidence Bounds for SubGaussian Variables to UCB

Corollary 5.5 [Lattimore and Szepesvari, Bandit Algorithms]. Assume that  $X_i - \mu$  are independent,  $\sigma$ -sub-Gaussian random variables. Let  $\hat{\mu} = \frac{1}{n} \sum_{t=1}^n X_t$ . Then for any  $\varepsilon \geq 0$ ,

$$\mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right), \quad \mathbb{P}(\hat{\mu} \leq \mu - \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right),$$

Therefore: For any  $\delta \in (0, 1]$ , with probability at least  $1 - \delta$ ,

$$\mu \leq \hat{\mu} + \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}.$$

**(Post class):** Note, we often may want to get an upper and lower bound so that we can bound  $\hat{\mu} - \mu$ . We can do this with a union bound of defining  $\delta' = \delta/2$  such that with probability at least  $1 - \delta$

$$|\mu - \hat{\mu}| \leq \sqrt{\frac{2\sigma^2 \log(2/\delta)}{n}}.$$

- This leads to the UCB1 algorithm (assume rewards are 1-sub-Gaussian ( $\sigma^2 = 1$ ))

$$a_t = \arg \max_{a \in \mathcal{A}} \left[ \hat{Q}(a) + \sqrt{\frac{2 \log \frac{1}{\delta}}{N_t(a)}} \right]$$

# Toy Fake Example: Ways to Treat Broken Toes, Thompson Sampling

- True (unknown) parameters for each arm (action) are
  - surgery:  $Q(a^1) = \theta_1 = .95$
  - buddy taping:  $Q(a^2) = \theta_2 = .9$
  - doing nothing:  $Q(a^3) = \theta_3 = .1$
- Optimism under uncertainty, UCB1 (Auer, Cesa-Bianchi, Fischer 2002)
  - 1 Sample each arm once

# Toy Fake Example: Ways to Treat Broken Toes, Optimism

- True (unknown) parameters for each arm (action) are
  - surgery:  $Q(a^1) = \theta_1 = .95$
  - buddy taping:  $Q(a^2) = \theta_2 = .9$
  - doing nothing:  $Q(a^3) = \theta_3 = .1$
- UCB1 (Auer, Cesa-Bianchi, Fischer 2002)
  - 1 Sample each arm once
    - Take action  $a^1$  ( $r \sim \text{Bernoulli}(0.95)$ ), get +1,  $\hat{Q}(a^1) = 1$
    - Take action  $a^2$  ( $r \sim \text{Bernoulli}(0.90)$ ), get +1,  $\hat{Q}(a^2) = 1$
    - Take action  $a^3$  ( $r \sim \text{Bernoulli}(0.1)$ ), get 0,  $\hat{Q}(a^3) = 0$

# Toy Fake Example: Ways to Treat Broken Toes, Optimism

- True (unknown) parameters for each arm (action) are
  - surgery:  $Q(a^1) = \theta_1 = .95$
  - buddy taping:  $Q(a^2) = \theta_2 = .9$
  - doing nothing:  $Q(a^3) = \theta_3 = .1$
- UCB1 (Auer, Cesa-Bianchi, Fischer 2002)
  - 1 Sample each arm once
    - Take action  $a^1$  ( $r \sim \text{Bernoulli}(0.95)$ ), get +1,  $\hat{Q}(a^1) = 1$
    - Take action  $a^2$  ( $r \sim \text{Bernoulli}(0.90)$ ), get +1,  $\hat{Q}(a^2) = 1$
    - Take action  $a^3$  ( $r \sim \text{Bernoulli}(0.1)$ ), get 0,  $\hat{Q}(a^3) = 0$
  - 2 Set  $t = 3$ , Compute upper confidence bound on each action

$$UCB(a) = \hat{Q}(a) + \sqrt{\frac{2 \log \frac{1}{\delta}}{N_t(a)}}$$

# Toy Fake Example: Ways to Treat Broken Toes, Optimism

- True (unknown) parameters for each arm (action) are
  - surgery:  $Q(a^1) = \theta_1 = .95$
  - buddy taping:  $Q(a^2) = \theta_2 = .9$
  - doing nothing:  $Q(a^3) = \theta_3 = .1$
- UCB1 (Auer, Cesa-Bianchi, Fischer 2002)
  - 1 Sample each arm once
    - Take action  $a^1$  ( $r \sim \text{Bernoulli}(0.95)$ ), get +1,  $\hat{Q}(a^1) = 1$
    - Take action  $a^2$  ( $r \sim \text{Bernoulli}(0.90)$ ), get +1,  $\hat{Q}(a^2) = 1$
    - Take action  $a^3$  ( $r \sim \text{Bernoulli}(0.1)$ ), get 0,  $\hat{Q}(a^3) = 0$
  - 2 Set  $t = 3$ , Compute upper confidence bound on each action

$$UCB(a) = \hat{Q}(a) + \sqrt{\frac{2 \log \frac{1}{\delta}}{N_t(a)}}$$

- 3  $t = 3$ , Select action  $a_t = \arg \max_a UCB(a)$ ,
- 4 Observe reward 1
- 5 Compute upper confidence bound on each action

# Toy Fake Example: Ways to Treat Broken Toes, Optimism

- True (unknown) parameters for each arm (action) are
  - surgery:  $Q(a^1) = \theta_1 = .95$
  - buddy taping:  $Q(a^2) = \theta_2 = .9$
  - doing nothing:  $Q(a^3) = \theta_3 = .1$
- UCB1 (Auer, Cesa-Bianchi, Fischer 2002)
  - 1 Sample each arm once
    - Take action  $a^1$  ( $r \sim \text{Bernoulli}(0.95)$ ), get +1,  $\hat{Q}(a^1) = 1$
    - Take action  $a^2$  ( $r \sim \text{Bernoulli}(0.90)$ ), get +1,  $\hat{Q}(a^2) = 1$
    - Take action  $a^3$  ( $r \sim \text{Bernoulli}(0.1)$ ), get 0,  $\hat{Q}(a^3) = 0$
  - 2 Set  $t = 3$ , Compute upper confidence bound on each action

$$UCB(a) = \hat{Q}(a) + \sqrt{\frac{2 \log \frac{1}{\delta}}{N_t(a)}}$$

- 3  $t = t + 1$ , Select action  $a_t = \arg \max_a UCB(a)$ ,
- 4 Observe reward 1
- 5 Compute upper confidence bound on each action

# Toy Example: Ways to Treat Broken Toes, Optimism, Assessing Regret

- True (unknown) parameters for each arm (action) are
  - surgery:  $Q(a^1) = \theta_1 = .95$
  - buddy taping:  $Q(a^2) = \theta_2 = .9$
  - doing nothing:  $Q(a^3) = \theta_3 = .1$
- UCB1 (Auer, Cesa-Bianchi, Fischer 2002)

Action	Optimal Action	Regret
$a^1$	$a^1$	
$a^2$	$a^1$	
$a^3$	$a^1$	
$a^1$	$a^1$	
$a^2$	$a^1$	

# Confidence Level $\delta$

- Subtle
- Union bound:  $P(\cup E_i) \leq \sum_i P(E_i)$

# Regret Bound for UCB (Sketch)

Let  $\Delta_i = \mu^* - \mu_i$ , where  $\mu_i = \mathbb{E}[a_i]$  and  $\mu^*$  is the mean reward of the optimal arm. Assume arm  $a_1$  is optimal. Let  $N_i(n)$  denote the number of times arm  $i$  is pulled up to time  $n$ .

$$R_n = \sum_{i:\Delta_i>0} \Delta_i \mathbb{E}[N_i(n)].$$

Note: updated post class  
One shown in class had errors

Define the "good event" where the true mean of the optimal arm is always below its UCB and after  $U_i$  samples of action  $i$ , its UCB is below  $\mu_1$ :

$$G_i := \left\{ \mu_1 \leq \min_t \text{UCB}_1(t, \delta) \right\}, \& \left\{ \hat{\mu}_{i, U_i} + \sqrt{2 \log(2/\delta)/U_i} < \mu_1 \right\},$$

**Claim.** If  $G_i$  holds, then arm  $i \neq a^*$  is played at most  $U_i$  times.

**Proof (by contradiction).** Assume  $G_i$  holds. Then there exists a time when arm  $i$ 's UCB must have been highest and it was pulled again for some  $N + t(i) > U_i$ .

$$\text{UCB}_i(t, \delta) \leq \mu_i + \sqrt{2 \log(1/\delta)/U_i} \leq \mu_1 \leq \text{UCB}_1(t, \delta)$$

where the first inequality holds because  $N + t(i) > U_i$  and the second and third hold due to the definition of the good event holding. But this implies arm  $i$  cannot be selected at time  $t$ , a contradiction. Hence  $N_t(i) \leq U_i$  when  $G_i$  holds.

Therefore,

$$\mathbb{E}[N_i(n)] \leq U_i + \Pr(G_i^c) n.$$

If we choose

$$U_i := 2 \log(1/\delta) / \Delta_i^2.$$

and set  $\delta = 1/n^2$  and applying a union bound ( $\Pr(G_i^c) \leq \sum_n \delta = n\delta = \frac{1}{n}$ ) yields

$$R_n \lesssim \sum_{i:\Delta_i>0} \frac{4 \log n}{\Delta_i} + \sum_i \Delta_i.$$

The full proof requires more details on how to derive  $U_i$ : see Theorem 7.1 attimore and Szepesvari, Bandit Algorithms for a formal proof.

# Optional Check Your Understanding

- An alternative would be to always select the arm with the highest lower bound
- Does this ensure low regret?
- Consider a two arm case for simplicity

- Setting: Introduction to multi-armed bandits & Approach: greedy methods
- Framework: Regret
- Approach:  $\epsilon$ -greedy methods
- Approach: Optimism under uncertainty
- Note: bandits are a simpler place to see these ideas, but these ideas will extend to MDPs
- Next time: more fast learning