- CME 241 Notes: Multi-armed bandit

... Digital notes start at 30 minute mark.

- Decaying $\epsilon_t$ - Greedy Algorithm

  ↳ Schedule $\epsilon$ to gradually decrease over time

  ↳ Consider schedule:

  $$c > 0$$

  $$d = \min_{a | \Delta_a > 0} \Delta_a$$

  $$\epsilon_t = \min\left(1, \frac{c|A|}{d^2 t}\right)$$

  ↳ Logarithmic Total regret (proof not shown)

- Lower Bound:

  ↳ Goal: Find an algorithm with sublinear total regret for any multi-armed bandit (without any prior knowledge of R)

  ↳ Lai-Robbins Theorem: You can never build an algorithm with less than logarithmic total regret

- The more uncertain we are about an action, the more that action would like to be explored.

  ↳ Balance mean of Q (exploit) and variance of Q (explore)

- Upper confidence $\hat{U}_t(a)$ for each action value

  ↳ Upper confidence bound algorithm:

  $$a_t = \underset{a \in A}{\operatorname{argmax}} \left\{ \hat{Q}_t(a) + \hat{U}_t(a) \right\}$$

  Depends on # of time steps and number of times you've tried a. See slide 19

↳ UCB algorithm achieves logarithmic total regret

- **Bayesian Bandits**

$$P[R \mid h_t]$$

$h_t$: history

$R$: Probability distribution of arms

This is a probability distribution over distributions.

- We are really interested in the mean of distributions.
- Bayesian bandit algorithms work well at fitting known functional forms of distributions.

- **E.g.** See slide 21 for fitting independent Gaussians

- **Probability Matching** selects action $a$ according to probability that $a$ is the optimal action.

↳ From slide 22:

$$D_t \sim P[R \mid h_t]$$

$D_t$: Some distribution at time $t$

Sampled from distribution of distributions

↳ Probability of choosing action is computed as the probability that action is optimal

↳ **Thompson Sampling** implements probability matching

... See slide 23

1. Sample
2. Calculate Posterior
3. Find argmax

... Skipping Gradient Bandit Algorithm

↳ Skipped in lecture

- Value of Information

↳ Information gain is higher for uncertain situations

↳ It makes sense to explore uncertain arms more

↳ Value of Information lets us trade off exploration and exploitation optimally.

↳ Information state $\tilde{s}$ is a function of history.

$$\tilde{s}_t = f(h_t)$$

↳ See slide 29 for example on Bernoulli bandit

↳ Essentially a vector of counters.

↳ Counters of each outcome from each arm

↳ Bayesian Model-based RL helps you find the balance between explore and exploit.

- In advertising, Bandits can be expressed as "contextual Bandits"

↳ A contextual Bandit is a 3-tuple $(A, S, R)$