Lecture 5: Multi-Armed Bandits (MAB)

Guest Lecturer

Thodoris Lykouris
(Microsoft Research NYC)
Learning objective: Intro to exploration

Previously on CS 6789

• Planning via Bellman equations: known underlying MDP known
• Generative model: ability to reset from anywhere
Learning objective: Intro to exploration

Previously on CS 6789

• Planning via Bellman equations: known underlying MDP known
• Generative model: ability to reset from anywhere

Today: Exploration

• Maximize expected reward w/o known underlying MDP or ability to reset!
Learning objective: Intro to exploration

Previously on CS 6789
• *Planning via Bellman equations:* known underlying MDP known
• *Generative model:* ability to reset from anywhere

Today: Exploration
• Maximize expected reward w/o known underlying MDP or ability to reset!

Focus: Multi-Armed Bandits
• Simplest setting capturing *explore-exploit* trade-off
• Key ideas extend to richer RL settings
Multi-Armed-Bandits: High-level picture

Setting

• Set of alternatives (arms)
• Each arm has a reward distribution

• Learner adaptively selects arms
• Challenge: Distributions not known

Images from:
https://towardsdatascience.com/beyond-a-b-testing-multi-armed-bandit-experiments-1493f709f804
Multi-Armed-Bandits: High-level picture

Setting

• Set of alternatives (arms)
• Each arm has a reward distribution
• Learner adaptively selects arms
• Challenge: Distributions not known

Application: Online advertising

• Arms are advertisers
• Each arm has click-through-rate (CTR) probability of getting clicked
• Platform adaptively selects ads
• Challenge: CTRs are not known

Images from:
https://towardsdatascience.com/beyond-a-b-testing-multi-armed-bandit-experiments-1493f7098104
MAB Protocol

Arm $a \in [k]$ has distribution $F(a)$ with mean $\mu(a)$ and support $[0,1]$

At round $t = 1 \ldots T$:

1. Learner selects arm $a^t$ (possibly in randomized manner)

2. Reward for arm $a$: $r^t(a) \sim F(a)$

3. Learner earns (and only observes) reward $r^t(a^t)$
Probabilistic Approximate Correct (PAC)

**Benchmark:** Best arm had we known the distributions: $a^* = \max_a \mu(a)$

*Fix $\epsilon, \delta > 0$*

How many samples to identify an $\epsilon$-optimal arm $a$ w.p. $1 - \delta$?

$\mu(a^*) - \mu(a) < \epsilon$
Regret Objective

**Explore-exploit version:**
Average cumulative mean:

\[ \text{ALG} = \frac{1}{T} \sum_t \mu(a^t) \]

**Benchmark (no exploration):**
Mean of best arm:

\[ \text{OPT} = \mu(a^*) \]

\[ \text{Regret} = \text{OPT} - \text{ALG} \]
Greedy algorithm

Pick each arm once; then highest empirical mean
Greedy algorithm

Pick each arm once; then highest empirical mean

Action 1:
Reward is always 0.4

Action 2:
Reward is Bernoulli
1 w.p. 60% and 0 else
Greedy algorithm

Pick each arm once; then highest empirical mean

Action 1:
Reward is always 0.4

Action 2:
Reward is Bernoulli
1 w.p. 60% and 0 else

$\epsilon < 0.4, \delta < 0.2$: Greedy does not achieve PAC
Greedy algorithm

Pick each arm once; then highest empirical mean

Action 1:
Reward is always 0.4

Action 2:
Reward is Bernoulli
1 w.p. 60% and 0 else

\[ \epsilon < 0.4, \delta < 0.2: \]
Greedy does not achieve PAC

\[ \text{Regret} = 0.4 \cdot 0.2 \cdot T = 0.08 \cdot T \]
Regret linear in time-horizon
Explore-Then-Commit (ETC)

Pick each arm \( N(\epsilon) = \frac{\log(\frac{kT}{\delta})}{\epsilon^2} \) times; then highest empirical mean \( \tilde{\mu}(\alpha) \)
Explore-Then-Commit (ETC)

Pick each arm $N(\epsilon) = \frac{\log (kT/\delta)}{\epsilon^2}$ times; then highest empirical mean $\tilde{\mu}(a)$

Actual reward means
Explore-Then-Commit (ETC)

Pick each arm $N(\epsilon) = \frac{4\log(k/\delta)}{\epsilon^2}$ times; then highest empirical mean $\tilde{\mu}(a)$

Hoeffding inequality:
$X_1, X_2, ..., X_n$ r.v. in $[0,1]$ with mean $\mu$

$$ Pr \left[ \left| \frac{1}{n} \sum_{i} X_i - \mu \right| \geq \rho \right] \leq 2 \cdot \exp(-2n\rho^2) $$
Explore-Then-Commit (ETC)

Pick each arm \( N(\varepsilon) = \frac{4\log \left( \frac{k}{\delta} \right)}{\varepsilon^2} \) times; then highest empirical mean \( \tilde{\mu}(a) \)

By Hoeffding, \( \forall a \in [k] \) after \( N(\varepsilon) \) plays of \( a \), with probability \( \geq 1 - \delta/k \), it holds:

\[
|\tilde{\mu}(a) - \mu(a)| \leq \varepsilon/2
\]
Explore-Then-Commit (ETC)

Pick each arm $N(\epsilon) = \frac{4\log (k/\delta)}{\epsilon^2}$ times; then highest empirical mean $\tilde{\mu}(a)$

By Hoeffding, $\forall a \in [k]$ after $N(\epsilon)$ plays of $a$, with probability $\geq 1 - \delta/k$, it holds:

$$|\tilde{\mu}(a) - \mu(a)| \leq \frac{\epsilon}{2}$$

By union bound, after $N(\epsilon)$ plays of every arm, with probability $\geq 1 - \delta$, it holds $\forall a \in [k]$: $|\tilde{\mu}(a) - \mu(a)| \leq \frac{\epsilon}{2}$
Explore-Then-Commit (ETC)

Pick each arm $N(\epsilon) = \frac{4\log(k/\delta)}{\epsilon^2}$ times; then highest empirical mean $\tilde{\mu}(a)$

PAC bound: $k \cdot N(\epsilon) = k \cdot \frac{4\log(k/\delta)}{\epsilon^2}$

$UCB = \tilde{\mu}(a_2) + \epsilon/2$

$LCB = \tilde{\mu}(a_2) - \epsilon/2$

Actual reward means
Explore-Then-Commit (ETC)

Pick each arm $N(\varepsilon) = \frac{4\log(k/\delta)}{\varepsilon^2}$ times; then highest empirical mean $\hat{\mu}(a)$

PAC bound: $k \cdot N(\varepsilon) = k \cdot \frac{4\log(k/\delta)}{\varepsilon^2}$

Proof: For selected arm $a: \hat{\mu}(a) \geq \tilde{\mu}(a^*), \mu(a) \leq (\hat{\mu}(a^*) + \frac{\varepsilon}{2}) - (\hat{\mu}(a) - \frac{\varepsilon}{2}) \leq UCB(a^*) - LCB(a)$
Explore-Then-Commit (ETC)

Pick each arm $N(\epsilon) = \frac{4\log \left( \frac{k}{\delta} \right)}{\epsilon^2}$ times; then highest empirical mean $\tilde{\mu}(a)$

PAC bound: $k \cdot N(\epsilon) = k \cdot \frac{4\log \left( \frac{k}{\delta} \right)}{\epsilon^2}$

Regret bound: $k \cdot \frac{4\log \left( \frac{k}{\delta} \right)}{\epsilon^2} + \epsilon \cdot T + \delta \cdot T$

Actual reward means

$\mu(a^*) - \mu(a) \geq \frac{1}{2} \epsilon$
Explore-Then-Commit (ETC)

Pick each arm $N(\epsilon) = \frac{4\log(k/\delta)}{\epsilon^2}$ times; then highest empirical mean $\tilde{\mu}(a)$

PAC bound: $k \cdot N(\epsilon) = k \cdot \frac{4\log(k/\delta)}{\epsilon^2}$

Regret bound: $\mathbb{E}[\text{Regret}] = k \cdot \frac{4\log(k/\delta)}{\epsilon^2} + \epsilon \cdot T + \frac{\delta \cdot T}{3}$

Setting $\epsilon = (k \cdot 4 \log(kT/\delta))^{1/3} \cdot T^{-1/3}$ and $\delta = 1/T$

Regret $\leq O \left( (k \cdot \log(kT))^{1/3} \cdot T^{2/3} \right)$. 

Actual reward means
Active Arm Elimination (AAE)

1. Keep adaptive Upper/Lower Confidence Bounds and active set $A^t$
2. Play round-robin across arms in $A^t$
3. Remove $a$ from $A^t$ if $UCB^t(a) < \max_{a' \in A^{t-1}} LCB^t(a')$
Active Arm Elimination (AAE)

1. Keep adaptive Upper/Lower Confidence Bounds and active set $A^t$
2. Play round-robin across arms in $A^t$
3. Remove $a$ from $A^t$ if $UCB^t(a) < \max_{a' \in A^{t-1}} LCB^t(a')$

\[
UCB^t(a) = \tilde{\mu}^t(a) + \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}}
\]

\[
LCB^t(a) = \tilde{\mu}^t(a) - \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}}
\]
Active Arm Elimination (AAE)

1. Keep adaptive Upper/Lower Confidence Bounds and active set $A^t$
2. Play round-robin across arms in $A^t$
3. Remove $a$ from $A^t$ if $UCB^t(a) < \max_{a' \in A^{t-1}} LCB^t(a')$

$$
UCB^t(a) = \tilde{\mu}^t(a) + \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}}
$$

$$
LCB^t(a) = \tilde{\mu}^t(a) - \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}}
$$

By Hoeffding and union bound, with probability $\geq 1 - \delta$, it holds $\forall a \in [k], t \in [T]$:

$$
\mu(a) \in [LCB^t(a), UCB^t(a)]
$$
Active Arm Elimination (AAE)

1. Keep adaptive Upper/Lower Confidence Bounds and active set $A^t$
2. Play round-robin across arms in $A^t$
3. Remove $a$ from $A^t$ if $UCB^t(a) < \max_{a' \in A^{t-1}} LCB^t(a')$

$UCB^t(a) = \bar{\mu}^t(a) + \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}}$

$LCB^t(a) = \bar{\mu}^t(a) - \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}}$

Claim A: If confidence intervals hold, i.e. $\forall a, t: \mu(a) \in [LCB^t(a), UCB^t(a)]$, the best arm is never eliminated, i.e., $\forall t: a^* \in A^t$
Active Arm Elimination (AAE)

1. Keep adaptive Upper/Lower Confidence Bounds and active set $A^t$
2. Play round-robin across arms in $A^t$
3. Remove $a$ from $A^t$ if $U CB^t(a) < \max_{a' \in A^{t-1}} L CB^t(a')$

$U CB^t(a) = \bar{\mu}^t(a) + \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}}$

$L CB^t(a) = \bar{\mu}^t(a) - \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}}$

**Claim A**: If confidence intervals hold, i.e. $\forall a, t$: $\mu(a) \in [L CB^t(a), U CB^t(a)]$, the best arm is never eliminated, i.e., $\forall t: a^* \in A^t$

**Proof**: 
$\forall a \neq a^*: U CB^t(a^*) \geq \mu(a^*) \geq \mu(a) \geq L CB(a)$
Active Arm Elimination (AAE)

1. Keep adaptive Upper/Lower Confidence Bounds and active set $A^t$
2. Play round-robin across arms in $A^t$
3. Remove $a$ from $A^t$ if $UCB^t(a) < \max_{a' \in A^{t-1}} LCB^t(a')$

$$UCB^t(a) = \bar{\mu}^t(a) + \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}}$$

$$LCB^t(a) = \bar{\mu}^t(a) - \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}}$$

**Claim B:** In that event, arm $a$: $\mu(a^*) - \mu(a) = \epsilon(a)$ is eliminated after $N(a) = \frac{2 \log(2kT/\delta)}{(\epsilon(a))^2}$ plays
Active Arm Elimination (AAE)

1. Keep adaptive Upper/Lower Confidence Bounds and active set $A^t$
2. Play round-robin across arms in $A^t$
3. Remove $a$ from $A^t$ if $UCB^t(a) < \max_{a' \in A^{t-1}} LCB^t(a')$

\[
UCB^t(a) = \bar{\mu}^t(a) + \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}}
\]
\[
LCB^t(a) = \bar{\mu}^t(a) - \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}}
\]

**Claim B:** In that event, arm $a$: $\mu(a^*) - \mu(a) = \epsilon(a)$ is eliminated after $N(a) = \frac{2 \log(2kT/\delta)}{(\epsilon(a))^2}$ plays.

**Proof:** Let $\tau(a)$ be that time. By Claim A: $a^* \in A^{\tau(a)}$.

\[
UCB^{\tau(a)}(a) \leq \mu(a) + \frac{\epsilon(a)}{2}
\]
\[
LCB^{\tau(a)}(a^*) \geq \mu(a^*) - \frac{\epsilon(a)}{2}
\]

\[
UCB^{\tau(a)}(a) - LCB^{\tau(a)}(a^*) \leq \mu(c) - \mu(a^*) + \frac{\epsilon(a)}{2} \leq 0
\]
Active Arm Elimination (AAE)

1. Keep adaptive Upper/Lower Confidence Bounds and active set $A^t$
2. Play round-robin across arms in $A^t$
3. Remove $a$ from $A^t$ if $UCB^t(a) < \max_{a' \in A^{t-1}} LCB^t(a')$

$$UCB^t(a) = \tilde{\mu}^t(a) + \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}}$$
$$LCB^t(a) = \tilde{\mu}^t(a) - \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}}$$

Claim B: In that event, arm $a^*: \mu(a^*) - \mu(a) = \epsilon(a)$ is eliminated after $N(a) = \frac{2 \log(2kT/\delta)}{(\epsilon(a))^2}$ plays.

PAC bound:
$$\sum_{a: \epsilon(a) > \epsilon} N(a) + \sum_{a: \epsilon(a) \leq \epsilon} N(a) < 2$$

Actual reward means
Active Arm Elimination (AAE)

1. Keep adaptive Upper/Lower Confidence Bounds and active set $A^t$
2. Play round-robin across arms in $A^t$
3. Remove $a$ from $A^t$ if $UCB^t(a) < \max_{a' \in A^{t-1}} LCB^t(a')$

$$UCB^t(a) = \widetilde{\mu}^t(a) + \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}}$$

$$LCB^t(a) = \widetilde{\mu}^t(a) - \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}}$$

Claim B: In that event, arm $a$: $\mu(a^*) - \mu(a) = \epsilon(a)$ is eliminated after $N(a) = \frac{2 \log(2kT/\delta)}{(\epsilon(a))^2}$ plays

Actual reward means

PAC bound
$$\sum_{a: \epsilon(a) > \epsilon} N(a) + \cdots$$

Regret bound:
$$\sum_a \min(N(\alpha), T) \cdot \epsilon(\alpha) + \delta \cdot T$$
Active Arm Elimination (AAE)

1. Keep adaptive Upper/Lower Confidence Bounds and active set $A^t$
2. Play round-robin across arms in $A^t$
3. Remove $a$ from $A^t$ if $UCB^t(a) < \max_{a' \in A^{t-1}} LCB^t(a')$

Setting $\delta = 1/T$

**Regret bound:** $\sum_a \min \left( \frac{4 \log (kT)}{\epsilon(a)}, \epsilon(a) \cdot T \right) + 1$

$\mu(a^*) - \mu(a) = \epsilon(a)$
Active Arm Elimination (AAE)

1. Keep adaptive Upper/Lower Confidence Bounds and active set $A^t$
2. Play round-robin across arms in $A^t$
3. Remove $a$ from $A^t$ if $UCB^t(a) < \max_{a' \in A^{t-1}} LCB^t(a')$

Setting $\delta = 1/T$

Regret bound: $\sum_a \min \left( \frac{A(\epsilon)}{\epsilon(a)}, \epsilon(a) \cdot T \right) + 1$

For worst-case choice of $\epsilon(a) = \sqrt{\frac{k \cdot \log(kT)}{T}}$:

Regret = $O\left(\sqrt{k \cdot T \cdot \log(kT)}\right)$

\[ \mu(a^*) - \mu(a) = \epsilon(a) \]
Upper Confidence Bound (UCB)

Pick arm with highest Upper Confidence Bound

\[ UCB^t(a) = \tilde{\mu}^t(a) + \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}} \]

\[ LCB^t(a) = \tilde{\mu}^t(a) - \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}} \]
Upper Confidence Bound (UCB)

Pick arm with highest Upper Confidence Bound

\[ UCB^t(a) = \tilde{\mu}^t(a) + \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}} \]

\[ LCB^t(a) = \tilde{\mu}^t(a) - \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}} \]

Actual reward means

By Hoeffding and union bound, with probability \( \geq 1 - \delta \), it holds \( \forall a \in [k], t \in [T] : \)

\[ \mu(a) \in [LCB^t(a), UCB^t(a)] \]
Upper Confidence Bound (UCB)

Pick arm with highest **Upper Confidence Bound**

By Hoeffding and union bound, with probability $\geq 1 - \delta$, it holds $\forall a \in [k], t \in [T]$:

$$\mu(a) \in [LCB^t(a), UCB^t(a)]$$

**Claim**: In the event that all confidence intervals hold, the regret is at most

$$\sum_t (UCB^t(a_t) - LCB^t(a_t)) + \delta \cdot T$$
**Upper Confidence Bound (UCB)**

Pick arm with highest *Upper Confidence Bound*

By Hoeffding and union bound, with probability $\geq 1 - \delta$, it holds $\forall a \in [k], t \in [T]$: 

$$\mu(a) \in [L_{CB}^t(a), U_{CB}^t(a)]$$

**Claim:** In the event that all confidence intervals confidence intervals hold, the regret is at most 

$$\sum_t (UCB^t(a^t) - LCB^t(a^t)) + \delta \cdot T$$

**Proof:** 

$$Reg_t^t = \mu(a^*) - \mu(a^t)$$

$$\leq UCB^t(a^*) - LCB^t(a^t)$$

$$\leq UCB^t(a^t) - LCB^t(a^t)$$
Upper Confidence Bound (UCB)

\[ UCB^t(a) = \tilde{\mu}^t(a) + \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}} \]

\[ LCB^t(a) = \tilde{\mu}^t(a) - \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}} \]

**Claim:** In the event that all confidence intervals confidence intervals hold, the regret is at most

\[ \sum_t (UCB^t(a^t) - LCB^t(a^t)) + \gamma \delta \cdot T \]
Upper Confidence Bound (UCB)

\[
UCB^t(a) = \tilde{\mu}^t(a) + \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}}
\]

\[
LCB^t(a) = \tilde{\mu}^t(a) - \sqrt{\frac{\log(2kT/\delta)}{2n^t(a)}}
\]

**Claim**: In the event that all confidence intervals hold, the regret is at most

\[
\sum_t (UCB^t(a^t) - LCB^t(a^t)) + \delta \cdot T
\]

**Regret bound by confidence sum**

\[
\sum_t (UCB^t(a^t) - LCB^t(a^t)) \leq 2 \cdot \sum_t \sqrt{\frac{\log \left( \frac{2kT}{\delta} \right)}{2n^t(a^t)}} = \sum_a \sum_{j=1}^{N(a)} \sqrt{\frac{\log \left( \frac{2kT}{\delta} \right)}{2 \cdot j}}
\]

\[
\leq \sum_a \sum_{j=1}^{T} \sqrt{\frac{\log \left( \frac{2kT}{\delta} \right)}{2 \cdot j}} \leq k \cdot \sqrt{\frac{\log \left( \frac{2kT}{\delta} \right)}{k}} \cdot T = O \left( \sqrt{T \cdot k \cdot \log \left( \frac{kT}{\delta} \right)} \right)
\]
Upper Confidence Bound (UCB)

Resulting guarantee similar to the one of AAE

Confidence sum analysis:
1. Extends to RL (see next lecture)
2. Gap-dependent guarantees
   • Small modification in analysis
3. Allows for anytime guarantees (unknown horizon)
   • Small modification in confidence bounds
Stochastic MAB Protocol

Arm $a \in [k]$ has distribution $F(a)$ with mean $\mu(a)$ and support $[0, 1]$

At round $t = 1 \ldots T$:

1. Learner commits to a distribution $p^t$ across arms

2. Reward for arm $a$: $r^t(a) \sim F(a)$

3. Learner draws arm $a^t \sim p^t$

4. Learner earns (and only observes) reward $r^t(a^t)$
Adversarial MAB Protocol

At round $t = 1 \ldots T$:

1. Learner commits to a distribution $p^t$ across arms

2. Reward for arm $a$: $r^t(a) \in [0, 1]$ adversarially selected

3. Learner draws arm $a^t \sim p^t$

4. Learner earns (and only observes) reward $r^t(a^t)$
Stochastic and Adversarial worlds

**Stochastic world**

- If arms have a gap in their means, i.e., $\mu(a^*) - \mu(a) = \epsilon(a)$ then regret of the order of:

$$\sum_a \min \left( \frac{4 \log (kT)}{\epsilon(a)}, \epsilon(a) \cdot T \right)$$

- If not then regret of the order of $\sqrt{kT}$

- If rewards are not stochastic, stochastic MAB algs: linear regret
Stochastic and Adversarial worlds

**Stochastic world**
- If arms have a gap in their means, i.e., \( \mu(a^*) - \mu(a) = \epsilon(a) \) then regret of the order of:
  \[
  \sum_a \min \left( \frac{4 \log (kT)}{\epsilon(a)}, \epsilon(a) \cdot T \right)
  \]
- If not then regret of the order of \( \sqrt{kT} \)
- If rewards are not stochastic, stochastic MAB algs: linear regret

**Adversarial world**
- Regret of the order of \( \sqrt{kT} \) without assuming stochasticity (e.g., EXP3)
- If rewards are stochastic, adversarial MAB algs: no enhanced bounds
Stochastic and Adversarial worlds

**Stochastic world**
- If arms have a gap in their means, i.e., $\mu(a^*) - \mu(a) = \epsilon(a)$ then regret of the order of:
  $$\sum_a \min \left( \frac{4 \log (kT)}{\epsilon(a)}, \epsilon(a) \cdot T \right)$$
- If not then regret of the order of $\sqrt{KT}$
- If rewards are not stochastic, stochastic MAB algs: linear regret

**Adversarial world**
- Regret of the order of $\sqrt{KT}$ without assuming stochasticity (e.g., EXP3)
- If rewards are stochastic, adversarial MAB algs: no enhanced bounds

**Question: Best of both worlds?**
- Single algorithm with logarithmic guarantee when input *stochastic* and square-root when input *adversarial*!

[Bubeck, Slivkins, COLT ’12]
Best of both worlds

**Question: Best of both worlds?**

- Single algorithm with logarithmic guarantee when input stochastic and square-root when input adversarial!  
  [Bubeck, Slivkins, COLT ’12]
Best of both worlds

Question: Best of both worlds?

• Single algorithm with logarithmic guarantee when input stochastic and square-root when input adversarial! [Bubeck, Slivkins, COLT ’12]

  Answer: Yes!

• Approach 1: Start from AAE and test for non-consistency; if identified then switch to EXP3 [Bubeck, Slivkins, COLT ’12] [Auer, Chiang, ICML’ 16]

• Approach 2: Start from adversarial with aggressive “learning rate”; adapt it over time [Seldin, Slivkins, ICML’14] [Seldin, Lugosi, COLT ’17] [Wei, Luo, COLT ’18] [Zimmert, Seldin, AISTATS ’19]
Best of both worlds

Question: Best of both worlds?

• Single algorithm with logarithmic guarantee when input stochastic and square-root when input adversarial! [Bubeck, Slivkins, COLT ’12]

   Answer: Yes!

• Approach 1: Start from AAE and test for non-consistency; if identified then switch to EXP3 [Bubeck, Slivkins, COLT ’12] [Auer, Chiang, ICML’ 16]

• Approach 2: Start from adversarial with aggressive “learning rate”; adapt it over time [Seldin, Slivkins, ICML’14] [Seldin, Lugosi, COLT ’17] [Wei, Luo, COLT ’18] [Zimmert, Seldin, AISTATS ’19]

RL: Only very preliminary results for known transitions [Jin, Luo, working’ 20]
Corrupted MAB

Arm \( a \in [k] \) has distribution \( F(a) \) with mean \( \mu(a) \) and support \([0, 1]\)

At round \( t = 1 \ldots T \):

1. Learner commits to a distribution \( p^t \) across arms

2. Reward for arm \( a \): \( r^t(a) \sim F(a) \)

3. Adversary corrupts rewards \( r^t(a) \) \( (total \ corruption \ budget \ of \ C) \)

4. Learner draws arm \( a^t \sim p^t \)

5. Learner earns uncorrupted (or corrupted) reward & observes only corrupted
Corrupted MAB

**Question**

- Algorithm with logarithmic guarantee when stochastic and gracefully degrades with corruption budget  
  [Lykouris, Mirrokni, Paes Leme, STOC ’18]
Corrupted MAB

**Question**

- Algorithm with logarithmic guarantee when stochastic and gracefully degrades with corruption budget
  
  \[\text{Lykouris, Mirrokni, Paes Leme, STOC ’18}\]

**Answer:**

- Initial algorithm based on a multi-layering version of AAE and lower bound for high-probability
  
  \[\text{Lykouris, Mirrokni, Paes Leme, STOC ’18}\]

- Improved algorithm using a phase scheme and the Improved-UCB algorithm of Otner and Auer’10
  
  \[\text{Gupta, Koren, Talwar, COLT ’19}\]

- For expectations and corrupted: Between both worlds
  
  \[\text{Zimmert, Seldin ’20}\]
Corrupted MAB

**Question**

- Algorithm with logarithmic guarantee when stochastic and gracefully degrades with corruption budget [Lykouris, Mirrokni, Paes Leme, STOC ’18]

**Answer:**

- Initial algorithm based on a multi-layering version of AAE and lower bound for high-probability [Lykouris, Mirrokni, Paes Leme, STOC ’18]
- Improved algorithm using a phase scheme and the Improved-UCB algorithm of Otner and Auer’10 [Gupta, Koren, Talwar, COLT ’19]
- For expectations and corrupted: Between both wolds [Zimmert, Seldin ’20]

**RL:** Multi-layering version of UCBVI enhanced with appropriate active sets [Lykouris, Simchowitz, Slivkins, Sun’ 19]
From MAB to episodic RL

Best of both worlds and corrupted MAB: Examples of MAB informing RL
From MAB to episodic RL

Best of both worlds and corrupted MAB: Examples of MAB informing RL

Other MAB-informing-RL settings

1. MAB with feedback graphs (captures side-information)
   [Dann, Mansour, Mohri, Sekhari, Sridharan '20]
From MAB to episodic RL

Best of both worlds and corrupted MAB: Examples of MAB informing RL

Other MAB-informing-RL settings

1. MAB with feedback graphs (captures side-information)
   [Dann, Mansour, Mohri, Sekhari, Sridharan ’20]

2. MAB with constraints
   [Brantley, Dudik, Lykouris, Miryoosefi Simchowitz, Slivkins, Sun ’20]
From MAB to episodic RL

Best of both worlds and corrupted MAB: Examples of MAB informing RL

Other MAB-informing-RL settings
1. MAB with feedback graphs (captures side-information)
   [Dann, Mansour, Mohri, Sekhari, Sridharan ’20]
2. MAB with constraints
   [Brantley, Dudik, Lykouris, Miryoosefi Simchowitz, Slivkins, Sun ’20]
3. MAB with continuous actions
   [Sinclair, Banerjee, Lee Yu, POMACS ’20]
From MAB to episodic RL

Best of both worlds and corrupted MAB: Examples of MAB informing RL

Other MAB-informing-RL settings

1. MAB with feedback graphs (captures side-information)
   [Dann, Mansour, Mohri, Sekhari, Sridharan ’20]

2. MAB with constraints
   [Brantley, Dudik, Lykouris, Miryoosefi Simchowitz, Slivkins, Sun ’20]

3. MAB with continuous actions
   [Sinclair, Banerjee, Lee Yu, POMACS ’20]

Typically poses interesting complications requiring RL advancement
From MAB to episodic RL

Best of both worlds and corrupted MAB: Examples of MAB informing RL

Other MAB-informing-RL settings

1. MAB with feedback graphs (captures side-information)
   [Dann, Mansour, Mohri, Sekhari, Sridharan ’20]

2. MAB with constraints
   [Brantley, Dudik, Lykouris, Miryoosefi Simchowitz, Slivkins, Sun ’20]

3. MAB with continuous actions
   [Sinclair, Banerjee, Lee Yu, POMACS ’20]

Typically poses interesting complications requiring RL advancement
Summary

Today: Exploration

• Maximize expected reward w/o known underlying MDP or ability to reset!
Today: Exploration

• Maximize expected reward w/o known underlying MDP or ability to reset!

Focus: Multi-Armed Bandits

• Simplest setting capturing *explore-exploit* trade-off

• Key ideas extend to richer RL & tackle complexities not understood in RL
Summary

**Today: Exploration**
- Maximize expected reward w/o known underlying MDP or ability to reset!

**Focus: Multi-Armed Bandits**
- Simplest setting capturing *explore-exploit* trade-off
- Key ideas extend to richer RL & tackle complexities not understood in RL

**Algorithms**
- Greedy: Not PAC / Linear regret
- Explore-Then-Commit: Regret of $T^{2/3}$
- Active Arm Elimination: Regret logarithmic for arms separated and $\sqrt{T}$ else
- Upper Confidence Bound: Same regret; analysis extends to RL