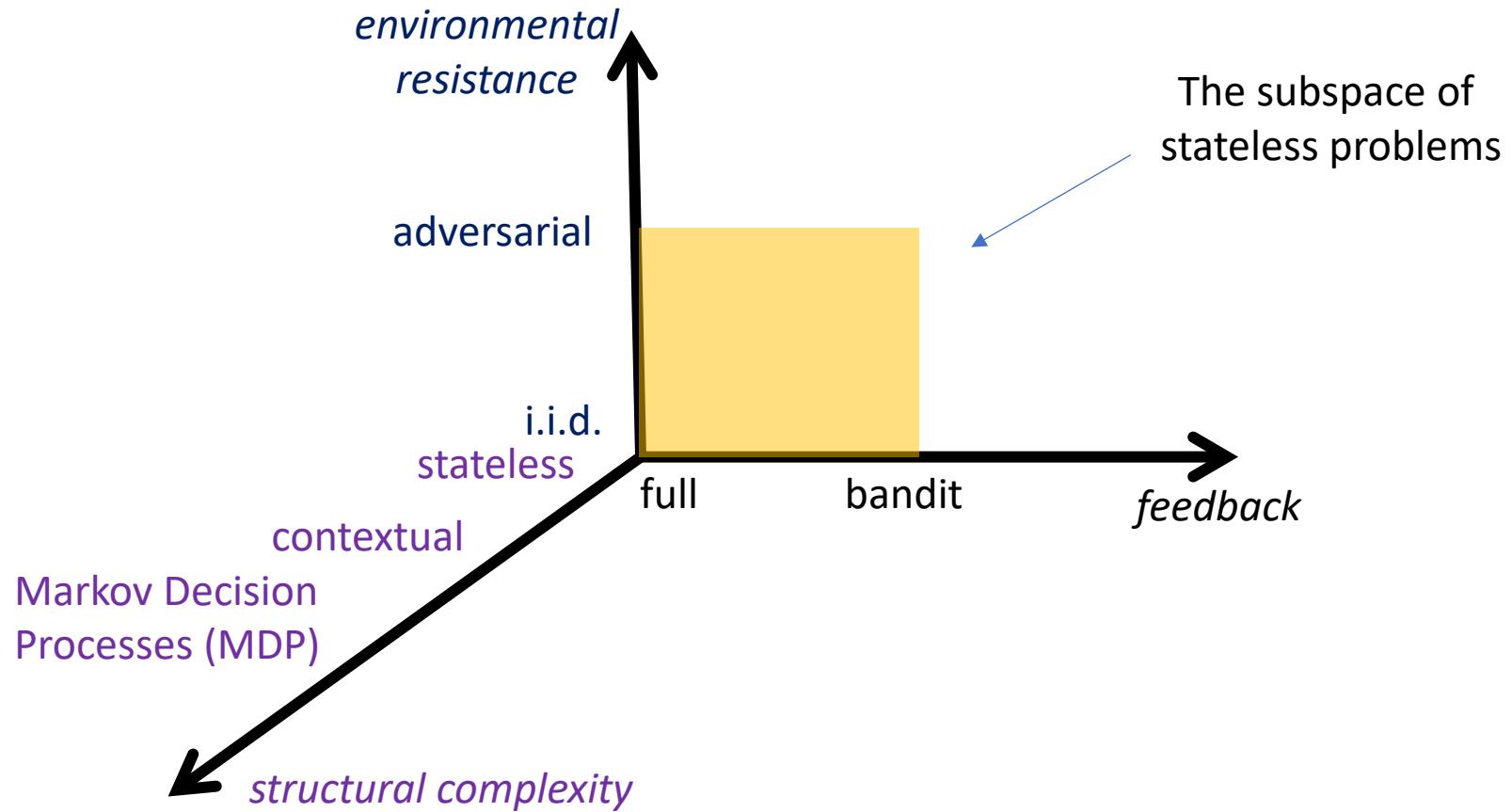


Online Learning Setup and Stochastic Bandits

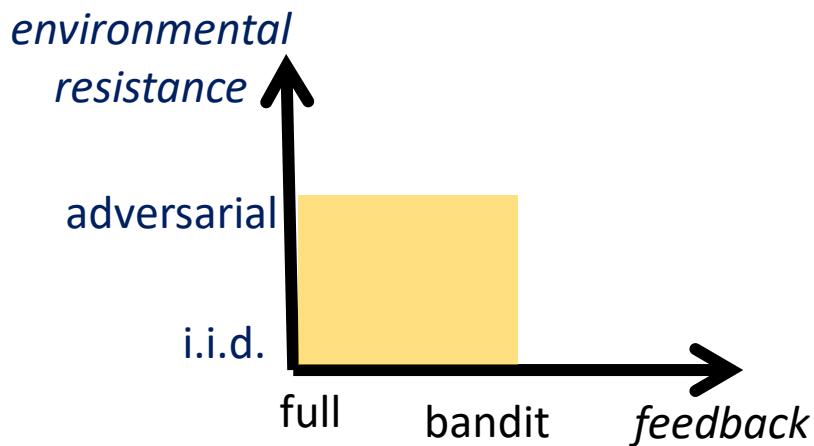
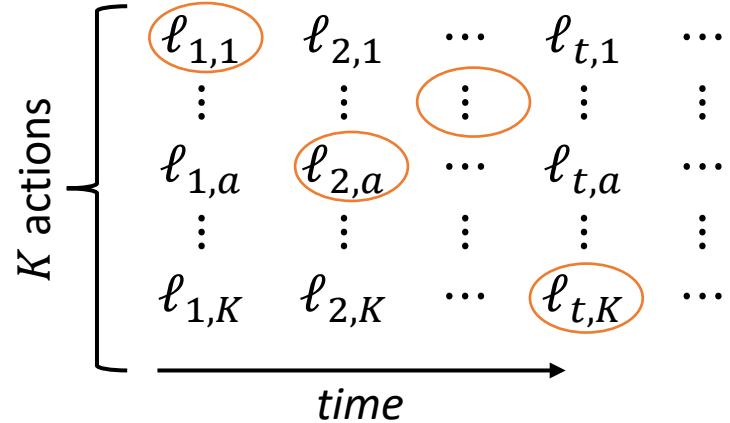
Yevgeny Seldin

Online Learning Setup

The space of online learning problems



The stateless setting



Game protocol:

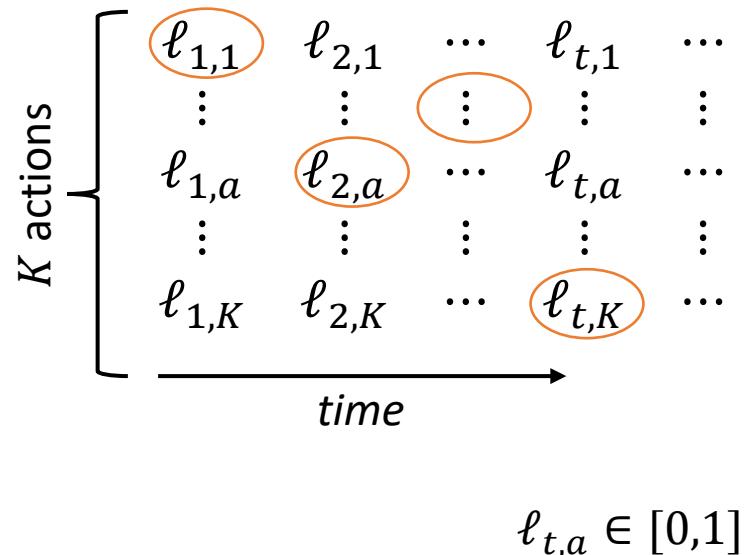
For $t = 1, 2, \dots :$

1. Pick a row A_t
2. Suffer the loss ℓ_{t,A_t}
3. Observe ...

Observations	Full: $\ell_{t,1}$ \vdots $\ell_{t,K}$	Bandit: ℓ_{t,A_t}
Generation of $\ell_{t,a}$		
Adversarial: $\ell_{t,a}$ arbitrary		
I.I.D.: $\ell_{t,a}$ sampled i.i.d., such that $\mathbb{E}[\ell_{t,a}] = \mu(a)$		

Performance measure

- Regret: $R_T = \underbrace{\sum_{t=1}^T \ell_{t,A_t}}_{\text{Loss of the algorithm}} - \underbrace{\min_a \sum_{t=1}^T \ell_{t,a}}_{\text{Loss of the best action in hindsight}}$
 - Regret of order T means no learning
 - The loss of A_t stays at the same distance from the loss of the optimal action as the game proceeds
 - The aim is to achieve sublinear regret
 - Why do we compare to the best fixed action in hindsight and not to the best path in hindsight?
 - “The best path in hindsight” is an overly strong competitor – we cannot guarantee sublinear regret
 - Show that the regret relative to the best path in hindsight can be as large as $\frac{K-1}{K} T$



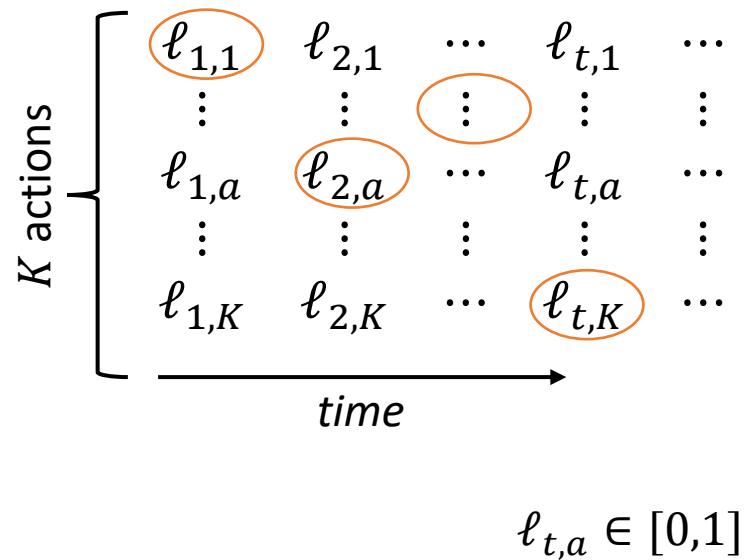
Example: $\begin{matrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{matrix} \dots$

Performance measure

- Regret: $R_T = \underbrace{\sum_{t=1}^T \ell_{t,A_t}}_{\text{Loss of the algorithm}} - \underbrace{\min_a \sum_{t=1}^T \ell_{t,a}}_{\text{Loss of the best action in hindsight}}$
- Regret of order T means no learning
 - The loss of A_t stays at the same distance from the loss of the optimal action as the game proceeds
- The aim is to achieve sublinear regret
- Why do we compare to the best fixed action in hindsight and not to the best path in hindsight?
 - “The best path in hindsight” is an overly strong competitor – we cannot guarantee sublinear regret
 - Show that the regret relative to the best path in hindsight can be as large as $\frac{K-1}{K} T$

Performance measures

- Regret: $R_T = \underbrace{\sum_{t=1}^T \ell_{t,A_t}}_{\text{Loss of the algorithm}} - \underbrace{\min_a \sum_{t=1}^T \ell_{t,a}}_{\text{Loss of the best action in hindsight}}$
- Expected regret: $\mathbb{E}[R_T] = \mathbb{E}\left[\sum_{t=1}^T \ell_{t,A_t}\right] - \mathbb{E}\left[\min_a \sum_{t=1}^T \ell_{t,a}\right]$
 $\qquad\qquad\qquad \stackrel{\text{oblivious}}{\equiv} \sum_{t=1}^T \ell_{t,A_t} - \min_a \sum_{t=1}^T \ell_{t,a}$
- Oblivious adversary:
 - $\ell_{t,a}$ is independent of A_1, \dots, A_{t-1}
 - The losses can be written down before the game starts
- Adaptive adversary:
 - $\ell_{t,a}$ may depend on A_1, \dots, A_{t-1}



Performance measures

- Pseudo-regret (stochastic setting):

$$\bar{R}_T = \mathbb{E} \left[\sum_{t=1}^T \ell_{t,A_t} \right] - \min_a \mathbb{E} \left[\sum_{t=1}^T \ell_{t,a} \right]$$

$$= \mathbb{E} \left[\sum_{t=1}^T \ell_{t,A_t} \right] - T \underbrace{\min_a \mu(a)}_{\mu^*}$$

$$= \mathbb{E} \left[\sum_{t=1}^T (\ell_{t,A_t} - \mu^*) \right]$$

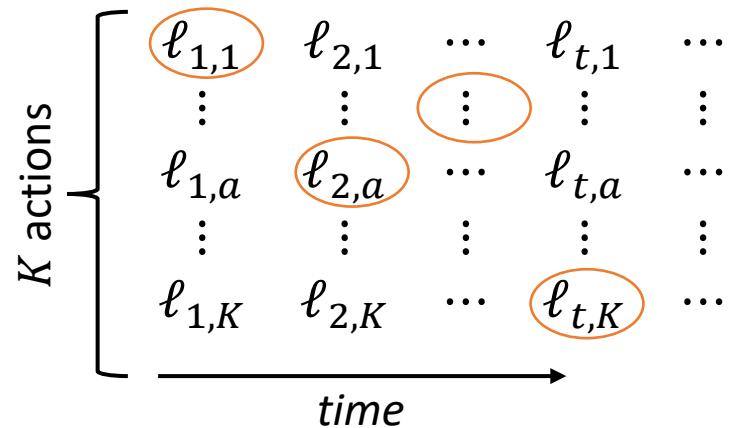
$$= \mathbb{E} \left[\sum_{t=1}^T \Delta(A_t) \right]$$

$$= \mathbb{E} \left[\sum_{a=1}^K \Delta(a) N_T(a) \right]$$

$$= \sum_{a=1}^K \Delta(a) \mathbb{E}[N_T(a)]$$

- $\mathbb{E}[\ell_{t,a}] = \mu(a)$
- $\mu^* = \min_a \mu(a)$
- $a^* \in \arg \min_a \mu(a)$
 - An optimal arm (may be multiple optimal arms with the same μ^*)
- $\Delta(a) = \mu(a) - \mu^*$ suboptimality gap

- $\mathbb{E}[\ell_{t,A_t} - \mu^*] = \mathbb{E} \left[\mathbb{E}[\ell_{t,A_t} - \mu^* | A_1, \dots, A_t] \right] = \mathbb{E}[\mu(A_t) - \mu^*] = \mathbb{E}[\Delta(A_t)]$



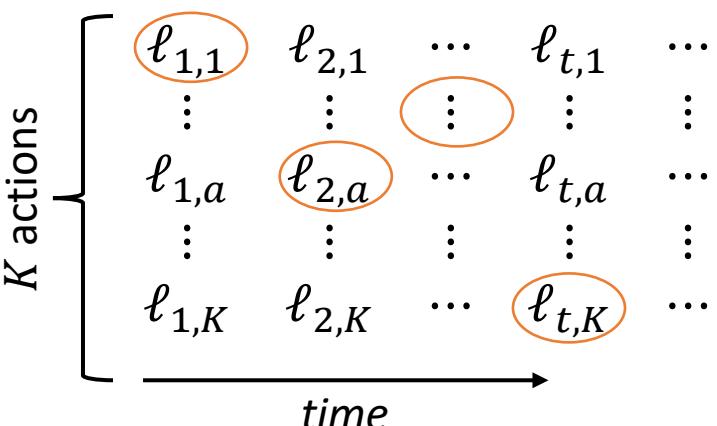
Regret: $R_T = \underbrace{\sum_{t=1}^T \ell_{t,A_t}}_{\text{Loss of the algorithm}} - \underbrace{\min_a \sum_{t=1}^T \ell_{t,a}}_{\text{Loss of the best action in hindsight}}$

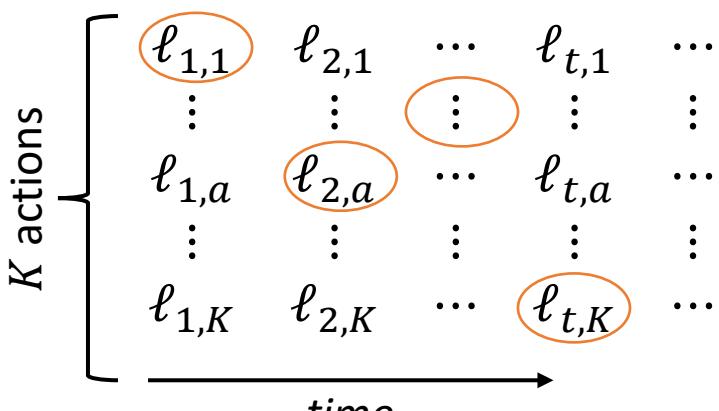
Expected regret:

$$\mathbb{E}[R_T] = \mathbb{E} \left[\sum_{t=1}^T \ell_{t,A_t} \right] - \mathbb{E} \left[\min_a \sum_{t=1}^T \ell_{t,a} \right]$$

Expected regret vs. Pseudo regret

- Expected regret: $\mathbb{E}[R_T] = \mathbb{E}\left[\sum_{t=1}^T \ell_{t,A_t}\right] - \mathbb{E}\left[\min_a \sum_{t=1}^T \ell_{t,a}\right]$
- Pseudo-regret: $\bar{R}_T = \mathbb{E}\left[\sum_{t=1}^T \ell_{t,A_t}\right] - \min_a \mathbb{E}\left[\sum_{t=1}^T \ell_{t,a}\right] = \mathbb{E}\left[\sum_{t=1}^T \ell_{t,A_t}\right] - T\mu^*$
- $\mathbb{E}\left[\min_a f(a, B)\right] \leq \min_a \mathbb{E}[f(a, B)] \Rightarrow \bar{R}_T \leq \mathbb{E}[R_T]$
- Oblivious adversarial setting:
 - $\ell_{t,a}$ are deterministic and the two notions of regret coincide
 - $\mathbb{E}\left[\min_a \sum_{t=1}^T \ell_{t,a}\right] = \min_a \mathbb{E}\left[\sum_{t=1}^T \ell_{t,a}\right] = \min_a \sum_{t=1}^T \ell_{t,a}$

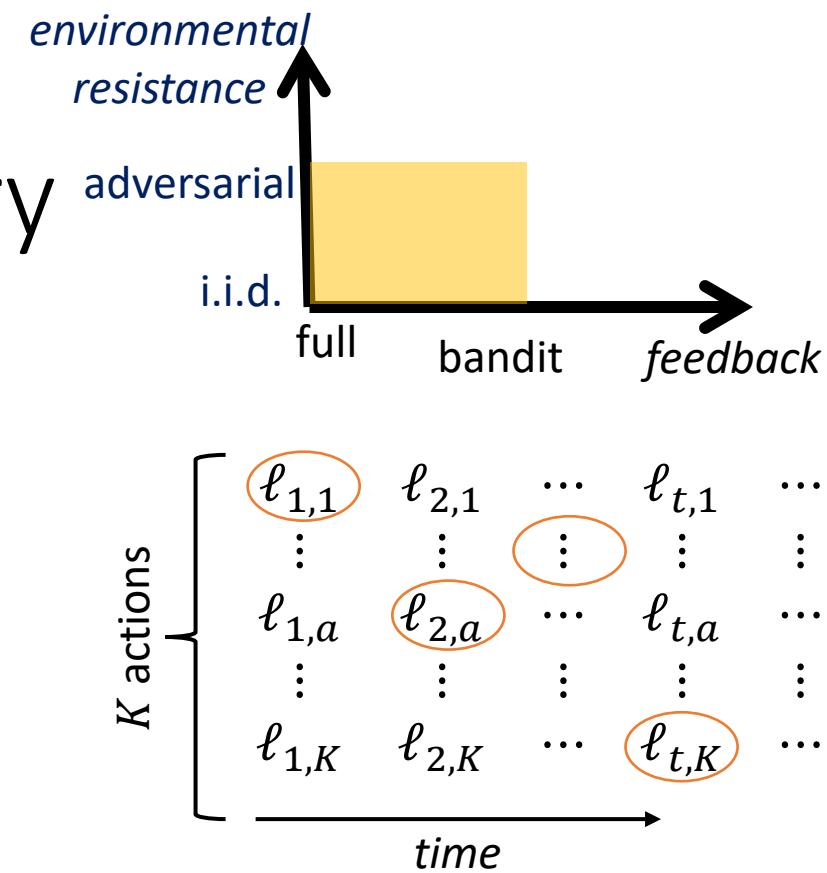
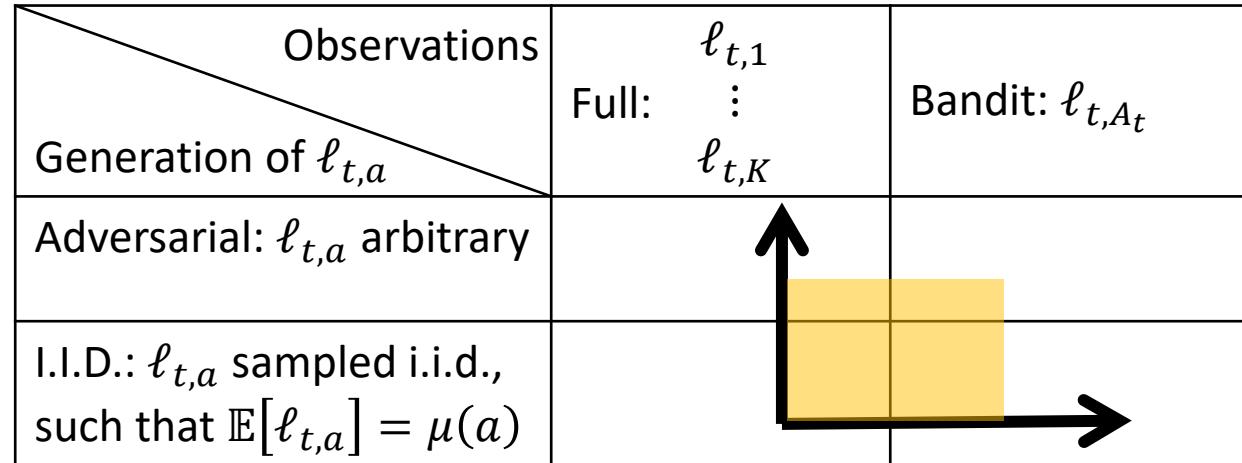




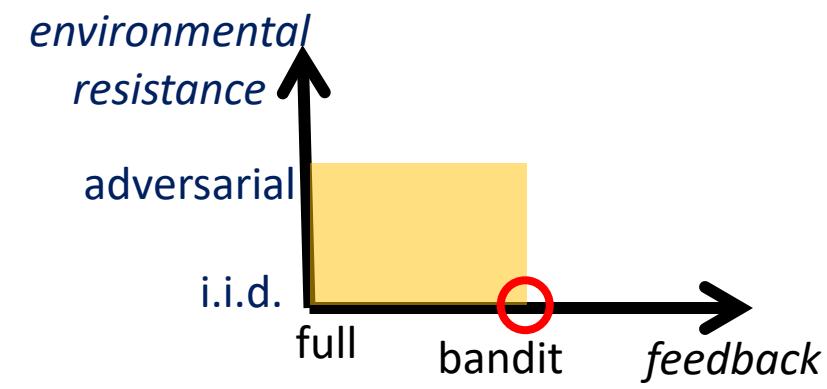
Expected regret vs. Pseudo regret

- Expected regret: $\mathbb{E}[R_T] = \mathbb{E}\left[\sum_{t=1}^T \ell_{t,A_t}\right] - \mathbb{E}\left[\min_a \sum_{t=1}^T \ell_{t,a}\right]$
- Pseudo-regret: $\bar{R}_T = \mathbb{E}\left[\sum_{t=1}^T \ell_{t,A_t}\right] - \min_a \mathbb{E}\left[\sum_{t=1}^T \ell_{t,a}\right] = \mathbb{E}\left[\sum_{t=1}^T \ell_{t,A_t}\right] - T\mu^*$
- $\mathbb{E}\left[\min_a f(a, B)\right] \leq \min_a \mathbb{E}[f(a, B)] \Rightarrow \bar{R}_T \leq \mathbb{E}[R_T]$
- Stochastic setting: imagine that $\mu(a) = \frac{1}{2}$ for all a . Then
 - $\mathbb{E}\left[\sum_{t=1}^T \ell_{t,A_t}\right] = \frac{1}{2}T$
 - $\mathbb{E}\left[\sum_{t=1}^T \ell_{t,a}\right] = \frac{1}{2}T$ for all a
 - $\bar{R}_T = 0$
 - $\mathbb{E}\left[\min_a \sum_{t=1}^T \ell_{t,a}\right] \approx \frac{1}{2}T - \sqrt{\frac{1}{2}T \ln K}$
 - $\mathbb{E}[R_T] \approx \sqrt{\frac{1}{2}T \ln K}$
 - Pseudo-regret is a more reasonable quantity to look at
 - Expected regret provides an artificial advantage to the competitor due to their ability to select out of K trials

Online Learning Setup - Summary



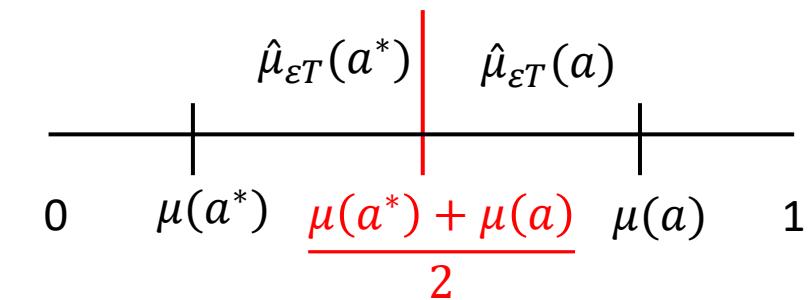
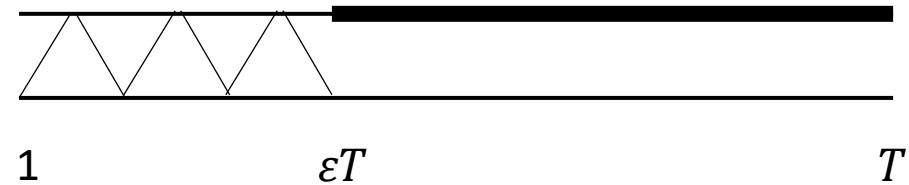
- Regret: $R_T = \sum_{t=1}^T \ell_{t,A_t} - \min_a \sum_{t=1}^T \ell_{t,a}$
 - Expected regret: $\mathbb{E}[R_T] = \mathbb{E}\left[\sum_{t=1}^T \ell_{t,A_t}\right] - \mathbb{E}\left[\min_a \sum_{t=1}^T \ell_{t,a}\right]$
 - Pseudo-regret: $\bar{R}_T = \mathbb{E}\left[\sum_{t=1}^T \ell_{t,A_t}\right] - \min_a \mathbb{E}\left[\sum_{t=1}^T \ell_{t,a}\right] = \mathbb{E}\left[\sum_{t=1}^T \ell_{t,A_t}\right] - T\mu^*$



Stochastic (i.i.d.) bandits

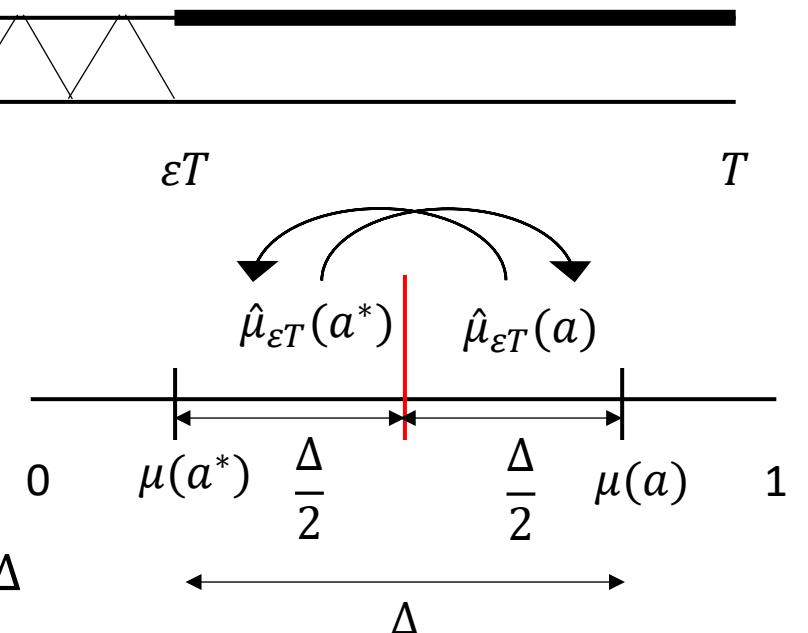
Exploration-Exploitation trade-off: a simple approach

- Setting:
 - Two actions
 - Bandit feedback
 - T is known
 - Δ is known
- Approach:
 - Explore 50/50 for εT rounds
 - Exploit for the remaining rounds
- Analysis approach:
 - Take a separation line at $\frac{\mu(a^*) + \mu(a)}{2}$
 - If at time εT the empirical means are on the “correct” side of the separation line, the arm selection for exploitation will be correct
 - Bound the probability that at εT the empirical means are estimated incorrectly



Analysis

- Let $\delta(\varepsilon) = \mathbb{P}(\hat{\mu}_{\varepsilon T}(a) \leq \hat{\mu}_{\varepsilon T}(a^*))$ be the prob. of confusion
- $$\bar{R}_T = \sum_{t=1}^T \Delta(A_t) \leq \underbrace{\frac{1}{2}\varepsilon T \Delta}_{\text{Exploration}} + \underbrace{\delta(\varepsilon)(1-\varepsilon)T\Delta}_{\text{Exploitation}} \leq \left(\frac{\varepsilon}{2} + \delta(\varepsilon)\right)T\Delta$$



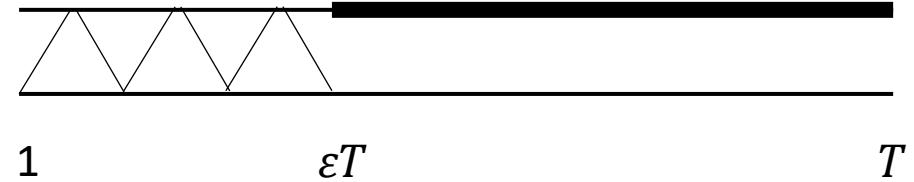
$$\begin{aligned}
 \delta(\varepsilon) &= \mathbb{P}(\hat{\mu}_{\varepsilon T}(a) \leq \hat{\mu}_{\varepsilon T}(a^*)) \\
 &\leq \mathbb{P}\left(\hat{\mu}_{\varepsilon T}(a^*) \geq \mu(a^*) + \frac{1}{2}\Delta\right) + \mathbb{P}\left(\hat{\mu}_{\varepsilon T}(a) \leq \mu(a) - \frac{1}{2}\Delta\right) \\
 &\leq 2e^{-2\frac{\varepsilon T}{2}\left(\frac{1}{2}\Delta\right)^2} = 2e^{-\varepsilon T \Delta^2 / 4}
 \end{aligned}$$

Hoeffding:

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n Z_i - \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n Z_i\right] \geq \alpha\right) \leq e^{-2n\alpha^2}$$

- Minimization of $\frac{\varepsilon}{2} + 2e^{-\varepsilon T \Delta^2 / 4}$ with respect to ε gives $\varepsilon^* = \frac{4 \ln(T\Delta^2)}{T\Delta^2}$
- With exploration phase of length $\varepsilon^* T$, we get $\bar{R}_T \leq \frac{2(\ln(T\Delta^2) + 1)}{\Delta}$

Reflection



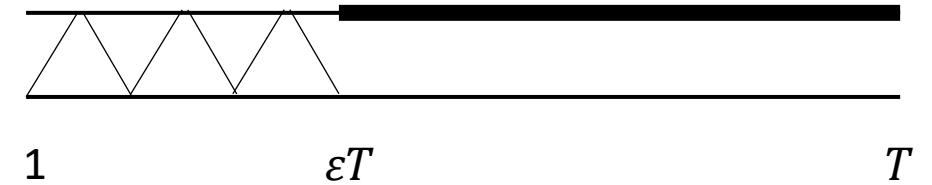
- $\varepsilon^* = \frac{4 \ln(T\Delta^2)}{T\Delta^2}$
- Exploration phase: $\varepsilon^* T = \frac{4 \ln(T\Delta^2)}{\Delta^2}$
- $\delta(\varepsilon^*) = 2e^{-\varepsilon^* T \Delta^2 / 4} = \frac{1}{T\Delta^2}$
- $\bar{R}_T \leq \frac{2(\ln(T\Delta^2) + 1)}{\Delta}$

- It takes $\sim \frac{\ln T}{\Delta^2}$ rounds to identify the best arm with confidence $\frac{1}{T\Delta^2}$
- At each exploration round we pay Δ
- Total regret order:

$$\bar{R}_T \approx \underbrace{\frac{\ln T}{\Delta^2} \Delta}_{\text{Exploration}} + \underbrace{\frac{1}{T\Delta^2} T\Delta}_{\text{Exploitation}} \approx \frac{\ln T}{\Delta}$$

- Small $\Delta \Rightarrow$ Harder problem (Larger \bar{R}_T)

Limitations

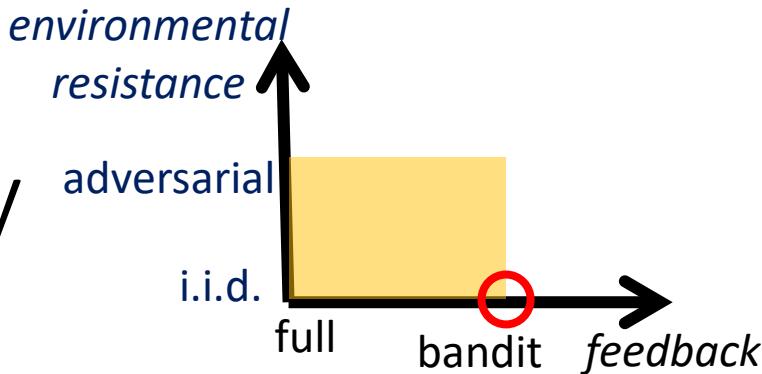


- Assumes knowledge of T
- Assumes knowledge of Δ
- Generalization to more than two actions is not straightforward

Lower Confidence Bound (LCB) algorithm for losses (Originally Upper Confidence Bound (UCB) for rewards) ("Optimism in the face of uncertainty" approach)

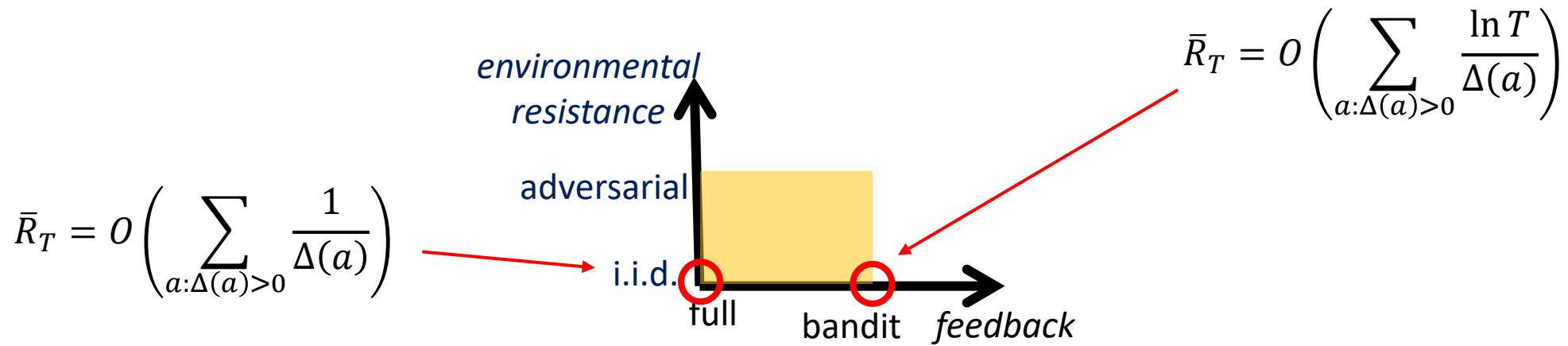
- Define $L_t^{CB}(a) = \hat{\mu}_{t-1}(a) - \sqrt{\frac{3 \ln t}{2N_{t-1}(a)}}$ lower confidence bound
 - (We will show that with high probability $L_t^{CB}(a) \leq \mu(a)$ for all t)
- LCB Algorithm:
 - Play each arm once
 - For $t = K + 1, K + 2, \dots$:
 - Play $A_t = \arg \min_a L_t^{CB}(a)$
- Theorem:
$$\bar{R}_T \leq 6 \sum_{a: \Delta(a) > 0} \frac{\ln T}{\Delta(a)} + \left(1 + \frac{\pi^2}{3}\right) \sum_a \Delta(a)$$
Rewards \leftrightarrow Losses
 - $\ell_{t,a} = 1 - r_{t,a}$
 - $r_{t,a} = 1 - \ell_{t,a}$
- Proof: next time

Stochastic bandits – mid-summary



- It takes $\sim \frac{\ln T}{\Delta^2}$ rounds to identify the best arm with confidence $\frac{1}{T\Delta^2}$
- Each exploration round costs Δ , but their number grows as $\frac{1}{\Delta^2}!$
- $\bar{R}_T \approx \underbrace{\frac{\ln T}{\Delta^2} \Delta}_{\text{Exploration}} + \underbrace{\frac{1}{T\Delta^2} T\Delta}_{\text{Exploitation}} \approx \frac{\ln T}{\Delta}$
- Problems with small Δ are **harder** than problems with large Δ !

Home Assignment



- In full information there is no need for exploration
- In T factor is the cost of exploration (the cost of bandit feedback) in i.i.d.