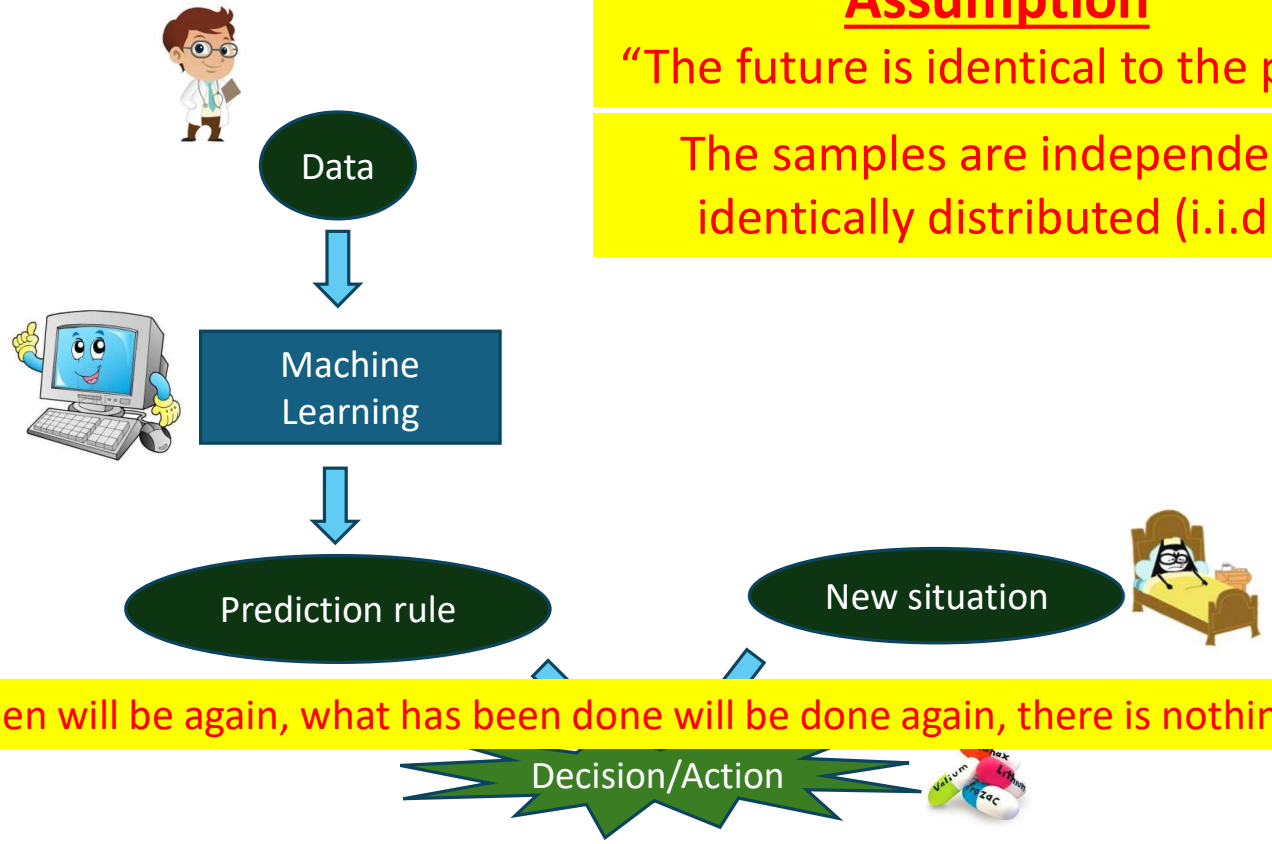
Introduction to Online Learning and Bandits

Yevgeny Seldin
University of Copenhagen

Learning Theory Summer School 2026

“Classical” (Batch) Machine Learning



Assumption

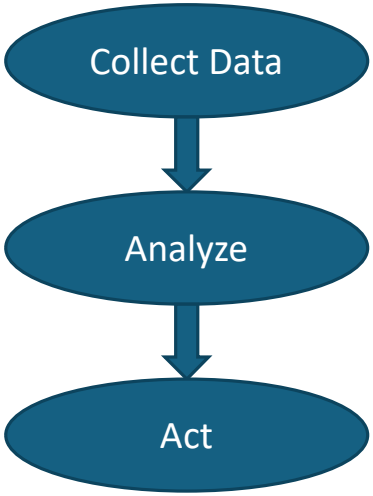
“The future is identical to the past”

The samples are independent identically distributed (i.i.d.)

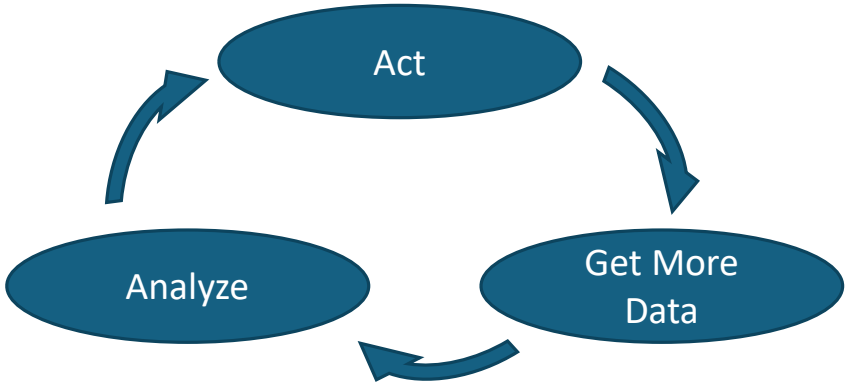
“What has been will be again, what has been done will be done again, there is nothing new under the sun.”

How Online is different from “batch”?

- Batch Learning



- Online Learning



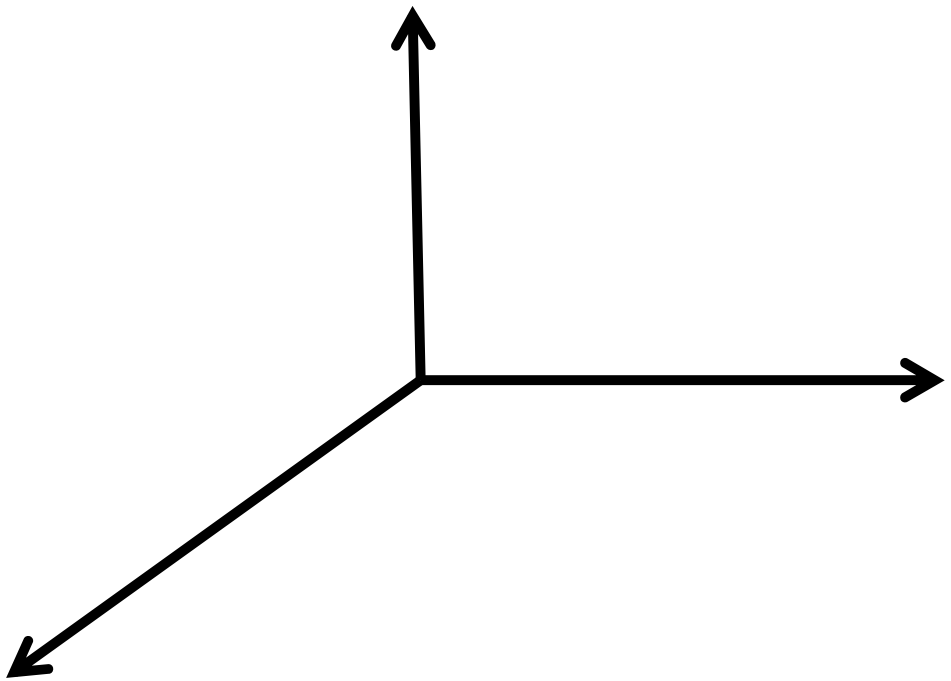
When do we need Online Learning?

- Intelligent data collection
- Interactive learning
- Large-scale data analysis
- “Adversarial” game-theoretic settings
  - No assumption on similarity of past and future

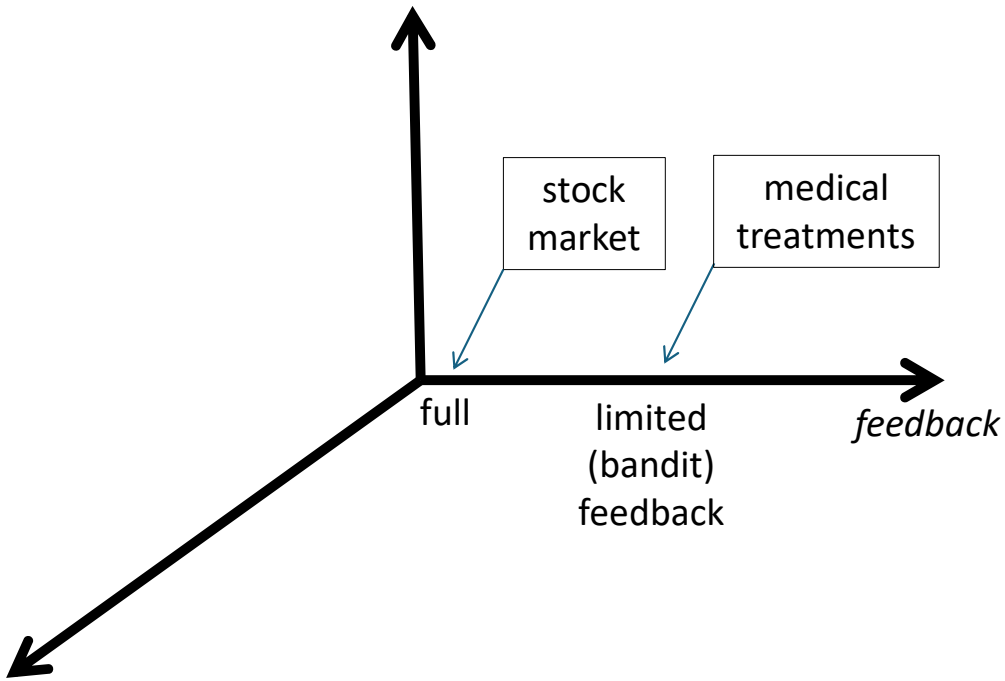
there can be no objection to the use of data, however meagre, as a guide to action required before more can be collected

(Thompson, 1933)

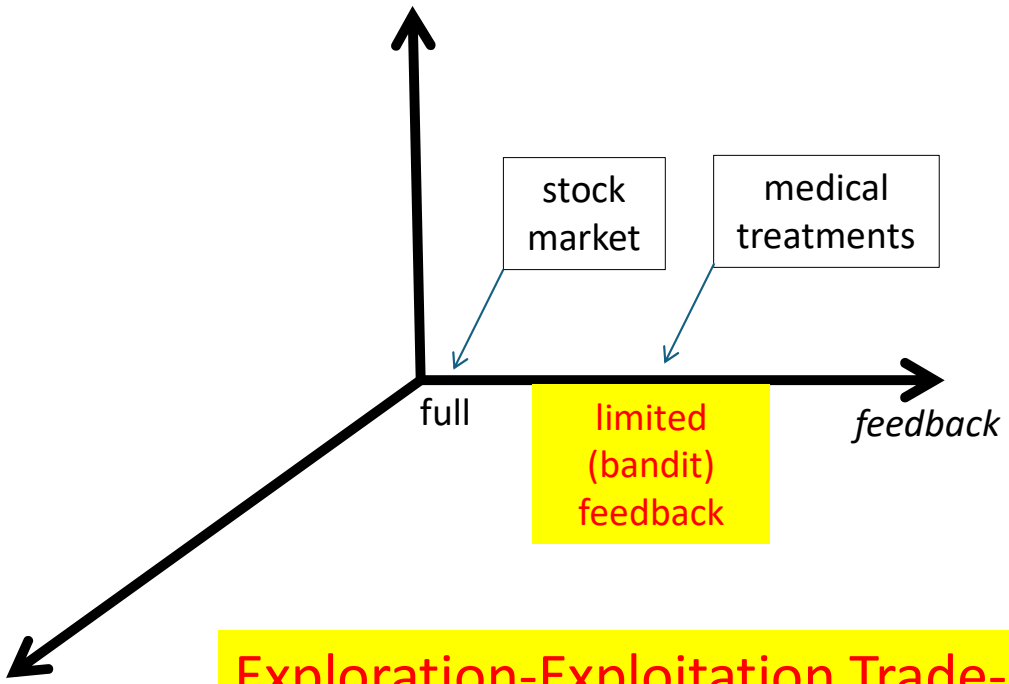
The Space of Online Learning Problems



The Space of Online Learning Problems



The Space of Online Learning Problems

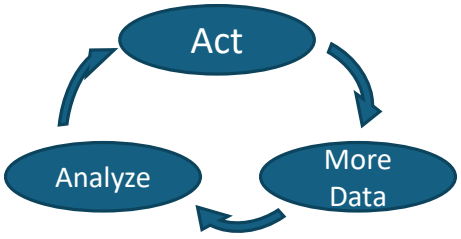


Exploration-Exploitation Trade-off

Exploration-Exploitation Trade-off

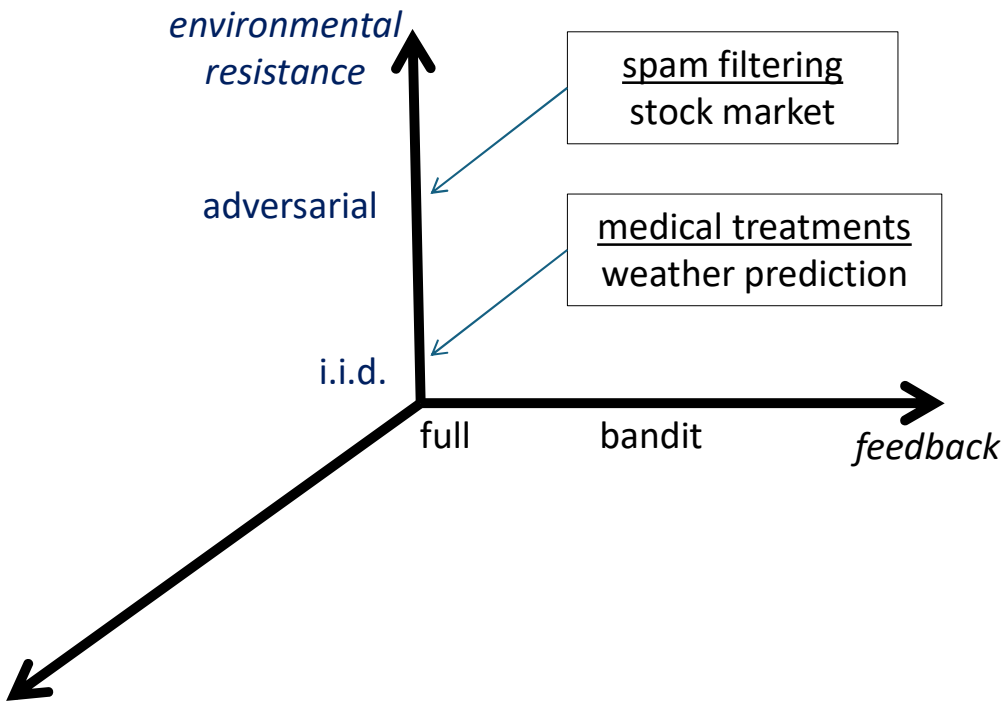
What drug to give to a new patient

When there are more patients to come…

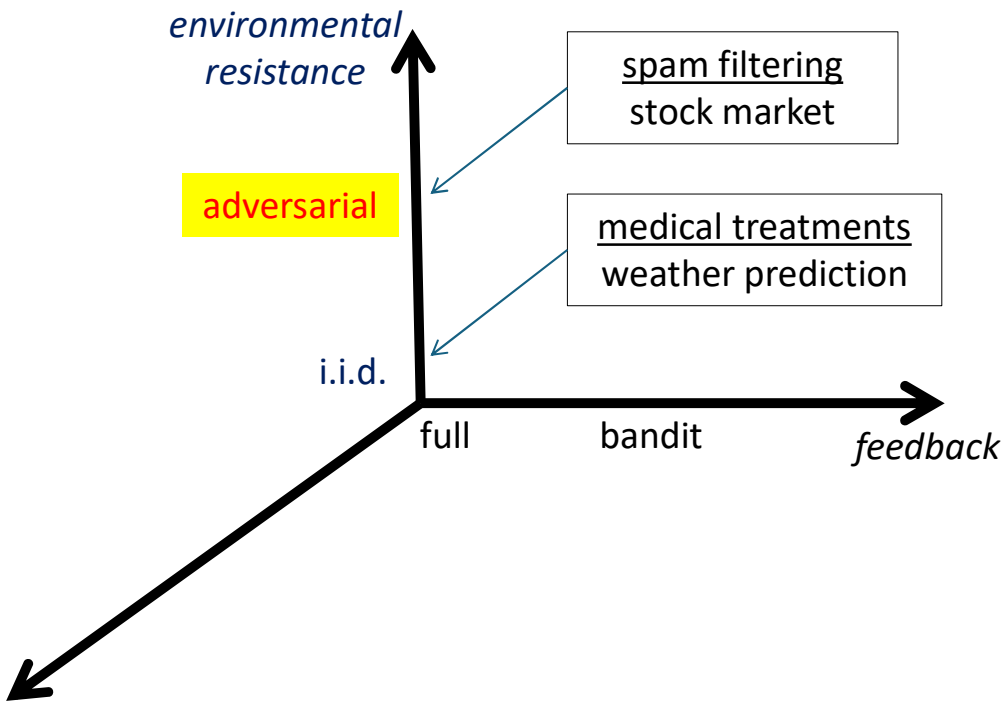


We are building the dataset for ourselves

The Space of Online Learning Problems



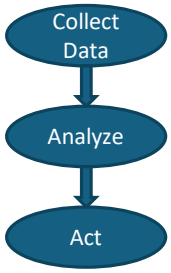
The Space of Online Learning Problems

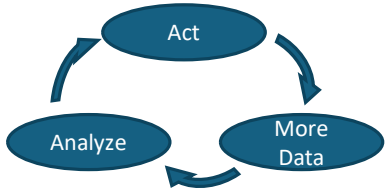


Learning in Adversarial Environments

- Game theoretic setting

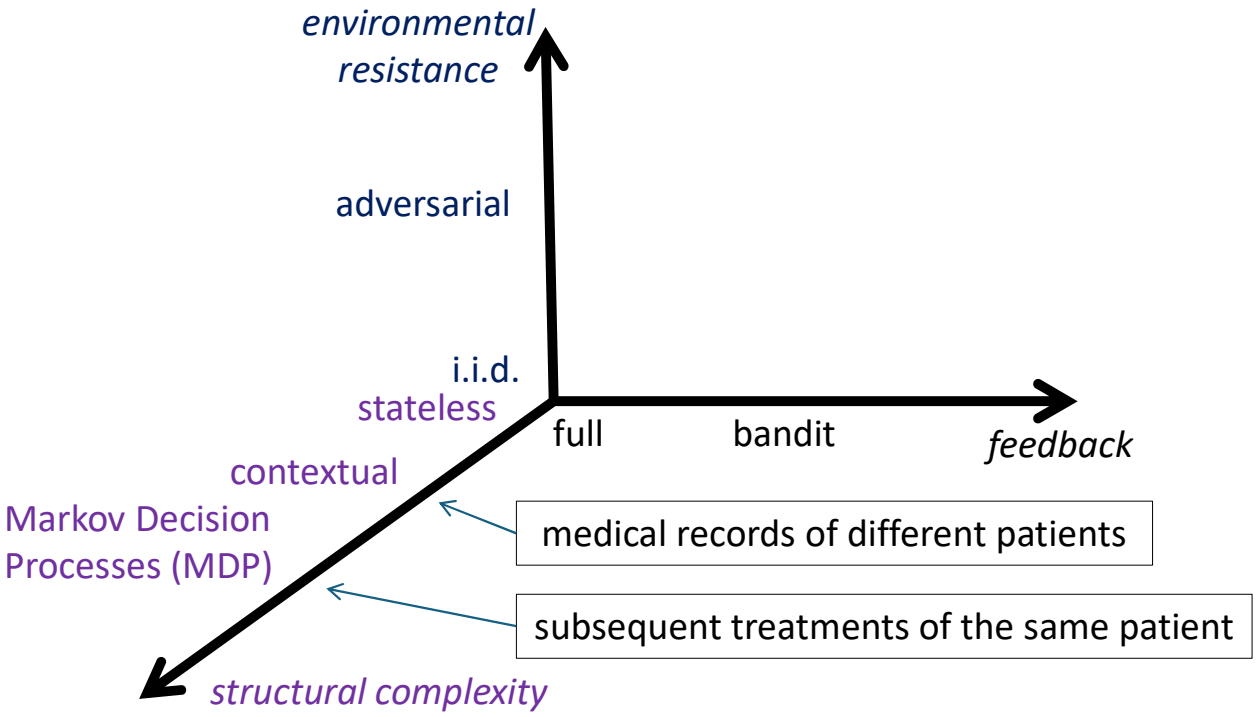
- Cannot be treated in batch learning



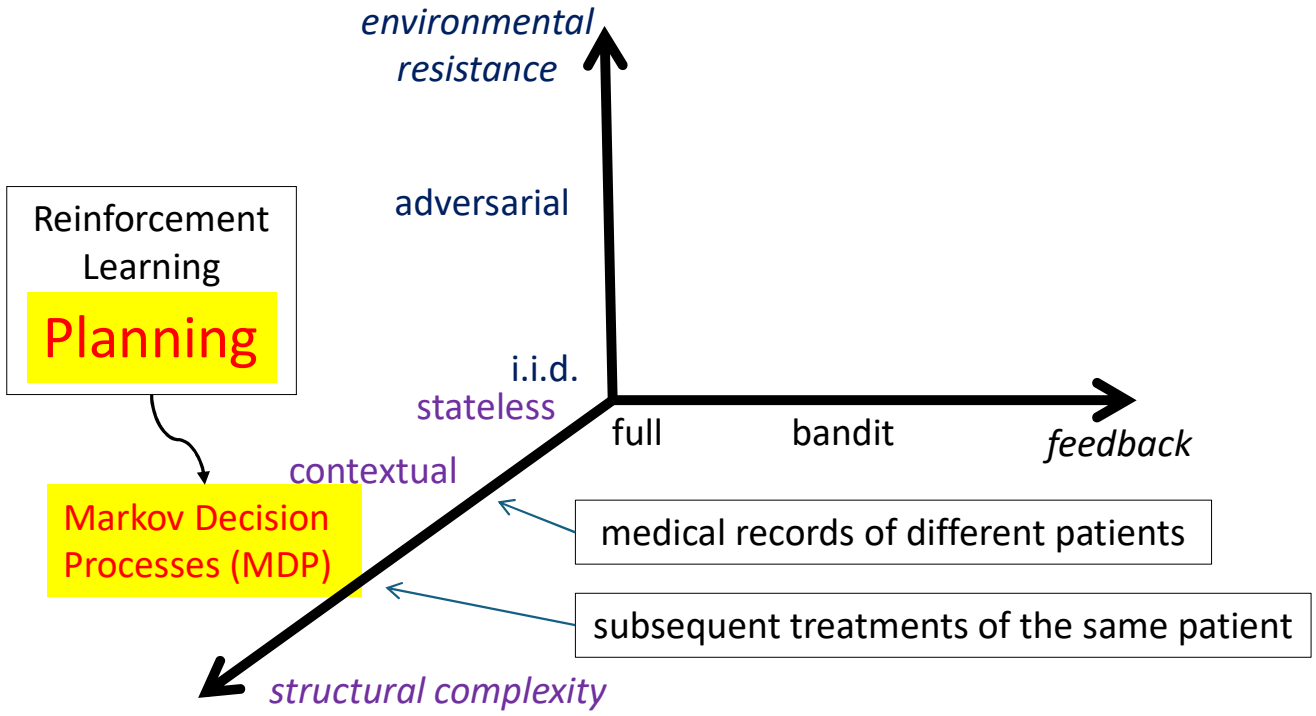


- Evaluation measure: *regret*
  - Difference in performance compared to the best choice in hindsight (out of a limited set)
  - E.g. investment revenue vs. the best stock in hindsight

The Space of Online Learning Problems

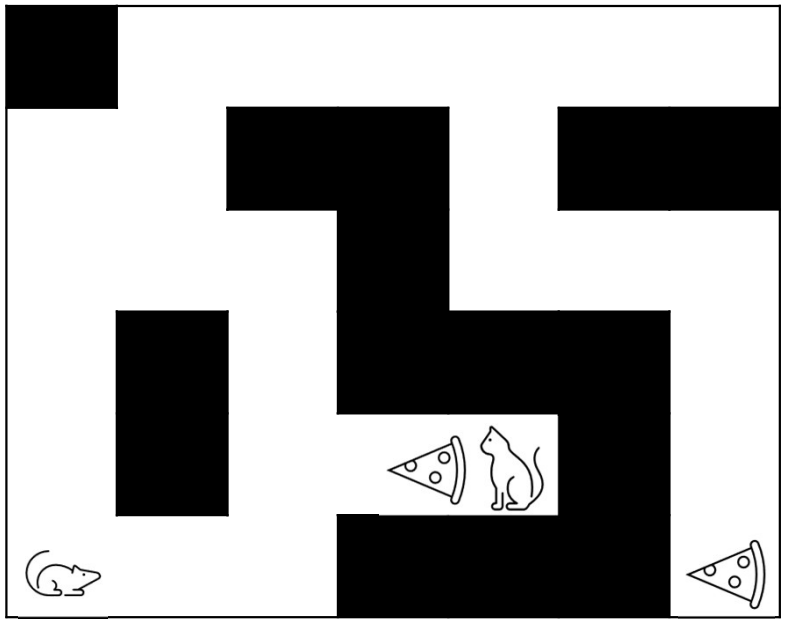


The Space of Online Learning Problems

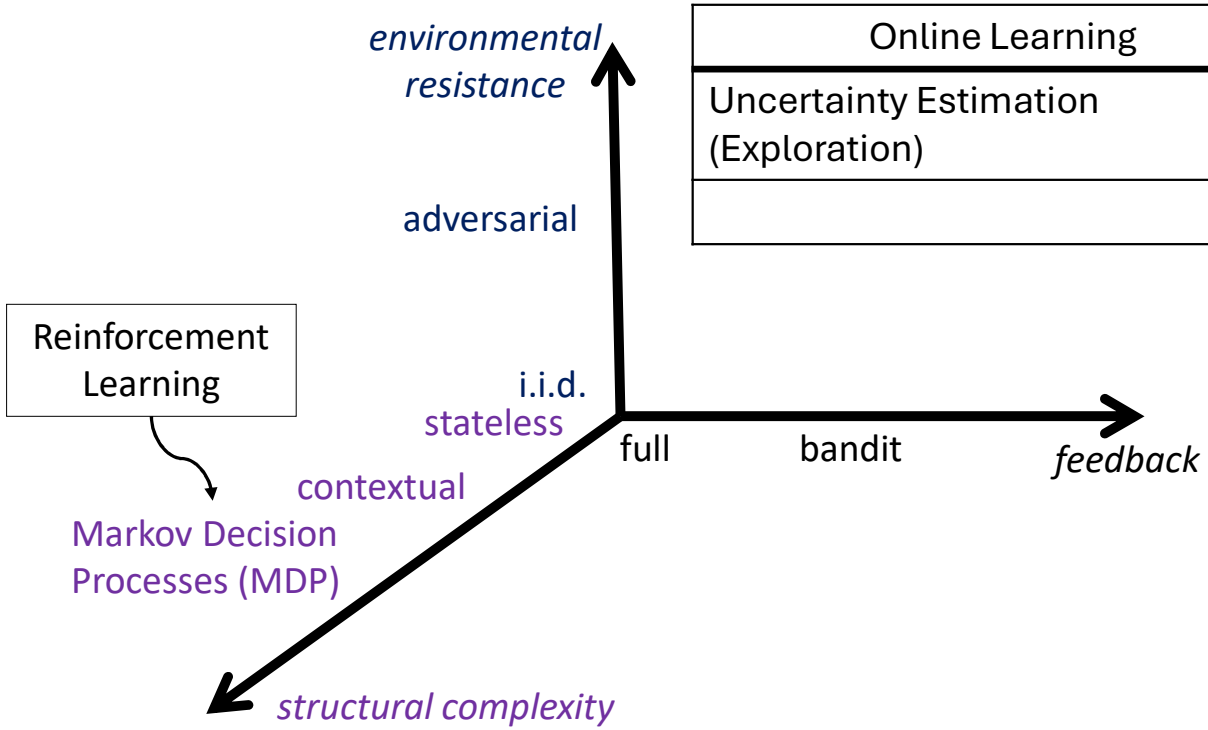


Planning

- Even if the immediate outcomes are known, long-term goals require planning

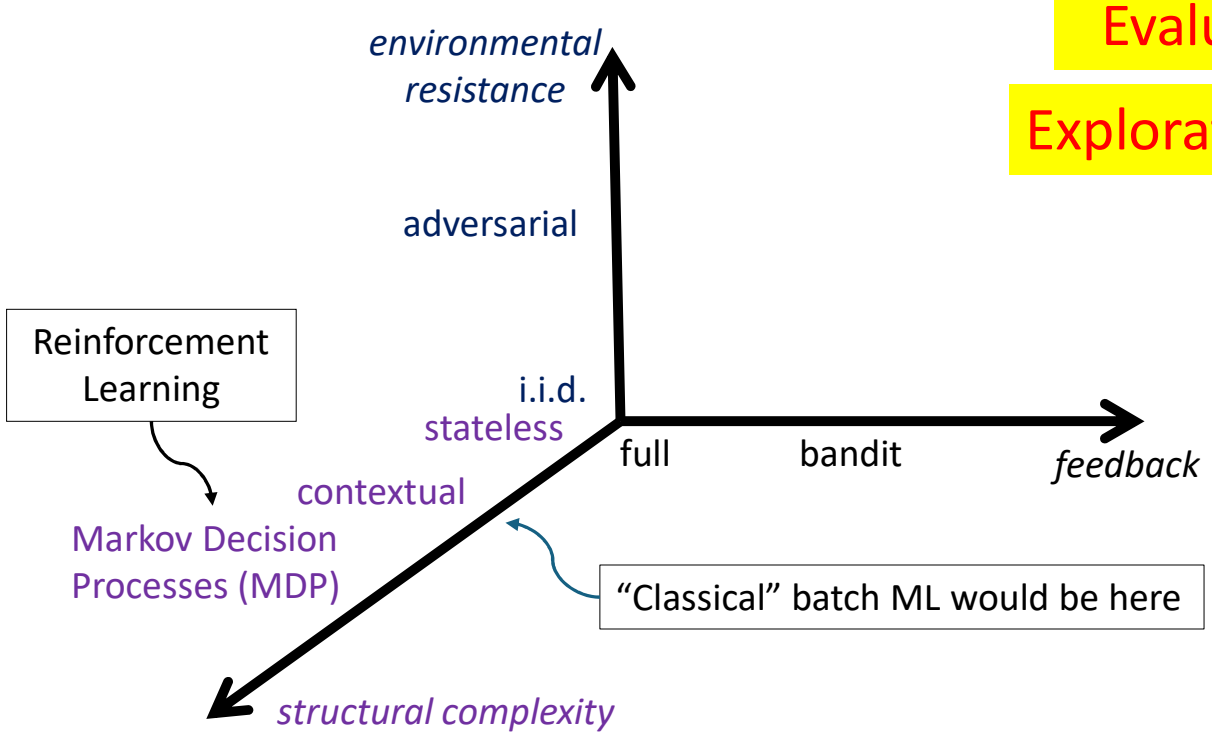


The Space of Online Learning Problems



Online Learning	Reinforcement Learning
Uncertainty Estimation (Exploration)	Uncertainty Estimation (Exploration)
	Planning

The Space of Online Learning Problems



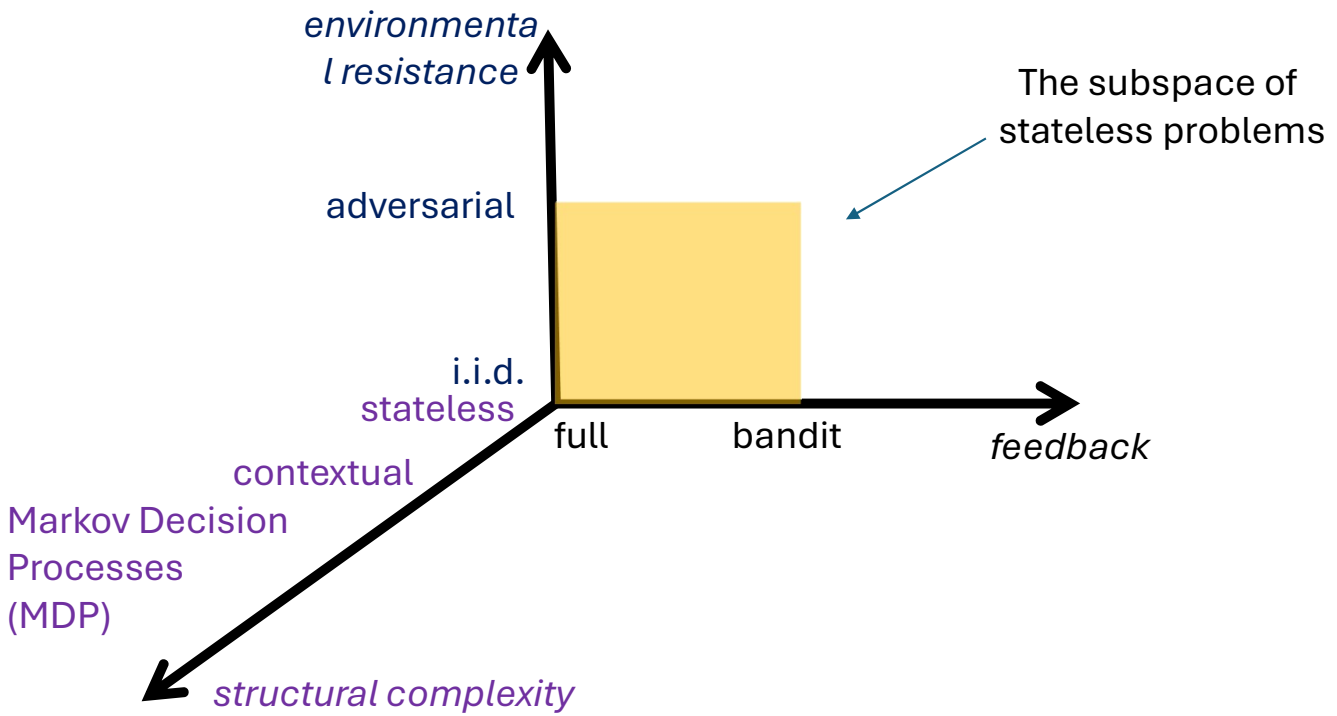
Evaluation measure: regret

Exploration-Exploitation Trade-off

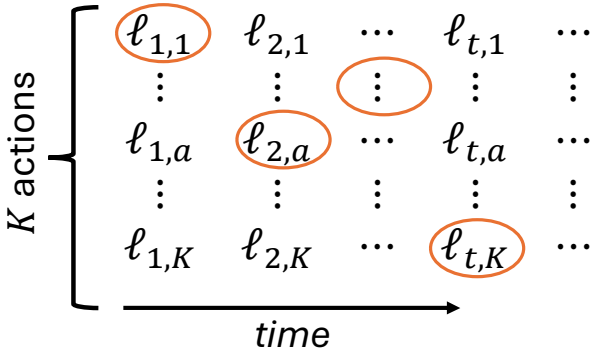
Planning

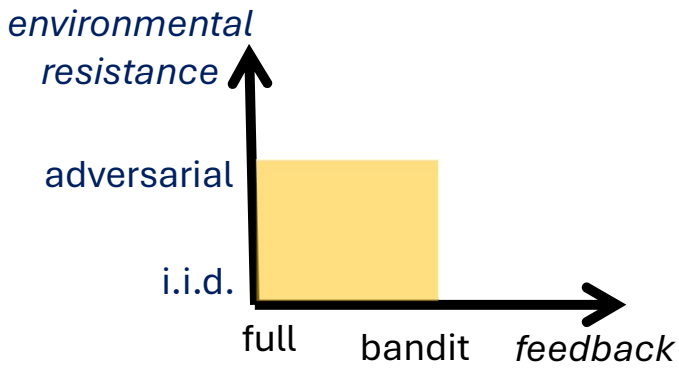
Online Learning Setup

The space of online learning problems



The stateless setting





$\ell_{t,a} \in [0,1]$

Game protocol:

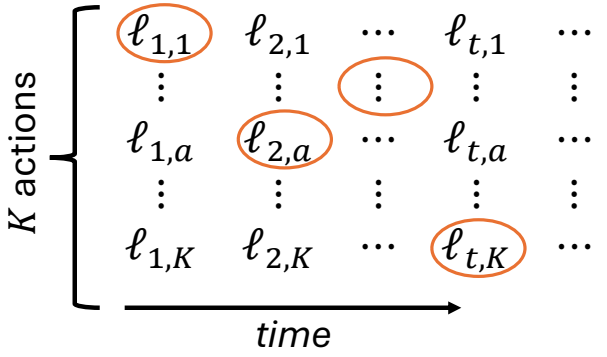
For $t = 1,2, \ldots$:

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

Generation of $\ell_{t,a}$ \ Observations	Full: $\ell_{t,1} \ldots \ell_{t,K}$	Bandit: ℓ_{t,A_t}
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: $R_T = \underbrace{\sum_{t=1}^{T} \ell_{t,A_t}}_{\text{Loss of the algorithm}} - \underbrace{\min_a \sum_{t=1}^{T} \ell_{t,a}}_{\text{Loss of the best action in hindsight}}$



$\ell_{t,a} \in [0,1]$

- Regret of order T means no learning
  - On average 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 measure

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

Example:

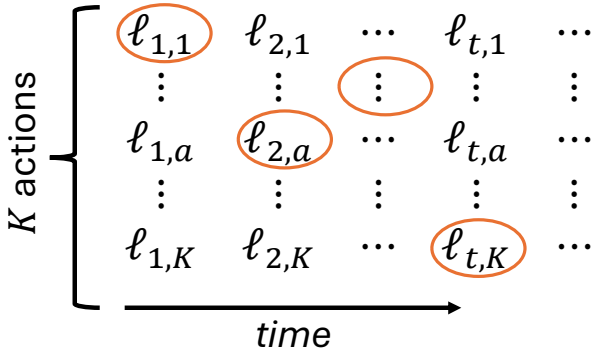
$$\begin{matrix} 1 & 0 & 1 & \\ 1 & 1 & 0 & \cdots \\ 0 & 1 & 1 & \end{matrix}$$

All entries are 1 except one entry selected uniformly at random that is 0.
Loss of the best path: 0
Expected loss of *any* algorithm: $\frac{K-1}{K}T$

- Regret of order T means no learning
  - On average 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: $R_T = \underbrace{\sum_{t=1}^{T} \ell_{t,A_t}}_{\text{Loss of the algorithm}} - \underbrace{\min_a \sum_{t=1}^{T} \ell_{t,a}}_{\text{Loss of the best action in hindsight}}$

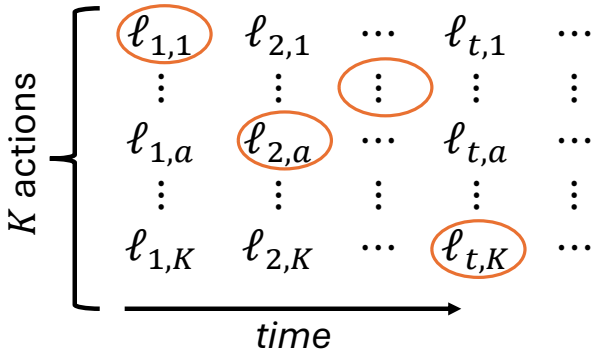
$\ell_{t,a} \in [0,1]$

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

$$\underbrace{=}_{\text{oblivious adversary}} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,A_t}\right] - \min_a \sum_{t=1}^{T} \ell_{t,a}$$

- Oblivious adversary:
  - $\ell_{t,a}$ is independent of $A_1, \ldots, A_{t-1}$
  - The losses can be written down before the game starts
- Adaptive adversary:
  - $\ell_{t,a}$ may depend on $A_1, \ldots, A_{t-1}$

Performance measures



- Pseudo-regret (stochastic setting):

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

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

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

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

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

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

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

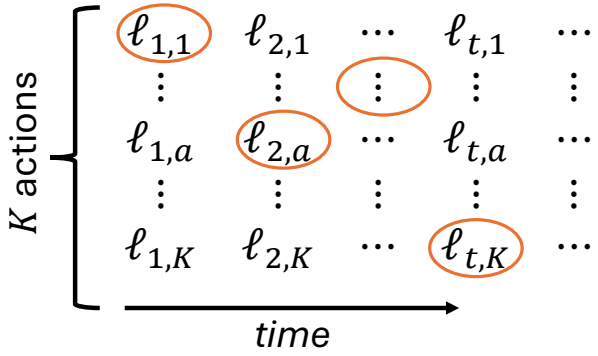
Expected regret:

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

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

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

Expected regret vs. Pseudo regret

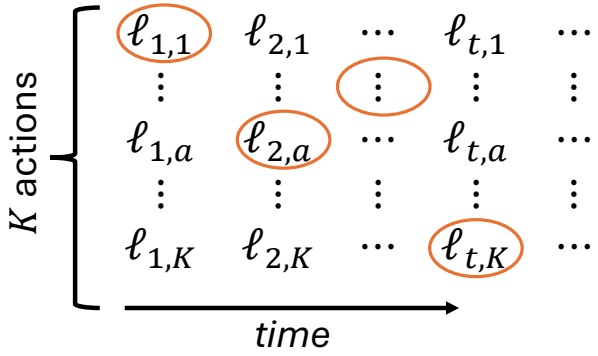


- Expected regret: $\mathbb{E}[R_T] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,A_t}\right] - \mathbb{E}\left[\min_a \sum_{t=1}^{T} \ell_{t,a}\right]$
- Pseudo-regret: $\bar{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,A_t}\right] - \min_a \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,a}\right] = \mathbb{E}\left[\sum_{t=1}^{T}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

- $\mathbb{E}\left[\min_a f(a, B)\right] \leq \min_a \mathbb{E}[f(a, B)] \;\Rightarrow\; \bar{R}_T \leq \mathbb{E}[R_T]$

- 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

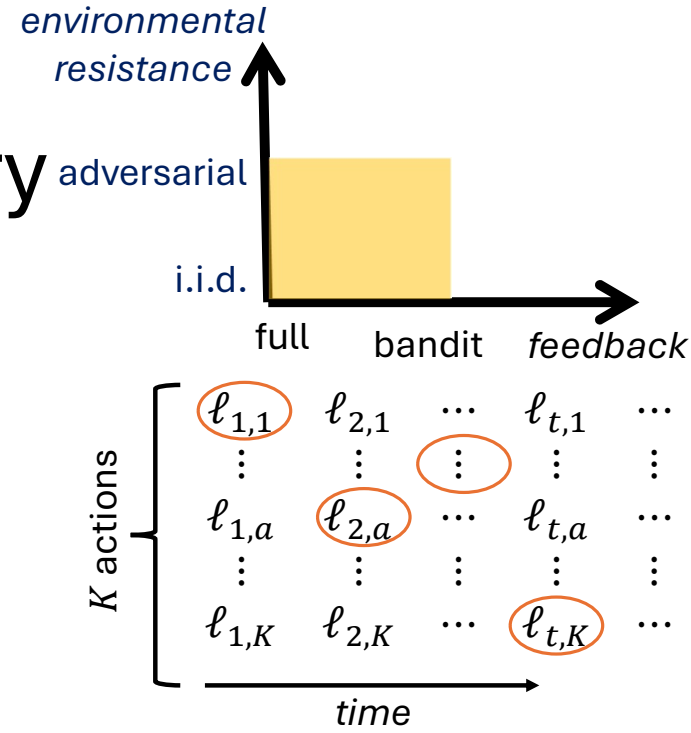


- Expected regret: $\mathbb{E}[R_T] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,A_t}\right] - \mathbb{E}\left[\min_a \sum_{t=1}^{T} \ell_{t,a}\right]$
- Pseudo-regret: $\bar{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,A_t}\right] - \min_a \mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,a}\right] = \mathbb{E}\left[\sum_{t=1}^{T}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

- $\mathbb{E}\left[\min_a f(a, B)\right] \leq \min_a \mathbb{E}[f(a, B)] \;\Rightarrow\; \bar{R}_T \leq \mathbb{E}[R_T]$

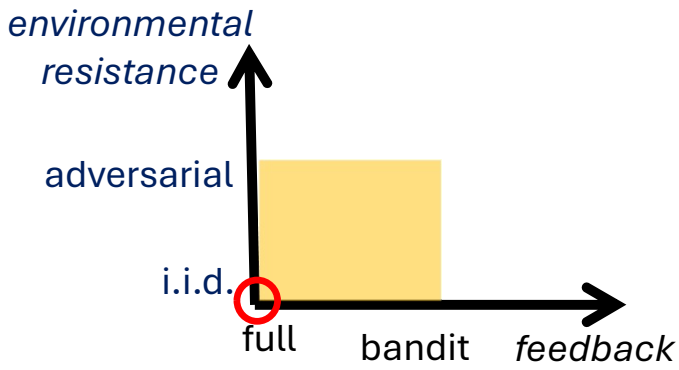
- Stochastic setting: imagine that $\mu(a) = \frac{1}{2}$ for all a. Then
  - $\mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,A_t}\right] = \frac{1}{2}T$
  - $\mathbb{E}\left[\sum_{t=1}^{T} \ell_{t,a}\right] = \frac{1}{2}T$ for all a
  - $\bar{R}_T = 0$
  - $\mathbb{E}\left[\min_a \sum_{t=1}^{T} \ell_{t,a}\right] \approx \frac{1}{2}T - \sqrt{\frac{1}{2}T \ln K}$
  - $\mathbb{E}[R_T] \approx \sqrt{\frac{1}{2}T \ln K}$
  - 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



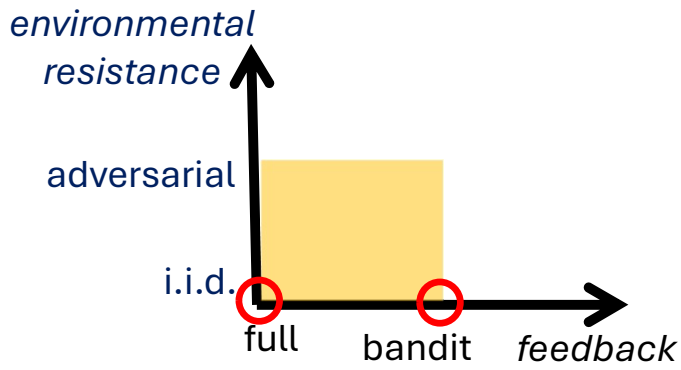
Generation of $\ell_{t,a}$ \ Observations	Full: $\ell_{t,1} \dots \ell_{t,K}$	Bandit: ℓ_{t,A_t}
Adversarial: $\ell_{t,a}$ arbitrary		
I.I.D.: $\ell_{t,a}$ sampled i.i.d., such that $\mathbb{E}[\ell_{t,a}] = \mu(a)$		

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



i.i.d. full info

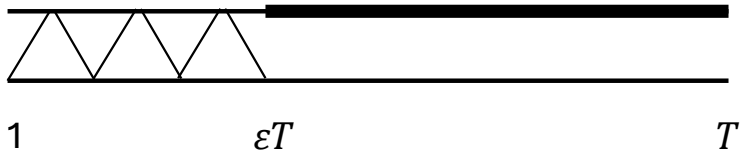
- Algorithm: Follow the Leader (FTL)
  - $A_t = \arg\min_a \sum_{s=1}^{t-1} \ell_{t,a}$
  - (ties resolved arbitrarily)
- $\bar{R}_T = O\left(\sum_{a:\Delta(a)>0} \frac{1}{\Delta(a)}\right)$
  - Pseudo-regret upper bound does not grow with time!
- Proof: exercise



Stochastic (i.i.d.) bandits

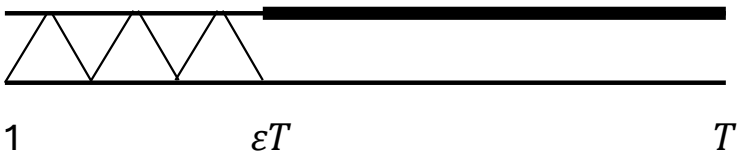
Exploration-Exploitation trade-off: a simple approach

- Setting:
  - Two actions
  - Bandit feedback
  - T is known
  - Δ is known



- Approach:
  - Explore for εT rounds
  - Exploit for the remaining rounds

Analysis



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

  $= \mathbb{P}\left(\Delta - \underbrace{\left(\hat{\mu}_{\varepsilon T}(a) - \hat{\mu}_{\varepsilon T}(a^*)\right)}_{\frac{2}{\varepsilon T}\sum_{t=1}^{\varepsilon T/2}(\ell_{2t-1,a} - \ell_{2t,a^*})} \ge \Delta\right)$

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

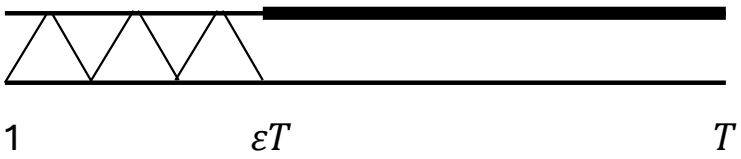
Hoeffding's inequality:

Let $Z_1, \ldots, Z_n$ be i.i.d. within interval of length β ($Z_i \in [c, c+\beta]$), then

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

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

Analysis



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

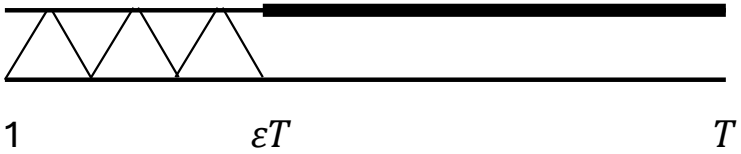
  $= \mathbb{P}\left(\Delta - \underbrace{\left(\hat{\mu}_{\varepsilon T}(a) - \hat{\mu}_{\varepsilon T}(a^*)\right)}_{\frac{2}{\varepsilon T}\sum_{t=1}^{\varepsilon T/2}(\ell_{2t-1,a} - \ell_{2t,a^*})} \ge \Delta\right)$

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

Reflection

- Exploration phase required to identify a^* with sufficient confidence ($\delta(\varepsilon^*) \approx \frac{1}{T\Delta^2}$) is $\varepsilon^* T \approx \frac{\ln}{\Delta^2}$
- Each exploration round costs Δ
- $\bar{R}_T \approx \Delta\frac{\ln}{\Delta^2} = \frac{\ln}{\Delta}$
- Problems with **small Δ are harder** than problems with large Δ
  - up to a limit, because $\bar{R}_T = O\left(\max\left(\Delta T, \frac{\ln}{\Delta}\right)\right)$

Analysis



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

  $= \mathbb{P}\left(\Delta - \underbrace{\left(\hat{\mu}_{\varepsilon T}(a) - \hat{\mu}_{\varepsilon T}(a^*)\right)}_{\frac{2}{\varepsilon T}\sum_{t=1}^{\varepsilon T/2}(\ell_{2t-1,a} - \ell_{2t,a^*})} \ge \Delta\right)$

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

Limitations

- Assumed knowledge of T
- Assumed knowledge of Δ
- Generalization beyond 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 $t = K+1, K+2, \ldots$:
    - Play $A_t = \arg\min_a L_t^{CB}(a)$

- No knowledge of T
- No knowledge of Δ
- Works for any K

Rewards ↔ Losses

$\ell_{t,a} = 1 - r_{t,a}$

$r_{t,a} = 1 - \ell_{t,a}$

- Theorem:

$$\bar{R}_T \leq 6 \sum_{a:\Delta(a)>0} \frac{\ln T}{\Delta(a)} + \left(1 + \frac{\pi^2}{3}\right) \sum_a \Delta(a)$$

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 $t = K+1, K+2, \ldots$:
    - Play $A_t = \arg\min_a L_t^{CB}(a)$

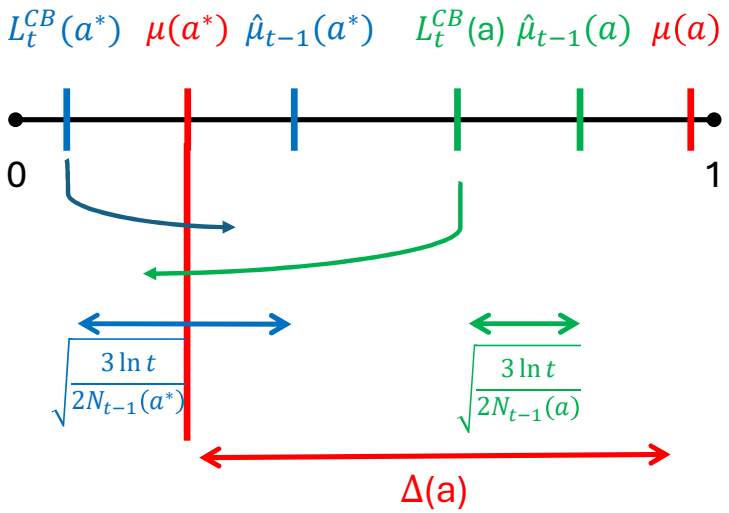
- Theorem:

$$\bar{R}_T \leq 6 \sum_{a:\Delta(a)>0} \frac{\ln T}{\Delta(a)} + \left(1 + \frac{\pi^2}{3}\right) \sum_a \Delta(a)$$

- Proof:
  - $\bar{R}_T = \sum_{a=1}^K \Delta(a) \mathbb{E}[N_T(a)]$
  - When can we play $a \neq a^*$?
  - Bound the expected number of times $L_t^{CB}(a) \leq L_t^{CB}(a^*)$

Proof

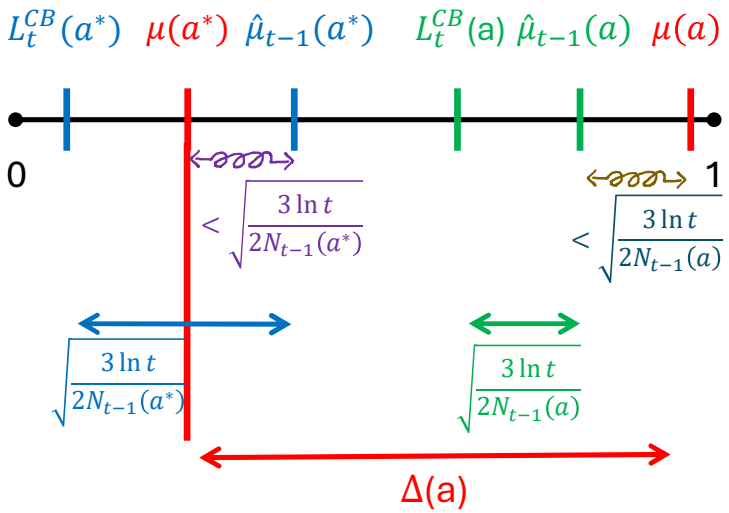
- $\bar{R}_t(a) = \sum_a \Delta(a)\mathbb{E}[N_T(a)]$
- $L_t^{CB}(a) = \hat{\mu}_{t-1}(a) - \sqrt{\frac{3 \ln t}{2N_{t-1}(a)}}$



- Bound the expected number of times $L_t^{CB}(a) \leq L_t^{CB}(a^*)$

- The expected number of times $L_t^{CB}(a) \leq L_t^{CB}(a^*)$ is bounded by
  1. The expected number of times $L_t^{CB}(a^*) \geq \mu(a^*)$
  2. Plus expected the number of times $L_t^{CB}(a) \leq \mu(a^*)$

Proof continued



1. The expected number of times $L_t^{CB}(a^*) \geq \mu(a^*)$ is bounded by

   The expected number of times $\hat{\mu}_{t-1}(a^*) \geq \mu(a^*) + \sqrt{\frac{3 \ln t}{2N_{t-1}(a^*)}}$

   (↜"breaks" – the confidence bound fails)

2. The expected the number of times $L_t^{CB}(a) \leq \mu(a^*)$ is bounded by

   2.1 The expected number of times $\hat{\mu}_{t-1}(a) \leq \mu(a) - \sqrt{\frac{3 \ln t}{2N_{t-1}(a)}}$

   (↜"breaks" – the confidence bound fails)

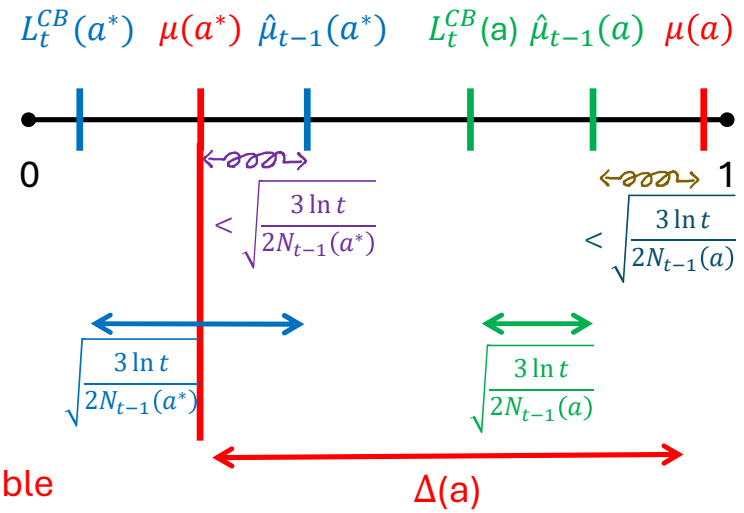
   2.2 Plus the number of times $\sqrt{\frac{3 \ln t}{2N_{t-1}(a)}} + \sqrt{\frac{3 \ln t}{2N_{t-1}(a)}} = \sqrt{\frac{6 \ln t}{N_{t-1}(a)}} \geq \Delta(a)$

   (The confidence bound holds, but it is not tight enough)

   $\Rightarrow \quad N_{t-1}(a) \leq \frac{6 \ln t}{\Delta(a)^2} \leq \frac{6 \ln T}{\Delta(a)^2}$

- Mid-summary: $\mathbb{E}[N_T(a)] \leq \frac{6 \ln T}{\Delta(a)^2} + 1 + \mathbb{E}[1.] + \mathbb{E}[2.1]$

Proof continued



- $\mathbb{E}[N_T(a)] \leq \frac{6 \ln T}{\Delta(a)^2} + 1 + \mathbb{E}[\leftrightsquigarrow] + \mathbb{E}[\leftrightsquigarrow]$
- Let $F(a^*)$ be the expected number of times $\hat{\mu}_{t-1}(a^*) \geq \mu(a^*) + \sqrt{\frac{3 \ln t}{2N_{t-1}(a^*)}}$
- Bound $\mathbb{P}\left(\hat{\mu}_{t-1}(a^*) - \mu(a^*) \geq \sqrt{\frac{3 \ln t}{2N_{t-1}(a^*)}}\right)$ ← $N_{t-1}(a^*)$ is a random variable dependent on $\hat{\mu}_t(a^*)$!

- Idea: break dependent events into independent events and take a union bound
- Introduce $X_1, \ldots, X_T$ r.v. with the same distribution as ℓ_{t,a^*}
- Let $\bar{\mu}_s = \frac{1}{s}\sum_{i=1}^{s} X_i$
- $\mathbb{P}\left(\hat{\mu}_{t-1}(a^*) - \mu(a^*) \geq \sqrt{\frac{3 \ln t}{2N_{t-1}(a^*)}}\right) \leq \mathbb{P}\left(\exists s \in \{1, \ldots, t\}: \bar{\mu}_s - \mu(a^*) \geq \sqrt{\frac{\ln\ ^3}{2s}}\right)$

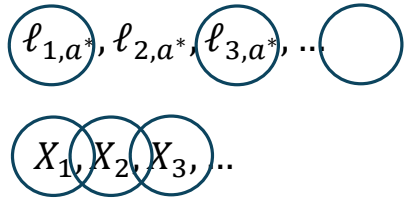
$$\overset{\text{union}}{\leq} \sum_{s=1}^{t} \mathbb{P}\left(\bar{\mu}_s - \mu(a^*) \geq \sqrt{\frac{\ln t^3}{2s}}\right)$$

$$\overset{\text{Hoeffding}}{\leq} \sum_{s=1}^{t} \frac{1}{t^3} = \frac{1}{t^2}$$

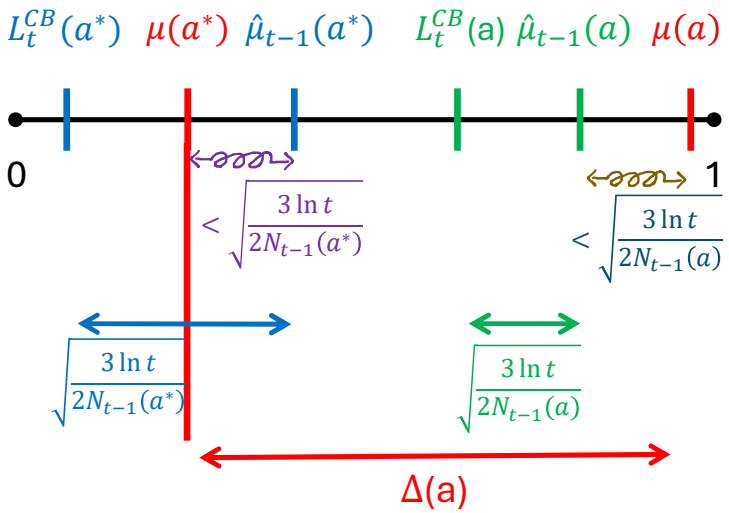
- $\mathbb{E}[F(a^*)] = \sum_{t=1}^{\infty} \mathbb{P}\left(L_t^{CB}(a^*) \geq \mu(a^*)\right) \leq \sum_{t=1}^{\infty} \frac{1}{t^2} \leq \frac{\pi^2}{6}$

Hoeffding:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu \geq \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}\right) \leq \delta$$

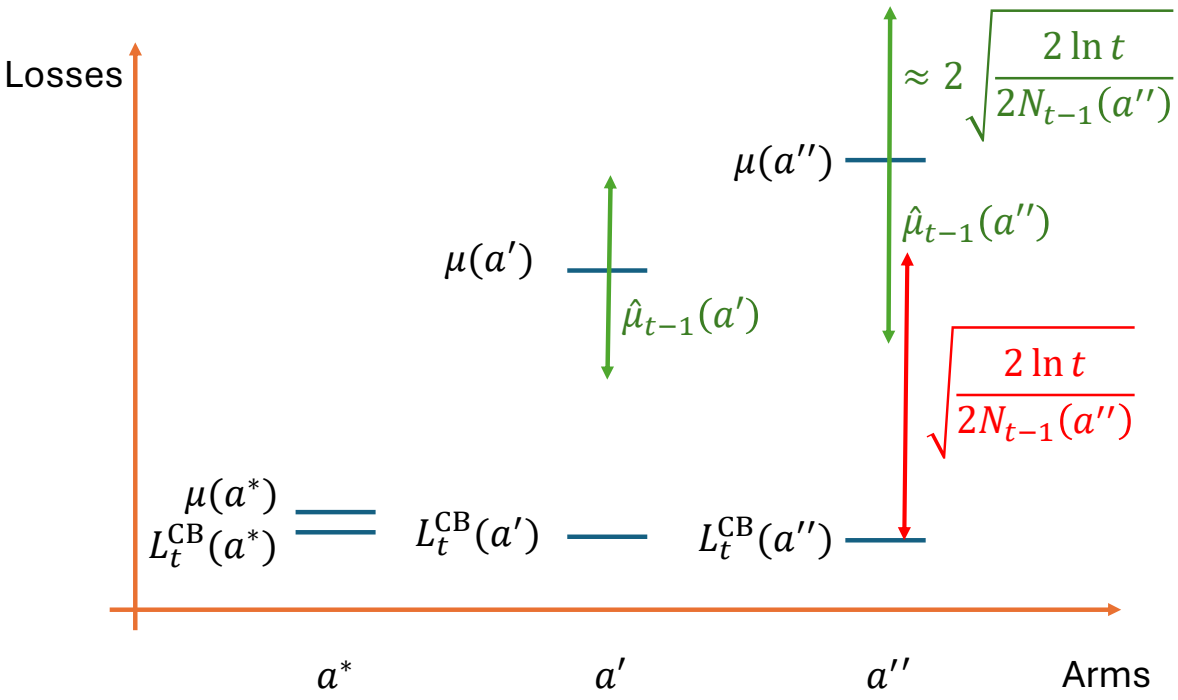


Proof summary



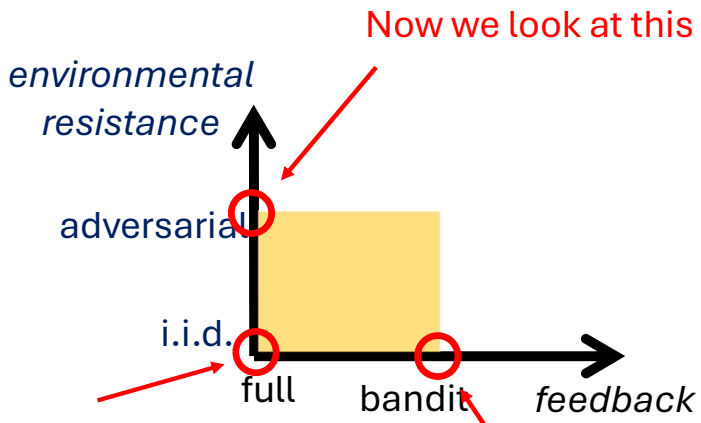
- $\bar{R}_t(a) = \sum_a \Delta(a) \mathbb{E}[N_T(a)]$
- $\mathbb{E}[N_T(a)] \leq \underbrace{\frac{6 \ln T}{\Delta(a)^2}}_{\text{The time it takes the confidence interval to start working}} + 1 + \underbrace{\frac{\pi^2}{6} + \frac{\pi^2}{6}}_{\text{The expected number of times confidence intervals fail}}$
- $\bar{R}_T \leq 6 \sum_{a:\Delta(a)>0} \frac{\ln}{\Delta(a)} + \left(1 + \frac{\pi^2}{3}\right) \sum_a \Delta(a)$
- Exercise:
  - Take $L_t^{CB}(a) = \hat{\mu}_{t-1}(a) - \sqrt{\frac{2 \ln t}{2N_{t-1}(a)}}$ (instead of $L_t^{CB}(a) = \hat{\mu}_{t-1}(a) - \sqrt{\frac{3 \ln t}{2N_{t-1}(a)}}$; i.e. confidence $\frac{1}{t^2}$ instead $\frac{1}{t^3}$)
  - Show $\bar{R}_T \leq 4 \sum_{a:\Delta(a)>0} \frac{\ln}{\Delta(a)} + (2 \ln T + 3) \sum_a \Delta(a)$

LCB algorithm dynamics (with $L_t^{CB}(a) = \hat{\mu}_{t-1}(a) - \sqrt{\frac{2 \ln t}{2N_{t-1}(a)}}$)



- Confidence interval of the played arm shrinks ($N_{t-1}(a)$ grows)
- Confidence intervals of all other arms grow ($\ln t$ grows)
- ⇒ all LCBs are roughly at the same level
- Most of the time $L_t^{\mathrm{CB}}(a^*) \leq \mu(a^*)$
- a^* is played a lot, so $L_t^{\mathrm{CB}}(a^*)$ is very close to $\mu(a^*)$
- All other arms are played just enough to keep $\sqrt{\frac{2 \ln t}{2N_{t-1}(a)}} = \theta(\Delta(a))$, i.e. $N_t(a) = \theta\left(\frac{\ln}{\Delta(a)^2}\right)$

So far



Algorithm: FTL

$$\bar{R}_T = O\left(\sum_{a:\Delta(a)>0} \frac{1}{\Delta(a)}\right)$$

not growing with T

In full info there is no need for exploration!

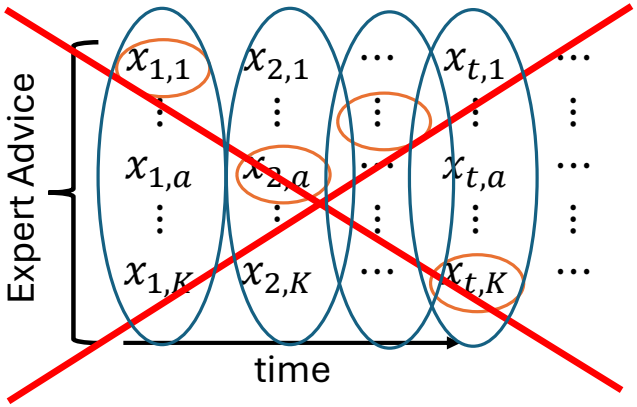
Algorithm: UCB / LCB

$$\bar{R}_T = O\left(\sum_{a:\Delta(a)>0} \frac{\ln T}{\Delta(a)}\right)$$

$\ln T$ factor - the cost of exploration

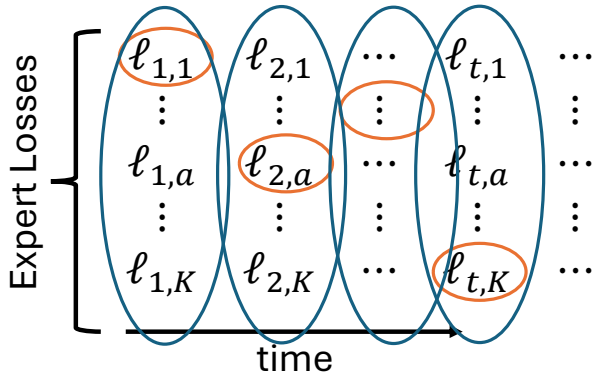
Prediction with Expert Advice (Adversarial full info game)

Prediction with Expert Advice (Adversarial full info game)



- Performance measures
- Regret:

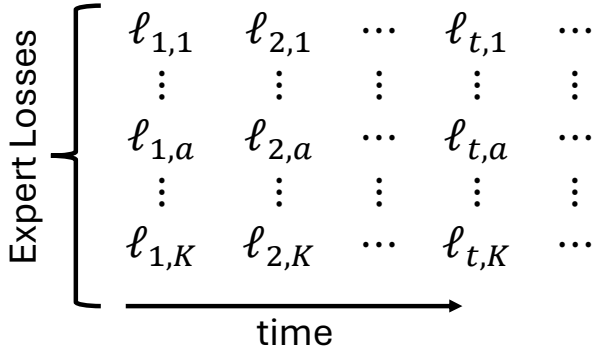
$$R_T = \sum_{t=1}^{T} \ell_{t,A_t} - \min_a \sum_{t=1}^{T} \ell_{t,a}$$



- Expected regret (oblivious setting):

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

General observation



- Deterministic algorithms are not suitable for adversarial environments
  - Why?
  - What deterministic algorithms have we seen so far?
- We need to randomize
- Playing uniformly at random will not work
- We need to balance between randomizing and giving preference to better actions

Algorithm for adversarial full info: Hedge / Exponential weights

- $\forall a: L_0(a) = 0$
- For $t = 1, 2, \ldots$
  - $\forall a: p_t(a) = \frac{e^{-\eta_t L_{t-1}(a)}}{\sum_{a'} e^{-\eta_t L_{t-1}(a')}}$
  - $A_t \sim p_t$
  - [Observe $\ell_{t,1}, \ldots, \ell_{t,K}$]
  - $\forall a: L_t(a) = L_{t-1}(a) + \ell_{t,a}$

- Some intuition:
  - In the early versions $p_t(a) \propto p_{t-1}(a)(1-\varepsilon)^{\ell_{t,a}}$
    - $\ell_{t,a} \in \{0,1\}$
  - In Hedge: $p_t(a) \propto p_{t-1}(a)e^{-\eta_t \ell_{t,a}}$

Numerically stable calculation of p_t

- $p_t(a) = \frac{e^{-\eta_t L_{t-1}(a)}}{\sum_{a'} e^{-\eta_t L_{t-1}(a')}}$

- For large t, $L_{t-1}(a)$ can be large and $e^{-\eta_t L_{t-1}(a)}$ numerically become zero, leading to $\frac{0}{0}$ numerical instability

- Remedy:
  - $\frac{e^{-\eta_t L_{t-1}(a)}}{\sum_{a'} e^{-\eta_t L_{t-1}(a')}} = \frac{e^{-\eta_t \left(L_{t-1}(a) - \min_{a''} L_{t-1}(a'')\right)}}{\sum_{a'} e^{-\eta_t \left(L_{t-1}(a') - \min_{a''} L_{t-1}(a'')\right)}}$
  - In the expression on the right the denominator is at least 1, resolving numerical instability

$$p_t(a) = \frac{e^{-\eta_t L_{t-1}(a)}}{\sum_{a'} e^{-\eta_t L_{t-1}(a')}}$$

Analysis

- Lemma: For any sequence of non-negative $\ell_{t,a}$ and $p_t(a)$ as in Hedge

$$\underbrace{\sum_{t=1}^{T}\sum_{a=1}^{K} p_t(a)\ell_{t,a}}_{\text{The expected loss of Hedge at round } t} - \underbrace{\min_a L_T(a)}_{\text{The best loss in hindsight}} \leq \frac{\ln K}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{a=1}^{K} p_t(a)\underbrace{\left(\ell_{t,a}\right)^2}_{\leq 1}$$

≤ 1

$\leq T$

The expected loss of Hedge

The expected regret of Hedge $\mathbb{E}[R_T]$

- Corollary: $\mathbb{E}[R_T] \leq \frac{\ln}{\eta} + \frac{\eta}{2}T$
- Take $\eta = \sqrt{\frac{2\ln K}{T}}$, then $\mathbb{E}[R_T] \leq \sqrt{2T\ln K}$

Proof of the lemma

$$p_t(a) = \frac{e^{-\eta_t L_{t-1}(a)}}{\sum_{a'} e^{-\eta_t L_{t-1}(a')}}$$

$$\sum_{t=1}^{T}\sum_{a=1}^{K} p_t(a)\ell_{t,a} - \min_a L_T(a) \le \frac{\ln K}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{a=1}^{K} p_t(a)\left(\ell_{t,a}\right)^2$$

- Define $W_t = \sum_a e^{-\eta L_t(a)}$

$$\frac{W_t}{W_{t-1}} = \frac{\sum_a e^{-\eta L_t(a)}}{\sum_{a'} e^{-\eta L_{t-1}(a')}}$$

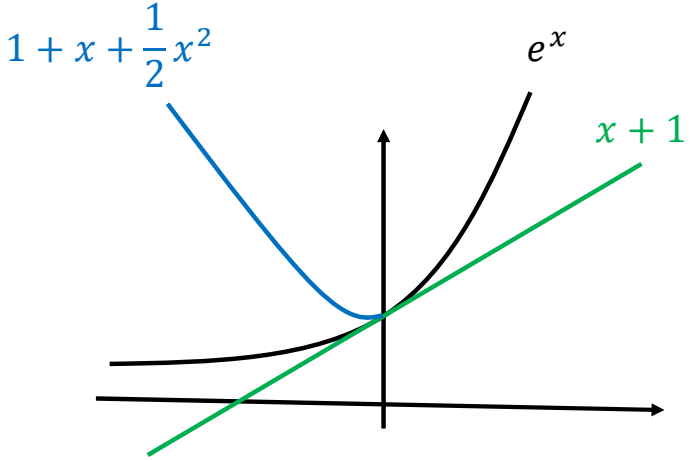
$$= \sum_a e^{-\eta \ell_{t,a}} \underbrace{\frac{e^{-\eta L_{t-1}(a)}}{\sum_{a'} e^{-\eta L_{t-1}(a')}}}_{p_t(a)}$$

$$= \sum_a e^{-\eta \ell_{t,a}}\, p_t(a)$$

$$\le \sum_a \left(1 - \eta\ell_{t,a} + \frac{1}{2}\eta^2\left(\ell_{t,a}\right)^2\right) p_t(a)$$

$$= 1 - \eta\sum_a \ell_{t,a} p_t(a) + \frac{\eta^2}{2}\sum_a \left(\ell_{t,a}\right)^2 p_t(a)$$

$$\le e^{-\eta\sum_a \ell_{t,a} p_t(a) + \frac{\eta^2}{2}\sum_a \left(\ell_{t,a}\right)^2 p_t(a)}$$



- For $x \le 0$:

$$e^x \le 1 + x + \frac{1}{2}x^2$$

- For any x:

$$1 + x \le e^x$$

Proof continued

$$W_t = \sum_a e^{-\eta L_t(a)}$$

$$\frac{W_t}{W_{t-1}} \le e^{-\eta \sum_a \ell_{t,a} p_t(a) + \frac{\eta^2}{2} \sum_a \left(\ell_{t,a}\right)^2 p_t(a)}$$

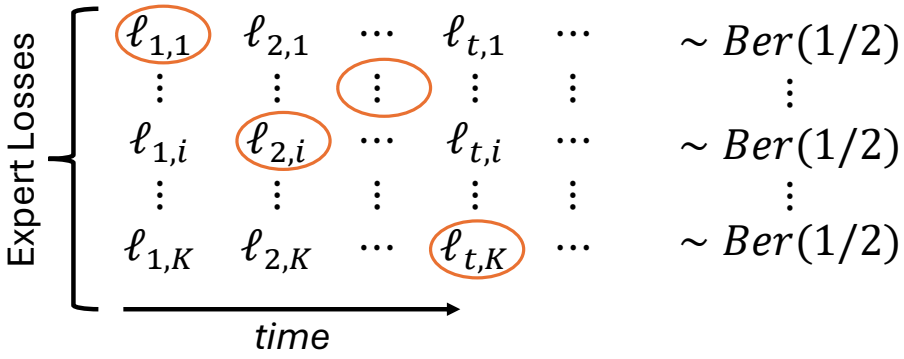
$$\frac{W_T}{W_0} = \frac{W_1}{W_0}\frac{W_2}{W_1}\frac{W_3}{W_2}\cdots\frac{W_T}{W_{T-1}} \le e^{-\eta \sum_{t=1}^{T} \sum_a \ell_{t,a} p_t(a) + \frac{\eta^2}{2} \sum_{t=1}^{T} \sum_a \left(\ell_{t,a}\right)^2 p_t(a)}$$

$$\frac{W_T}{W_0} = \frac{\sum_a e^{-\eta L_T(a)}}{K} \ge \frac{\max_a e^{-\eta L_T(a)}}{K} = \frac{e^{-\eta \min_a L_T(a)}}{K}$$

Put the two sides together, take a logarithm and normalize by η:

$$\sum_{t=1}^{T}\sum_{a=1}^{K} p_t(a)\ell_{t,a} - \min_a L_T(a) \le \frac{\ln K}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{a=1}^{K} p_t(a)\left(\ell_{t,a}\right)^2$$

Full information lower bound



$$\forall a:\ \mathbb{E}[L_T(a)] = \frac{T}{2}$$

$$\mathbb{E}\left[\sum_t^T \ell_{t,A_t}\right] = \frac{T}{2}$$

- Lemma

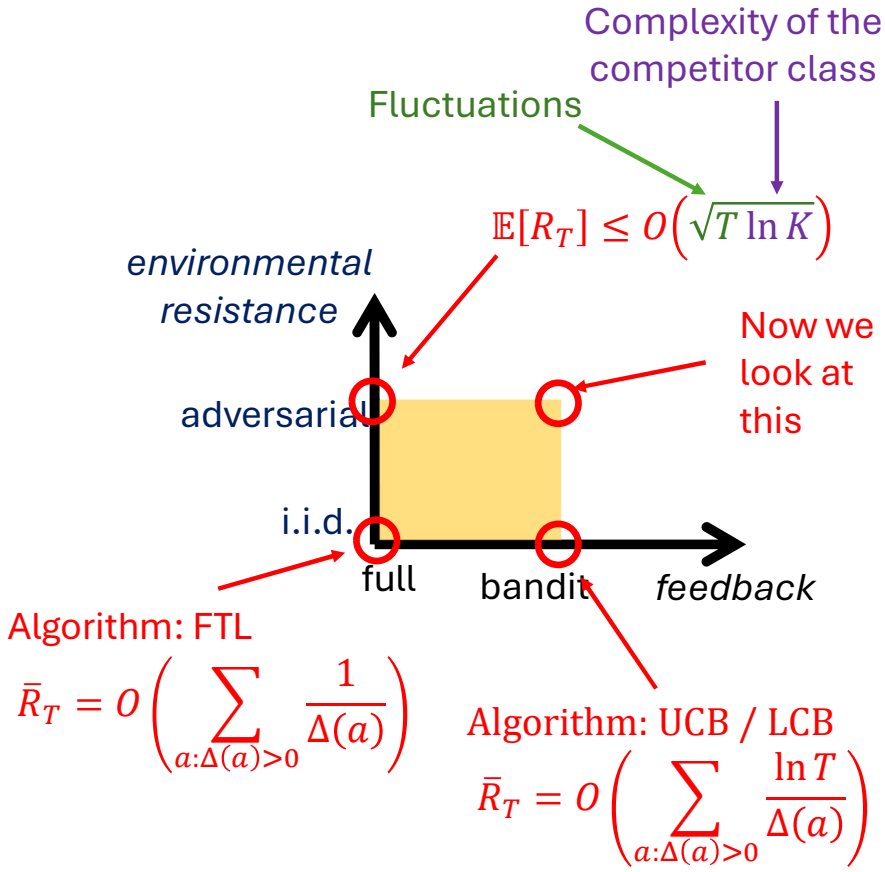
$$\lim_{T\to\infty}\lim_{K\to\infty}\frac{\frac{T}{2} - \mathbb{E}\left[\min_a L_T(a)\right]}{\sqrt{\frac{1}{2}T\ln K}} = 1$$

- In the limit of large T and K:

$$\underbrace{\underbrace{\frac{T}{2}}_{\mathbb{E}\left[\sum_t^T \ell_{t,A_t}\right]} - \mathbb{E}\left[\min_a L_T(a)\right]}_{\mathbb{E}[R_T]} \approx \sqrt{\frac{1}{2}T\underbrace{\ln K}_{\text{Complexity of the competitor (Amount of selection)}}}$$

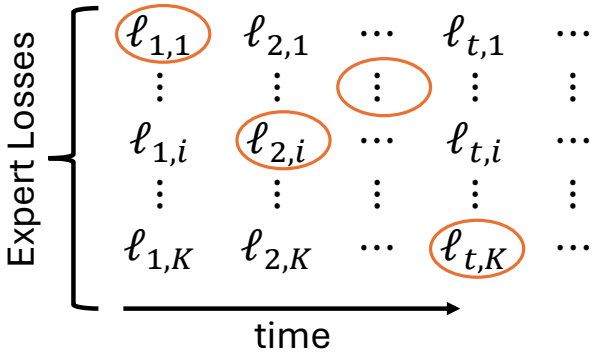
Summary

- Hedge:
  - $p_t(a) = \frac{e^{-\eta_t L_{t-1}(a)}}{\sum_{a'} e^{-\eta_t L_{t-1}(a')}}$
- Analysis:
  - Evolution of the potential function $W_t = \sum_a e^{-\eta L_t(a)}$
- Matching upper and lower bounds $\mathbb{E}[R_T] = \theta\left(\sqrt{T \ln K}\right)$



Adversarial Bandits

Adversarial bandits



- Performance measures
- Regret:

$$R_T = \sum_{t=1}^{T} \ell_{t,A_t} - \min_a \sum_{t=1}^{T} \ell_{t,a}$$

- Expected regret (oblivious setting):

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

Algorithm for adversarial bandits: EXP3

(Exponential Exploration Exploitation)

Hedge → EXP3

- $\forall a: L_0(a) = 0$
- For $t = 1, 2, \ldots$
  - $\forall a:\ p_t(a) = \frac{e^{-\eta_t L_{t-1}(a)}}{\sum_{a'} e^{-\eta_t L_{t-1}(a')}}$
  - $A_t \sim p_t$
  - ~~[Observe $\ell_{t,1}, \ldots, \ell_{t,K}$]~~
  - [Observe ℓ_{t,A_t}]
  - ~~$\forall a: L_t(a) = L_{t-1}(a) + \ell_{t,a}$~~
  - $\forall a: L_t(a) = L_{t-1}(a) + \frac{\ell_{t,a}\mathbb{I}(A_t = a)}{p_t(a)}$

- Importance-weighted loss estimate

$$\tilde{\ell}_{t,a} = \frac{\ell_{t,a}\mathbb{I}(A_t = a)}{p_t(a)}$$

- Defined for all a

Properties of importance-weighted samples $\tilde{\ell}_{t,a} = \frac{\ell_{t,a}\mathbb{I}(A_t=a)}{p_t(a)}$

- **Not** independent!
  - $\tilde{\ell}_{t,1}, \ldots \tilde{\ell}_{t,K}$ are dependent (only one is nonzero)
  - $p_t(a)$ is a random variable dependent on $A_1, \ldots, A_{t-1}$
- $\tilde{\ell}_{t,a}$ is an unbiased estimate of $\ell_{t,a}$ (meaning $\mathbb{E}\left[\tilde{\ell}_{t,a}\right] = \ell_{t,a}$)

$$\mathbb{E}\left[\tilde{\ell}_{t,a}\right] = \mathbb{E}\left[\frac{\ell_{t,a}\mathbb{I}(A_t = a)}{p_t(a)}\right]$$
$$= \ell_{t,a}\mathbb{E}\left[\frac{\mathbb{I}(A_t=a)}{p_t(a)}\right]$$
$$= \ell_{t,a}\mathbb{E}_{\mathrm{A}_1,\ldots,A_{t-1}}\left[\frac{1}{p_t(a)}\mathbb{E}_{A_t}[\mathbb{I}(A_t = a)|\mathrm{A}_1, \ldots, A_{t-1}]\right]$$
$$= \ell_{t,a}\mathbb{E}_{\mathrm{A}_1,\ldots,A_{t-1}}\left[\frac{1}{p_t(a)}p_t(a)\right]$$
$$= \ell_{t,a}$$

- $\ell_{t,a} \in [0,1] \;\Rightarrow\; \tilde{\ell}_{t,a} \in \left[0, \frac{1}{p_t(a)}\right]$

Properties continued

$$\tilde{\ell}_{t,a} = \frac{\ell_{t,a}\mathbb{I}(A_t = a)}{p_t(a)}$$

- The second moment of $\tilde{\ell}_{t,a}$ is considerably smaller than the second moment of a general random variable with the same range:

$$\mathbb{E}\left[\left(\tilde{\ell}_{t,a}\right)^2\right] = \mathbb{E}\left[\left(\frac{\ell_{t,a}\mathbb{I}(A_t = a)}{p_t(a)}\right)^2\right]$$

$$= \mathbb{E}\left[\frac{\overbrace{\left(\ell_{t,a}\right)^2}^{\leq 1}\overbrace{\left(\mathbb{I}(A_t=a)\right)^2}^{=\mathbb{I}(A_t=a)}}{p_t(a)^2}\right]$$

$$\leq \mathbb{E}\left[\frac{\mathbb{I}(A_t=a)}{p_t(a)^2}\right]$$

$$= \mathbb{E}\left[\frac{1}{p_t(a)}\right]$$

- "The bandit magic":

$$\mathbb{E}\left[\sum_a p_t(a)\left(\tilde{\ell}_{t,a}\right)^2\right]$$

$$\leq \mathbb{E}\left[\sum_a p_t(a)\frac{1}{p_t(a)}\right]$$

$$= K$$

Importance weighted sampling - summary

- $\tilde{\ell}_{t,a} = \frac{\ell_{t,a}\mathbb{I}(A_t=a)}{p_t(a)}$
- Defined for all a
- Unbiased estimates of the losses: $\mathbb{E}[\tilde{\ell}_{t,a}] = \ell_{t,a}$
- Dependent
- Large range $\tilde{\ell}_{t,a} \in \left[0, \frac{1}{p_t(a)}\right]$
- Second moment proportional to the range $\mathbb{E}\left[\left(\tilde{\ell}_{t,a}\right)^2\right] \leq \mathbb{E}\left[\frac{1}{p_t(a)}\right]$
  - rather than the square of the range
- The bandit magic: $\mathbb{E}\left[\sum_a p_t(a)\left(\tilde{\ell}_{t,a}\right)^2\right] \leq K$

EXP3: Expected regret bound

- By the Hedge lemma ($\tilde{\ell}_{t,a}$ satisfy $\tilde{\ell}_{t,a} \geq 0$):

$$\sum_{t=1}^{T}\sum_{a=1}^{K} p_t(a)\tilde{\ell}_{t,a} - \min_a \tilde{L}_T(a) \leq \frac{\ln K}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{a=1}^{K} p_t(a)\left(\tilde{\ell}_{t,a}\right)^2$$

- Taking expectations on both sides:

$$\sum_{t=1}^{T}\mathbb{E}\left[\sum_{a=1}^{K} p_t(a)\ell_{t,a}\right] - \mathbb{E}\left[\min_a \tilde{L}_T(a)\right] \leq \frac{\ln K}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{a=1}^{K} p_t(a)\left(\tilde{\ell}_{t,a}\right)^2\right]$$

- $\mathbb{E}[\min[\cdot]] \leq \min \mathbb{E}[\cdot]$:

$$\sum_{t=1}^{T}\mathbb{E}\left[\sum_{a=1}^{K} p_t(a)\ell_{t,a}\right] - \min_a \underbrace{\mathbb{E}[\tilde{L}_T(a)]}_{=L_T(a)} \leq \frac{\ln K}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathbb{E}\left[\sum_{a=1}^{K} p_t(a)\left(\tilde{\ell}_{t,a}\right)^2\right]$$

Expected regret bound

$$\underbrace{\underbrace{\sum_{t=1}^{T} \underbrace{\mathbb{E}\left[\sum_{a=1}^{K} p_t(a)\ell_{t,a}\right]}_{\text{Expected loss at round } t} - \underbrace{\min_a L_T(a)}_{\text{Best in hindsight}}}_{\text{Expected loss of EXP3}} \leq \frac{\ln K}{\eta} + \frac{\eta}{2}\underbrace{\sum_{t=1}^{T} \underbrace{\mathbb{E}\left[\sum_{a=1}^{K} p_t(a)\left(\tilde{\ell}_{t,a}\right)^2\right]}_{\leq K}}_{\leq KT}}_{\text{Expected regret}}$$

- Expected regret bound:

$$\mathbb{E}[R_T] \leq \frac{\ln K}{\eta} + \frac{\eta}{2}KT$$

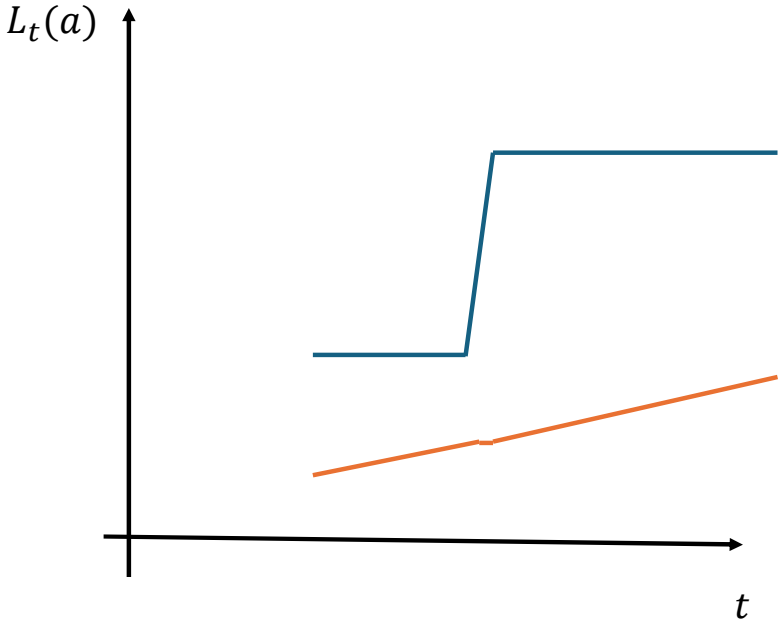
- Optimize with respect to η:
  - $\eta = \sqrt{\frac{2 \ln K}{KT}}$
  - $\mathbb{E}[R_T] \leq \sqrt{2KT \ln K}$

Algorithm's dynamics

EXP3:

- $\forall a: L_0(a) = 0$
- For $t = 1, 2, \ldots$
  - $\forall a: p_t(a) = \frac{e^{-\eta_t L_{t-1}(a)}}{\sum_{a'} e^{-\eta_t L_{t-1}(a')}}$
  - $A_t \sim p_t$
  - [Observe ℓ_{t,A_t}]
  - $\forall a: L_t(a) = L_{t-1}(a) + \frac{\ell_{t,a}\mathbb{I}(A_t=a)}{p_t(a)}$

- Algorithm's dynamics ensures exploration!



Lower bound for adversarial bandits

Game 0

$$\begin{matrix} \ell_{1,1} & \ell_{2,1} & \cdots & \ell_{t,1} & \cdots & \sim Ber(1/2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \ell_{1,a} & \ell_{2,a} & \cdots & \ell_{t,a} & \cdots & \sim Ber(1/2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \ell_{1,K} & \ell_{2,K} & \cdots & \ell_{t,K} & \cdots & \sim Ber(1/2) \end{matrix}$$

Game 1

$$\begin{matrix} \ell_{1,1} & \ell_{2,1} & \cdots & \ell_{t,1} & \cdots & \sim Ber(1/2-\varepsilon) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \ell_{1,a} & \ell_{2,a} & \cdots & \ell_{t,a} & \cdots & \sim Ber(1/2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \ell_{1,K} & \ell_{2,K} & \cdots & \ell_{t,K} & \cdots & \sim Ber(1/2) \end{matrix}$$

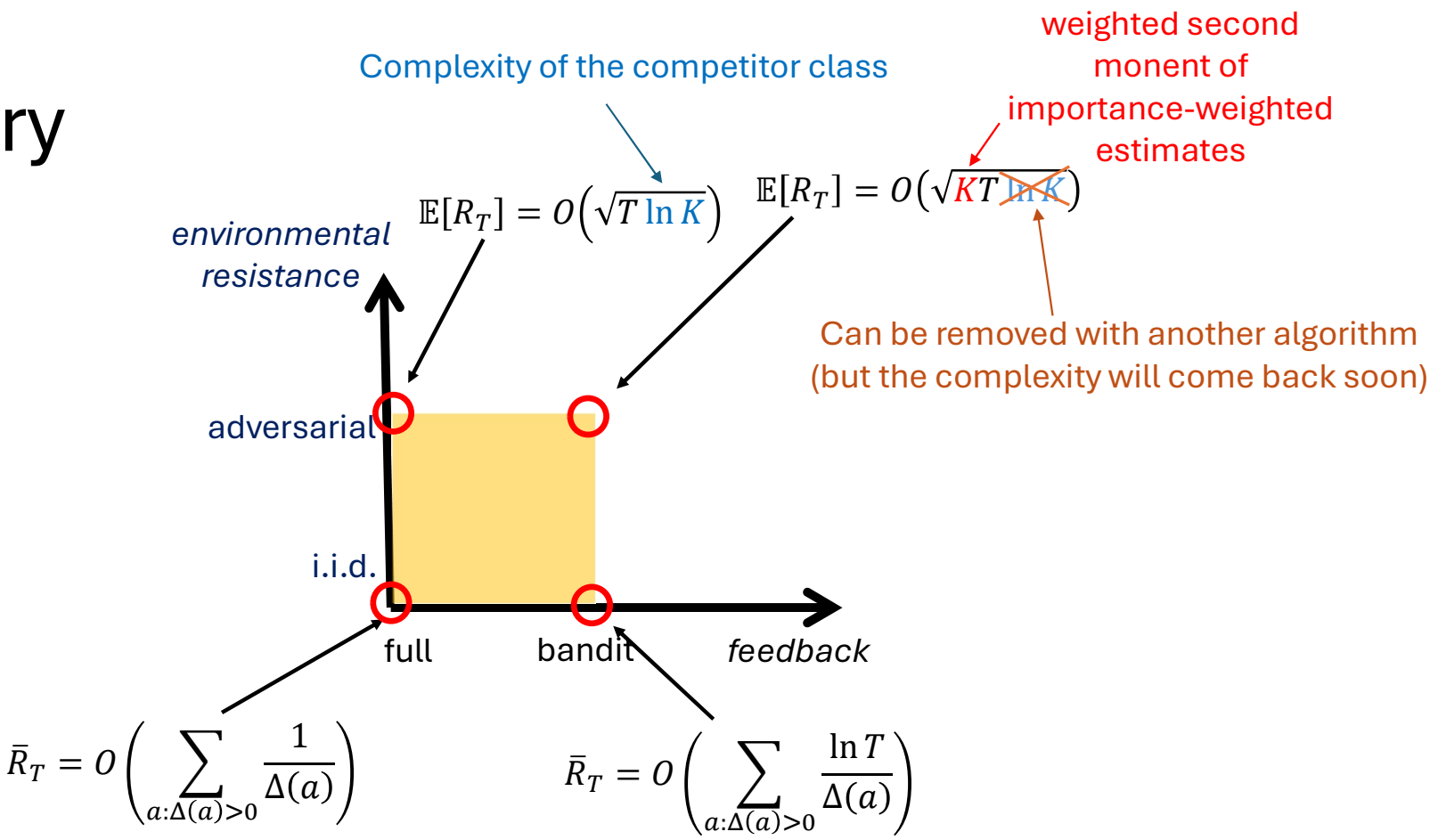
$$\vdots$$

Game K

$$\begin{matrix} \ell_{1,1} & \ell_{2,1} & \cdots & \ell_{t,1} & \cdots & \sim Ber(1/2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \ell_{1,a} & \ell_{2,a} & \cdots & \ell_{t,a} & \cdots & \sim Ber(1/2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \ell_{1,K} & \ell_{2,K} & \cdots & \ell_{t,K} & \cdots & \sim Ber(1/2-\varepsilon) \end{matrix}$$

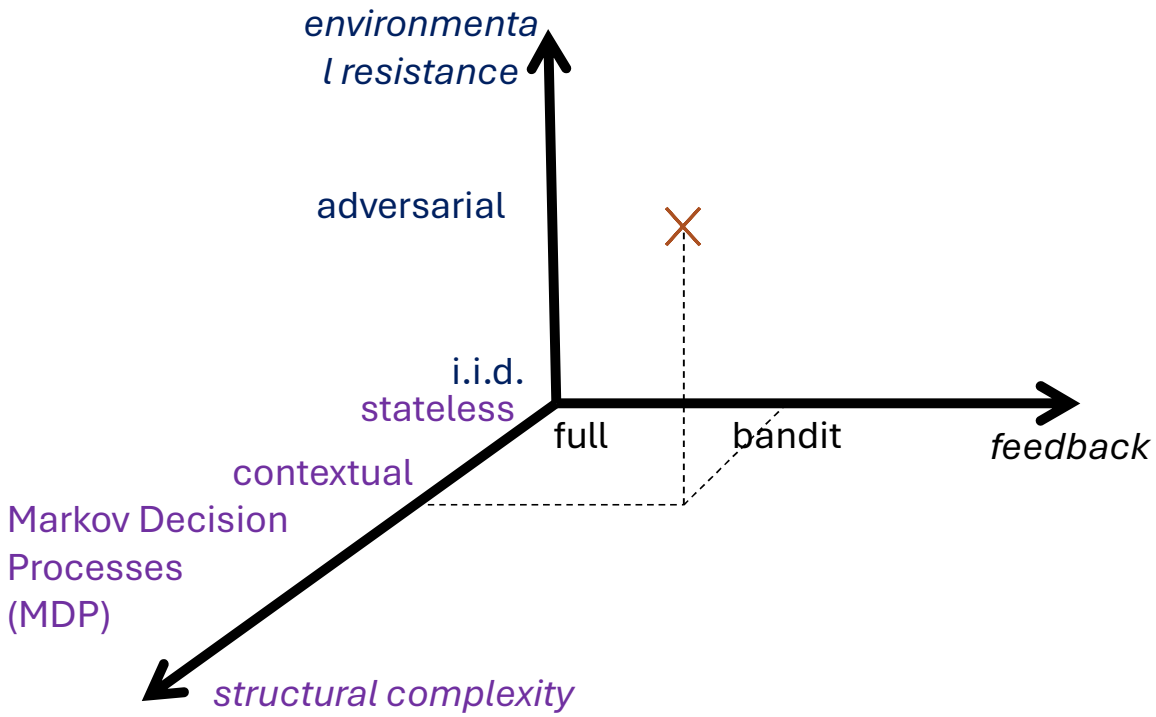
- At least one action is played at most T/K times
- For that action it is impossible to distinguish between $Ber(1/2)$ and $Ber\left(1/2-1/\sqrt{T/K}\right) = Ber\left(1/2-\sqrt{K/T}\right)$
- $\mathbb{E}[R_T] \geq T\sqrt{K/T} = \sqrt{KT}$

Summary

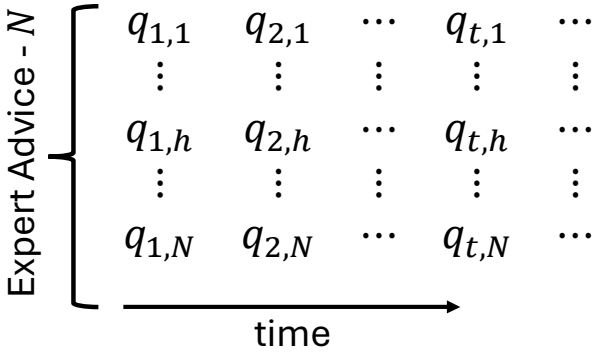

Complexity of the competitor class
weighted second monent of importance-weighted estimates
$\mathbb{E}[R_T] = O\left(\sqrt{T \ln K}\right)$
$\mathbb{E}[R_T] = O\left(\sqrt{KT \ln K}\right)$
environmental resistance
adversarial
i.i.d.
full
bandit
feedback
Can be removed with another algorithm (but the complexity will come back soon)
$\bar{R}_T = O\left(\sum_{a:\Delta(a)>0} \frac{1}{\Delta(a)}\right)$
$\bar{R}_T = O\left(\sum_{a:\Delta(a)>0} \frac{\ln T}{\Delta(a)}\right)$

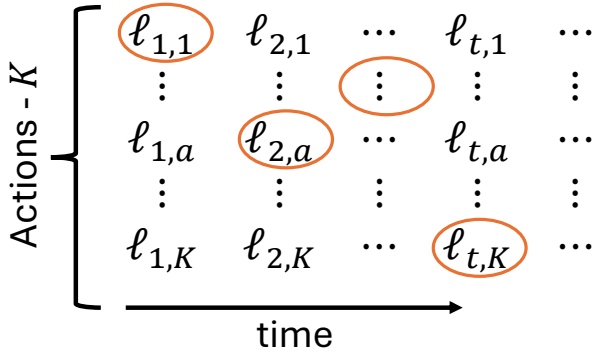
Contextual Bandits

Contextual Bandits



Version #1: Bandits with Expert Advice





Game definition:

- For $t = 1, 2, \ldots$
  - Observe advice of N experts $q_{t,1}, \ldots, q_{t,N}$
    - where $q_{t,h}$ is a distribution on actions $\{1, \ldots, K\}$
  - Play an action A_t
  - Suffer and observe ℓ_{t,A_t}

Performance measure – regret:

$$R_T = \underbrace{\sum_{t=1}^{T} \ell_{t,A_t}}_{\text{Loss of the algorithm}} - \min_h \underbrace{\sum_{t=1}^{T} \sum_a q_{t,h}(a)\ell_{t,a}}_{\text{(Expected)loss of expert } h}$$

Deterministic $q_{t,h}$ models a path through loss matrix

$\ln_*(x) = \max(1, \ln x)$

Algorithm: EXP4

(Exponential Exploration Exploitation with Expert Advice)

- $\forall h: \tilde{L}_0(h) = 0$
- For $t = 1, 2, \ldots$
  - $\forall a: p_t(a) = \sum_h \underbrace{q_{t,h}(a)}_{\text{Advice}} \underbrace{\frac{e^{-\eta_t \tilde{L}_{t-1}(h)}}{\sum_{h'} e^{-\eta_t \tilde{L}_{t-1}(h')}}}_{\text{Weight of expert } h}$
  - $A_t \sim p_t$
  - [Observe ℓ_{t,A_t}]
  - $\forall a: \tilde{\ell}_{t,a} = \frac{\ell_{t,a} \mathbb{I}(A_t = a)}{p_t(a)}$
  - $\forall h: \tilde{\ell}_{t,h} = \sum_a q_{t,h}(a)\, \tilde{\ell}_{t,a}$
  - $\forall h: \tilde{L}_t(h) = \tilde{L}_{t-1}(h) + \tilde{\ell}_{t,h}$

- EXP4 Expected regret upper bound:

$$\mathbb{E}[R_T] \leq \sqrt{2 \underbrace{K}_{\text{Price of bandit feedback}} T \underbrace{\ln N}_{\text{Complexity of the comparator class}}}$$

- A different algorithm (based on regularization achieving $O(\sqrt{KT})$ regret in stateless bandits) achieves

$$\mathbb{E}[R_T] = O\left(\sqrt{KT \ln_* \frac{N}{K}}\right)$$

- And there is a matching lower bound

Possible application

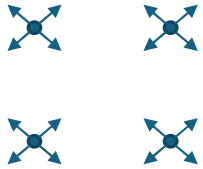
- The experts are various prediction algorithms and/or various parametrizations of the same prediction algorithm
- EXP4 can be used to weigh the predictions and achieve performance converging to performance of the best expert in the set

- No assumptions on stationarity (i.i.d.-ness) of the data stream!

Version #2: Bandits with side information

Game definition:

- For $t = 1, 2, \ldots$
  - Observe side info (state) S_t
  - Play an action A_t
  - Suffer and observe $\ell(A_t, S_t)$



Regret:

$$R_T = \underbrace{\sum_{t=1}^{T} \ell_t(A_t, S_t)}_{\text{Loss of the algorithm}} - \underbrace{\sum_{s \in \mathcal{S}} \underbrace{\min_a \sum_{t:S_t=s} \ell_t(a, s)}_{\text{Loss of the best action in hindsight in state } s}}_{\text{Total loss assuming the best action in hindsight in each state}}$$

- Regret upper bound:
  - Run the optimal $O(\sqrt{KT})$ bandit algorithm in each state
  - T_s - the number of times state s appears in the sequence

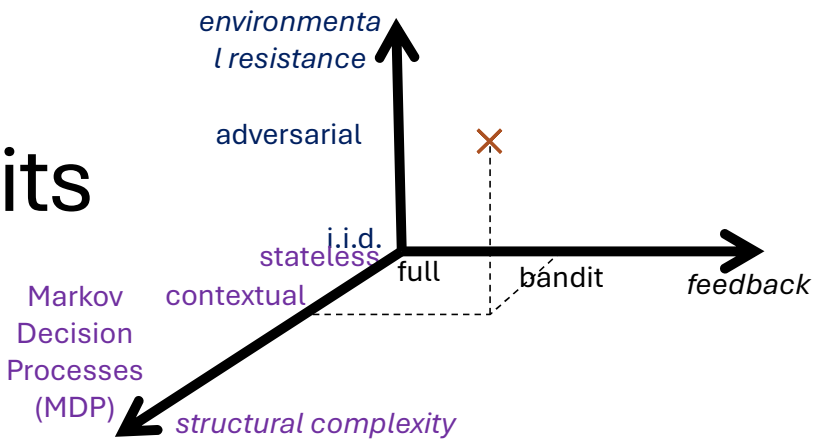
$$\mathbb{E}[R_T] \leq \sum_{s \in \mathcal{S}} O\left(\sqrt{KT_s}\right) \leq O\left(|\mathcal{S}|\sqrt{KT/|\mathcal{S}|}\right) = O\left(\sqrt{KT|\mathcal{S}|}\right)$$

- Lower bound:
  - Construct $|\mathcal{S}|$ independent worst-case bandits played $T/|\mathcal{S}|$ times each

$$\mathbb{E}[R_T] = \Omega\left(\underbrace{|\mathcal{S}|}_{\#(bandits)} \underbrace{\sqrt{KT/|\mathcal{S}|}}_{\substack{regret\ of \\ each\ bandit}}\right) = \Omega\left(\sqrt{KT|\mathcal{S}|}\right)$$

- $|\mathcal{S}|$ - structural complexity

Summary – Contextual Bandits



- Bandits with Expert Advice
  - EXP4 algorithm
    - $p_t(a) = \sum_h \underbrace{q_{t,h}(a)}_{\text{Advice}} \underbrace{\frac{e^{-\eta_t \tilde{L}_{t-1}(h)}}{\sum_{h'} e^{-\eta_t \tilde{L}_{t-1}(h')}}}_{\text{Weight of expert } h}$
    - $\mathbb{E}[R_T] \leq \sqrt{2KT \ln N}$
  - Optimal algorithm: $\mathbb{E}[R_T] = \theta\left(\sqrt{2KT \ln_* \frac{N}{K}}\right)$
  - Application: meta-selection/tuning of algorithms

- Bandits with side information
  - $\mathbb{E}[R_T] = \theta\left(\sqrt{KT|\mathcal{S}|}\right)$

Evaluation of Bandit Algorithms

Evaluation of bandit algorithms in practice

- Challenge: previously unobserved actions or (state,action) pairs

- Deployment
  - Risky and time-consuming

- Environment simulation
  - Requires a good simulator
    - This may be very hard or even impossible to produce
    - If we have a good simulator, we probably already have a solution to the problem

Evaluation of bandit algorithms in practice

- Offline evaluation for i.i.d. problems
  1. Use full information data where possible and relevant
  2. "Importance-weighting" of logged limited feedback data
     - Requires randomized sampling in the logging policy with non-zero probability for taking all the (potentially relevant) actions
     - Requires logging the sampling distribution (to do importance-weighting)
     - Variance of the estimates scales with $\frac{1}{p_{\text{logging}}(a)}$
       - So for every potentially relevant action $p_{\text{logging}}(a)$ should not be too small
     - Exercise: experiment with this approach!
- Evaluation in the adversarial regime
  - Generally impossible
  - Sparring

Alternative algorithms for bandits

Alternative algorithms for i.i.d. bandits

- UCB-style algorithms
  - kl-UCB (based on kl inequality)
  - UCB-V (based on Empirical Bernstein or Unexpected Bernstein inequality)
- Thompson sampling (Bayesian approach)
- Subsampling
- Best-of-both-worlds algorithms

Variations of EXP3 – high probability regret bound

- EXP3
  - $p_t(a) = \frac{e^{-\eta_t L_{t-1}(a)}}{\sum_{a'} e^{-\eta_t L_{t-1}(a')}}$
  - $\tilde{\ell}_{t,a} = \frac{\ell_{t,a}\mathbb{I}(A_t=a)}{p_t(a)}$
  - $\mathbb{E}[R_T] = O\left(\sqrt{KT \ln K}\right)$

- EXP3-IX: high-probability regret guarantee
  - $\tilde{\ell}_{t,a} = \frac{\ell_{t,a}\mathbb{I}(A_t=a)}{p_t(a)+\frac{\eta_t}{2}}$
  - $\mathbb{P}\left(R_T \geq O\left(\sqrt{KT \ln K} \ln \frac{1}{\delta}\right)\right) \leq \delta$

Variations of EXP3 – best-of-both-worlds

$$\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$$

$$\sum_{t=1}^{T} \frac{1}{t} \leq 1 + \ln T$$

$$\sum_{t=1}^{T} \frac{1}{t\Delta(a)^2} \leq \frac{1 + \ln T}{\Delta(a)^2}$$

- EXP3
  - $p_t(a) = \frac{e^{-\eta_t L_{t-1}(a)}}{\sum_{a'} e^{-\eta_t L_{t-1}(a')}}$
  - $\tilde{\ell}_{t,a} = \frac{\ell_{t,a} \mathbb{I}(A_t = a)}{p_t(a)}$
  - $\mathbb{E}[R_T] = O\left(\sqrt{KT \ln K}\right)$
- EXP3++: best-of-both-worlds
  - $\tilde{p}_t(a) = (1 - \sum_a \varepsilon_t(a)) p_t(a) + \varepsilon_t(a)$
  - $\varepsilon_t(a) = \theta \left(\frac{\ln}{t\, \widehat{\Delta}_t(a)^2} \right)$, where $\widehat{\Delta}_t(a)$ is a lower confidence bound on the gap
  - $\mathbb{E}[R_T] = O\left(\sqrt{KT \ln K}\right)$
  - $\bar{R}_T = O\left(\sum_{a:\Delta(a)>0} \frac{(\ln \quad)^2}{\Delta(a)} \right)$

$$\langle p, L_{t-1} \rangle = \sum_a p_a L_{t-1}(a)$$

Adversarial bandits: alternative regularization

- EXP3
  - $p_t = \frac{e^{-\eta_t L_{t-1}(a)}}{\sum_{a'} e^{-\eta_t L_{t-1}(a')}} = \arg\min_p \langle p, L_{t-1} \rangle + \underbrace{\frac{1}{\eta_t} \sum_a p_a \ln p_a}_{\text{Regularization: Negative entropy}}$

- Tsallis-INF – the ultimate algorithm: Best-of-both-worlds and minimax optimal
  - $p_t = \arg\min_p \langle p, L_{t-1} \rangle \underbrace{- \frac{1}{\eta_t} \sum_a \sqrt{p_a}}_{\text{Regularization: Tsallis entropy}}$
  - Adversarial: $\mathbb{E}[R_T] = O\left(\sqrt{KT}\right)$
  - I.I.D.: $\bar{R}_T = O\left(\sum_{a:\Delta(a)>0} \frac{\ln}{\Delta(a)}\right)$

Tsallis-INF

[Zimmert & Seldin, JMLR 2021; Masoudian & Seldin, COLT 2021; Ito, COLT 2021]

Tsallis-INF Algorithm:

- $\forall a: \hat{L}_0(a) = 0$
- For $t = 1, 2, \ldots$
  - $\forall a:\ p_t(a) = \arg\min_{p \in \Delta^{K-1}} \left(\langle \hat{L}_t, p \rangle - \frac{1}{\eta_t} \sum_{a'} \sqrt{p(a')} \right)$
  - $A_t \sim p_t$
  - $\forall a: \hat{L}_t(a) = \hat{L}_{t-1}(a) + \frac{\ell_{t,a} \mathbb{I}(A_t = a)}{p_t(a)}$

- Original analysis [Audibert & Bubeck, COLT 2009]:

$$\bar{R}_T \leq c \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \mathbb{E}\left[\sum_{a} \sqrt{p_t(a)} \right] = O\left(\sqrt{KT}\right)$$

- Observation: $p_t(a^*) = 1 - \sum_{a \neq a^*} p_t(a)$

$$\bar{R}_T \leq c \sum_{t=1}^{T} \frac{1}{\sqrt{t}} \mathbb{E}\left[\sum_{a \neq a^*} \sqrt{p_t(a)} \right] = O\left(\sqrt{KT}\right)$$

- Self-bounding analysis:

  If $\bar{R}_T = \sum_{t=1}^{T} \sum_{a} \Delta(a) \mathbb{E}[p_t(a)]$ then

$$\begin{aligned} \bar{R}_T &= 2\bar{R}_T - \sum_{t=1}^{T} \sum_{a \neq a^*} \Delta(a) \mathbb{E}[p_t(a)] \\ &\leq \sum_{t=1}^{T} \mathbb{E}\left[\sum_{a \neq a^*} \frac{2c\sqrt{p_t(a)}}{\sqrt{t}} - \Delta(a) p_t(a) \right] \\ &\leq \sum_{t=1}^{T} \sum_{a \neq a^*} \frac{c'}{t\Delta(a)} = O\left(\sum_{a \neq a^*} \frac{\ln}{\Delta(a)} \right) \end{aligned}$$

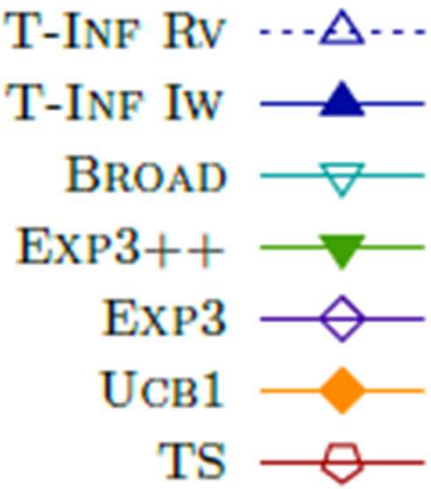
Tsallis-INF – summary

[Zimmert & Seldin, JMLR 2021; Masoudian & Seldin, COLT 2021; Ito, COLT 2021]

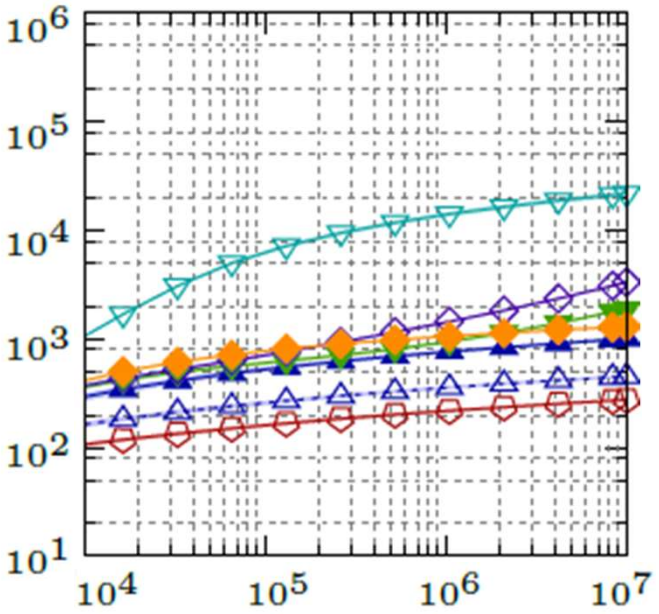
- Adversarial: $\bar{R}_T = O\left(\sqrt{KT}\right)$
- Stochastic: $\bar{R}_T = O\left(\sum_{a \neq a^*} \frac{\ln T}{\Delta(a)}\right)$
- Bonus: refined regret guarantees in adversarial regimes with (Δ, C, T) self-bounding constraints $\bar{R}_T \geq \sum_t^T \sum_a \Delta(a)\mathbb{E}[p_t(a)] - C$
  - Stochastically constrained adversarial bandits:
    - The gaps $\Delta(a)$ are fixed, but the means $\mu(a)$ can fluctuate
    - $\bar{R}_T = \sum_{t=1}^T \sum_a \Delta(a)\mathbb{E}[p_t(a)]$ (the same as in the stochastic bandits)
    - $\bar{R}_T = O\left(\sum_{a \neq a^*} \frac{\ln}{\Delta(a)}\right)$ (the same as in the stochastic bandits)
  - Stochastic bandits with adversarial corruptions: $\bar{R}_T = O\left(\sum_{a \neq a^*} \frac{\ln}{\Delta(a)} + \sqrt{C \sum_{a \neq a^*} \frac{\ln}{\Delta(a)}}\right)$

Tsallis-INF in practice

[Zimmert & Seldin, JMLR, 2021]

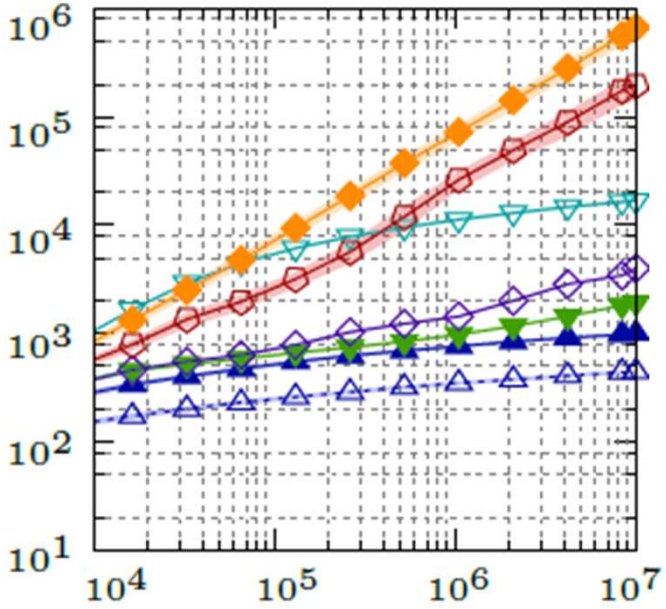


$K = 8, \Delta = 0.125$



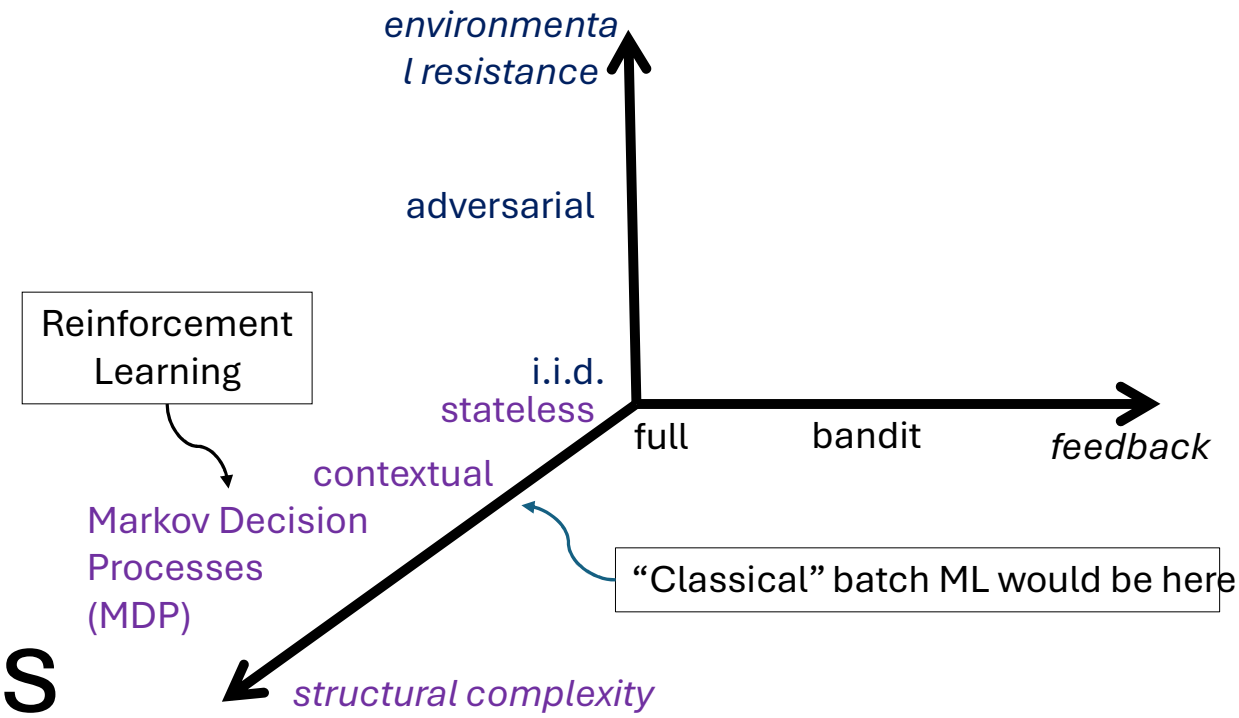
Stochastic Regime

$K = 8, \Delta = 0.125$



Stochastically Constrained Adversarial Regime

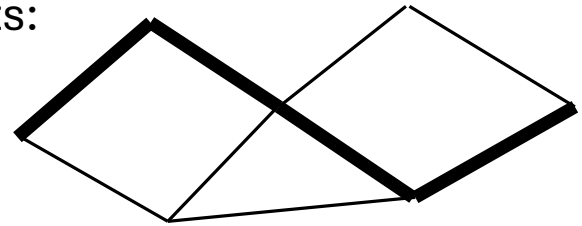
Other Settings



Structure forms: (Generalized) Linear Bandits

Linear Bandits:

- $r_t = \langle \bar{A}_t, \bar{\theta}_* \rangle + \xi_t$
- $\bar{A}_t \in \mathcal{D}_t$
- Special cases:
  - $\mathcal{D} = \{(1,0,\dots,0),\dots,(0,\dots,0,1)\}$ - multiarmed bandits
  - $\mathcal{D}_t = \{\phi(c_t, a): a \in \{1,\dots,K\}\}$ - contextual bandits
  - Combinatorial (semi-)bandits:
  - Cascading bandits



Generalized Linear Bandits:

- $r_t = f\left(\langle \bar{A}_t, \bar{\theta}_* \rangle\right) + \xi_t$

Feedback forms

- From full to limited: paid observations, decoupled exploration, graph feedback, …
- Dueling Bandits
  - Relative comparison of pairs arms, but not their true value
    - Would you like fish or chicken?
  - Very useful for implicit information collection from user feedback
- Ranking
  - Selection from a ranked list
- Partial Monitoring
  - Separation between observations and losses
  - Example: dynamic pricing

Environment forms

- Contaminated stochastic
- Stochastically constrained adversarial

Bandit variations

- Bandits with switching costs
- Recharging/recovering bandits
- Rotting bandits
- Bandits with knapsacks
- ….

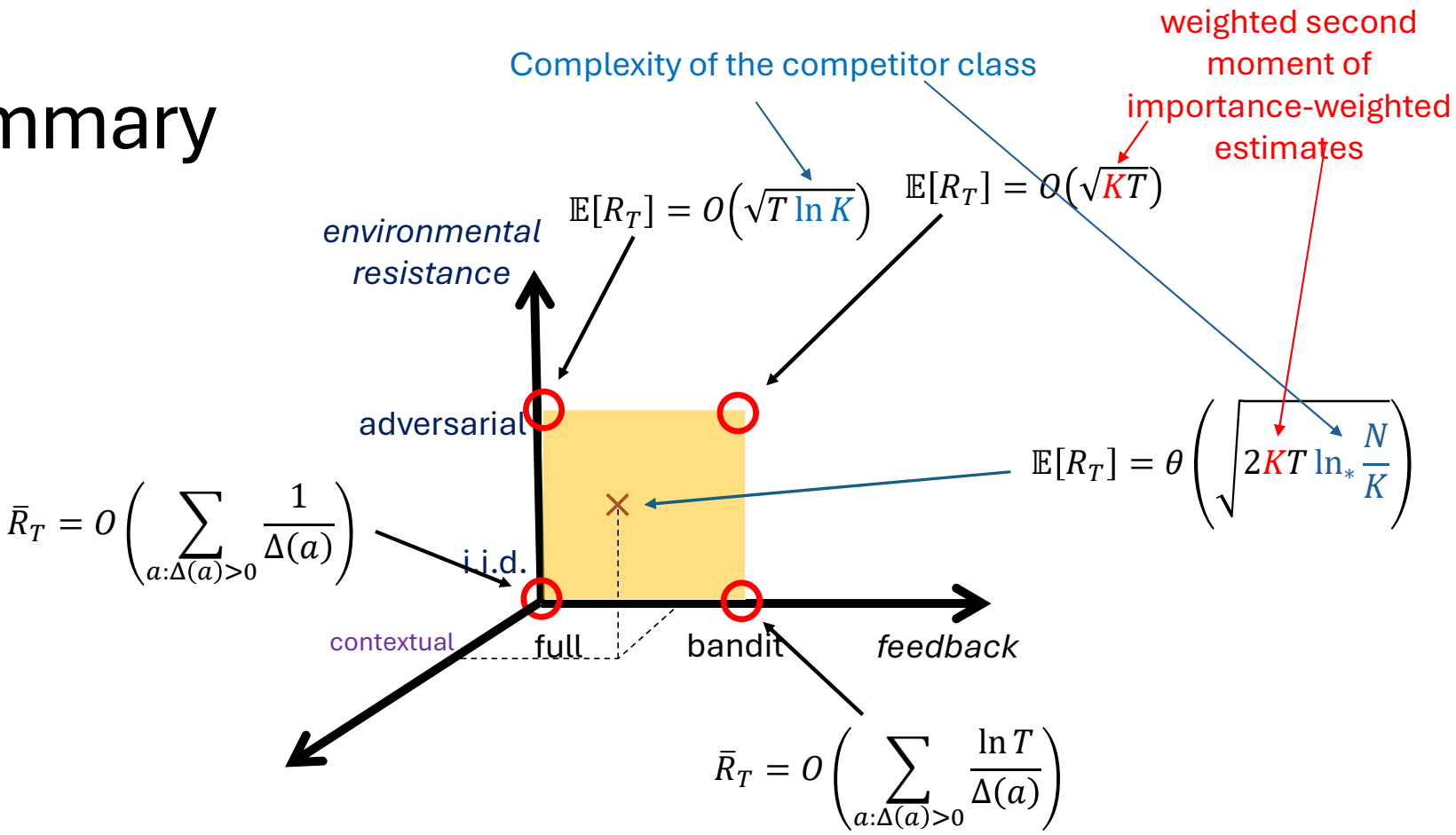
Delayed feedback

Alternative objectives

- We have studied regret minimization
  - Cumulative loss of actions along the way

- Pure Exploration / Best arm identification / Experiment design
  - Find the best action as fast as possible
  - Losses along the way are not counted

Summary


Complexity of the competitor class
weighted second moment of importance-weighted estimates
$\mathbb{E}[R_T] = O\left(\sqrt{T \ln K}\right)$
$\mathbb{E}[R_T] = O\left(\sqrt{KT}\right)$
environmental resistance
adversarial
$\mathbb{E}[R_T] = \theta\left(\sqrt{2KT \ln_* \frac{N}{K}}\right)$
$\bar{R}_T = O\left(\sum_{a:\Delta(a)>0} \frac{1}{\Delta(a)}\right)$
i.i.d.
contextual
full
bandit
feedback
$\bar{R}_T = O\left(\sum_{a:\Delta(a)>0} \frac{\ln T}{\Delta(a)}\right)$

Summary – key techniques covered

- i.i.d.
  - Optimism + confidence intervals
- Adversarial
  - Randomization + regularization
  - Analysis based on evolution of potential functions
  - Importance-weighting to cope with limited feedback
  - Best-of-both-worlds results for i.i.d. by using self-bounding!

Summary

- An infinite world of exciting problem formulations

- Further reading + exercises:
  - Yevgeny Seldin. Machine Learning – the science of selection under uncertainty. https://arxiv.org/abs/2509.21547, 2025. (Chapter 7)