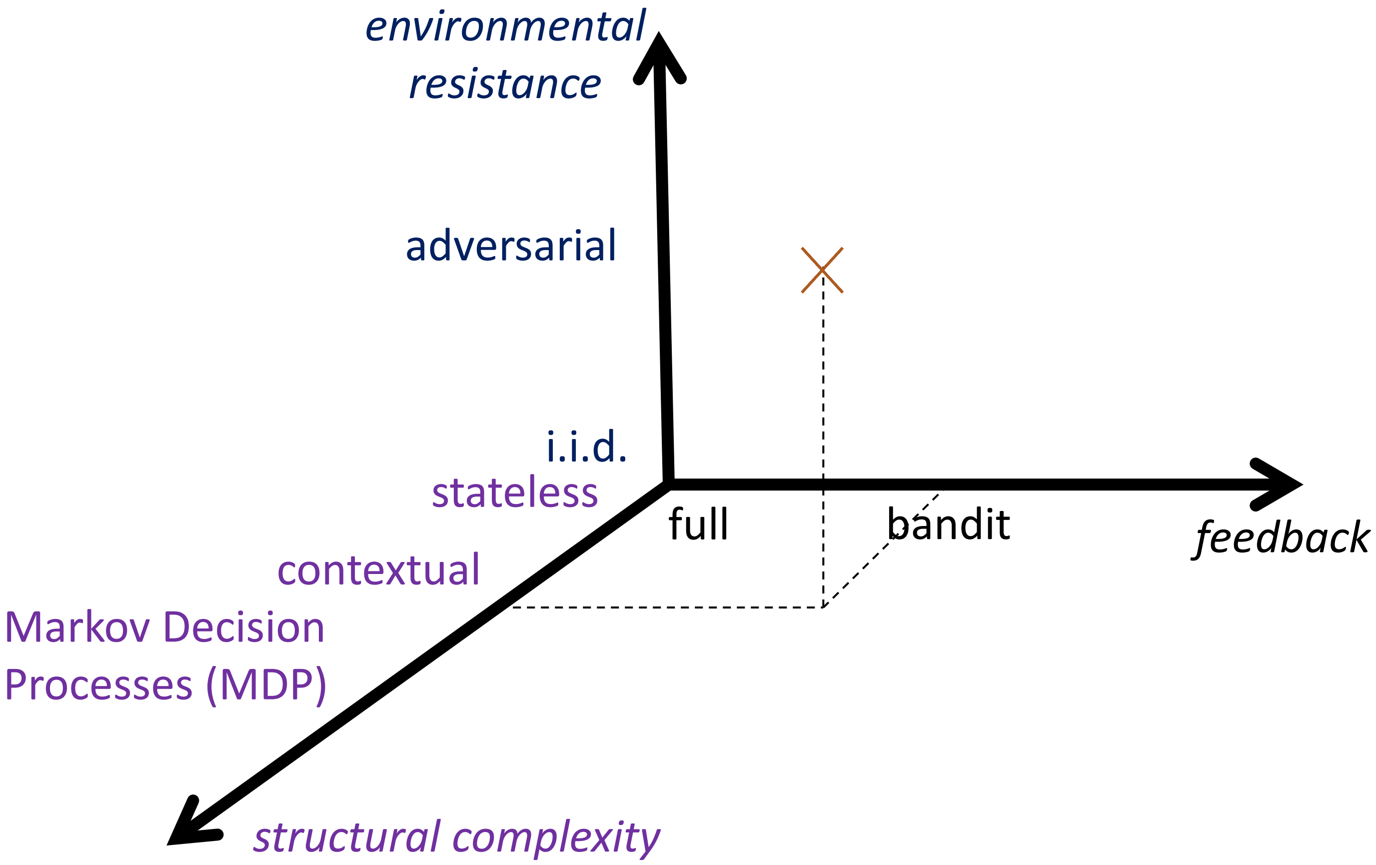
Contextual Bandits

Yevgeny Seldin

Contextual Bandits



Version #1: Bandits with Expert Advice

Expert Advice - N

$$\left\{\begin{array}{ccccc} q_{1,1} & q_{2,1} & \cdots & q_{t,1} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ q_{1,h} & q_{2,h} & \cdots & q_{t,h} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ q_{1,N} & q_{2,N} & \cdots & q_{t,N} & \cdots \end{array}\right.$$

time

Actions - K

$$\left\{\begin{array}{ccccc} \ell_{1,1} & \ell_{2,1} & \cdots & \ell_{t,1} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \ell_{1,a} & \ell_{2,a} & \cdots & \ell_{t,a} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \ell_{1,K} & \ell_{2,K} & \cdots & \ell_{t,K} & \cdots \end{array}\right.$$

time

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

Algorithm: EXP4

(Exponential Exploration Exploitation with Expert Advice)

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

- EXP4 Expected regret upper bound:

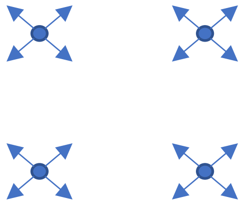
$$\mathbb{E}[R_T] \le \sqrt{2\underbrace{K}_{\text{Price of bandit feedback}}T\underbrace{\ln N}_{\text{Size of the comparator class}}}$$

EXP4: $\mathbb{E}[R_T] \le \sqrt{2KT \ln N}$

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

Reduction (assuming expert advice is deterministic – each expert recommends just 1 action):

- #2→#1: Experts → all possible mappings $h: \mathcal{S} \to \{1, \ldots, K\}$
  - $N = K^{|\mathcal{S}|}$

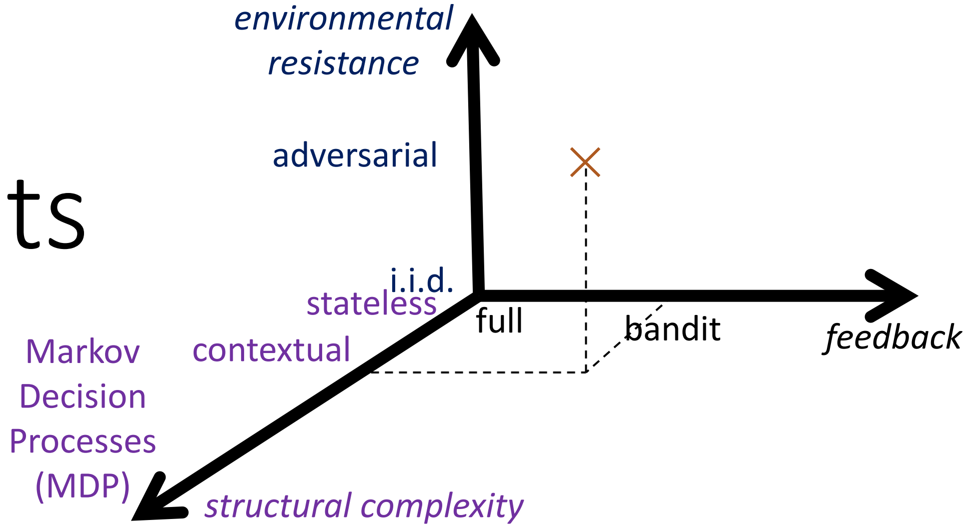
Regret bounds:

- Running EXP4:
$$\mathbb{E}[R_T] = O\left(\sqrt{KT|\mathcal{S}| \ln K}\right)$$

- Lower bound in a nutshell:
  - Generate $|\mathcal{S}|$ independent bandit problems
  - Take $N = K^{|\mathcal{S}|}$ experts – all possible ways of assigning best actions to states
  - Reminder: Lower bound for a single bandit is $\Omega(\sqrt{KT})$
  - Each bandit is played $T/|\mathcal{S}|$ times, so its regret is $\Omega\left(\sqrt{K\frac{T}{|\mathcal{S}|}}\right)$
  - Total regret $\mathbb{E}[R_T] = \Omega\left(\underbrace{|\mathcal{S}|}_{\#(bandits)} \underbrace{\sqrt{K\frac{T}{|\mathcal{S}|}}}_{regret\ of\ each\ bandit}\right) = \Omega\left(\sqrt{KT|\mathcal{S}|}\right)$

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

Summary – Contextual Bandits



- Bandits with Expert Advice
  - EXP4 algorithm
    - $p_t(a) = \sum_h \underbrace{q_{t,h}(a)}_{\text{Advice}} \underbrace{\frac{e^{-\eta_t \tilde{L}_{t-1}(h)}}{\sum_{h'} e^{-\eta_t \tilde{L}_{t-1}(h')}}}_{\text{Weight of expert } h}$
  - $\mathbb{E}[R_T] \leq \sqrt{2KT \ln N}$

- Bandits with side information
  - Reduction to prediction with expert advice
  - $\mathbb{E}[R_T] = O\left(\sqrt{KT|\mathcal{S}| \ln K}\right)$
  - $\mathbb{E}[R_T] = \Omega\left(\sqrt{KT|\mathcal{S}|}\right)$