

Online Markov Decision Processes under Bandit Feedback

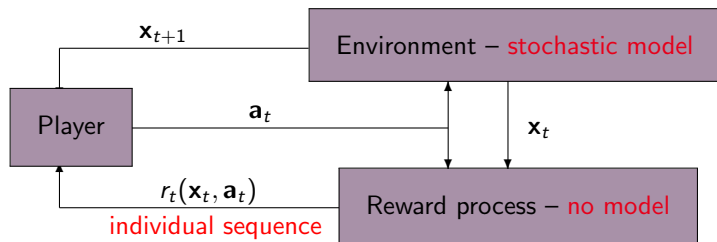
Gergely Neu^{1,2} András György² Csaba Szepesvári³
András Antos²

¹Department of Computer Science and Information Theory, Budapest University of Technology and Economics, Hungary

²Machine Learning Research Group, MTA SZTAKI Institute for Computer Science and Control, Hungary

³Department of Computing Science, University of Alberta, Canada

Online Markov Decision Processes



- ▶ **Goal:** minimize regret relative to the best fixed policy

$$\hat{L}_T = \max_{\pi} R_T^{\pi} - \hat{R}_T = \max_{\pi} \mathbb{E} \left[\sum_{t=1}^T r_t(\mathbf{x}'_t, \mathbf{a}'_t) \right] - \mathbb{E} \left[\sum_{t=1}^T r_t(\mathbf{x}_t, \mathbf{a}_t) \right].$$

- ▶ **Earlier work:**

- ▶ Full information: $\hat{L}_T = O(\sqrt{T})$ (Even-dar et al., 2005).
- ▶ Bandit information: $\hat{L}_T = o(T)$ (Yu et al., 2009).
- ▶ Bandit information for episodic loop-free MDPs: $\hat{L}_T = O(\sqrt{T})$ (Neu et al., 2010).

Online learning with bandit information: the algorithm

- ▶ Define unbiased estimates of rewards

$$\hat{\mathbf{r}}_t(x, a) = \begin{cases} \frac{r_t(x, a)}{\mathbf{p}_t^N(x, a | \mathbf{x}_{t-N}, \mathbf{a}_{t-N})} & \text{if } (x, a) = (\mathbf{x}_t, \mathbf{a}_t) \\ 0 & \text{otherwise,} \end{cases}$$

where

$$\mathbf{p}_t^N(x, a | \mathbf{x}_{t-N}, \mathbf{a}_{t-N}) = \mathbb{P}[\mathbf{x}_t = x, \mathbf{a}_t = a | \mathbf{x}_{1:t-N}, \mathbf{a}_{1:t-N}].$$

- ▶ Let $\hat{\rho}_t = \mathbb{E}[\hat{\mathbf{r}}_t(x, a) | x \sim \mu^{\pi_t}, a \sim \pi_t]$ and $\hat{\mathbf{q}}_t$ be the solution to the Bellman equations

$$\hat{\mathbf{q}}_t(x, a) = \hat{\mathbf{r}}_t(x, a) - \hat{\rho}_t + \sum_{x', a'} P(x' | x, a) \pi_t(a' | x') \hat{\mathbf{q}}_t(x', a').$$

- ▶ Feed an instance of **Exp3** with the computed values of $\hat{\mathbf{q}}_t(x, a)$ in each state x .

Online learning in MDPs with bandit information

► **Assume:**

- General MDP with cycles.
- Every policy π induces a stationary distribution μ^π over the states.
- Every policy mixes fast (with mixing time τ).
- $\mu^\pi(x) \geq \alpha > 0$ for all π .

► **Result:** sublinear regret relative to the best fixed policy

$$\hat{L}_T = \mathcal{O} \left(\tau T^{2/3} \left(\frac{|\mathcal{A}| \log |\mathcal{A}| \log T}{\alpha} \right)^{1/3} \right).$$

► **Proof idea:**

$$\hat{L}_T = \underbrace{\left(R_T^\pi - \sum_{t=1}^T \rho_t^\pi \right)}_{\leq 2\tau+2} + \underbrace{\left(\sum_{t=1}^T \rho_t^\pi - \sum_{t=1}^T \rho_t^{\pi^*} \right)}_{\mathcal{O}(T^{2/3})} + \underbrace{\left(\sum_{t=1}^T \rho_t^{\pi^*} - \hat{R}_T \right)}_{\mathcal{O}(T^{2/3})}$$

See you at poster 95!