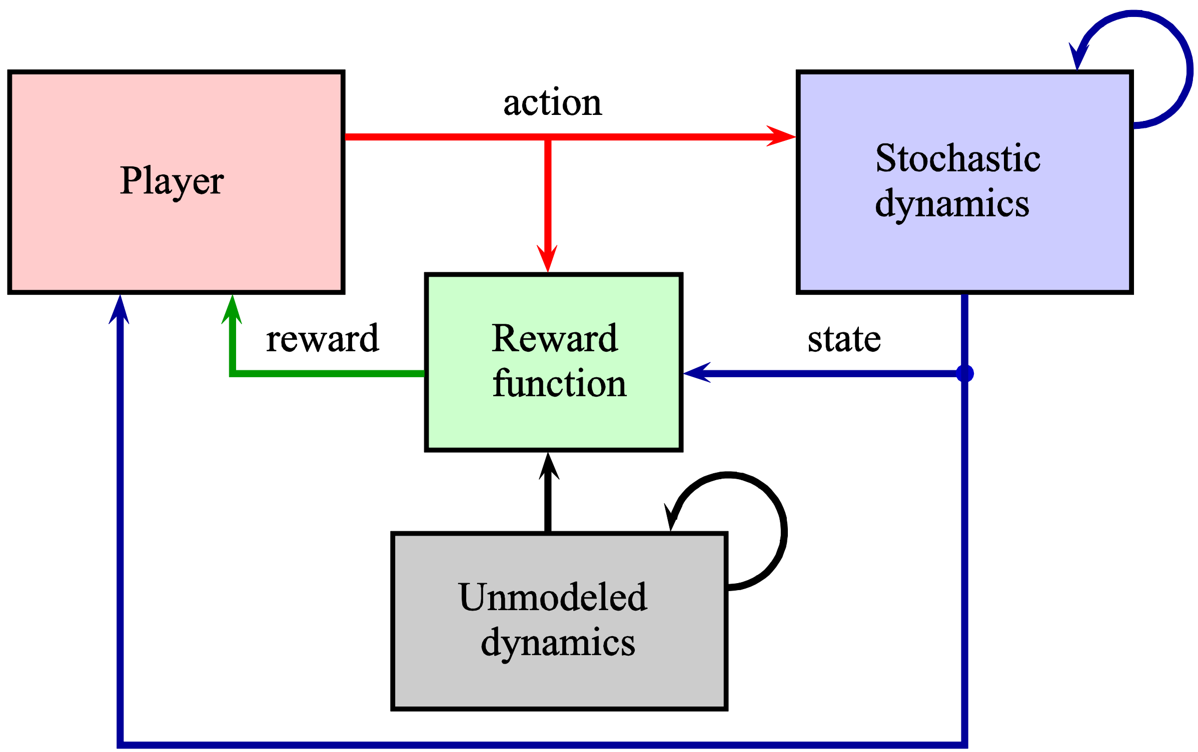
# Online Learning in Non-Stationary Markov Decision Processes


Gergely Neu
gergely.neu@gmail.com
MTA SZTAKI, Hungary

Csaba Szepesvári
szepesva@ualberta.ca
University of Alberta, Canada

András György
gyorgy@ualberta.ca
University of Alberta, Canada

# The learning problem

Goal: minimize total expected regret:

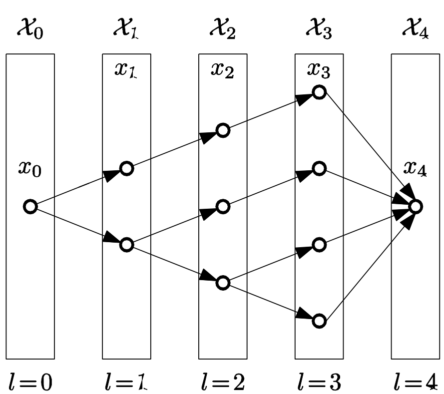
$$\hat{L}_T = \max_{\pi} E\left[R_T^{\pi} - \hat{R}_T\right] \rightarrow \min$$

Total reward of policy $\pi$:

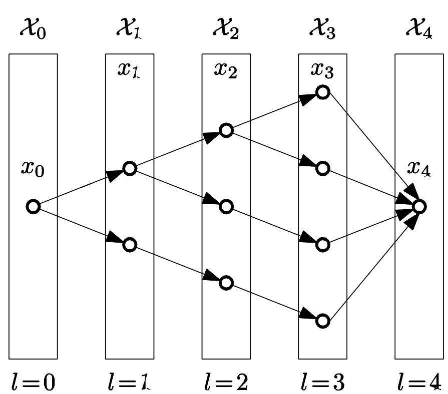
$$R_T^{\pi} = \sum_{t=1}^{T} r_t(x_t', \pi(x_t'))$$

Total reward of agent:

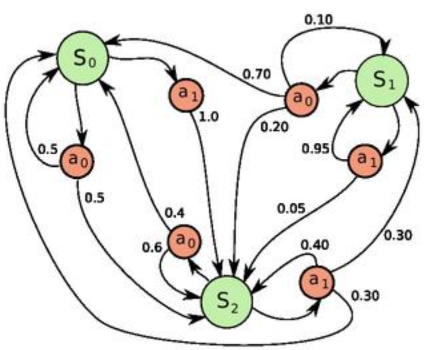
$$\hat{R}_T = \sum_{t=1}^{T} r_t(x_t, a_t)$$




$a_1$




$a_2$



# Results

| | Whole $r_t$ observed | Only $r_t(x_t, a_t)$ observed |
|---|---|---|
| $P$ **known** | • Even-Dar et al. (2006,2009)<br>  • Unichain MDP<br>  • $\hat{L}_T = O\left(\tau^2\sqrt{T \log\lvert A\rvert}\right)$ | • Neu et al. (2010a, 2013a)<br>  • Loop-free SSP<br>  • $\hat{L}_T = O\left(L^2\sqrt{T\lvert A\rvert/\alpha}\right)$<br>• Neu et al. (2010b ,2013b)<br>  • Unichain MDP<br>  • $\hat{L}_T = O\left(\tau^{3/2}\sqrt{T\lvert A\rvert/\alpha'}\right)$ |
| $P$ **unknown** | • Jaksch et al. (2010)<br>  • **Stochastic** rewards<br>  • Connected MDP<br>  • $\hat{L}_T = \tilde{O}\left(D\lvert X\rvert\sqrt{T\lvert A\rvert}\right)$<br>• Neu et al. (2012)<br>  • Loop-free SSP<br>  • $\hat{L}_T = \tilde{O}\left(L\lvert X\rvert\lvert A\rvert\sqrt{T}\right)$ |  |