Optimistic Linear Programming gives Logarithmic Regret for Irreducible MDPs

Anonymous Author(s)
Affiliation
Address
email

Appendix

Proof of Proposition 2. Fix $(i, a) \in \mathrm{Crit}(P)$. Drop dependence of ϕ^*, K, $\mathrm{MakeOpt}$, H^* on i, a, P for readability. Also, let $J(\epsilon) = J_{i,a}(p_i(a); P, \epsilon)$. Since $|\mathrm{Crit}(P)| \leq |S||A|$, it suffices to show that $\phi^*/K \leq (H^*)^2/\phi^*$. Let $\epsilon < \phi^*$ be arbitrary. By definition of $\mathrm{MakeOpt}(\epsilon)$ and ϕ^*, we have, for all $q \in \mathrm{MakeOpt}(\epsilon)$, $\langle q, h^* \rangle - \langle p_i(a), h^* \rangle \geq \phi^* - \epsilon$. This implies $\|p_i(a) - q\|_1 \geq (\phi^* - \epsilon)/H^*$ since $H^* = \|h^*\|_\infty$. Thus, by definition of $J(\epsilon)$, we have

$$J(\epsilon) \geq \frac{(\phi^* - \epsilon)^2}{(H^*)^2} \qquad \Rightarrow \qquad \lim_{\epsilon \to 0} J(\epsilon) \geq \frac{(\phi^*)^2}{(H^*)^2} \, .$$

By definition, the left hand side is K. Thus, $\phi^*/K \leq (H^*)^2/\phi^*$. □

Proof of Proposition 3. Fix $\mu \in \mathcal{O}(P, \mathcal{A})$. Drop the dependence of $h^*, H^*, \lambda^*, T_\mu$ on P. It suffices to prove that $H^* \leq T_\mu$ for the result then follows from Proposition 2 and definition of $T(P)$. Since rewards and hence the gain λ^* are in $[0, 1]$, and λ^*, h^* satisfy, for all $i \in S$,

$$\lambda^* + h^*(i) = r(i, \mu(i)) + \langle p_i(\mu(i)), h^* \rangle \, , \tag{1}$$

we have, for all $i \in S$,

$$\langle p_i(\mu(i)), h^* \rangle \geq h^*(i) - 1 \, . \tag{2}$$

Start the policy μ in state j and define the random variables $Y_t := h(s_t)$. Clearly $Y_0 = h(j)$. Fix a state $i \neq j$. Define the stopping time

$$\tau := \min\{t > 0 \, : \, s_t = i\} \, .$$

Because of (2), we have

$$\mathbb{E}_j^{\mu,P}[Y_{t+1} \,|\, Y_t] \geq Y_t - 1 \, .$$

Adding $t+1$ to both sides we see that $Y_t + t$ is a submartingale and hence using the optional stopping theorem (see, for example, [1, p. 489]), we have

$$\mathbb{E}_j^{\mu,P}[Y_\tau + \tau] \geq Y_0 \, .$$

By definition of T_μ, we have

$$\mathbb{E}_j^{\mu,P}[\tau] \leq T_\mu \, .$$

Thus, noting that $Y_\tau = h^*(i)$ and $Y_0 = h^*(j)$, we have

$$h^*(i) + T_\mu \geq h^*(j) \, .$$

But this is true for all $i \neq j$. Also, there is some $i^* \in S$ such that $h^*(i^*) = 0$. Therefore, $H^* = \|h^*\|_\infty \leq T_\mu$. □

References

[1] Grimmett, G.R. & Stirzaker, D.R. (2001) *Probability and Random Processes*. Oxford: Oxford University Press, third edition.