Correction des travaux dirigés d'
Apprentissage par renforcement
Université Paris–Saclay

Joon Kwon

lundi 25 novembre 2024

Exercice 1 (*Bornes d'erreur et critère d'arrêt*). —

1) On écrit
$$v - v_\pi = v - \mathrm{B}_\pi v_\pi = v - \mathrm{B}_\pi v + \mathrm{B}_\pi v - \mathrm{B}_\pi v_\pi.$$
Puis, en prenant la norme ℓ^∞, et en utilisant l'inégalité triangulaire, et le fait que v_π est l'unique point fixe de l'opérateur B_π qui est une γ-contraction, on obtient
$$\|v - v_\pi\|_\infty \leqslant \|v - \mathrm{B}_\pi v\|_\infty + \gamma \|v - v_\pi\|_\infty ,$$
d'où le résultat. On procède de même pour les fonctions état-valeur, et pour les fonctions valeur optimales.

2) Pour une itération valeur $v_{k+1} = \mathrm{B}_\pi v_k$, on peut vérifier le critère d'arrêt
$$\|v_{k+1} - \mathrm{B}_\pi v_k\|_\infty \leqslant \varepsilon(1-\gamma),$$
ce qui implique, d'après ce qui précède, que $\|v_k - v_\pi\|_\infty \leqslant \varepsilon$.

3) On considère une itération action-valeur pour le contrôle, autrement dit $q_{k+1} = \mathrm{B}_* q_k$. On pose
$$\varepsilon = \frac{1}{3} \cdot \min_{s \in \mathcal{S}} \min_{\substack{a,a' \in \mathcal{A} \\ q_*(s,a) \neq q_*(s,a')}} |q_*(s,a) - q_*(s,a')| .$$

Puisque $q_k \to q_*$, il existe un rang k_0 à partir duquel $\|q_k - q_*\|_\infty \leqslant \varepsilon$. Soit $k \geqslant k_0$ et $\pi \in \Pi_g[q_k]$. Montrons que $\pi \in \Pi_g[q_*]$, cela montrera qu'il s'agit d'une politique optimale. Soit $s \in \mathscr{S}$. Montrons que $\pi(s) \in \operatorname{Arg\,min}_{a \in \mathscr{A}} q_*(s, a)$, autrement dit que pour tout $a \in \mathscr{A}$,

$$q_*(s, \pi(s)) \leqslant q_*(s, a).$$

Soit $a \in \mathscr{A}$. On a

$$q_*(s, \pi(s)) \leqslant q_k(s, \pi(s)) + \varepsilon \leqslant q_k(s, a) + \varepsilon \leqslant q_*(s, a) + 2\varepsilon,$$

où on a utilisé la définition de π pour la deuxième inégalité. Si $q_*(s, \pi(s)) \geqslant q_*(s, a)$, on a

$$0 \leqslant q_*(s, \pi(s)) - q_*(s, a) \leqslant 2\varepsilon < \min_{\substack{a', a'' \in \mathscr{A} \\ q_*(s,a') \neq q_*(s,a'')}} |q_*(s, a') - q_*(s, a'')|,$$

et donc nécessairement $q_*(s, \pi(s)) = q_*(s, a)$. Dans tous les cas, on a

$$q_*(s, \pi(s)) \leqslant q_*(s, a),$$

et finalement que $\pi \in \Pi_g[q_*]$. On a bien obtenu une politique optimale au bout d'un nombre fini d'itérations.

EXERCICE 2 (*Convergence en temps fini de l'itération de politiques*). — L'itération de politiques produit une suite $(\pi^{(k)})_{k \geqslant 0}$ de politiques stationnaires et déterministes. Le nombre de telles politiques est $|\mathscr{A}|^{|\mathscr{S}|}$.

Notons $v_k = v_{\pi_k}$ pour tout $k \geqslant 0$. Par propriété d'amélioration gloutonne de politique, on a $v_{k+1} \geqslant v_k$ pour tout $k \geqslant 0$. Soit $k_0 < |\mathscr{A}|^{|\mathscr{S}|}$ le plus petit entier $k \geqslant 0$ tel que $v_{k+1} = v_k$. Un tel entier existe car sinon les fonctions valeur v_k seraient toutes différentes pour $0 \leqslant k \leqslant |\mathscr{A}|^{|\mathscr{S}|}$, ce qui impliquerait que les politiques $(\pi^{(k)})_{0 \leqslant k \leqslant |\mathscr{A}|^{|\mathscr{S}|}}$ seraient toutes différentes, ce qui est impossible.

Puisque $v_{k_0+1} = v_{k_0}$, et qu'une fonction valeur inchangée après amélioration gloutonne implique l'optimalité, on a $v_{k_0} = v_*$. De plus, pour tout $k \geqslant k_0 + 1$, par définition de v_*,

$$v_* \geqslant v_k \geqslant v_{k_0} = v_*,$$

donc $v_k = v_*$ et $\pi^{(k)}$ est optimale.

Exercice 3 (*Amélioration gloutonne par rapport à plusieurs politiques*). — Pour tout $1 \leqslant m \leqslant \mathrm{M}$, on écrit

$$v_{\pi_m} = \mathrm{B}_{\pi_m} v_{\pi_m} \leqslant \max_{\pi' \in \Pi_0} \mathrm{B}_{\pi'} v_{\pi_m}.$$

En prenant le maximum sur $1 \leqslant m \leqslant \mathrm{M}$, on obtient

$$\begin{aligned}\left(\max_{1\leqslant m\leqslant \mathrm{M}} v_{\pi_m}\right) \leqslant \max_{1\leqslant m\leqslant \mathrm{M}} \max_{\pi' \in \Pi_0} \mathrm{B}_{\pi'} v_{\pi_m} &= \max_{\pi' \in \Pi_0} \mathrm{B}_{\pi'} \left(\max_{1\leqslant m\leqslant \mathrm{M}} v_{\pi_m}\right) \\ &= \mathrm{B}_* \left(\max_{1\leqslant m\leqslant \mathrm{M}} v_{\pi_m}\right) = \mathrm{B}_\pi \left(\max_{1\leqslant m\leqslant \mathrm{M}} v_{\pi_m}\right),\end{aligned}$$

où on a utilisé la définition de π. En appliquant l'opérateur B_π qui est monotone, on a

$$\mathrm{B}_\pi \left(\max_{1\leqslant m\leqslant \mathrm{M}} v_{\pi_m}\right) \leqslant \mathrm{B}_\pi^2 \left(\max_{1\leqslant m\leqslant \mathrm{M}} v_{\pi_m}\right).$$

Par récurrence immédiate, on obtient que pour tout $k \geqslant 1$,

$$\max_{1\leqslant m\leqslant \mathrm{M}} v_{\pi_m} \leqslant \mathrm{B}_\pi^k \left(\max_{1\leqslant m\leqslant \mathrm{M}} v_{\pi_m}\right) \xrightarrow[k\to+\infty]{} v_\pi.$$

D'où le résultat. On en déduit immédiatement l'inégalité correspondante pour les fonction action-valeur en appliquant l'opérateur monotone D.