Faut-il minimiser le résidu de Bellman ou maximiser la valeur moyenne?

Matthieu Geist[1,2,3], Bilal Piot[4 †], Olivier Pietquin[4 †]

[1] Université de Lorraine, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France.

[2] CNRS, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France.

[3] LORIA, CentraleSupélec, Université Paris-Saclay, 57070 Metz, France.

[4] Univ. Lille, CNRS, Centrale Lille, Inria UMR 9189 - CRIStAL, F-59000 Lille, France.

Résumé : Cet article a pour objectif la comparaison tant théorique qu'empirique de deux critères d'optimisation classiques en apprentissage par renforcement : (i) la maximisation de la valeur moyenne et (ii) la minimisation du résidu de Bellman. Pour cela, nous nous plaçons dans le cadre de la recherche directe dans un espace de politiques, le cadre naturel pour la maximisation de la valeur moyenne, et nous proposons une méthode minimisant le résidu $\|T_* v_\pi - v_\pi\|_{1,\nu}$ sur un espace de politiques. Une analyse théorique montre que cette approche bénéficie d'une borne de performance meilleure que la seule connue pour la maximisation de la valeur moyenne, et également meilleure que les bornes de programmation dynamique approchée, respectivement en termes de concentrabilité et d'horizon impliqués. Toutefois, des expériences sur des processus décisionnels de Markov générés aléatoirement, conçues pour étudier l'influence du coefficient de concentrabilité, montrent que le résidu de Bellman est généralement un mauvais substitut à l'optimisation de politique. Comparativement, maximiser la valeur moyenne semble insensible à ce problème. Ces résultats suggèrent que bien que la minimisation du résidu de Bellman permette d'obtenir de bonnes bornes de performance, maximiser directement la valeur moyenne est plus susceptible de produire des algorithmes d'apprentissage par renforcement robustes et efficaces, malgré le manque de compréhension théorique actuelle.

L'article complet, en anglais, est disponible en ligne sur arXiv (Geist *et al.*, 2016).

Remerciements

Matthieu Geist remercie le programme européen FEDER INTERREG VA (projet GRONE) et la Région Grand-Est pour leur soutien financier.

Références

GEIST M., PIOT B. & PIETQUIN O. (2016). Should one minimize the bellman residual or maximize the mean value ? *arXiv preprint arXiv :1606.07636*.

†. Bilal Piot et Olivier Pietquin sont actuellement chez Deepmind, Londres.