Action-Gap Phenomenon in Reinforcement Learning

Amir-massoud Farahmand

academic.SoloGen.net
Easy choice! Even if we don’t know the exact quality (value) of each choice (action)
Not a big deal if we choose the wrong one!
• **Setup:** Finite-action discounted MDP with general state space.

• **Question:** An estimate \( \hat{Q} \) of the optimal \( Q^* \) is given. What is the performance loss of following the greedy policy with respect to \( \hat{Q} \) (i.e., \( \| Q^* - \hat{Q}^\pi(\cdot; \hat{Q}) \|_{1, \rho} \))?

• **Answer – Part I:** It depends on the distribution of the action-gap function \( g_{Q^*}(x) \triangleq |Q^*(x, 1) - Q^*(x, 2)| \) (action-gap regularity).

• **Answer – Part II:** Favourable action-gap regularity implies faster convergence rate.

  – Simplified result:

    \[
    \| Q^* - \hat{Q}^\pi(\cdot; \hat{Q}) \|_\infty \leq c \| \hat{Q} - Q^* \|_\infty^{1+\zeta} \quad (\zeta \geq 0)
    \]

  – Interesting similarity with the low-noise (or margin) condition in classification problems.