UCB Momentum Q-learning: Correcting the bias without forgetting

Pierre Ménard[1], Omar Darwiche Domingues[2], Xuedong Shang[2,3] Michal Valko[2,4]

[1]Otto von Guericke University Magdeburg, [2]Inria Lille, [3]Université de Lille, [4]DeepMind

Overview

- Analyze the benefits of adding a momentum term to Q-learning in the episodic setting.
- `UCBMQ` algorithm with regret bound that scales *linearly* with the number of states S.

Setting

- Tabular MDP: H horizon, S states, A actions, $p_h(s'|s,a)$ unknown transitions, deterministic reward $r_h(s,a)$.
- Regret: $\sum_{t=1}^{T} V_1^\star(s_1) - V_1^{\pi^t}(s_1)$

Intuition

- Generic Q-learning step:

$$Q_h^n(s,a) = \alpha_n(r_h + p_h^n \overline{V}_h^{n-1})(s,a) + (1-\alpha_n)Q_h^{n-1}(s,a)$$

$$\overline{Q}_h^n(s,a) = Q_h^n(s,a) + b_h^n(s,a) \qquad \overline{V}_h^n(s) = \max_a \overline{Q}_h^n(s,a)$$

where the sample expectation $(p_h^n f)(s,a) = f(s_{h+1}^n)$

- *How to choose the learning rate α_n and the bonus b_h^n?*
- *learning rate* $\alpha_n \approx 1/n$, unfolding the formula for Q_h^n + Hoeffding inequality

$$Q_h^n(s,a) \approx r_h(s,a) + \frac{1}{n}\sum_{i=1}^{n} p_h^i \overline{V}_{h+1}^{i-1}(s,a)$$

$$\approx r_h(s,a) + p_h \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} \overline{V}_{h+1}^{i-1}\right)}_{:=V_{h,s,a}^n \text{ bias-value function}}(s,a) \pm \underbrace{\sqrt{\frac{H^2}{n}}}_{\text{variance term}\to\text{bonus}}$$

- *learning rate* $\alpha_n \approx H/n$ *(`OptQL` [Jin et al., 2018])*

$$Q_h^n(s,a) \approx r_h(s,a) + \frac{H}{n}\sum_{i\geq n-H/n}^{n} p_h^i \overline{V}_{h+1}^{i-1}(s,a)$$

$$\approx r_h(s,a) + p_h \underbrace{\left(\frac{H}{n}\sum_{i\geq n-n/H}^{n} \overline{V}_{h+1}^{i-1}\right)}_{:=V_{h,s,a}^n \text{bias-value function}}(s,a) \pm \underbrace{\sqrt{\frac{H^3}{n}}}_{\text{variance term}}.$$

UCB Momentum Q-learning

Idea add a (negative) momentum to correct the bias [Azar et al., 2011]

learning rate $\alpha_n \approx 1/n$ *and momentum rate* $\gamma_n \approx H/n$*:* `UCBMQ`

$$Q_h^n(s,a) = \alpha_n(r_h + p_h^n \overline{V}_{h+1}^{n-1})(s,a) + (1-\alpha_n)Q_h^{n-1}(s,a) + \gamma_n \underbrace{p_h^n(\overline{V}_{h+1}^{n-1} - V_{h,s,a}^{n-1})(s,a)}_{\leq 0,\ \text{momentum}}$$

where the bias-value function

$$V_{h,s,a}^n(s') = (\alpha_n + \gamma_n)\overline{V}_{h+1}^{n-1}(s') + (1-\alpha_n-\gamma_n)V_{h,s,a}^{n-1}(s')$$

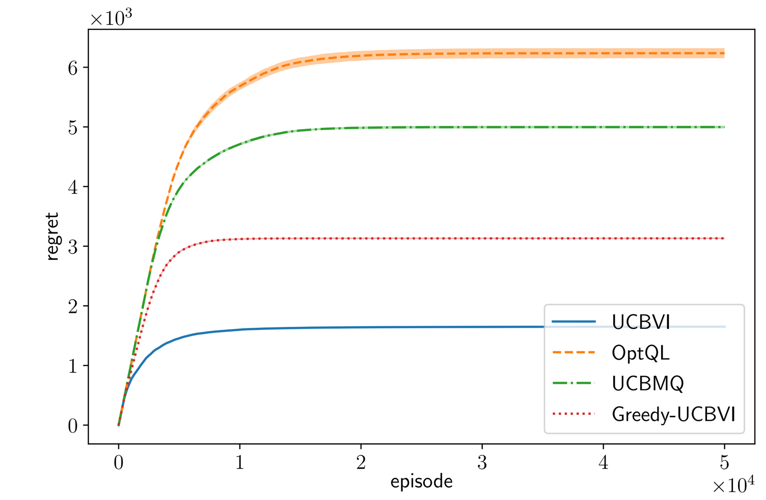
$$\approx \frac{H}{n}\sum_{i\geq n-n/H}^{n} \overline{V}_{h+1}^{i-1}(s')$$

Unfolding+ Hoeffding inequality

$$Q_h^n(s,a) \approx r_h(s,a) + \frac{1}{n}\sum_{i=1}^{n} p_h^i\left((H+1)\overline{V}_{h+1}^{i-1} - V_{s,a,h}^{i-1}\right)(s,a)$$

$$\approx r_h(s,a) + p_h \underbrace{\left(\frac{H}{n}\sum_{i\geq n-n/H}^{n} \overline{V}_h^{i-1}\right)}_{\approx V_{h,s,a}^n \text{ bias-value function}}(s,a) \pm \underbrace{\sqrt{\frac{H^2}{n}}}_{\text{variance term}} \pm \underbrace{\sqrt{\frac{H^3}{n}\sum_{i=1}^{n} p_h(V_{h,s,a}^{n-1} - \overline{V}_h^{n-1})(s,a)\frac{1}{n}}}_{\text{momentum variance term}}.$$

- keep only the last H/n fraction of the past targets: bound polynomial in H
- n samples to approximate the mean
- still an extra H in the bonus $\to$ Bernstein inequality instead of Hoeffding



Rates

Algorithm	Upper bound
`UCBVI` [Azar et al., 2017]	$\widetilde{\mathcal{O}}(\sqrt{H^3SAT} + H^3S^2A)$
`UBEV` [Dann et al., 2017]	$\widetilde{\mathcal{O}}(\sqrt{H^4SAT} + H^2S^3A^2)$
`EULER` [Zanette and Brunskill, 2019]	$\widetilde{\mathcal{O}}\left(\sqrt{H^3SAT} + H^3S^{3/2}A(\sqrt{S}+\sqrt{H})\right)$
`OptQL` [Jin et al., 2018] (Bernstein)	$\widetilde{\mathcal{O}}(\sqrt{H^4SAT} + H^{9/2}S^{3/2}A^{3/2})$
`UCB-Advantage` [Zhang et al., 2020]	$\widetilde{\mathcal{O}}(\sqrt{H^3SAT} + H^{33/4}S^2A^{3/2}T^{1/4})$
`UCBMQ` (this paper)	$\widetilde{\mathcal{O}}(\sqrt{H^3SAT} + H^4SA)$