UCB Momentum Q-learning: Correcting the bias without forgetting

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

[1]OvGU [2]Inria [3]Université de Lille [4]DeepMind

Markov Decision Process (MDP)

Tabular, episodic MDP: H horizon, S states, A actions.

Learning in MDP: at episode t, step h

- state s_h^t
- action a_h^t
- next state $s_{h+1}^t \sim p_h(\cdot|s_h^t, a_h^t)$
- reward $r_h(s_h^t, a_h^t)$ (known)

Bellman equation policy π

$$\begin{aligned} Q_h^\pi(s,a) &= (r_h + p_h V_{h+1}^\pi)(s,a) \\ V_h^\pi(s) &= Q_h^\pi(s, \pi_h(s)) \\ V_{H+1}^\pi(s) &= 0 \end{aligned}$$

where $p_h f = \sum_{s'} p_h(s'|s,a) f(s')$

Markov Decision Process (MDP)

Tabular, episodic MDP: H horizon, S states, A actions.

Learning in MDP: at episode t, step h

- state s_h^t
- action a_h^t
- next state $s_{h+1}^t \sim p_h(\cdot|s_h^t, a_h^t)$
- reward $r_h(s_h^t, a_h^t)$ (known)

Optimal Bellman equation

$$
\begin{aligned}
Q_h^\star(s,a) &= (r_h + p_h V_{h+1})(s,a) \\
V_h^\star(s) &= \max_a Q_h^\star(s,a) \\
V_{H+1}^\star(s) &= 0
\end{aligned}
$$

where $p_h f = \sum_{s'} p_h(s'|s,a) f(s')$

Regret after T episodes: $R^T = \sum_{t=1}^T V_1^\star(s_1) - V_1^{\pi^t}(s_1)$

Regret minimization

Lower bound $\mathbb{E}[R^T] \geq \Omega(\sqrt{H^3SAT})$ [?, ?]

Typical regret bound $R^T \leq \widetilde{\mathcal{O}}(\sqrt{H^3SAT} + \mathrm{poly}(H)S^2A)$

$\rightarrow$ optimal bound only for $T \geq \mathrm{poly}(H)S^2A$, bad when S large, continuous...

$\rightarrow$ non-trivial bound i.e. $R^T \leq TH$, for $\mathrm{poly}(H)S$ samples per state-actions

Algorithm	**Upper bound**
`UCBVI` [?]	$\widetilde{\mathcal{O}}(\sqrt{H^3SAT} + H^3S^2A)$
`UBEV` [?]	$\widetilde{\mathcal{O}}(\sqrt{H^4SAT} + H^2S^3A^2)$
`EULER` [?]	$\widetilde{\mathcal{O}}\left(\sqrt{H^3SAT} + H^3S^{3/2}A(\sqrt{S} + \sqrt{H})\right)$
`OptQL` [?] (Bernstein)	$\widetilde{\mathcal{O}}(\sqrt{H^4SAT} + H^{9/2}S^{3/2}A^{3/2})$
`UCB-Advantage` [?]	$\widetilde{\mathcal{O}}(\sqrt{H^3SAT} + H^{33/4}S^2A^{3/2}T^{1/4})$

Regret minimization

Lower bound $\mathbb{E}[R^T] \geq \Omega(\sqrt{H^3 SAT})$ [?, ?]

Wanted regret bound $R^T \leq \widetilde{\mathcal{O}}(\sqrt{H^3 SAT} + \mathrm{poly}(H) SA)$

$\rightarrow$ optimal bound only for $T \geq \mathrm{poly}(H) SA$

$\rightarrow$ non-trivial bound i.e. $R^T \leq TH$, for poly(H) samples per state-actions

Question: Regret first order optimal (in T) and at most linear in S?

Algorithm	**Upper bound**
UCBVI [?]	$\widetilde{\mathcal{O}}(\sqrt{H^3 SAT} + H^3 S^2 A)$
UBEV [?]	$\widetilde{\mathcal{O}}(\sqrt{H^4 SAT} + H^2 S^3 A^2)$
EULER [?]	$\widetilde{\mathcal{O}}\left(\sqrt{H^3 SAT} + H^3 S^{3/2} A(\sqrt{S} + \sqrt{H})\right)$
OptQL [?] (Bernstein)	$\widetilde{\mathcal{O}}(\sqrt{H^4 SAT} + H^{9/2} S^{3/2} A^{3/2})$
UCB-Advantage [?]	$\widetilde{\mathcal{O}}(\sqrt{H^3 SAT} + H^{33/4} S^2 A^{3/2} T^{1/4})$

Regret minimization

Lower bound $\mathbb{E}[R^T] \geq \Omega(\sqrt{H^3SAT})$ [?, ?]

Wanted regret bound $R^T \leq \widetilde{\mathcal{O}}(\sqrt{H^3SAT} + \mathrm{poly}(H)SA)$

$\rightarrow$ optimal bound only for $T \geq \mathrm{poly}(H)SA$

$\rightarrow$ non-trivial bound i.e. $R^T \leq TH$, for $\mathrm{poly}(H)$ samples per state-actions

Question: Regret first order optimal (in T) and at most linear in S? Yes!

Algorithm	**Upper bound**
UCBVI [?]	$\widetilde{\mathcal{O}}(\sqrt{H^3SAT} + H^3S^2A)$
UBEV [?]	$\widetilde{\mathcal{O}}(\sqrt{H^4SAT} + H^2S^3A^2)$
EULER [?]	$\widetilde{\mathcal{O}}\left(\sqrt{H^3SAT} + H^3S^{3/2}A(\sqrt{S} + \sqrt{H})\right)$
OptQL [?] (Bernstein)	$\widetilde{\mathcal{O}}(\sqrt{H^4SAT} + H^{9/2}S^{3/2}A^{3/2})$
UCB-Advantage [?]	$\widetilde{\mathcal{O}}(\sqrt{H^3SAT} + H^{33/4}S^2A^{3/2}T^{1/4})$
UCBMQ (this paper)	$\widetilde{\mathcal{O}}(\sqrt{H^3SAT} + H^4SA)$

Algorithms

Principle $a_h^n \in \operatorname{argmax}_a \overline{Q}_h^n(s,a)$, act greedily with respect to upper confidence bound on the optimal Q-values $Q^\star$

If p_h is known: dynamic Q-value iteration

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

Algorithms

Principle $a_h^n \in \operatorname{argmax}_a \overline{Q}_h^n(s,a)$, act greedily with respect to upper confidence bound on the optimal Q-values $Q^\star$

If p_h is known: dynamic Q-value iteration

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

If p_h unknown, approximate the expectation with samples: Q-learning

$$Q_h^n(s,a) = \alpha_n (r_h + p_h^n \overline{V}_h^{n-1})(s,a) + \left(1-\alpha_n\right) Q_h^{n-1}(s,a)$$
$$\overline{Q}_h^n(s,a) = Q_h^n(s,a) + b_h^n(s,a) \qquad \overline{V}_h^n(s) = \max_a \overline{Q}_h^n(s,a)$$

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

Algorithms

Principle $a_h^n \in \operatorname{argmax}_a \overline{Q}_h^n(s,a)$, act greedily with respect to upper confidence bound on the optimal Q-values $Q^\star$

If p_h is known: dynamic Q-value iteration

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

If p_h unknown, approximate the expectation with samples: Q-learning

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

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?

Q-learning

learning rate $\alpha_n \approx 1/n$, unfolding the formula for Q_h^n + Hoeffding inequality

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

→ no S to pay for passing from sample average p_h^i to true transition p_h

→ uniform average over the past targets $\overline{V}_{h+1}^{i-1}$: bound exponential in H

Q-learning

learning rate $\alpha_n \approx H/n$ (`OptQL` [?])

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

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

→ keep only the last H/n fraction of the past targets: bound polynomial in H

→ only n/H samples in the average: extra H in the bonus

UCB Momentum Q-learning

Idea Add a (negative) momentum to correct the bias [?]

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

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

UCB Momentum Q-learning

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

→ 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 → Bernstein inequality instead of Hoeffding

UCBMQ algorithm

Regret bound w.h.p.
$R^T \leq \widetilde{\mathcal{O}}(\sqrt{H^3SAT} + H^4SA)$

Time complexity per episode
$\mathcal{O}(HS)$

Space complexity $\mathcal{O}(HS^2A)$ (bias value function per state-action)
Model-free vs model-based?

Open problem

- linear in S regret bound for model-based algorithms? (UCBVI $\widetilde{\mathcal{O}}(\sqrt{H^3SAT} + H^3S^2A)$)
- Algorithm with bound $\widetilde{\mathcal{O}}(\sqrt{H^3SAT} + H^2SA)$?
- With time complexity $\mathcal{O}(H)$ per episode and space complexity $\mathcal{O}(HSA)$?

Thank you!

Bibliography I

- Azar, M. G., Munos, R., Ghavamzadeh, M., and Kappen, H. J. (2011). Speedy Q-learning. In Advances in Neural Information Processing Systems 24 (NIPS), pages 2411–2419.
- Azar, M. G., Osband, I., and Munos, R. (2017). Minimax regret bounds for reinforcement learning. In Proceedings of the 34th International Conference on Machine Learning (ICML), pages 405–433.
- Dann, C., Lattimore, T., and Brunskill, E. (2017). Unifying PAC and regret: Uniform PAC bounds for episodic reinforcement learning. In Advances in Neural Information Processing Systems 30 (NIPS), pages 5714–5724.
- Domingues, O. D., Ménard, P., Kaufmann, E., and Valko, M. (2021). Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited. In Proceedings of the 32nd International Conference on Algorithmic Learning Theory (ALT).
- Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is Q-learning provably efficient? In Advances in Neural Information Processing Systems 31 (NeurIPS), pages 4863–4873.
- Zanette, A. and Brunskill, E. (2019). Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In Proceedings of the 36th International Conference on Machine Learning (ICML), pages 12676–12684.
- Zhang, Z., Zhou, Y., and Ji, X. (2020). Almost optimal model-free reinforcement learning via reference-advantage decomposition. arXiv preprint arXiv:2004.10019.