Regret Minimization for Reinforcement Learning by Evaluating the Bias Function

Zihan Zhang
Tsinghua Unversity
zihan-zh17@mails.tsinghua.edu.cn

October 21, 2019

Organization

1. Problem Formulation
2. Related Works
3. Confidence Set for the Bias Function
4. OFU Framework
5. Regret Analysis of EBF
6. Conclusion

Problem Formulation

Markov Decision Process

Reinforcement learning is modeled as a Markov decision process. This work focus on finite discrete MDPs. A discrete finite MDP is a tuple $M = \langle \mathcal{S}, \mathcal{A}, r, P, s_1 \rangle$

- State space $\mathcal{S}$, $|\mathcal{S}| = S$
- Action space $\mathcal{A}$, $|\mathcal{A}| = A$
- Reward function $r(s, a) \in [0, 1]$ (which is assumed to be known)
- Transition Probability $P(\cdot|s, a) \in \Delta^{\mathcal{S}}$ (which is unknown)
- Initial state $s_1 \in \mathcal{S}$

Problem Formulation

Policy

A stationary policy π a mapping from $\mathcal{S}$ to $\Delta^{\mathcal{A}}$.

A non-stationary policy π is a mapping from history trajectory to $\Delta^{\mathcal{A}}$.

A trajectory $\mathcal{L}_T = \{s_1, a_1, r_1, s_2, ..., s_T, a_T, r_T, s_{T+1}\}$ is generated with $a_t \sim \pi(\mathcal{L}_{t-1})$ $r_t = r(s_t, a_t)$ and $s_{t+1} \sim P(\cdot|s_t, a_t)$, $t = 1, 2, ..., T$.
For a distribution γ over $\mathcal{A}$, let $r(s, \gamma) = \sum_a \gamma(a) r(s, a)$ and $P(\cdot|s, \gamma) = \sum_a \gamma(a) P(\cdot|s, a)$

$$r(\pi, s) = r(\pi(\cdot|s), s), \quad P(\cdot|s, \pi) = P(\cdot|s, \pi(\cdot|s))$$

Problem Formulation

Average Reward and Bias

For a stationary policy π, the average reward is defined as:

$$\rho(\pi, s) = \lim_{T \to \infty} \frac{1}{T} \mathbb{E}_\pi[\sum_{t=1}^{T} r_t | s_1 = s],$$

In the case $\rho(\pi, s)$ is independent of s, the bias function for $s \in \mathcal{S}$ is defined as

$$h(\pi, s) = \lim_{T \to \infty} \frac{1}{T} \sum_{i=1}^{T} \mathbb{E}[\sum_{t=1}^{i} (r_t - \rho(\pi)) | s_1 = s]$$

Problem Formulation

Optimal Policy for Weak-communicating MDP [Bartlett and Tewari, 2009]

In a weak-communicating MDP, the state space $\mathcal{S}$ decomposes into two sets: in the first, each state is reachable from every other state in the set under some policy; in the second, all states are transient under all policies.

An optimal policy π^* for a weak-communicating MDP M satisfies that:

$$\rho(\pi^*) = max_\pi \rho(\pi)$$

$$h(\pi^*, \cdot) + \rho(\pi^*)\mathbf{1} = P(\pi)h(\pi^*, \cdot) + r(\cdot, \pi^*) = \max_\pi P(\pi)h(\pi^*, \cdot) + r(\cdot, \pi)$$

The optimal bias function and optimal average reward are given by

$$h^* = h^*(M) := h(\pi^*, \cdot)$$

$$\rho^* = \rho^*(M) := \rho(\pi^*)$$

Span of the optimal bias function

$$sp(h^*) = \max_{s,s'} h^*(s) - h^*(s') \leq H$$

Problem Formulation

Regret

For a distribution γ over action space , the gap in $s \in \mathcal{S}$ is defined as:

$$reg_{s,\gamma} = h^*(s) + \rho^* - r(s,\gamma) - P(\cdot|,s,\gamma)^T h^*$$

$$reg_{s,a} := reg_{s,\mathbf{1}_a}$$

The regret in T steps could be defined in the following two ways

$$\overline{R}_T := \sum_{t=1}^{T} reg_{s_t,a_t}$$

$$R_T = T\rho^* - \sum_{t=1}^{T} r_{s_t,a_t}$$

Related Works

Algorithm	Regret Bound
Lower Bound	$\Omega(\sqrt{SAHT})(\Omega(\sqrt{SADT}))$
UCRL2	$\tilde{O}(DS\sqrt{AT})$
REGAL.C	$\tilde{O}(HS\sqrt{AT})$
SCAL	$\tilde{O}(S\sqrt{HAT})$
TSDE	$\tilde{O}(HS\sqrt{AT})$
EBF(ours)	$\tilde{O}(\sqrt{SAHT})(\tilde{O}(\sqrt{SADT}))$

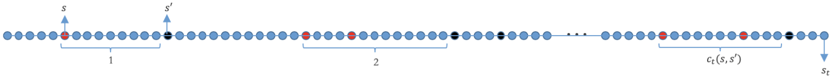
Confidence Set for the Bias Function

$M_{vir} = \langle \mathcal{S}, \mathcal{A}, r_{vir}, P, s_1 \rangle$ where $r_{vir}(s,a) = r(s,a) + reg_{s,a}(r_{vir}(s,a)$ may exceed $[0,1]$).

$h_{vir} := h^*(M_{vir}) = h^*, \; rho_{vir} := \rho^*(M_{vir}) = \rho^*$.

Let $\mathcal{L} = \{s_1, a_1, r_{vir,1}, ..., s_T, a_T, r_{vir,T}, s_{T+1}\}$ be the roll-out trajectory.

For two different states s, s', let
$Y_i(s,s') = \sum_{j \le i, I_j(s,s')=1}(r_{vir,j} - rho^*)$ and
$Y'_i(s,s') = \sum_{j \le i, I_j(s,s')=1}(r_{s_j,a_j} - rho^*)$, where $I_j(s,s') = 1$ iff the j-th step is in the underbraceed part as below.



Confidence Set for the Bias Function

Let $U_i = \{j | s_j \in \{s, s'\}, 1 \leq j \leq i\}$ and $j*$ be the maximal element of U_i if U_i is not empty. $I_i(s, s')$ is defined formally as:

$$I_i(s, s') = \begin{cases} 0 & i \geq t+1, U_i = \emptyset \text{ or } s_{j^*} = s' \\ 1 & \text{else} \end{cases}$$

$W_i(s, s') := \sum_{j=1}^{i} I_j(s, s')(r_{vir,j} - h^*_{s_j} + h^*_{s_{j+1}} - \rho^*) =$
$Y_i(s, s') - c_i(s, s')(h^*(s) - h^*(s'))$ is a martingale difference, since
$I_j(\mathcal{L}, s, s')$ is measurable w.r.t.
$\sigma(s_1, a_1, r_{vir,1}, ..., s_{j-1}, a_{j-1}, r_{vir,j-1}, s_j)$

Based on Azuma's inequality

$$|W_i(s, s')| \underset{w.h.p.}{\leq} \tilde{O}(H\sqrt{T}) \Longrightarrow$$

$$|c_i(s, s')(h^*(s) - h^*(s')) - Y_i(s, s')| \underset{w.h.p.}{\leq} \tilde{O}(H\sqrt{T}) \Longrightarrow$$

$$|c_i(s, s')(h^*(s) - h^*(s')) - Y_i'(s, s')| \underset{w.h.p.}{\leq} \tilde{O}(H\sqrt{T}) + \overline{R}_T$$

Confidence Set for the Bias Function

Under the framework of REGAL.C [Bartlett and Tewari, 2009], $\overline{R}_T$ is bounded by $\tilde{O}(HS\sqrt{AT})$ by induction. The final confidence set for the bias function at the t-th step is given by

$$\begin{aligned}\mathcal{H}_t =&\{h \in [0, H]^S \,|\, |c_t(s, s')(h(s) - h(s')) - Y'_t(s, s')| \\ &\leq \tilde{O}(HS\sqrt{AT}), \forall s, s', s \neq s'\}\end{aligned}$$

OFU Framework

Extended MDP

$$\begin{aligned}\mathcal{M}(t) = \{M' = \langle \mathcal{S}, \mathcal{A}, r, P', s_1\rangle | h^*(M') \in \mathcal{H}_t,\\ P'(\cdot|s,a) \in B^P_{t,1}(s,a) \cap B^P_{t,2}(s,a) \cap B^P_{t,3}(s,a), \forall (s,a) \in \mathcal{S}\times\mathcal{A}\}\end{aligned}$$

$$B^P_{t,1}(s,a) := \{P'(\cdot|s,a) || P'(s'|s,a) - \hat{P}(s'|s,a)| \le C_1\sqrt{\frac{\hat{P}_t(s'|s,a)\log(2/\delta)}{N_t(s,a)}}\}$$

$$B^P_{t,2}(s,a) := \{P'(\cdot|s,a) || P'(\cdot|s,a) - \hat{P}_t(s,a)||_1 \le C_2\sqrt{\frac{S\log(2/\delta)}{N_t(s,a)}}\}$$

$$B^P_{t,3}(s,a) := \{P'(\cdot|s,a) || (P'(\cdot|s,a) - \hat{P}_t(\cdot|s,a))^T h^*(M')| \le C_3\sqrt{\frac{V_(\hat{P}(\cdot|s,a), h^*(M'))\log(2/\delta)}{N_t(s,a)}}$$

$$V(p,h) = \sum_s p(s)h^2(s) - (\sum_s p(s)h(s))^2$$

Remark1: In this work, the reward is assumed to be known. It is not difficult to extend the proof to the original case.
Remark2: In the definition of confidence set above, the lower order terms in the RHS of these constraints are omitted for simplicity.

OFU Framework

EBF Algorithm

Algorithm 1 EBF

Input parameter: H, T

for episodes $k = 1, 2, \dots$ **do**

  $t_k \leftarrow$current time

  $\mathcal{M}_k \leftarrow \mathcal{M}(t_k)$

  Choose $M_k \in \mathcal{M}_k$ to maximize $\rho^*(M_k)$.

  $\pi_k \leftarrow$ deterministic optimal policy for M_k;

  Follow π_k until the visit count of some (s, a) pair doubles or the T steps are finished.

end for

Regret Analysis of EBF

$$v_k(s,a) := N_{t_{k+1}}(s,a) - N_{t_k}(s,a)$$

$$h_k := h^*(M_k), \quad P'_k := P_{M_k}(\pi_k), \quad P_k := P_M(\pi_k), \quad \overline{P}_k := \hat{P}_{t_k}(\pi_k)$$

The regret in the k-th episode R_k can be written in a vector form for simplicity:

$$\begin{aligned} R_k = v_k^T(\rho^* - r) &\underset{w.h.p.}{\leq} v_k^T(\rho^*(M_k) - r) \\ &= \underbrace{v_k^T(P_k - I)^T h_k}_{①_k} + \underbrace{v_k^T(\overline{P}_k - P_k)^T h^*}_{②_k} \\ &+ \underbrace{v_k^T(P'_k - \hat{P}_k)^T h_k}_{③_k} + \underbrace{v_k^T(\overline{P}_k - P_k)^T(h_k - h^*)}_{④_k} \end{aligned} \tag{1}$$

Regret Analysis of EBF

The first two terms can be bounded by Bernstein Inequality.

Lemma (Lemma3)

When $T \geq S^2AH^2\log(2/\delta)$, *with probability* $1-3\delta$, *it holds that*

$$\sum_k ①_k = \tilde{O}(\sqrt{TH})$$

Lemma (Lemma4)

When $T \geq S^2AH^2\log(2/\delta)$, *with probability* $1-\delta$, *it holds that*

$$\sum_k ②_k \leq \sum_{k,s,a} v_k(s,a)\sqrt{\frac{V(P(\cdot|s,a),h^*)\log(2/\delta)}{\max\{N_{t_k}(s,a),1\}}} = \tilde{O}(\sqrt{SAHT})$$

Regret Analysis of EBF

Let $\delta_k(s,s') = h_k(s) - h_k(s')$ and $\delta^*(s,s') = h^*(s) - h^*(s')$.

The thrid and fourth term can be bounded as:

$$\sum_k ③_k = \sum_k ②_k + \sum_k (③_k - ②_k)$$

$$\underset{w.h.p.}{\leq} \tilde{O}(\sqrt{SAHT}) + O(\sum_{k,s,a} v_k(s,a) \frac{\sqrt{|V(\hat{P}_{t_k}(\cdot|s,a),h_k) - V(P(\cdot|s,a),h^*)| \log(2/\delta)}}{\sqrt{\max\{N_{t_k}(s,a),1\}}})$$

$$\leq \tilde{O}(\sqrt{SAHT}) + \underbrace{O(\sqrt{H\log(2/\delta)} \sum_{k,s,a} v_k(s,a) \sum_{s'} \sqrt{\frac{\hat{P}_{t_k}(s'|s,a)|\delta_k(s,s') - \delta^*(s,s')|}{\max\{N_{t_k}(s,a),1\}}})}_{X_1}$$

$$\sum_k ④_k \underset{w.h.p.}{\leq} \sum_{k,s,a} v_k(s,a) \sum_{s'} (\hat{P}_{t_k}(s'|s,a) - P(s'|s,a)) |\delta_k(s,s') - \delta^*(s,s')|$$

$$\leq \underbrace{O(\sqrt{H\log(2/\delta)} \sum_{k,s,a} v_k(s,a) \sum_{s'} \sqrt{\frac{\hat{P}_{t_k}(s'|s,a)|\delta_k(s,s') - \delta^*(s,s')|}{\max\{N_{t_k}(s,a),1\}}})}_{X_1}$$

Regret Analysis of EBF

Lower Order Term

$$N_{t_k}(s'|s,a) := N_{t_k}(s,a)\hat{P}_{t_k} = \#\{j < t_k | s_j = s, a_j = a, s_{j+1} = s'\} \leq c_{t_k}(s,s')$$

$$\begin{aligned}
X_1 :&= O\Big(\sqrt{H\log(2/\delta)}\sum_{k,s,a} v_k(s,a)\sum_{s'}\sqrt{\frac{\hat{P}_{t_k}(s'|s,a)|\delta_k(s,s') - \delta^*(s,s')|}{\max\{N_{t_k}(s,a),1\}}}\Big) \\
&= O\Big(\sqrt{H\log(2/\delta)}\sum_{k,s,a}\frac{v_k(s,a)}{\max\{N_{t_k}(s,a),1\}}\sum_{s'}\sqrt{N_{t_k}(s'|s,a)|\delta_k(s,s') - \delta^*(s,s')|}\Big) \\
&\underset{w.h.p.}{\leq} \tilde{O}(S^{\frac{7}{2}}A^{\frac{5}{4}}HT^{\frac{1}{4}})
\end{aligned}$$

The total regret is bounded by:

$$\begin{aligned}
R_T &\underset{w.h.p.}{\leq} \sum_k (①_k + ②_k + ③_k + ④_k) \\
&\underset{w.h.p.}{\leq} \tilde{O}(\sqrt{TH}) + \tilde{O}(\sqrt{SAHT}) + \tilde{O}(S^{\frac{7}{2}}A^{\frac{5}{4}}HT^{\frac{1}{4}}) \\
&= \tilde{O}(\sqrt{SAHT})
\end{aligned}$$

Regret Analysis of EBF

Final Result

Theorem (Theorem 1)

With probability $1-\delta$, for any weak-communicating MDP M and any initial state $s_{ini} \in S$, when $T \geq p_1(S, A, H, \log(\frac{1}{\delta}))$ and $S, A, H \geq 20$ where p_1 is a polynomial function, the regret of EBF algorithm is bounded by

$$R_T \leq 490\sqrt{SAHT\log(\frac{40S^2A^2T\log(T)}{\delta})},$$

whenever an upper bound of the span of optimal bias function H is known. By setting $\delta = \frac{1}{T}$, we get that $\mathbb{E}[R_T] = \tilde{O}(SAHT)$.

Regret Analysis of EBF

Final Result

When such an upper bound for $sp(h^*)$ is not given, EBF algorithm can still provide an $\tilde{O}(\sqrt{SADT})$ regret bound by estimating the diameter directly.

Corollary (Corollay 1)

For MDP M with finite unknown diameter D and any initial state $s_{ini} \in S$, with probability $1-\delta$, when $T \geq p_2(S, A, D, \log(\frac{1}{\delta}))$ and $S, A, D \geq 20$ where p_2 is a polynomial function, the regret can be bounded by

$$R_T \leq 490\sqrt{SADT\log(\frac{S^3A^2T\log(T)}{\delta})}$$

Conclusion

In this work the open problems proposed in [Jiang and Agarwal, 2018] are partly solved by an OFU based algorithm EBF, which provides a regret bound of $\tilde{O}(\sqrt{SAHT})$ whenever H, an upper bound on $sp(h^*)$ is known.

This regret upper bound matches the corresponding lower bound up to a logarithmic factor and outperform the best previous known bound by an $\sqrt{S}$ factor.

References

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. *arXiv preprint arXiv:1703.05449*, 2017.

Peter L Bartlett and Ambuj Tewari. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 35–42. AUAI Press, 2009.

A. N. Burnetas and M. N. Katehakis. *Optimal Adaptive Policies for Markov Decision Processes*. 1997.

Ronan Fruit, Matteo Pirotta, and Alessandro Lazaric. Near optimal exploration-exploitation in non-communicating markov decision processes. In *Advances in Neural Information Processing Systems*, pages 2998–3008, 2018a.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. *arXiv preprint arXiv:1802.04020*, 2018b.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

Nan Jiang and Alekh Agarwal. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*, pages 3395–3398, 2018.

Sham Kakade, Mengdi Wang, and Lin F Yang. Variance reduction methods for sublinear reinforcement learning. *arXiv preprint arXiv:1802.09184*, 2018.

References

Sham Kakade, Mengdi Wang, and Lin F Yang. Variance reduction methods for sublinear reinforcement learning. *arXiv preprint arXiv:1802.09184*, 2018.

Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. A finite-time analysis of multi-armed bandits problems with kullback-leibler divergences. In *Proceedings of the 24th annual Conference On Learning Theory*, pages 497–514, 2011.

Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? *arXiv preprint arXiv:1607.00215*, 2016.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, pages 3003–3011, 2013.

Yi Ouyang, Mukul Gagrani, Ashutosh Nayyar, and Rahul Jain. Learning unknown markov decision processes: A thompson sampling approach. In *Advances in Neural Information Processing Systems*, pages 1333–1342, 2017.

M L Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. 1994.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Georgios Theocharous, Zheng Wen, Yasin Abbasi-Yadkori, and Nikos Vlassis. Posterior sampling for large scale reinforcement learning. *arXiv preprint arXiv:1711.07979*, 2017.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. *arXiv preprint arXiv:1901.00210*, 2019.