Provably Efficient Offline RL with Options

Extended Abstract

Xiaoyan Hu
The Chinese University of Hong Kong
Hong Kong SAR, China
xyhu21@cse.cuhk.edu.hk

Ho-fung Leung
The Chinese University of Hong Kong
Hong Kong SAR, China
lhf@cuhk.edu.hk

ABSTRACT

Temporal abstraction helps to reduce the sample complexity in long-horizon planning in reinforcement learning (RL). One powerful approach is the *options* framework, where the agent interacts with the environment using closed-loop policies, i.e., options, instead of primitive actions. Recent works show that in the *online* setting, where the agent can continuously explore the environment, lower PAC-like sample complexity or regret can be attained by learning with options. However, these results are no longer applicable in scenarios where collecting data in an online manner is impossible, e.g., automated driving and healthcare. In this paper, we provide the first analysis of the sample complexity for *offline* RL with options, where a dataset is provided and no further interaction with the environment is allowed. Two procedures of the data collecting process are considered, which adapt to different scenes of applications and are of great importance to study. Inspired by previous works on offline RL, we propose **PE**ssimistic **V**alue **I**teration for Learning with **O**ptions (PEVIO) and derive suboptimality bounds for both datasets, which are near-optimal according to a novel information-theoretic lower bound for offline RL with options. Further, the suboptimality bound shows that learning with options can be more sample-efficient than learning with primitive actions in the offline setting.

KEYWORDS

learning with options; offline reinforcement learning; sample complexity; sample-efficient learning

ACM Reference Format:
Xiaoyan Hu and Ho-fung Leung. 2023. Provably Efficient Offline RL with Options: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

Long-horizon planning is a great challenge in reinforcement learning (RL) [2, 5, 11]. Rather than learning with primitive actions, exploiting the hierarchical structures in RL and planning with temporally-extended action has long been explored [3, 4, 6, 13, 16, 17, 19]. One powerful and popular approach is the *options* framework [21, 24]. In this framework, the agent is provided with a set of *options*, i.e., closed-loop policies for taking action over a period of time. Upon arriving at a state, the option used in the previous timestep is terminated with a pre-specified probability, and (if terminated) the agent selects a new option according to a hierarchical policy. She then takes primitive actions according to the chosen option in the following steps until the option is replaced. Empirical success [22, 23] shows that options help to accelerate learning and achieve sample-efficient performance.

Recent works analyze the sample complexity of RL with options in the *online* setting, where the agent continuously explores the environment and learns a near-optimal policy. Brunskill and Li [1] derive a PAC-like sample complexity of RL with options in the semi-Markov decision processes (SMDPs). They show that the use of options may reduce the sample complexity of a lifelong learning agent. However, these results cannot be immediately translated into a reduction of the sample complexity of learning with options in *Markov decision processes* (MDPs) due to the fundamental difference in the formulation between SMDPs and MDPs (see the discussion of Sutton, Precup, and Singh [21, Section 0]). Fruit and Larzaric [7] provide the first regret analysis of RL with options in MDPs. They show that an SMDP-variant of the UCRL algorithm [12] attains a sublinear regret. Nonetheless, the algorithm requires prior knowledge of the environment, which is not usually available in practice. This problem is addressed later by Fruit, Pirotta, Lazaric, and Brunskill [8]. They propose an algorithm that does not require prior knowledge, yet achieving a near-optimal regret bound. While these results report the regret bounds when options are used in learning in an unknown environment, they are applicable only when dealing with scenarios where online exploration is possible. However, in many real-world applications, having to learn in an online manner is undesirable. For examples, it has been argued that online learning in healthcare [10] and automated driving [20] is risky and costly. In such scenarios, *offline* learning, where a dataset is provided and the agent is then asked to learn a near-optimal policy by only using the dataset, is preferred. We note that there has been a rich literature on regret bound analysis for offline RL with primitive actions only (i.e., without the use of options) [9, 15, 18]. Unfortunately, to the best of our knowledge, there have been no results reported on the analysis of regret bounds for offline RL under the options framework.

In this paper, we provide the first analysis of the sample complexity of offline RL with options in episodic MDP. We consider two procedures to collect the offline dataset. The first dataset $\mathcal{D}_1$ contains state transitions and cumulative rewards by interacting with the environment using options. The second dataset $\mathcal{D}_2$ contains state transitions and (single-timestep) rewards by taking primitive actions. The distinctive structure yields advantages in different real-world applications. In short, $\mathcal{D}_1$ enables direct evaluation of options and requires smaller storage, while $\mathcal{D}_2$ provides more information about the environment and allows the design of new options. Hence, both datasets are of great importance to study. Inspired by the PEVI algorithm [14], we propose the **PE**ssimistic **V**alue **I**teration for Learning with **O**ptions (PEVIO) algorithm. To

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

analyze the suboptimality of the hierarchical policy output by PEVIO, we extend important results in previous works on RL with primitive actions. Compared to offline RL with primitive actions, our suboptimality bound enjoys a better dependence on the length of each episode and the size of state space. This shows that the options help to reduce the sample complexity and facilitate more efficient learning not only in the online setting but also the offline setting. To our best knowledge, this is the first study to provide such theoretical guarantees for the options framework in offline RL.

2 PRELIMINARIES

An episodic MDP with options is a sextuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{O}, H, \mathcal{P}, r)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the (primitive) action set, $\mathcal{O}$ the finite set of options, H the length of each episode, $\mathcal{P} = \{P_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})\}_{h\in[H]}$ the transition kernel,[1] $r = \{r_h : \mathcal{S}\times\mathcal{A} \to [0,1]\}_{h\in[H]}$ the deterministic reward function.[2] We define $S := |\mathcal{S}|$, $A := |\mathcal{A}|$, and $O := |\mathcal{O}|$. A (Markov) option [21, 24] $o \in \mathcal{O}$ is a pair (π^o, β^o) where $\pi^o = \{\pi^o_h : \mathcal{S} \to \Delta(\mathcal{A})\}_{h\in[H]}$ is the option's policy and $\beta^o = \{\beta^o_h : \mathcal{S} \to [0,1]\}_{h\in[H]}$ is a probability of the option's *termination*. Upon arriving at state s_h at timestep h, if $h = 1$ (at the beginning of an episode), the agent selects option $o_1 \sim \mu_1(\cdot|s_1)$, where $\mu = \{\mu_h : \mathcal{S} \to \Delta(\mathcal{O})\}_{h\in[H]}$ is a hierarchical policy to select option at each state. Otherwise ($h \geq 2$), the agent first terminates option o_{h-1} with probability $\beta^{o_{h-1}}_h(s_h)$. If option o_{h-1} is terminated, she then selects a new option $o_h \sim \mu_h(\cdot|s_h)$ according to the hierarchical policy μ. If option o_{h-1} is not terminated, the agent keeps using option o_{h-1} at timestep h, i.e., $o_h = o_{h-1}$. After that, the agent takes action $a_h \sim \pi^{o_h}_h(\cdot|s_h)$, receives a reward of $r_h(s_h, a_h)$, and transits to the next state $s_{h+1} \sim P_h(\cdot|s_h, a_h)$. An episode terminates at timestep $H + 1$. The value function of hierarchical policy μ is denoted by $V^\mu_h(s) := \mathbb{E}_\mu[\sum_{h'=h}^H r_{h'}|s_h = s]$ for any $(h, s) \in [H] \times \mathcal{S}$. There exists an optimal (and deterministic) hierarchical policy $\mu^* = \{\mu^*_h : \mathcal{S} \to \mathcal{O}\}_{h\in[H]}$ that attains the optimal value function, which is denoted by $V^* = \{V^*_h\}_{h\in[H]}$. We denote by $\theta^\mu = \{\theta^\mu_h : \mathcal{S}\times\mathcal{O} \to [0,1]\}_{h\in[H]}$ the state-option occupancy measure of the hierarchical policy μ, where $\theta^\mu_h(s,o)$ is the probability that the agent selects a particular option o at state s at timestep h. Further, we define $Z^\mu_\mathcal{O} := \sum_{h\in[H]}\sum_{s\in\mathcal{S}}\sum_{o\in\mathcal{O}}\theta^\mu_h(s,o)$ and $\overline{Z}^\mu_\mathcal{O} := \sum_{h\in[H]}\sum_{s\in\mathcal{S}}\sum_{o\in\mathcal{O}}\mathbb{I}[\theta^\mu_h(s,o) > 0]$ for any hierarchical policy μ, where $\mathbb{I}[\cdot]$ is the indicator function. Note that for a deterministic hierarchical policy $\mu = \{\mu_h : \mathcal{S} \to \mathcal{O}\}_{h\in[H]}$, it holds that $Z^\mu_\mathcal{O} \leq H$ and $\overline{Z}^\mu_\mathcal{O} \leq HS$.

In the offline setting, a dataset collected by an experimenter through interacting with the environment is provided, which consists of trajectories of state transitions and rewards. We consider two procedures to collect the dataset and the details are presented in Section 3. Given any such dataset $\mathcal{D}$, the algorithm is asked to learn a near-optimal hierarchical policy. Let $\widehat{\mu}$ denote the hierarchical policy output from the algorithm. We aim to minimize its suboptimality at state s, which is given by

$$\text{SubOpt}_\mathcal{D}(\widehat{\mu}, s) = V^*_1(s) - V^{\widehat{\mu}}_1(s)$$

3 MAIN RESULTS

3.1 Offline Learning with options Using Dataset $\mathcal{D}_1$

We consider dataset $\mathcal{D}_1 := \{\{(s^k_h, o^k_h, u^k_h)\}_{h\in\mathcal{H}^k}\}_{k\in[K]}$ collected by the first procedure, where $\mathcal{H}^k := \{t^k_i | 1 = t^k_1 < t^k_2 < \cdots < t^k_{j^k} \leq H\}_{i=1}^{j^k}$ is a set of timesteps within the kth episode. Particularly, at timestep t^k_i of the kth episode, the experimenter randomly selects a new option $o^k_{t^k_i}$ according to the hierarchical behavior policy ρ, uses it for $(t^k_{i+1} - t^k_i)$ timesteps, collects a cumulative reward of $u^k_{t^k_i}$ within these $(t^k_{i+1} - t^k_i)$ timesteps, and finally terminates this option at state $s^k_{t^k_{i+1}}$ at timestep t^k_{i+1}. Define the concentrability

$$C^{\text{option}}_1 := \max_{(h,s,o)\in[H]\times\mathcal{S}\times\mathcal{O}} \frac{\theta^{\mu^*}_h(s,o)}{\theta^\rho_h(s,o)}$$

for dataset $\mathcal{D}_1$. We show that the PEVIO algorithm attains the following suboptimality bound.

$$\text{SubOpt}_{\mathcal{D}_1}(\widehat{\mu}, s) \leq \tilde{O}\left(\sqrt{\frac{C^{\text{option}}_1 H^3 Z^{\mu^*}_\mathcal{O} \overline{Z}^{\mu^*}_\mathcal{O}}{K}}\right)$$

Compared to the suboptimality bound $\tilde{O}(\sqrt{H^5 S C^*/K})$ attained by the VI-LCB algorithm that learns with primitive actions [25, Theorem 1], where C^* is defined in [25, Assumption A], our suboptimality bound is smaller w.r.t the horizon H and state space S when $Z^*_\mathcal{O} \ll H$ and $\overline{Z}^*_\mathcal{O} \ll HS$. This shows that the options can accelerate learning by temporal abstraction (i.e., $Z^*_\mathcal{O} \ll H$) and the reduction of the state space (i.e., $\overline{Z}^*_\mathcal{O} \ll HS$), as pointed out by Fruit and Lazaric [7]. While recent works [7, 8] testify this statement in online RL, our results provide theoretical guarantee for offline RL.

3.2 Offline Learning with options Using Dataset $\mathcal{D}_2$

We also consider dataset $\mathcal{D}_2 := \{(s^k_h, a^k_h, r^k_h)\}_{k,h=1}^{K,H}$ collected by the second procedure. Particularly, the experimenter randomly takes action a^k_h at state s^k_h at timestep h of the kth episode according to the behaviral policy ρ, receives a reward of r^k_h, and transits to state s^k_{h+1}. Define the concentrability

$$C^{\text{option}}_2 := \max_{(h,s,o)\in[H]\times\mathcal{S}\times\mathcal{O}} \max_{(s',a')\in\mathcal{S}^m_{h,s,o}, h\leq m\leq H} \frac{\theta^{\mu^*}_h(s,o)}{d^\rho_m(s',a')}$$

for dataset $\mathcal{D}_2$, where $d^\rho_h(s,a)$ is the probability that the agent takes action a at state s at timestep h. We show that the PEVIO algorithm attains the following suboptimality bound.

$$\text{SubOpt}_{\mathcal{D}_2}(\widehat{\mu}, s) \leq \tilde{O}\left(\sqrt{\frac{C^{\text{option}}_2 H^4 S^2 A Z^{\mu^*}_\mathcal{O} \overline{Z}^{\mu^*}_\mathcal{O}}{K}} + \frac{H^5 S^5 A O C^{\text{option}}_2}{K}\right)$$

[1] We use the notation $\Delta(\mathcal{X})$ to denote the probability simplex on space $\mathcal{X}$ throughout the paper, i.e., each element in $\Delta(\mathcal{X})$ is a probability distribution on space $\mathcal{X}$. For any positive integer N, we define $[N] = \{1, \cdots, N\}$.

[2] While we assume deterministic rewards for simplicity, our results can be directly generalized to stochastic rewards, as the major difficulty is in learning the transitions rather than learning the rewards.

ACKNOWLEDGMENTS

The work presented in this paper is partially supported by a research grant from the Research Grants Council, Hong Kong, China (RGC Ref. No. CUHK 14206820) and a CUHK Research Committee Direct Grant for Research.

REFERENCES

[1] Emma Brunskill and Lihong Li. 2014. PAC-inspired Option Discovery in Lifelong Reinforcement Learning. In *Proceedings of the 31st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 32)*, Eric P. Xing and Tony Jebara (Eds.). PMLR, Bejing, China, 316–324. https://proceedings.mlr.press/v32/brunskill14.html

[2] John Co-Reyes, YuXuan Liu, Abhishek Gupta, Benjamin Eysenbach, Pieter Abbeel, and Sergey Levine. 2018. Self-Consistent Trajectory Autoencoder: Hierarchical Reinforcement Learning with Trajectory Embeddings. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 1009–1018. https://proceedings.mlr.press/v80/co-reyes18a.html

[3] Gary L Drescher. 1991. *Made-up minds: a constructivist approach to artificial intelligence.* MIT press.

[4] Akram Erraqabi, Marlos C. Machado., Mingde Zhao, Sainbayar Sukhbaatar, Alessandro Lazaric, Denoyer Ludovic, and Yoshua Bengio. 2022. Temporal abstractions-augmented temporally contrastive learning: An alternative to the Laplacian in RL. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence (Proceedings of Machine Learning Research, Vol. 180)*, James Cussens and Kun Zhang (Eds.). PMLR, 641–651. https://proceedings.mlr.press/v180/erraqabi22a.html

[5] Ben Eysenbach, Russ R Salakhutdinov, and Sergey Levine. 2019. Search on the Replay Buffer: Bridging Planning and Reinforcement Learning. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/5c48ff18e0a47baaf81d8b8ea51eec92-Paper.pdf

[6] Richard E Fikes, Peter E Hart, and Nils J Nilsson. 1972. Learning and executing generalized robot plans. *Artificial Intelligence* 3 (1972), 251–288. https://doi.org/10.1016/0004-3702(72)90051-3

[7] Ronan Fruit and Alessandro Lazaric. 2017. Exploration-Exploitation in MDPs with Options. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 54)*, Aarti Singh and Jerry Zhu (Eds.). PMLR, 576–584. https://proceedings.mlr.press/v54/fruit17a.html

[8] Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Emma Brunskill. 2017. Regret Minimization in MDPs with Options without Prior Knowledge. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/90599c8fdd2f6e7a03ad173e2f535751-Paper.pdf

[9] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. 2020. D4RL: Datasets for Deep Data-Driven Reinforcement Learning. https://doi.org/10.48550/ARXIV.2004.07219

[10] Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. 2019. Guidelines for reinforcement learning in healthcare. *Nature Medicine* 25, 1 (2019), 16–18.

[11] Christopher Hoang, Sungryull Sohn, Jongwook Choi, Wilka Carvalho, and Honglak Lee. 2021. Successor Feature Landmarks for Long-Horizon Goal-Conditioned Reinforcement Learning. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 26963–26975. https://proceedings.neurips.cc/paper/2021/file/e27c71957d1e6c223e0d48a165da2ee1-Paper.pdf

[12] Thomas Jaksch, Ronald Ortner, and Peter Auer. 2010. Near-optimal Regret Bounds for Reinforcement Learning. *Journal of Machine Learning Research* 11, 51 (2010), 1563–1600. http://jmlr.org/papers/v11/jaksch10a.html

[13] YiDing Jiang, Shixiang (Shane) Gu, Kevin P Murphy, and Chelsea Finn. 2019. Language as an Abstraction for Hierarchical Deep Reinforcement Learning. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/0af787945872196b42c9f73ead2565c8-Paper.pdf

[14] Ying Jin, Zhuoran Yang, and Zhaoran Wang. 2021. Is Pessimism Provably Efficient for Offline RL?. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 5084–5096. https://proceedings.mlr.press/v139/jin21e.html

[15] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. https://doi.org/10.48550/ARXIV.2005.01643

[16] Marlos C. Machado, Andre Barreto, and Doina Precup. 2021. Temporal Abstraction in Reinforcement Learning with the Successor Representation. arXiv:2110.05740 [cs.LG]

[17] Ofir Nachum, Haoran Tang, Xingyu Lu, Shixiang Gu, Honglak Lee, and Sergey Levine. 2019. Why Does Hierarchy (Sometimes) Work So Well in Reinforcement Learning? https://doi.org/10.48550/ARXIV.1909.10618

[18] Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. 2021. Bridging Offline Reinforcement Learning and Imitation Learning: A Tale of Pessimism. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 11702–11716. https://proceedings.neurips.cc/paper/2021/file/60ce36723c17bbac504f2ef4c8a46995-Paper.pdf

[19] Earl D. Sacerdoti. 1973. Planning in a Hierarchy of Abstraction Spaces. *Artif. Intell.* 5 (1973), 115–135.

[20] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. 2016. Safe, Multi-Agent, Reinforcement Learning for Autonomous Driving. https://doi.org/10.48550/ARXIV.1610.03295

[21] Richard S. Sutton, Doina Precup, and Satinder Singh. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence* 112, 1 (1999), 181–211. https://doi.org/10.1016/S0004-3702(99)00052-1

[22] Chen Tessler, Shahar Givony, Tom Zahavy, Daniel Mankowitz, and Shie Mannor. 2017. A Deep Hierarchical Approach to Lifelong Learning in Minecraft. *Proceedings of the AAAI Conference on Artificial Intelligence* 31, 1 (Feb. 2017). https://doi.org/10.1609/aaai.v31i1.10744

[23] Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. 2017. FeUdal Networks for Hierarchical Reinforcement Learning. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 3540–3549. https://proceedings.mlr.press/v70/vezhnevets17a.html

[24] Yi Wan, Abhishek Naik, and Rich Sutton. 2021. Average-Reward Learning and Planning with Options. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 22758–22769. https://proceedings.neurips.cc/paper/2021/file/c058f544c737782deacefa532d9add4c-Paper.pdf

[25] Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. 2021. Policy Finetuning: Bridging Sample-Efficient Offline and Online Reinforcement Learning. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 27395–27407. https://proceedings.neurips.cc/paper/2021/file/e61eaa38aed621dd776d0e67cfeee366-Paper.pdf