P2. *The claim "we seek to elucidate the conditions under which model-based Q-learning excels in terms of sample efficiency compared to its model-free counterpart." is vague. After reading this paper, the reviewer cannot find an answer to this question.*

A2. Thank you for pointing out the vagueness of the expression. We agree with the reviewer, and the sentence indicated by the reviewer will be corrected in the revision as follows: in this paper, we will prove that the proposed online model-based Q-learning excels the standard Q-learning in terms of sample efficiency.

P3. *The claims "a natural method to improve the sample efficiency of RL algorithms is to incorporate the model into the learning phase. Both in theoretical and experimental sense, leveraging the knowledge of learned or known models has been shown to improve over the model-free methods." are not well-supported. The reviewer checked the cited paper Kidambi et al. [2020], it didn't claim that model-based methods outperform model-free methods in terms of sample complexity. Actually, this claim is hard to be proved through several small experiments.*

A3. Thank you for raising the concerns. We will provide a clarification on the claim in the sequel: In page 2 of Kidambi et al. [2020], it is stated that "*MBRL algorithms have been highly sample efficient for online RL*". The experimental results in Table 5 of the paper show that the proposed model-based RL algorithms outperform SAC, which is a model-free algorithm. In theoretical aspect, it was shown in [4] that the model-based Q-value iteration achieves optimal sample complexity in an offline learning setting. These works support that model-based approach can be better in terms of sample efficiency than model-free approach under particular scenarios. However, as the reviewer mentioned, it is not a simple problem to whether to argue that model-based approach always outperforms its model-free counterpart. We will incorporate this discussion in the revised manuscript.

P4. *The second and the third points in the contribution part are not contributions. They are simply describing the online setting.*

A4. Thank you for pointing out the issue. We will remove second and third contributions in the revised manuscript.

P6. *In the related work section, you might also add discussions to some Q-learning papers on finite-time analysis in the function approximation setting. Examples are "A finite-time analysis of Q-learning with neural network function approximation", "On the Convergence and Sample Complexity Analysis of Deep Q-Networks with ϵ-Greedy Exploration", etc.*

A6. Thank you for the suggestion. We will add the papers and related discussions in the revision.

P7. *The reviewer is wondering if you have some assumption on the behavior policy, say, the distribution d. The proposed method does not explicitly deal with the exploration and exploitation trade-off. One could question how the proposed method explores the state-action space efficiently to get the current theoretical guarantee.*

A7. Thank you for the insightful feedback. The difference with exploration-exploitation trade-off setting is as follows: We assume that we can sample each state-action pair from an i.i.d. distribution rather than following a single trajectory generated from a Markov chain. Therefore, we do not have to worry about exploration issue given that $d(s,a) > 0$ for all $s, a \in \mathcal{S} \times \mathcal{A}$, which is the only assumption we pose on the behavior policy. The assumption that sampling from an i.i.d. distribution is stronger assumption than following a single trajectory. However, our bound shows better sample complexity than Szita and Szepesvari [2010] and Lattimore and Hutter [2014], which is the best achievable sample complexity result **when using model-based approach in online learning setting**. (Please refer to P12 for more discussion on the results under exploration-exploitation dilemma)

P8. *This work focuses on the sample complexity to get the ε-optimal value function rather than the ε-optimal policy. In practice, we actually care about the performance of the policy learned. Can your result be translated to a performance guarantee of the learned policy?*

A8. Thank you for raising the interesting concern about near-optimal policy. First of all, we note that in most existing works on Q-learning and its variants, ε-optimal value function

is the main concern. However, as the reviewer mentioned, it seems that it is possible to convert it into a performance guarantee of the learned policy using possibly two ways: 1) we can use the so-called performance difference lemma; 2) we can use the fact that the learned policy is expressed as a greedy w.r.t. the learned value. We appreciate suggesting interesting question and topics, which will be seriously considered in the revision.

P10. *The algorithm is in two stages. The first stage is purely collecting data. This operation is rarely seen in the online setting. Usually, the estimated transition kernel is initialized to a reasonable value. The reviewer doubts the necessity of this design.*

A10. Thank you for the detailed comment. As suggested by the reviewer, even if we randomly initialize the transition matrix, the strong law of large numbers suggests that it will converge close to the true transition matrix. The reason why we introduced the first stage is for the simplicity of the proof to remove the error caused by wrong initialization. We will incorporate the discussion in the revised manuscript.

P11. *In the paper, the authors use the terms "off-line" and "on-line". They are not of the same meaning as "offline" and "online". The reviewer never heard of these terms, please provide citations or clear definitions.*

A11. Thank you for the valuable comment. We agree that we have used wrong terminologies. In particular, we used the terminology "off-line" and "on-line" as same as "offline" and "online", respectively. In the revision, we will correct "off-line" and "on-line" to "offline" and "online", respectively.

P13. *In theorem 3.7, what is the meaning of the range of ϵ? It seems that the upper bound is far larger than the maximum value of Q functions, $\frac{1}{1-\gamma}$?*

A13. Thank you for raising the question. The bound on ϵ has been used for technical simplicity in the proof of Lemma B.2 in the Appendix. As the authors stated, the bound is larger than the maximum value of Q functions, and given the boundedness of the iterate Q_k, the requirement on ϵ can be removed when γ is close to one. According to the reviewer's feedback, the above points will further be clarified in the revision.

P14. *When translating the result into $|\mathcal{S}|, |\mathcal{A}|, \frac{1}{1-\gamma}$ and the authors suppose $d_{min} = \frac{1}{|\mathcal{S}||\mathcal{A}|}$. Without further assumption, we only know $d_{\min} \leq \frac{1}{|\mathcal{S}||\mathcal{A}|}$. Only in the case of uniform exploration, we can get $d_{\min} = \frac{1}{|\mathcal{S}||\mathcal{A}|}$. While one could question if the uniform exploration is efficient in exploration. And how well does uniform exploration solve the exploration and exploitation trade-off.*

A14. Thank you for raising the question. We note that we can consider $d_{\min} = \frac{1}{|\mathcal{S}||\mathcal{A}|}$ only for the purpose of theoretical comparisons of the sample complexity of the proposed method with those based on the generative model settings. This is because with $d_{\min} = \frac{1}{|\mathcal{S}||\mathcal{A}|}$, one gets the best the sample complexity in terms of the derived bound. This also implies that the uniform sampling of the state-action pair is potentially one of the best exploration strategies. Regarding the exploration-exploitation trade-off, $d_{\min} = \frac{1}{|\mathcal{S}||\mathcal{A}|}$ implies that actions are selected uniformly, and hence, it solves the exploration problem, while exploitation is no issue in Q-learning setting because it is an off-policy learning. We will clarify the confusions in the revised manuscript.

P15. *The lower bound in Li et al. [2023] is an algorithm specific lower bound, which only suits their specific algorithm. If the authors want to claim their result matches the lower bound, an algorithm-specific lower bound should be derived.*

A15. We appreciate your valuable feedback. The comparison with Li et al. [2023] is as follows: As the authors stated, we did not claim to have proven a lower bound for our algorithm. However, given that our algorithm is an extension of model-free Q-learning, and our result aligns with its lower bound, we expect the sample efficiency of our algorithm to be at least as sample-efficient as model-free Q-learning under relaxed conditions. We will clarify the confusions in the revised manuscript.

Refrences

[1] Zhang, Zihan, Yuan Zhou, and Xiangyang Ji. "Almost optimal model-free reinforcement learningvia reference-advantage decomposition." Advances in Neural Information Processing Systems 33 (2020): 15198-15207.

[2] Shalev-Shwartz, Shai. "Online learning and online convex optimization." Foundations and Trends® in Machine Learning 4.2 (2012): 107-194.

[3] Lee, Donghwan. "Final Iteration Convergence Bound of Q-Learning: Switching System Approach." IEEE Transactions on Automatic Control (2024).

[4] Mohammad Gheshlaghi Azar, Rémi Munos, and Bert Kap- pen. On the sample complexity of reinforcement learning with a generative model. arXiv preprint arXiv:1206.6461, 2012.

[5] Zhang, Zihan, Yuan Zhou, and Xiangyang Ji. "Model-free reinforcement learning: from clipped pseudo-regret to sample complexity." International Conference on Machine Learning. PMLR, 2021.