Title: Learning a Diffusion Model Policy from Rewards via Q-Score Matching

Reviewer:
Raman Kumar Jha
rj2712

Summary:
The paper solves the challenge of optimizing policies in off-policy reinforcement learning (RL) using diffusion models. Existing approaches fail to fully utilize the score-based structure of diffusion models, relying instead on simpler methods like behavior cloning, which limits their effectiveness in actor-critic settings. The authors introduce Q-Score Matching (QSM), a novel policy update algorithm that aligns the score of the diffusion model policy with the action gradient of a learned Q-function. This approach leverages the unique structure of diffusion models and avoids the computational overhead of differentiating through the entire diffusion model evaluation, focusing only on its denoising component. The resulting multi-modal and explorative policies make them well-suited for continuous action spaces. This innovation makes sense as it addresses both computational efficiency and policy expressiveness. Experiments demonstrate that QSM achieves or surpasses state-of-the-art performance (e.g., SAC, TD3) on various continuous control tasks while sample-efficient.

Strengths:

1. The paper reframes policy optimization in RL by linking diffusion model scores to Q-function gradients, offering a new geometric perspective.
2. QSM introduces a unique method that only requires differentiating through the denoising model, reducing computational complexity while maintaining expressiveness.
3. Extensive experiments show that QSM outperforms or matches popular baselines like SAC and TD3 across diverse tasks. It is particularly effective in sample efficiency and learning multi-modal policies.

4. The approach inherently supports multi-modal policy learning, enabling better exploration in continuous domains compared to Gaussian-based methods

Yes, the paper proposes a new problem formulation by introducing optimization policies in off-policy reinforcement learning (RL) using diffusion models. It incorporates diffusion models and uses Q-Score Matching (QSM) to take a new approach to the problem. The model and the experiments seem ingenuine to me, as the paper has provided a clear theoretical explanation of all the sections. The results remarkably stand out as it surpasses state-of-the-art performance on various continuous control tasks.

Weaknesses:

Please describe the weaknesses of the paper, and explain why. Does the premise make sense? Is the methodology comprehensive? Are there any experiment comparisons that are missing to prove the point? What are the potential limitations of the approach? Be critical but back up your point.

1. While comparisons with SAC and TD3 are thorough, the evaluation against other diffusion-based RL methods (e.g., Diffusion-QL) could be more detailed to highlight relative advantages beyond computational efficiency12.
2. Although computationally lighter than alternatives, scaling QSM to high-dimensional state-action spaces or deeper diffusion models might introduce challenges not fully addressed in this work1.
3. The paper briefly mentions alternative exploration strategies but does not delve into their potential impact on performance, leaving room for further investigation

Yes, the premise makes sense to me, as it is novel in terms of off-policy reinforcement learning (RL). Yes, I find the methodology comprehensive. In the paper, The noise model is fixed throughout the experiments, which may limit adaptability in environments, so this experiment is missing.

Possible Future Extensions:

1. Extension of the QSM can be performed, so that it can optimize both the score and noise components of the diffusion model for more adaptive exploration strategies. The method can increase computational complexity as a risk.

2. Another extension is to Investigate how QSM performs in high-dimensional state-action spaces, like robotics, by introducing hierarchical or factorized diffusion models. The risk in this method is the loss of sample efficiency.
3. Exploration of the connections between QSM and maximum entropy RL frameworks could be performed to explicitly balance exploration and exploitation. The risk includes additional theoretical development.

Conclusion:

The paper makes significant contributions by introducing QSM, which significantly improves the algorithm for training diffusion model policies in off-policy RL settings. Its strengths lie in leveraging the structural advantages of diffusion models while maintaining computational efficiency and supporting multi-modal policies. However, limitations such as scalability concerns and limited exploration strategies leave room for improvement. If encountered during a review process, I would give this paper a strong acceptance due to its strong theoretical foundation, innovative approach, and empirical success despite some areas for further development.