To Reviewer #3: Thank you for your careful reading and thoughtful reviews.

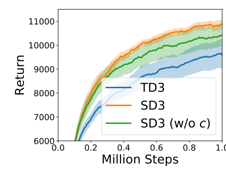
Q1: Theorems 3 and 4. (i) Theorem 3: Theorem 3 shows that SD2 helps to reduce the overestimation bias compared with DDPG. We empirically show that SD2 does not underestimate and can reduce the absolute bias in Figure 4. It will be interesting to further study the theoretical problem in future work. (ii) Clarification of Theorem 4: We clarify the correctness of Theorem 4 as below. The left-hand side in Eq. (19) equals to $\mathbb{E}\left[\mathcal{T}_{\mathrm{TD3}}(s')\right]$. Since TD3 uses target policy smoothing (which adds a sampled noise to the action) when estimating the value of s' in implementation, $\mathbb{E}\left[\mathcal{T}_{\mathrm{TD3}}(s')\right]$ is exactly the averaged Q-value and Eq. (19) holds.

Q2: The rate in example 1. We present example 1 to show that our bound is tight when β is large. We will clarify this in the revised version to make this point clear.

Q3: What if $\ln F$ is negative in Theorem 2? In the case where $\ln F$ is negative, the last term on the right-hand side of Theorem 2 still converges to 0, as $\ln F$ is bounded for any given positive ϵ. Thus, the error between the value function induced by the softmax operator and the optimal one can still be bounded and controlled.

Q4: Clarifications of condition and definitions. (i) Condition on the action set $\mathcal{A}$. It is required to be bounded, and we will clarify this point in the paper. (ii) Definition of bias. As defined in line 188 in the text, it denotes the difference between the estimated value of the next state induced by the operator $\mathcal{T}$ and the true value of the next state. (iii) Definition of $\mathcal{T}_{\mathrm{SD3}}$. The definition is given in Eq. (18) in the appendix, and we will formally define it in the main text.

Q5: How is the performance of the proposed approximation method? The results in continuous control tasks can validate that the proposed practical approximation method achieves good performance. We also conduct an ablative experiment which studies the effect of noise clipping that we proposed for a robust estimate of the softmax Q-value in HalfCheetah-v2. As shown in the figure, SD3 outperforms its counterpart without using noise clipping as expected. We will discuss it in the paper.



To Reviewer #4: Thank you for your careful reading and valuable comments, and we greatly appreciate your suggestions! We will clarify the details, include a high-level figure for the structure, and polish the figures to make the fonts larger in the revised version.

Q1: A more fundamental view why softmax alleviates both over- and under-estimation problem. We appreciate the suggestion, and it is an interesting direction to unify SD2 and SD3 into a same framework that leverages softmax and single or double critics to study the effects on value estimations. We will try to further investigate it in future research.

Q2: Related works about ensemble methods. Thanks for the suggestion, we will definitely incorporate the discussion and connection with ensemble methods in the paper.

To Reviewer #6: Thank you for your detailed evaluation of our paper and thoughtful reviews, and the comments are greatly appreciated! We will restructure Section 4.2 to make it more clear.

Q1: About the action space. (i) Requirement: Yes, the action space is required to be bounded. Thanks for pointing this out, and we will clarify this in the paper. (ii) Large action space: For large action space, the gap will also approach to $\epsilon/(1-\gamma)$ as β increases. It will be an interesting direction to further improve the theoretical bound.

Q2: Clarifications of notations and the algorithm. Thanks for pointing these out, and we will clarify them in the revised version. (i) The term c in Theorem 3. Yes, it refers to the noise clipping in Section 4.2 (line 179), based on which we defined the SD2 operator. Theorem 3 proves that the SD2 operator defined in the paper with this form helps to reduce the overestimation bias compared with DDPG. (ii) The $(1-d)$ term in the algorithm box. The notation d refers to the boolean type done signal, i.e., whether the step is the end of an episode. (iii) Importance sampling in the algorithm box. We will elaborate the details for computing softmax with importance sampling in the algorithm box.

To Reviewer #7: Thank you for your careful reading and thoughtful reviews.

Q1: Clarification of the significance of Theorems 1 and 2. Thanks for the question. The reason why the theorem in Song et al. shows that the bound converges to 0 (which considers the discrete case) while the bound in our paper does not, is due to the critical difference between continuous and discrete action spaces, where we have discussed the difference in Appendix A.1.1. We show that for any $\epsilon > 0$, our bound can converge to $\epsilon/(1-\gamma)$, which can be arbitrarily close to 0.

Q2: Does the first term in the bias definition depends on θ^{true}? Please note that $\mathbb{E}\left[\mathcal{T}(s')\right]$ is determined by the target policy network and the target value network with parameter θ^{-}, and does not depend on θ^{true}.

Q3: How to choose the parameter β? For implementation, we use grid search to find the best value of β to trade-off between the bias and variance of value estimates, as discussed in lines 296-299 in the text. It is also interesting to study an adaptive scheduling strategy of β, and we leave it as a future work.