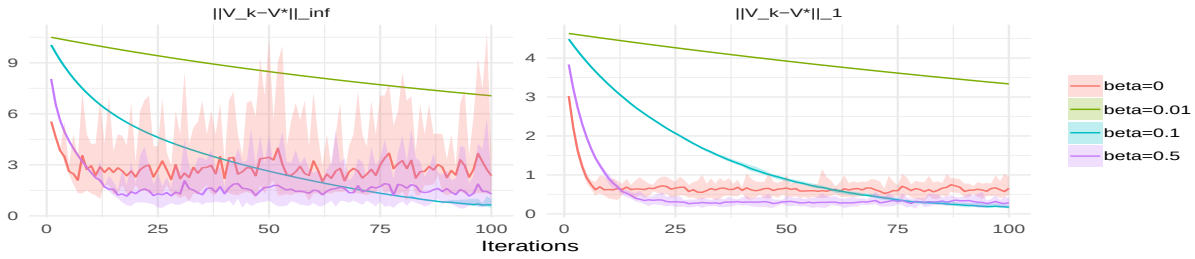
We thank the reviewers for their valuable feedback. In this rebuttal, we address the reviewers' comments and questions.

[R1] ***"Line 122: it's not obvious to me that the smoothed bound is more stable"*** The bound (8) stabilizes training by multiplying the error term E_N by a smoothing factor $\beta \in (0, 1]$ at a given initialization. If β is sufficiently small, then the bound is dominated by the initialization term that decreases at rate $\tilde{\gamma}$, slower than the standard AMPI scheme since $\tilde{\gamma} \geq \gamma$. The bound (8) does not contain γ in the numerator – that was a typo in paper that we corrected in the Appendix, so the Eq. (26-27) are correct. The reason is that in the proof of Lemma 1, line 366 contains the loss at $N - 1$ step.

[R1] ***"Very simple experiment, would greatly help the reader"*** We include in this rebuttal a numerical illustration of the smoothing technique (these results have also been integrated in the manuscript). We run experiments on a toy stochastic gridworld problem with the evaluation step error due to the sampling of state-transitions. We plot the average performance loss over 30 runs with varying values of smoothing factor β. As can be seen from the figure below, smaller values of β result in tighter confidence intervals, but slower convergence speed.



[R1] ***"The above also applies for the discussion on overestimation"***. The advantage of the bound (16) is that the term E_N could be made small by appropriately tuning the temperature parameter α_t to get the regularization gap $\Omega_t^*(A_{V_t})$ to match the overestimation error ϵ_t. In exchange, the A_N term remains and results in a different regularized fixed point.

[R1] ***"The above applies for the combined smoothness + regularization algorithm"*** The combined bound (19) is beneficial since (a) it downweights terms $\tilde{E}_N$ and $\tilde{A}_N$ by a factor of $\beta \in (0, 1]$, (b) allows to adjust the temperature parameter α_t to match the noise level, making the term $\tilde{E}_N$ even smaller.

[R1] On (24) and the approximation $K(t) = K + \mathcal{O}(m^{-1/2})$. We use the standard NTK arguments: if we taylor-expand $K(t)$ (defined in (24)) around the initialization parameters $\theta(0)$, then all terms of order 1 and higher are $\mathcal{O}(m^{-1/2})$. Thus, $K(t) = K + \mathcal{O}(m^{-1/2})$, where $K = K(0)$. The use of $\mathcal{O}(\dots)$ is in the following sense: the constants are absolute (don't depend on m). This is because we suppose the initial NN parameters $\theta(0)$ are made of random matrices with iid $\mathcal{N}(0, m^{-1})$ entries.

[R1] On the proof of Theorem 2. The disappearance of u_j in the first inequality in lines 482 – 483 is a typo; the stray term $+e^{-\lambda_{\min}(K)}$ is a typo too; these will be corrected in the manuscript. We repeat the corrected argument for lines 482 – 483 here. Define $v_j(t) := u_j^T(V(t) - b_{k+1})$, where $V(t) := V_{\theta(t)}$ as usual. Then, the ODE just before line 482 can be rewritten as $\frac{d}{dt}v_j(t) = -\lambda_j v_j(t)$. Integrating this ODE w.r.t time t gives $v_j(t) = e^{-\lambda_j t}v_j(0)$ $\forall t \in \mathbb{R}$. Taking absolute values gives $|v_j(t)| = e^{-\lambda_j t}|v_j(0)| \leq e^{-\lambda_{\min}(K)t}|v_j(0)|$, and plugging-in the definition of $v_j(t)$ above then gives $|u_j^T(V(t) - b_{k+1})| \leq e^{-\lambda_{\min}(K)t}|u_j^T(V(0) - b_{k+1})|$, $\forall j \in [\![S]\!]$, $t \in \mathbb{R}$, and large m. Because the eigenvectors $u_1, u_2, \dots$ are pairwise orthogonal, we deduce that $\|V(t) - b_{k+1}\|_2 \leq e^{-\lambda_{\min}(K)t}\|V(0) - b_{k+1}\|_2$, and so $V(t)$ converges to b_{k+1} (in any norm, since all norms are equivalent in finite-dim. spaces) exponentially fast.

[R1] ***"Missing work"***. We agree that other perspectives on the entropy regularization are worth mentioning, such as the ones you suggest on the loss landscape smoothing and value averaging effect of the KL divergence regularization.

[R1] ***What is "overwhelming probability"?*** Our use of this term was a misnomer. Indeed, those approximation statements are to be understood in an *almost-sure* sense i.e $\mathbb{P}(\text{statement holds for large } m) = 1$, over all random initializations of the parameters of the neural network V_θ. We note that we might be able to have more quantitative (nonasymptotic) statements using the results from [1], as suggested by **[R2]**.

[R2] ***"What is the norm in Cor. 1 and how that is derived from Thm. 1 with the projection error?"*** Cor. 1 bounds the norm of the approximation error vector ϵ_{k+1}^a that appears in Thm. 1. Due to the norm equivalence in finite dimensions, Cor. 1 holds in any norm (at the price of changing the constants in the $\mathcal{O}(m^{-1/2})$; this constant is $\sqrt{S}$ for the ℓ_∞-norm). Cor. 1 follows from Thm. 2, and should be understood in an almost-sure sense: $\mathbb{P}(\|V_{\theta_{k+1}} - b_{k+1}\|_\infty = \mathcal{O}(m^{-1/2})) = 1$, over random initializations of the neural net V_θ.

[R3, R4] Indeed, we analyze the smoothing update on state values instead of weights. As R4 mentioned, this can be argued away using Lipschitz constants. We will add a discussion about that.