We thank all reviewers for their insightful comments and the time they have spent carefully reviewing the paper. Consistent among all reviewers is the comment that the paper could be improved with further experiments. In response to this, we reiterate that our aim was to provide a novel framework with a theoretically sound interpretation of RL as inference that simultaneously identifies and addresses the shortcomings of existing work while opening up new classes of algorithms within this space that others can build upon. Our experiments were designed to provide empirical evidence that our approach does not harm performance compared to state of the art, to support our theoretical claims and demonstrate acceptable performance even when the most extreme approximations are used. While we feel the submitted version already contains more than a conference paper's worth of material, we are already running some additional experiments which, time and space permitting, we will include in the final version.

We will now address individual reviewer comments:

In response to Reviewer 1's third comment about modelling of entire trajectories in MERLIN, the algorithms for MERLIN can all be obtained by considering the joint objective:

$$\mathcal{L}(\omega,\theta) \coloneqq \mathbb{E}_{s\sim d(s)}\left[\mathbb{E}_{a\sim\pi_\theta(a|s)}\left[\frac{\hat{Q}_{\omega,soft}(h)}{\alpha}\right]\right],$$

where the variational distribution is $q_\theta(h) \coloneqq d(s)\pi_\theta(a|s)$, the temperature constant is α and $\hat{Q}_{\omega,soft}(h)$ is the parametrised approximation for the soft action value function. The above is equivalent to max-entropy formulation in [Reinforcement Learning and Control as Probabilistic Inference: Tutorial and Review, Levine 2018] with a variational approximation for the policy. However as the variational policy trains towards the Boltzmann distribution of the soft action values with temperature α, this indirectly takes into account the dependence of the entire trajectory on the policy via $\hat{Q}_{\omega,soft}(h)$, and thus inadvertently ends up modelling entire trajectories despite grounding the MDP dynamics.

In response to Reviewer 2's comment regarding the recursive definition of ϵ_ω, as we discuss in Appendix C, the definition of ε_ω is recursive but only if using the simple Bellman operator for the Boltzmann policy. We introduce and detail more complex operators in the set $\mathbb{T}$ that don't give a recursive definition in the Appendix. An example is the optimal Bellman operator, which results in a Q-learning algorithm. In Appendix F.2, we introduce another operator that recovers the optimal Bellman operator in a limit of sequences. Exploring further operators and investigating whether for any flexible Q there exist one (or many) consistent softmax temperatures $\varepsilon_\omega > 0$ when using the simple Bellman operator for the Boltzmann policy is an exciting line of theoretical research for us, but one we feel is best saved for future work.

In response to Reviewer 2's comment regarding comparisons to schemes where the adaptive entropy coefficient is annealed according to a schedule or optimisation scheme, as we discuss in Appendix B, our formulation has a unique Bayesian interpretation in that the entropy penalty is annealed according to the model uncertainty in the optimality of $\hat{Q}_\omega(h)$. We thank the reviewer for drawing our attention to the references [A Theory of Regularized Markov Decision Processes, Geist et al. 19] and [Soft Actor-Critic Algorithms and Applications, Haarnoja et al. 19], the former only having been published since submitting to NeurIPS, and will extend the discussion accordingly.

Addressing Reviewer 2's second comment under the Quality section about function approximation for variational policies, we implied that function approximation offers the choice to obtain arbitrary rich classes of variational distributions that can, in principle, model the posterior conditional of action given the state exactly [Variational Inference with Normalizing Flows, Rezende, 2015], instead of the simpler parametrisation involving Gaussian transformations. We would like to clarify that the class of variational policies used in our experiments are the same for SAC and VIREL (both using multi dimensional independent Gaussians), thus the experiments indeed demonstrate performance gains from adaptive regularisation. We will clarify this difference in the paper.

Finally, we would like to thank Reviewer 3 for their careful analysis of the paper and will consider their sensible suggestion of moving details from Appendix F3 into the main paper if the NeurIPS format permits. They are correct in pointing out a reference typo in Section 5; we will update the paper accordingly.