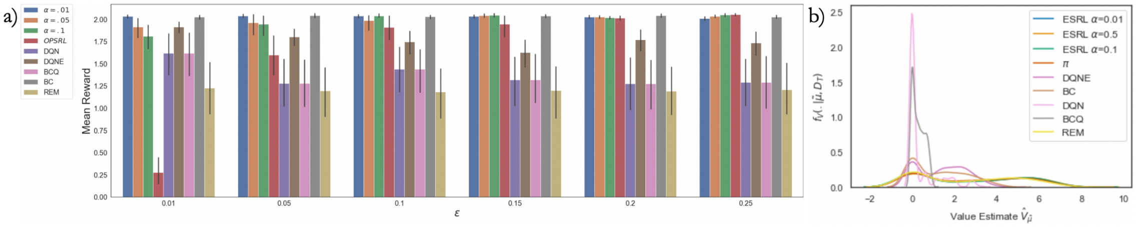
We thank the reviewers for their constructive criticism, we address the common and particular points below:

Baselines: Following the suggestions, we have added Taxi-v2 and FrozenLake-v0 environments as they have discrete state and actions spaces. Results will be shown in the main manuscript appendix. Additionally, we added the following offline methods for comparison: a) an ensemble of K DQN models (DQNE); b) behavior cloning (BC), c) Batch Constrained deep Q-learning (BCQ), and d) Random Ensemble Mixture (REM). We also include discussion on these and imitation learning methods to the updated manuscript. Fig. 1 a) Shows mean reward on Riverswim for offline training on T=200 episodes, we vary stochasticity in behavior policy π. Fig 1 b) shows posterior distribution for $\hat{V}(\hat{\mu})$ on Sepsis data. ESRL performs well across quality in π in a) and b). BC does well in Riverswim but not on Sepsis, on the contrary BCQ, and REM do better on Sepsis: a more complex and higher variance setting with more training data.



Reviewer 1: Thank you for pointing out the typos and hard to follow notation, we have corrected and simplified these to make derivations easier to follow. **1)** Regarding the p-value in Algorithm 1, we apologize for an unfortunate typo ($\mu_{\mathbb{K}}$ should be $\hat{\mu}_k$) and the confusion this caused. To clarify, we first estimate $\hat{\mu}(s,t)$ with sample $\mathcal{I}_1$ and use that to compare $Q_t^{(k)}(s, \hat{\mu}(s,t))$ and $Q_t^{(k)}(s, \pi(s,t))$ where $k \in \mathcal{I}_2$. It is correct that we are simply counting, but the subtlety is that the cross sample strategy allows for $Q_t^{(k)}(s, \hat{\mu}(s,t)) < Q_t^{(k)}(s, \pi(s,t))$. **2)** With respect to the combination of the expert and learned policy, the proposed strategy of using confidence intervals (CI) is appealing. We argue that both this strategy, and our method which relies on the p-value are equivalently optimal. As any $(1-\alpha)\%$ CI can be inverted to obtain an associated α-level hypothesis test, it can be shown that measuring CI overlap is equivalent to a hypothesis test. **3)** You mention an important example: when the expert action is optimal but presented in a high variance setting. The information theoretical lower bound tells us that in this high variance settings, the high noise-to-signal ratio will lead towards failure of rejecting H_0 due to lack of statistical power. ESRL would reject H_0 and choose the expert's action, same decision would come from a CI comparison, which in this case would be highly overlapping.

Reviewer 2: Thank you very much for the thorough suggestion on baselines, please see the results above for the methods you proposed. We cite the relevant papers in the main text. **1)** You raised an important question on access to $f(\cdot|\mathbf{D}_T)$ for sampling MDPs in real-world problems and Atari. As data increases, dependency of results on the prior decreases, so we can use any *working model* to approximate the MDP. Several models are computationally simple to sample from, and can be used for learning. For example, we use the Dirichlet/multinomial, and normal-gamma/normal conjugates which work well for all settings in the paper. Chosen priors should be flexible enough to capture the dynamics and easy to sample from efficiently. **2)** Regarding the Expert supervision in the title, we allude to the healthcare setting where physicians are experts, but contrary to imitation learning, their actions might be far from optimal.

Reviewer 3: Thank you for the careful look at the theoretical results. **1)** We certainly agree that the prior specification involves a strong assumption. However this is usual the case with all model based RL, and in particular the Bayesian aspect of our method alleviates this issue as the posterior distribution is robust to prior model specification. This follows as the posterior distribution of samples M_k will concentrate around M^* as sample size increases. The error induced from using M_k instead of M^* vanishes, as shown in the proof of Theorem 3.4. Please see response to Reviewer 2.1 for discussion on choosing a model. **2)** As you mention, ESRL works for stochastic behavior policies as well, thank you for pointing this out. **3)** The $\mu_{\mathbb{K}}$ in Algorithm 1 should be $\hat{\mu}$. **4)** Thank you for suggesting OPPE benchmarks, we added two non-parametric approximate-model comparisons to the manuscript, plots are not shown here due to lack of space. **5)** For Sepsis we visualize the $\hat{V}(\hat{\mu})$ in figure 4(c) in the manuscript, and have changed the caption to make this clearer.

Reviewer 4: Thank you for the suggestions on adding more comparison methods and environments, please see Baseline section for results, we cited the suggested papers in the main manuscript. **1)** Since our main motivation is the healthcare domain, we chose the Sepsis data set as it is standard in a lot of prior work on RL for healthcare (please see [1],[2],[26] in the manuscript). The common baseline is the physician's policy π. We chose Riverswim as it is has discrete state and action spaces, and requires deep exploration to reach high rewards. We added openAI's Taxi-v2 and FrozenLake-v0 to the paper. **2)** We added more discussion on Theorem 3.4, mainly we highlight how it shows ESRL is sample efficient, and flexible to risk aversion and stochasticity level of π. **3)** We believe that the risk aversion parameter α is actually a benefit; it allows the method to be tailored to different scientific and business applications where one might have different tolerance towards risk in search for more reward. **4)** We have improved readability and captions of the figures.