Thank you for your in-depth and constructive reviews, they will give us an excellent chance to further improve the paper. We address the reviewers concerns individually below. We will not address all the typos and editing catches, but will fix all of these in the final draft.

Reviewer 1:

- The goal here is to compare model-free methods, augmented with a buffer. A careful comparison between model-free and model-based approaches for OPE would be interesting and extremely valuable for the community, but is beyond the scope of this paper.

- It is a good suggestion to provide intuition for the proofs. We will include such a discussion in the camera-ready, which allows for an additional page.

Minor Concerns: We will address these concerns in the revision. We will improve notation consistency and clarity and include diagrams of the maze in the appendix.

Reviewer 2:

- SIR is a general strategy; in fact, there are a number of similar approaches with different names [Smith and Gelfand, 1992]. The main novelty here is investigating its use in RL, where the online setting requires us to consider a moving window dataset—rather than a fixed batch.

- The theoretical comparison of the bias of IR and WIS-Optimal is natural, because we show they are equal. IS is unbiased, so that comparison is not interesting. In practice, though, we cannot actually use WIS-Optimal, as it is a full batch approach. Empirically, then, it makes sense to compare to other mini-batch methods, like IS. We did not compare to WIS-minibatch due to the poor empirical performance, likely due to the additional bias of that estimator.

- $\bar{\rho} \approx \mathbb{E}[\rho(a|s)] = \mathbb{E}[\frac{\pi(a|s)}{\mu(a|s)}] = \sum_{s,a} \frac{\pi(a|s)}{\mu(a|s)} \mu(a|s) d_\mu(s) = 1.$

- Assumption 1 is common for analyzing OPE estimators. The idea is that we are effectively sampling from the stationary distribution, even though we know we are in Markov settings. An important next step is to consider alternative noise assumptions in sampled data.

- The result in line 203 is that we can directly use the prior results to look at the expected difference in variances over many buffers (i.e. these statements say our result holds across buffers of smaller sizes).

- Because $\bar{\rho} \approx 1$. When ρ is lower than the average it will make the rhs a large number, but when ρ is greater than the average we expect it to lower the rhs of the equation. As learning progresses, we expect the samples w/ high ρ to learn more quickly (thus having lower error). 'mean' is the correct one.

- We also looked at "softer" target policies, where similar conclusions can be drawn (see appendix). All the results presented in the appendix are qualitatively similar.

- The parameter sensitivity provides more information, because it gives some sense of how these might perform in practice for realistically chosen parameters, rather than optimal parameters. We will include the learning curves in the appendix for completeness.

- MARE is Mean Absolute Return Error. We use MARE when it is not tractable to compute the value function using dynamic programming or analytically (and otherwise MAVE).

Reviewer 3:

For O1, you are correct in your understanding. We will use some of the additional space in the camera-ready to include a brief discussion on O1 and O2.

Q1: The Steps corresponds to the Number of Interactions with the environment. The agent can update more or less frequently than every step. The Number of Updates for Figure 1(a) is once every 16 Steps. We also show performance for different Update frequencies in Figure 1(b).

Q3: Because the experiment is run 100 times, like in all the plots in figure 3, the error bars are not visible. The parameter sensitivity plots could provide some information about variance of the updates. If the variance of the updates is higher, we expect the magnitude of the largest updates to also be higher. This means a lower step size is needed to prevent divergence. A wider trough of the sensitivity curve could reflect lower variance in the updates, though as acknowledged in the paper, this is very much a proxy and we cannot make any strong conclusions based on it.

C1(f): See point 3 for Reviewer 2.

Q2, C1(a-e), C2, and C3: we will take all these points into consideration, and work to maximize clarity in the final revision. L287 - You are correct, this should be figure 3.