We thank the reviewers for thoughtful reviews and encouraging comments. We respond only to questions and concerns.

(R1) "In the introduction . . . ": Good suggestion. Will refer specifically to CDRD, CDRK in Table 1a. The point is extending the cited importance sampling methods to RL would lead to convergence rate that *deteriorates* in horizon.

(R1) "Something is said in Remark 2 . . . ": Our analysis handles known and unknown behavior policy and simply take as a condition the nuisance estimation rate. Knowing the behavior policy can help estimate nuisances. In experiments we consider known behavior policy as is common in offline RL. We still need to estimate $w_t^{\mathcal{K}}$ even if behavior is known.

(R1) "Out of the 9 combinations, . . . ": We focused on these as they represent the two extremes. We can easily provide in the supplement a general analysis of all combinations under the intersection of the assumptions need for each extreme.

(R1) "In the beginning of section 4, . . . ": Unfortunately, no, as smoothness conditions are necessary in continuous action space, else the finite data may be unrepresentative of the infinite possible unseen actions. This is the same as in density estimation. We will comment on this.

(R1) "Although half of the paper . . . ": With limited space, we thought actual *learning* would be of greatest interest. We will follow your suggestion and add to the supplement experiments for policy evaluation and varying H. Note this is easy with the submitted code – we will just run it and report the results.

(R1) "The paper is well written . . . In general, it would be interesting . . . ": The derivation is different and more complicated than [3] as in [3] the density ratio exists and is used directly. On the other hand we need to analyze the errors due to the kernelization, which complicates the analysis as it introduces slower leading terms with rate that depends on horizon dimensionality.

(R1) "Typos: . . . isn't the inverse?": Thanks for catching. Yes; the latter is a typo; it is the reverse.

(R2) "The experimental setting . . . ": The qualitative results are the same as we vary these. We will run the (submitted) code with a range of parameters and include additional plots in supplement.

(R2) "The overall structure are well planned and the paper is well written . . . ": We will add reminders of notation when used for first time much after problem setup and add short descriptions of steps in equations in appendix.

(R2) "Line 72–73 . . . ": Yes; it's a typo; "latter" and "former" should be exchanged.

(R2) "Line 81–82 . . . ": By Radon-Nikodym thm, exists f such that $\mathbb{E}_{\pi_e}[g(a) \mid s] = \mathbb{E}_{\pi_b}[f(a)g(a) \mid s]$ for all g measurable if and only if $\pi_e(\cdot \mid s)$ is absolutely continuous wrt $\pi_b(\cdot \mid s)$ (this is for each s). However, if $\pi_e(\cdot \mid s)$ is discrete, it is ***not*** enough for behavior to have positive density at its atoms. E.g., Dirac at 0.5 is ***not*** abs cts wrt the uniform distribution on $[0, 1]$ even though latter has density 1 at 0.5. Recall $\mu \ll \nu$ means $(\mu(A) > 0 \Rightarrow \nu(A) > 0,\ \forall A$ measurable), so if $\mu \ll \nu$ and $\mu(\{0.5\}) = 1$ then $\nu(\{0.5\}) > 0$, i.e., has an atom at 0.5. Will add this example.

(R3) "The novelty of . . . ": We respectfully disagree. Not only do we make it doubly robust *and* extend it to RL, we also analyze it and give rates under lax conditions and show the naïve extension yields very bad rates as horizon grows.

(R3) "The evaluation is . . . ": Indeed while we avoid curse of dimension in state space and horizon, we may suffer from it in action space, since we kernelize actions. In many practical offline RL settings, however, states are complex (e.g., many+rich health indicators) and actions simple and often one-dimensional (e.g., insulin dosing/timing). We will comment on this and run the (submitted) code on growing action dimension and add to supplement to visualize this.

(R4) "The expected return . . . ": In OPE, MSE is actually the metric of interest. In offline learning, policy value is indeed of interest and our policy gradient experiments (Sec 5) showcase how low-error gradients lead to high-value learning. Moreover, as Remark 8 mentions, error bound on gradients can be combined with standard gradient ascent analysis to get value regret guarantees: we simply replace the stochastic-policy gradient error bounds of Kallus & Uehara '20 with our new ones in Thms 11–13 therein; while this is straightforward use of existing work, we'll flesh this out more explicitly in supplement for completeness.

(R4) "The only concern . . . ": The primary contribution is a theoretical study of rates and the simple experiments are intended only to illustrate the new theory. Extensive experimentation is beyond the scope of such a short paper with so many new results already. We can nonetheless easily run our (submitted) code on the Warfarin dosing experiment of Kallus & Zhou '18 (the code is also public) and add to the supplement.

(R4) "The paper is well-organized and well-written . . . more intuition explanations . . . explain more on the comparison results in Table 1": Thank you; we will use this feedback to further improve the clarity. We will add *more* in-words explanations of each result and we will move Table 1.