We thank the reviewers for their close reading, detailed comments, and overall positive assessment. We address the questions raised by each reviewer separately.

Reviewer 1 Thanks for the appreciation and the concrete suggestions for paper refinement. We will include the discussion and summary section for the final version.

- **VC dimension in finite-sample CI.** Since the estimator relies on some function approximator, the finite-sample correction unavoidably relies on the complexity measures due to the union bound. We exploit the VC dimension due to its generality. In fact, the bound can be improved by considering a data-dependent measure, *e.g.*, Rademacher complexity, or by some function class dependent measure, *e.g.*, function norm in RKHS, for specific function approximators. We will include these discussions in the final version.
- **Finite-horizon CoinDICE.** While we mainly focus on infinite-horizon MDPs with a discounted factor, the dual method can be adapted to finite-horizon settings straightforwardly. For example, we have the finite-horizon d-LP as $\max_{d_h(s,a):S\times A\to\mathbb{R}_+} \sum_{h=1}^H \mathbb{E}_{d_h}\left[r_h\left(s,a\right)\right] \quad \text{s.t. } d_{h+1}\left(s,a\right) = \mathcal{P}_*^\pi d_h\left(s,a\right), d_0\left(s,a\right) = \mu_0\left(s\right)\pi\left(a|s\right),\ \forall h\in\{1,\dots,H\}.$ with the corresponding CoinDICE CI estimation as
$$[l_n, u_n] = \left[\min_{w\in\mathcal{K}_f} \min_{\beta_{h=1}^H\in\mathbb{R}^p} \max_{\tau_{h=1}^H\geqslant 0} \mathbb{E}_w\left[\ell_H\left(x;\tau_{h=1}^H,\beta_{h=1}^H\right)\right] \quad \max_{w\in\mathcal{K}_f} \max_{\tau_{h=1}^H\geqslant 0} \min_{\beta_{h=1}^H\in\mathbb{R}^p} \mathbb{E}_w\left[\ell_H\left(x;\tau_{h=1}^H,\beta_{h=1}^H\right)\right]\right],$$
where $x := \left\{(s,a,r,s',a',h)_{h=1}^H\right\}$, $\ell_H\left(x;\tau_{h=1}^H,\beta_{h=1}^H\right) := \sum_{h=1}^H \tau_h r_h + \sum_{h=1}^H \beta_h^\top \Delta_h\left(x;\tau_h,\phi\right)$, and $\Delta_h\left(x;\tau_h,\phi\right) := \tau_h\left(s,a\right)\phi\left(s',a'\right) - \tau_{h+1}\left(s',a'\right)\phi\left(s',a'\right)$.
- **Empirical comparison.** We compared with the WIS computed the same as in DualDICE/GenDICE papers and Liu et al. (2018), which is found to work best out of all IS variants. In preliminary experiments we had assessed the performance of other IS-based estimators as well as the clipped estimator from Thomas and Hoeffding's bound, but found them all to yield worse empirical coverage. For simplicity we did not present it in the paper. We did not include DR as a baseline, as the paper's focus is confidence interval estimation in infinite-horizon RL when marginalized importance ratios are used. DR requires extra information about the model and/or value function, and will make the comparison less clean. That said, an extension of CoinDICE to DR is indeed an interesting direction for future work.
- **True value** We did use the simulator to compute the groundtruth for the Bandit, FrozenLake, and Taxi environments. However, for Reacher, we use a large ensemble of learned networks in order to account for issues in approximation error.
- **Related work.** Regression IS: We focused comparisions with IS on scenarios where the behavior policy is known for simplicity. Our comparisons show that CoinDICE's performance is better even when compared to these strong baselines, suggesting that the IS ratio used in IS baselines likely leads to higher variance and thus looser bounds.

  Self-normalized Importance Weighting (SN) (Kuzborskij et al. (2020)): Thanks for pointing out this related work, which was released after the NeurIPS deadline. The major differences between CoinDICE and SN lies in three aspects: **i)**, the CoinDICE is designed for RL setting while the SN is proposed for contextual bandit setting. It is not clear how well a straightforward extension of SN to RL will work; **ii)**, CoinDICE is behavior-agnostic while SN needs to know the behavior policy; **iii)**, CoinDICE is asymptotically pivotal, meaning that there are no hidden quantities we need to estimate, while SN requires several unknown quantities, therefore, is not pivotal.

  We will include these discussions and comparisons in our final version.

Reviewer 2 We will provide further experimental details and empirical comparison in our final version, taking advantage of the extra page.

- **More empirical performances with different behavior policies:** With different behavior policies, the proposed CoinDICE achieves similar performances. The comparison with 100 samples from behavior policy choosing the optimal arm with probability 0.2 is illustrated in Figure 1 due to space limitation. We will include the complete comparison in our final version.

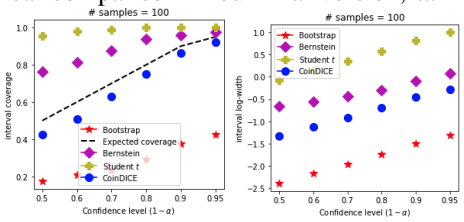

Figure 1: Additional comparison.

- **Behavior policies:** The behavior policy in FrozenLake is the optimal policy with 0.2 white noise, which reduces the policy value dramatically, from 0.74 to 0.24. For the behavior policies in Taxi and Reacher, we follow the same experiment setting for constructing the behavior policies to collect data as in DualDICE and Liu et al. (2018).

Reviewer 4 **Computation of the CoinDICE:** Although the derivation of CoinDICE is nontrivial (which is part of our major contributions), the optimal w is given in Eq. 11, and instantiated for various f-divergence choices in Appendix D.2. The final optimization problem is in a mini-max form (Eq. 14), which can be solved by stochastic gradient methods. Full details are provided in Appendix D.3 with pseudocode (Algorithm 1). We will add more descriptions in the final version using the extra page.