We thank the reviewers for their careful and valuable feedback. We address their main points in our comments below.

[R1] How the trade-off affects practical performance. This is a good point; we appreciate you raising it. Our paper does *not* seek to make a strong theoretical statement that trading off contraction & bias definitively leads to performance gains. Instead, we focus on identifying the potential operator trade-off with SIL and proposing a generalized n-step alternative which opens doors to further exploit the trade-offs. Though empirical evidence suggests that fast contraction tends to practical gains, there is little theoretical explanations, even with [Rowland et al, 2019] which motivates our work. We speculate that the theory is not straightforward because both us and [Rowland et al, 2019] focus on policy evaluation, while the the full algorithm interleaves with optimization, which greatly complicates the analysis. Therefore, we leave a more comprehensive study for future work. Empirically, we find in Table 1 (App. D) that n-step SIL with $n = 5$ outperforms $n = 1$. This is consistent with results from prior work that fast contraction tends to empirical gains.

[R1] Disconnect between theory & experiments. As an ideal operator, importance sampling (IS) achieves the fastest possible contraction (i.e. it contracts to the fixed point with one iteration) and is unbiased. However, its stochastic estimate has high variance and is rarely used in practice (see [Munos et al, 2016]). As a result, we believe that IS is not the best performing model in practical experiments. Consistent with this argument, most prior work also consider the high variance a major downside of IS [Munos et al, 2016; Rowland et al, 2019]. We will discuss more in the revisions.

[R1] Comparison to unbiased methods. For the continuous control tasks that we consider, the baseline algorithm TD3 adopts a deterministic policy and it is not feasible to apply Retrace (which requires stochastic policy to perform truncated IS). As an alternative unbiased baseline, the uncorrected $n = 1$ step generally underperforms $n = 5$-step SIL.

[R2] Performance. Thanks for raising this issue. From Fig 2, though SIL with $n = 5$ does not outperform *all* the baseline alternatives on *all* tasks, it consistently ranks as top two among the majority of tasks, indicating its more stable performance. For fair evaluations, we believe it is not reasonable to require n-step SIL to outperform the *best* among all other alternative baselines on every task. Instead, we believe it is more reasonable to compare n-step SIL with alternative baselines on a one-to-one basis – by such a metric, the improvement is clear. For example, n-step SIL clearly outperforms n-step uncorrected on 6/8 tasks while outperforms vanilla SIL on 7/8 tasks.

[R2] Random seeds & SAC. We agree that running more seeds potentially leads to more accurate assessments. However, we highlight that despite a relatively small number of seeds, in Fig 2 most curves are well separated, indicating statistically significant differences. Note also that the highly cited PPO paper uses 3 seeds across all experiments. Regarding SAC: We did not include SAC baseline for a few reasons: (1) Though we propose a maxent lower bound in Thm 1, all theories on the trade-offs of operators are exclusively derived in the conventional RL setup (including results from [Rowland et al, 2019]). As a result, we do not think comparing to SAC would offer much insights as to echo the theory; (2) We speculate that applying the n-step technique in Thm 1 to SAC might not lead to significant gains out of the box, as it might be sensitive to the entropy coeff. In fact, [Oh et al, 2018] derives the SIL formulation under maxent RL, but the entropy term is dropped when calculating the lower bounds in their implementation. In Fig 1, we provide SAC results, which mostly underperform n-step SIL, especially on DM suite. We speculate this is because SAC hyperparams have been commonly tuned on gym envs. This corroborates our speculation that SAC performance might be sensitive to the entropy coeff.

Tasks	SAC
DMWalkerRun	23 ± 1
DMWalkerWalk	87 ± 83
DMWalkerStand	440 ± 87
DMCheetah	3 ± 1
Ant	2645 ± 1462
HalfCheetah	11451 ± 406
Ant(B)	808 ± 29
HalfCheetah(B)	914 ± 251

Figure 1: The n-step SIL outperforms vanilla SAC on most tasks.

[R2] Montezuma. We did not include Montezuma as we initially could not replicate the results of [Oh et al, 2018]. We speculate that with proper tuning, n-step SIL should outperform typical baselines but might slightly underperform return-based SIL. This is partly because when rewards are sparse, using returns as lower bounds might be more accurate than using learned bootstrapped values. As a result, return-based SIL [Oh et al, 2018] might still be more suitable for tasks with highly sparse rewards as in Montezuma. However, we believe this does not undermine results in this paper, where we highlight the gains of n-step SIL on tasks with dense rewards & midly sparse rewards (delayed rewards).

[R3] Variance of the estimator & related work. This is a good point, we will discuss more details in the revisions. There a few reasons why the variance is not explicitly addressed: (1) Uncorrected n-step & SIL remove all IS ratios, which arguably greatly reduces the variance compared to IS based methods, e.g., Retrace. This is in line with arguments made in prior work such as [Rowland et al, 2019] where the variance is not addressed explicitly; (2) Though from each (x, a) pair there is only one trajectory, the bootstrapped values at the end of the n-step are learned and could interpolate between different pairs, which leads to more accurate estimates; (3) Particular to the continuous control tasks where both dynamics and policy are deterministic, one-sample estimate could have relatively low variance. See more discussions at line 263-275. Regarding related work: we are aware of the duality & state marginal method to off-policy evaluation. We will include them as related work and leave their combinations with SIL as future work.

[**ALL REVIEWERS**] We appreciate the other points you have raised that we cannot address in this one-page response; they improve our manuscript and we will adjust our text based on your comments.