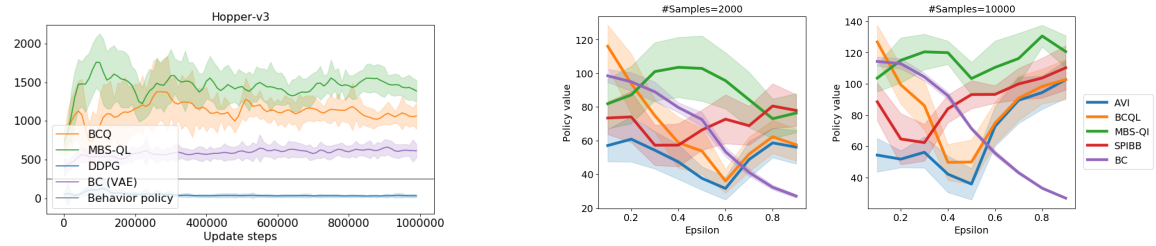

Figure 1: **Left**: Same setting as Figure 4, with 15 seeds. **Mid and Right**: Same setting as Figure 3 (left), with error bars and result in low data regime. (Results with $N = 1000, 5000$ are similar and will be included in revision.) Larger set of b for baselines: BCQ: {0, 0.05, 0.1, 0.2}, SPIBB: {1e-4, 5e-4, 1e-3, 5e-3, 1e-2}, MBS: {5e-4, 1e-3, 5e-3}

We thank the reviewers for the excellent feedback that helped us improve the manuscript. Our revision includes (1) additional experiments to address reviewer's concern about variability and rigor. (2) study of algorithms' behavior as the amount of data changes, (3) clearer description of the experiment setup, especially how baselines are tuned, (4) more emphasis on the theoretical contributions. We acknowledge that experiments with larger domains will be nicer, but note that (1) many related works have only tabular theory with heuristic extensions to function approximation. We give sound theory for the significantly harder function approximation setting, carefully accounting for all the error sources and these results are certainly the primary focus of this work. The experiments are intended to be illustrative proofs-of-concept. (2) Estimating state action visitation distributions is an active research area (e.g. Batch Stationary Distribution Estimation by J. Wen et al.) and MBS is composable with better estimators. We aim to study MBS in higher dimensions in a more empirically focused follow up.

(R1) Same as truncating conditional state-action probabilities in deterministic dynamics? No! Censoring based on conditional probabilities can be viewed as MBS with a very naive assumption that $\mu(s)$ is uniform. Using an empirical estimate $\hat{\mu}(s)$ through counts in discrete and density estimation in continuous domains is much nicer, and we capture the effect of modeling error through ϵ_μ in our bounds. Even for deterministic domains, many (s, a) pairs can lead to the same s'. Section 3 Figure 1(b) is an example where transitions are deterministic but truncating $\mu(a|s)$ can fail with any reasonable threshold ($\leq 0.5 = \mu_{max}$ in this case).
(R1) Comparison with Kumar et al. We discuss this in lines 222-224 (re the $f(\epsilon)$ term). There is no fundamental difference between the two notions of concentrability, and our improvement is algorithmic, not just in the analysis.
(R1) Comparison with residual gradient. We view residual gradient as a different optimization method of the AVI/API objectives, but still requires concentrability without additional work. Sure MBS is composable with it.

(R2) Are baselines in the experiments under-tuned? No! We use the "imperfect-imitation" experiment setting from the BCQ paper, their code as well as their hyper-parameter ranges. The only difference is that we run BCQ for more update steps (our BCQ learning curve shows the performance drops a lot after 300k steps). For comparison, the D4RL paper reports BCQ performance on their "medium" dataset close to our result ($1000 \sim 1500$). This observation shows that, without reliable stopping criteria for batch RL some non-conservative algorithms can eventually deteriorate. SPIBB and MBS use the same hyper-parameter ranges. Appendix E.2 may have caused confusion – that separate experiment for b-ablation searches over a larger set of b. We re-ran SPIBB for the larger b set, and BCQ with more b values as well and observe similar curves. We also replicated all experiments across more random seeds and different data regimes (see the figures above). Trends are the same as those reported in the paper.

(R3) What policies lie in the restricted policy set? Yes it's non-trivial to estimate it. However, rather than assuming that all available policies satisfy concentrability as in prior works, we provide an *adaptive* scheme which competes with all policies satisfying this assumption. Since the set is always non-trivial for small enough b, the algorithm can always output something sensible (In contrast, we don't know any guarantee of FQI when concentrability coefficient is infinite). Also our algorithm doesn't need to know Π_C^{all}. Corollary 1 covers the special case (learning $\pi^\star$) that shows the major theoretical advance in our result: only requires coverage for $\pi^\star$ instead of coverage for all $\pi \in \Pi$.

(R1) Can we compare with BEAR? Yes, for the final version. We saw persistent gradient explosion when running the authors' code in this BCQ experiment setting and have extensively debugged the issue with the authors. **(R2) D4RL?** If the included experiments are not adequate proofs-of-concept, we can include D4RL results.

(R2, R3) Missing baselines (SPIBB/BCQ) in Figure 2? Each example in Figure 2 is intended to demonstrate the weakness of one type of algorithm (BCQ/BEAR and SPIBB). In the other figure they will perform as well as MBS, we include both methods in both figures as per R2 in the revision.

(R2, R3) I.i.d. and positive rewards assumption? We (and prior theoretical works) use the i.i.d. assumption for cleaner analysis. In batch RL, data comes from a fixed Markov chain – so use of standard martingale concentration is feasible. If the minimum value V_{min} is known, we can set all blocked back-ups to V_{min} for negative reward settings.

We will incorporate other reviewer suggestions in the final version. Please consider raising the score if we addressed some of the concerns.