We thank the reviewers for their constructive feedback. **We have reformulated the theoretical statements to position them with previous work, clarified notation and surface-level inconsistencies in theoretical statements, added experiment baselines, and performed an additional comparison with SPIBB (Laroche et al. 2019).** The main contribution of our work is a *practical* offline deep RL algorithm, CQL, which attains strong results and outperforms prior methods by a large margin on a wide range of tasks. We have revised the paper to frame the theoretical results as providing motivation for our approach rather than a primary contribution.

Additional experiments & baselines: Fu et al. (2020) have added results for AWR, BCQ, REM and AlgaeDICE in the D4RL paper, which we now include in Tables 1 & 2. We have added an explicit mention that the baseline numbers are from Fu et al. 2020 **[R2, (R3, W4)]** and will add variance measurements **[R2]**. CQL outperforms AWR, BCQ, REM and AlgaeDICE **(R1)** in **26/29** tasks, often by a large margin. **(R3)** We also compared CQL to **SPIBB** on the Helicopter task from Laroche et al. 2019. With a dataset of size 10k, CQL outperforms SPIBB by attaining **4.11** mean return whereas SPIBB (w/ best $N_\wedge$) attains **3.22** return and soft-SPIBB (w/ best ϵ) attains **3.65** return.

R1/[R3, W1]: Policy improvement result, what is CQL doing. Based on R1/R3's requests, we have added a new theorem for policy improvement property of CQL that is more consistent with prior work. Similar to Thm. 1 in Laroche et al. 2019, we show that the CQL updates (Eqn. 2) converge to the optimal policy of a (empirical) penalized RL objective, i.e., $\pi_{\text{CQL}} = \arg\max_\pi \hat{J}_{\mathcal{D}}(\pi) - \alpha \mathbb{E}_{\mathbf{s}\sim\hat{d}^\pi}[D_{\text{CQL}}(\pi, \pi_\beta)](1-\gamma)^{-1}$, where $\hat{J}_{\mathcal{D}}(\pi)$ is an empirical estimate of expected discounted return and D_{CQL}, defined in lines 649-650, captures a notion of action distribution shift. Then, using the terminology from Laroche et al. 2019, we show that π_{CQL} is a ζ-approximate safe policy improvement over π_β (i.e. $J(\pi_{\text{CQL}}) \geq J(\pi_\beta) - \zeta$) with high probability $1-\delta$, where

$$\zeta = \frac{\gamma C \sqrt{\log(1/\delta)}}{(1-\gamma)^2} \mathbb{E}_{\mathbf{s}\sim\hat{d}^{\pi_{\text{CQL}}}(\mathbf{s})} \left[\frac{\sqrt{|\mathcal{A}|}}{\sqrt{|\mathcal{D}(\mathbf{s})|}} \sqrt{D_{\text{CQL}}(\pi_{\text{CQL}}, \pi_\beta)(\mathbf{s}) + 1} \right] - \overbrace{\left(\hat{J}_{\mathcal{D}}(\pi_{\text{CQL}}) - \hat{J}_{\mathcal{D}}(\pi_\beta) \right)}^{\geq \alpha \mathbb{E}_{\mathbf{s}\sim\hat{d}^{\pi_{\text{CQL}}}(\mathbf{s})}[D_{\text{CQL}}(\pi_{\text{CQL}}, \pi_\beta)(\mathbf{s})](1-\gamma)^{-1}} .$$

This follows from applying and extending the tools in Achiam et al. 2017. Note the similarity with Thm. 2 in (Laroche et al. 2019) and that we avoid the ∞-norm term in (Petrik et al. 2016). Intuitively, this indicates that the return of π_{CQL} is higher than the behavior policy π_β w.h.p. when the empirical policy improvement (i.e., the α term) exceeds the sampling error, which diminishes to 0 as the dataset size (i.e., $|\mathcal{D}(\mathbf{s})|$) grows.

R2: Behavior policy estimation. Since our submission, Nair et al. 2020 and Ghasemipour et al. 2020 have discussed the difficulty of behavior policy estimation at length. We will add extended discussion of this point in the paper.

R1/(R3, W2): Clarifying notation and resolving surface level inconsistencies. We have resolved the notational confusion and inconsistencies in the results. We discuss the main changes briefly:

- **(Also R3, C5)** We now explicitly indicate dimensions of vectors, matrices and scalars. Briefly, $[\mu(\mathbf{a}|\mathbf{s})/\pi_\beta(\mathbf{a}|\mathbf{s})]$ denotes a vector of size $|\mathcal{S}||\mathcal{A}|$ equal to elementwise ratio of μ and π_β and this is multiplied to the matrix $(I - \gamma P^\pi)^{-1}$.
- **(Also R3, C1)** $\pi_\beta(\mathbf{a}|\mathbf{s}) = 0$: We have added a discussion to indicate that when $\pi_\beta(\mathbf{a}|\mathbf{s})$ is zero in the tabular setting, the learned Q-values, $\hat{Q}(\mathbf{s}, \mathbf{a})$ can be $-\infty$. For Thm 3.1 and 3.2, we have edited these to include the assumption that $\pi_\beta(\mathbf{a}|\mathbf{s})$ has non-zero density on all actions to prevent $-\infty$ values, and clarified that this result holds only for $\mathbf{s} \in \mathcal{D}$.

R1/[R3, C3]: Tightened bounds with counts. For simplicity, we had omitted the count terms from the main text, instead deferring discussion to App. D.3. However, as the reviewers point out, they are necessary. We now include the count terms that scale inversely proportional to the square root of $|\mathcal{D}(\mathbf{s}, \mathbf{a})|$ in the main text. As expected, the sampling error and theoretically safe value of α both decrease as the dataset size increases. In practice, the theoretical value of α is overly-conservative, so as done in previous work, we use a smaller α rather than this theoretical value.

(R3, W3): Relation to prior work. We have added and positioned our ideas with respect to the suggested references on robust MDPs, CPI, safe PI, high confidence PI, batch RL **(Also R1)**, including **12** references from non-Alphabet authors. We suspect that CQL outperforms robust MDPs empirically due to the approximations necessary to adapt robust MDPs to a practical algorithm. We have also added an empirical comparison to SPIBB.

R1/[R3, C2]: Offline RL and state distribution shift. We have clarified these statements (e.g., lines 83-84) to say that *training procedure* for a Q-learning algorithm queries out-of-distribution actions which can lead to divergence.

(R3, C4): Lines 847-848: no guarantee $(\mathbf{s}', \mathbf{a}') \in \mathcal{D}$?. To clarify, the proof uses concentration of $\hat{T}(\cdot|\mathbf{s}, \mathbf{a})$ which depends on the counts $|\mathcal{D}(\mathbf{s}, \mathbf{a})|$. So while the inequalities involve $(\mathbf{s}', \mathbf{a}')$ through $\mathbb{E}[\hat{Q}^k(\mathbf{s}', \mathbf{a}')]$, we bound that term by $2R_{\text{max}}(1-\gamma)^{-1}$ avoiding concerns over the size of $|\mathcal{D}(\mathbf{s}', \mathbf{a}')|$. We explicitly mention this now.

R1: Intuition for gap expanding. We have elaborated on this in the paper now. By increasing the difference between Q-values at in-distribution actions (π_β) and under learned policy (μ_k), CQL constrains the resulting policy to lie in the support of $\mathcal{D}$. This controls the $D_{\text{CQL}}(\pi, \pi_\beta)$ term that appears in the sampling error component of ζ in our ζ-safe policy improvement result above. Empirically, this also makes CQL more robust to Q-function approx. error that makes OOD actions appear more optimal and results in better performance than policy constraint methods (App. B, Fig. 2).