Reviewer 1: We thank the reviewer for the thoughtful questions and comments, which we address below.

○ Adding a matrix to u_t can, in fact, help with the skewness in the projection. The obvious choice is $\phi(x_t, a_t)\phi^T(x_t, a_t)$, but the resulting algorithm does not retain the convergence guarantees. In that case, the convergence analysis presents the same difficulties of standard Q-learning. We also explored adding the matrix $\text{diag}(\phi(x_t, a_t))$, resulting in an algorithm that has tabular Q-learning as a particular case, concerning the provided limit solution. Again, convergence does not hold so generally.

○ We appreciate the reviewer's suggestion that the stability analysis of the fast o.d.e. (for v) can be simplified and will do so in the final version of the document. We also agree that the original Q-learning with function approximation (3) is not true gradient descent on the Bellman error and will correct the statement in the final version of the document.

○ Regarding the martingale difference sequences, we believe that our proof (Appendix B.3) establishes this assertion from Assumption (I). Nevertheless, we thank the suggestion of using Bhatnagar's recent works, which could provide an elegant alternative proof.

○ We break down the comparison between our assumptions and the ones from [13] in three, namely (i) technical assumptions; (ii) assumptions on the data; (iii) assumptions on the features. (i) Both works use standard conditions to facilitate the convergence analysis; (ii) [13] requires uniform ergodicity of the sampling policy, which is stronger than our assumption; (iii) while [13] requires features such that, in terms of $\Sigma_\mu = E_\mu[\phi\phi^T]$, the sampling policy is very close to the optimal policy, our assumption is met through normalization alone. For example, the assumptions of [13] do not hold on the $\theta \to 2\theta$ example in Section 4, contrarily to ours.

Reviewer 2: We thank the reviewer for the thoughtful questions and comments, which we address below.

We start by clarifying that Theorem 2 is not a convergence result, but rather an error bound. We will make this explicit in the final version of the document. Concerning the generality of Assumption (IV), particularly why we can assume that $\Sigma_\mu = E_\mu[\phi\phi^T]$ has equal non-zero elements in the diagonal, we note that Assumption (II) requires Σ_μ to be non-singular, which is equivalent to linear independence of the features. As such, we can assume that the features are orthogonal without loss of generality.[1] Each element in the diagonal of Σ_μ is thus the norm of the corresponding feature, in the norm induced by the inner product $\langle f, g\rangle = E_\mu[f^T g]$. We can thus scale all features appropriately to ensure (IV).

Reviewer 3: We thank the reviewer for the many thoughtful questions and comments, which we address below.

○ Regarding the advantages of our approach in comparison with Gradient Q-learning (greedy-GQ, [12]), our approach converges to a unique, well-defined solution that is independent of the initialization, whereas GQ may converge to any local optimum (see discussion in Section 1) and our assumptions are similar. If the reviewer refered instead to Gradient $Q(\sigma, \lambda)$ (GQ(σ, λ): A Unified Algorithm with Function Approximation for Reinforcement Learning), we notice, for instance, that their convergence result is established assuming the iterates remain bounded (Assumption 2(3) of their work), which we prove to be true in our case (Appendix C).

○ Although technically Assumptions (I), (II) and (III) are sufficient for convergence, the reviewer is right in that the policy used to collect the data is fundamental, since the limit solution directly depends on it through the distribution μ.

○ We acknowledge and will incorporate the clearer introduction to Assumption (IV) suggested by the reviewer.

○ The use of a two time-scale algorithm replicates the update structure of DQN: in DQN, a target network is mostly fixed and updated only infrequently; in CQL, the target network is updated on every time step, but very slowly (Section 2.1). The two time-scale formulation thus "mimics" the dynamics of the target and main networks in DQN, while being amenable to analysis using results from the stochastic approximation literature.

○ CQL builds directly on two key elements of DQN: experience replay (Assumption (I)) and a target network (u). Although the actual architecture of the two approaches is distinct (ours is linear, while DQN is non-linear; the target network is also not exactly the same), we still believe that our analysis of CQL provides theoretical insight regarding the two aforementioned elements of DQN.

○ Finally, regarding the experiments, we did perform an empirical sensitivity analysis to the learning rates, which showed CQL is robust to such variations. We can include those results in the supplementary material. Also, in the mountain car experiment we actually performed 10 runs, not 3. We thank the reviewer for bringing to our attention that this is not clearly phrased in the paper.

Reviewer 4: We gratefully acknowledge the reviewer's comments and suggestions, particularly with respect to the experimental section. We will incorporate these into the manuscript, since we agree that these help to make a clearer experimental section (namely with respect to the mountain car experiment).

○ Regarding the choice of parameters, we refer to our rebuttal to Reviewer 3. Adding to that, the choice of learning rates will be formalized through the complete data collection method, from the work the reviewer fittingly suggested (Jordan et al. (2020)).

○ We agree that the motivation for mountain car is inevitably weaker, given the focus of our work on convergence issues. We acknowledge the suggestion of presenting more detailed performance metrics, such as learning curves. We can add these to the supplementary material and, to the best of our ability and within space limitations, to the main paper.

○ Regarding the use of radial features, we use them out of their simplicity and natural compliance to Assumption (IV). On one hand, if their radius is small and the data is sampled uniform over states and actions, Assumption (IV) immediately follows. On the other hand, as the radius increases, the "violation" of Assumption (IV) increases accordingly, which facilitates assessing the impact of Assumption (IV) on the performance of CQL.

[1] If not, we can use Gram-Schmidt's method to obtain equivalent orthogonal features.