We thank all the reviewers for their helpful comments and suggestions which will substantially improve our manuscript. Before we respond to specific comments, we want to address a high level concern which was brought up by several reviewers regarding what we learn about RL from studying finite sample bounds for LQR. We note that most of the prior work in LQR has been focused on model-based methods. Motivated by the popularity of model-free methods in practice, our goal was to compare classic model-free algorithms and see how they measure up against the model-based methods on LQR. The results in this paper suggest that there is a substantial sample complexity hit when using model-free methods. We believe there are two high level takeaways: (a) we give a theoretical understanding of the trade-offs between policy evaluation and policy improvement for LQR, which is a core theoretical question in RL, and (b) we believe that there are small deltas to the problem setup studied in this paper which might result in model-free algorithms being more competitive with model-based algorithms, such as partial observability or introducing simple non-linearities; it is conceivable that analyzing model-free methods with these small deltas could heavily build on our tools and analysis. Finally, we hope that this line of research motivates further study into when to use model-based methods vs. model-free methods for more general RL.

Reviewer #2. Regarding the projection step, in practice we find that the projection step improves performance by helping to stabilize the algorithm in the beginning iterations. While the update defined by (2.10) would technically work as long as Q_{22} is invertible, a greedy policy improvement step with respect to a quadratic value function that is not positive definite is not well-defined.

Regarding the quantity μ, we note that μ does not actually need to be estimated since we simply set it to the minimum of $\lambda_{\min}(S)$ and $\lambda_{\min}(R)$, and the cost matrices are assumed to be known.

Next, we remark that LSPIv2 is defined in the main text in Algorithm 2 that starts on Line 122.

Finally, regarding whether $K^{(i+1)}$ is stable, we note that a non-trivial part of the regret proof is dedicated to ensuring that $K^{(i+1)}$ is stable. Assumption 1 is the main abstraction we use that allows us to analyze the meta-algorithm in Algorithm 5. It says that the batch learning algorithm EstimateK we use has a $O(\varepsilon)$ guarantee on sub-optimality using $O(1/\varepsilon^2)$ samples. The point of the online algorithm is to allow us to use any batch learning method which satisfies this guarantee and send ε to zero in a way that incurs sub-linear regret. Using Assumption 1, we carefully work through the perturbation analysis (c.f. perturbation related results in Section E) to ensure that enough data is collected to maintain stability. Finally, we show in Lines 718-724 that LSPIv2 satisfies Assumption 1 (by Theorem 2.2) and make the constants $C_{\mathrm{req}}, C_{\mathrm{err}}$ explicit.

Reviewer #4. We hope that our comments in the first paragraph address the main concerns with our work.

Reviewer #5. Regarding if the control signal u_t is noisy for both K_{play} and K_{eval}, the q parameter in Theorem 2.1 reflects the parameter of the Q-function in (2.3) where the policy $\pi(x) = Kx$. It is correct that K_{play} uses noise for exploration, but K_{eval} is not considered a stochastic policy (c.f. the definition of ψ_t in Line 86). We will make this more clear in the writing. As for exploration noise in K_{play}, if K_{play} had no exploration noise then the associated covariance matrices for LSTD-Q would be degenerate.

Regarding the generality of Theorem 2.1, Theorem 2.1 only applies to the situation where the data is actually coming from an LQR system and hence the solution to the Bellman equation actually has a quadratic function solution. It is not a general result for LSTD-Q, but specific to the estimator (2.5).

Regarding letting σ_w to zero, if we let σ_w go to zero, then the RHS of (2.9) goes to zero, meaning there will be perfect estimation of q. This is to be expected, since in a deterministic system we should be able to recover exactly the q parameter after $O((n+d)^2)$ samples (as long as the covariates are non-singular). Perhaps the reviewer meant to refer to sending σ_η, the exploration parameter, to zero? In which case yes, if we do not properly excite the system, then we should not expect to be able to recover the q parameter (the associated covariance matrices will be denegerate).

Regarding closeness of K_N to $K_\star$, please refer to Lemma C.3 (which is Lemma 12 in Fazel et al.), which relates the error in the controller to the error in the value functions. It says that an $O(\varepsilon)$ error in $\|K - K_\star\|$ translates to an $O(\varepsilon^2)$ error in the value functions. In this paper, we focus on the sub-optimality gap $J(K) - J_\star$, which is the main quantity of interest in prior work for model-based LQR (e.g. Dean et al). The reviewer is correct to point out that this is not quite the same as focusing on $\|P - P_\star\|$ as in Tu et al.

Regarding L_∞ bounds on value functions, thanks for bringing the papers to our attention. We will remark about more refined L_p bounds on value functions and include the relevant citations in our revision.

Regarding the comment about Lines 25-28, I think our wording here is unfortunately a bit unclear. What we meant to say was that the behavior of model-free methods on LQR is much less well understood compared to the behavior of model-based methods on LQR, and this paper is an attempt to address this. We will clean up this phrasing.