We thank reviewers for carefully reviewing our paper and providing constructive feedback.

Responses to Reviewer 1. We thank the reviewer for pointing to related control papers and providing a perspective from the control theory side. We will discuss these related works in the revised version. Stability and safety are indeed important, and it may be possible to incorporate such constraints into our framework, where our uncertainty estimates can characterize safety/stability constraints and enable more cautious exploration. We think this is an important future direction. We will soften some of our claims and state that we focus on optimality (in the no-regret sense).

Due to space limit, below we response to a few main questions from R1 while addressing all questions in the revised version: (1) We consider discrete time, continuous space control; (2) "Provably correct" means no-regret with respect to the best policy in Π on the real system; (3) We didn't explicitly assume boundness on states as under Gaussian noise, states' norm are not bounded—instead, we use Asmp 2 and info-Gain γ_T: our regret bound still holds as long as $V_{\max}$ and γ_T are bounded even though states could be large. (4) the reviewer's interpretation of "online" is correct; (5) Asmp 2 is actually weaker than the boundedness cost/reward assumption in almost all theoretical RL regret analysis; (6) Specializing to LQR, Asmp 2 implies that Π is a subset of stabilizing policies (e.g., see Cohen et.al 19 for LQR case).

Responses to Reviewer 2. KNR is fundamentally different from LQR, as it can capture nonlinear f such as Piecewise Affine System, i.e., hybrid system, and higher-order polynomials. The consequences of this generalization are:

1. Planning itself (even non-optimistic planning) is computationally intractable. This means there is really no hope for provably computationally efficient algorithms, and heuristics must be required in practice.
2. Smoothness, which is understood to be an important aspect of continuous controls, is not a prior inherent in KNR as it is in the LQR. Gaussian noise provides smoothness, although other explicit smoothness assumptions (e.g., Smoothness w.r.t value functions [Osband & Van Roy 14]) can generalize to sub-Gaussian noises.
3. LQR's value function is quadratic and smooth (under stable linear π) which is leveraged in the LQR's analysis. KNR's value functions can be complicated and we need to derive the self-bounding lemma 3.7 (with a novel application of optional stopping time argument) to use the second moment of the realized total cost.
4. Our analysis is far more than GP-bandit: as we do not assume the cost is bounded as most bandit/RL works did, this requires a key new technique: the self-bounding lemma (Lemma 3.7). This new lemma also enables us to get a first-order regret bound scaling with $J^\star$ (Thm 3.6) which was missing even in prior LQR works.
5. Full nonlinear $f(x, u)$ without further structural assumption is not tractable even in statistics, as this generalizes infinite arm bandit problems.

Episodic vs single-trajectory: we believe our results can extend to single trajectory setting under stronger assumptions such as strongly stable system with stable controllers (e.g., Cohen et.al 19), and this is a direction for future work. If the closed-loop system is strongly stable, then single trajectory setting is similar to episodic setting, as the dependency between the current state and the previous states are diminishing exponentially fast (e.g., see Hazan et.al 19).There are already challenges in the finite horizon, which is relevant for many practical settings work. This is related to Remark 3.4 which we will make precise in the final version (along with the d factors); we will make a more careful elaboration based on the system parameters as how these factors scale (based on the assumptions in [Simchowitz and Foster [2020]).

Instead of directly assuming strongly stable system (Cohen et.al 19), we assume that any policy in the policy class Π has bounded second moment of their realized total cost (Assumption 2). When specializing to LQR, such strongly stable LQR systems imply our assumption, and hence our assumption is more general.

Regarding γ_T: we will provide γ_T for popular kernels such as RBF and Matérn. for RBF, it scales $O(\log(HT)^{d_x+d_a+1})$.

Regarding Thompson sampling: we use TS in experiments as it's a simple alternative of UCB-based approach. We do believe that by leveraging the framework from Russo and Van Roy [2014], we can obtain a Bayesian regret bound. [2] presents a frequentist regret bound for 1-d LQR. Frequentist regret bound for KNR is challenging due to the value function of KNR can be complicated. A frequentist analysis of TS for KNR is an interesting future work.

Responses to Reviewer 3. The PILCO implementation from Wang et.al uses GP with RBF kernel while our implementation uses RFF feature corresponding to RBF kernel. So the model class capacity is similar.

We thank the reviewer for pointing us to a related work. The major difference is that [1] assumes Lipschitzness in the one-step future value function (same as Osband and Van Roy 14). As we do not assume boundness on the realized total cost, such Lipschitz constant can be unbounded making the theorem in [1] vacuous. Indeed, it is this relaxation forces us to develop a novel technique (Lemma 3.9) so that our results do not require such Lipschitzness assumption.

We use Thompson sampling (TS) in the experiments as a simple alternative of UCB-based approach. The mean model is used as a baseline to show exploration via TS (based on the uncertainty ball from LC^3) helps. We require any π in Π to satisfy assumption 2. Mapping to LQR, this means that Π could set to be a subset of all linear controllers that contains only stabilizing linear controllers which is a common assumption used in LQR analysis (e.g., Cohen et.al 19).