We would like to thank the reviewers for their helpful comments. We will revise accordingly in the revision.

Reviewer 1: (General RKHS) Our KOVI algorithm can be applied for any RKHS in generalized. As shown in the discussion below Theorem 4.3, we can set $\beta = O(H\sqrt{\log N_\infty})$ and obtain a $H^2\sqrt{\log N_\infty \gamma_T T}$ regret, where N_∞ is the ϵ^*-covering number of the value function class in Eq. (4.2) in the ℓ_∞-norm with $\epsilon^* = H/T$ and γ_T is the effective dimension of the RKHS. Here β is set in this way to ensure optimism and N_∞ appears due to a uniform concentration argument. See Lemma C.2. We will revise Lemma C.2 for handling general RKHS in the revision.

(Eigen-decay conditions) As discussed above, our analysis can be applied to general RKHS. For the general case, we only need to bound $\log N_\infty$ and γ_T. When the kernel has rank r, $\log N_\infty = \tilde{O}(r^2)$ and $\gamma_T = \tilde{O}(r)$, where $\tilde{O}(\cdot)$ omits $\log T$ terms. Then we obtain a $H^2\sqrt{r^3 T}$ regret, which recovers the linear case in Ref [33]. When the kernel has polynomial eigen-decay ($\sigma_j \lesssim j^{-\nu}$, $\nu > 1$), our Lemma D.4 gives an upper bound of γ_T and our Lemmas D.2 and D.3 can be modified to bound $\log N_\infty$. It can be shown that NOVI also achieves sublinear regret when ν is sufficiently large.

(Concrete examples of kernels) The Gaussian RBF kernel on the sphere $\mathbb{S}^{d-1}$ satisfies Assumption 4.2 for any fixed $\tau \in (0, 1)$. (See {1}). Moreover the NTK induced by sinusoidal activations recovers the Gaussian RBF (See {2}). The ReLU NTK satisfy the polynomial eigen-decay condition with $\nu = 1/d$. We will add concrete examples in the revision.

Reviewer 2: (Computational complexity) The computation complexity of KOVI is dominated by solving HT kernel ridge regression (KRR) problems, each with no more than T data points. Thus, the total computation needed is $H\text{poly}(T)$. Moreover, with sublinear regret, to achieve any fixed accuracy level ε, it suffices to set $T = \text{poly}(H, 1/\varepsilon)$. Thus, the total computation needed to achieve ε accuracy is polynomial in H and $1/\varepsilon$ and is thus **efficient**. Moreover, in the low-rank and exponential eigen-decay cases, the regret is $\tilde{O}(\sqrt{T})$, thus T depends on $1/\varepsilon$ only through ϵ^{-2}. For polynomial decay we can also obtain a sublinear regret, which gives a $\text{poly}(H, 1/\varepsilon)$ computation complexity. For NOVI, it is well-known that gradient descent converges linearly in training overparameterized NN (Refs [2,3,4,23]). Since width m is polynomial in T and H, the computation in the neural setting is also $\text{poly}(H, 1/\varepsilon)$ and thus **efficient**.

(Assumption 4.1) **(i)** We would like to emphasize that the Bellman rank of the MDP model we consider is infinity as we consider a infinite-dimensional function class. Our model only fall in the low Bellman-rank framework when the RKHS kernel is low-rank. In this case, we recover the result in linear setting (Ref [33]). **(ii)** Even when restricted to the linear case, it seems that [Zanette 2020] did not show that Assumption 4.1 is equivalent to having linear transition. Instead, their Proposition 2 prove that when the inherent Bellman error is zero for all linear functions with parameters in the unbounded set $\mathbb{R}^p$, the transition is linear. However, we require the Bellman operator maps any bounded function with values in $[0, H]$ to a linear function with bounded parameters. [Zanette 2020]'s result does not apply. **(iii)** Our assumption is implicit as it does not assume the transition to have a particular form and only assume Bellman operator maps bounded functions to a bounded RKHS ball. **(iv)** Such an assumption is required because without any structural assumption, the regret lower bound is $\sqrt{|S||A|H^3T}$, which is infinity when $S \times A$ is an uncountable set. To have meaningful result, we need to assume the target function $\mathbb{T}_h^\star f$ belongs to a function class with bounded capacity (in terms of ℓ_∞-norm), which is standard in supervised learning. Here we consider the class of infinite-dimensional RKHS-norm ball. **(v)** The complexity of RKHS is determined by its eigenvalues, and is fixed once the RKHS is specified. Thus, it seems impossible to "reduce the complexity of kernel space". We show that the regret bound depends on such intrinsic complexity through the covering number N_∞ and effective dimension γ_T, both can be computed using the eigenvalues. Moreover, γ_T is previously used in analyzing the regret of kernel bandit (Refs [16,34,49]) and N_∞ captures the temporal structure of MDP, which also appears in linear MDP (Ref [33]). Our work extends previous work on linear setting to the infinite-dimensional kernel and neural settings with a general framework of regret analysis.

Reviewer 3: (Impact on RL practice) Most of the existing deep RL approaches adopt heuristic exploration strategies. NOVI can be readily incorporated into the framework of neural fitted Q-learning (NFQ) in practice, which is the batch version of DQN. NOVI proposes to add a bonus term to each NFQ-iteration. When using overparameterized neural networks, we have proved that such an exploration scheme solves the deep RL problem with sample efficiency.

(KOVI v.s. NOVI) The regret of NOVI is worse than that of KOVI by $\beta TH \cdot \iota$, which is negligible when m is a polynomial of T and H. Thus, KOVI and NOVI essentially have the same regret when m is large. Besides, KOVI solves kernel ridge regressions, which requires the closed form of the solution. In contrast, NOVI solves the least-squares problems using gradient descent and can be applied to the case where we do not know the form of kernel function K.

Reviewer 4: (Long horizon setting) We consider the episodic setting where H is fixed. In this case, the regret lower bound is $\sqrt{H^3T}$ (Ref [32]) and our upper bound is $\sqrt{H}$-larger in terms of H. The $H^2\sqrt{T}$-upper bound also appear in various previous works with (generalized) linear function approximation (Refs [33,58,64,65,66], their T is equal to our TH). Moreover, our algorithms can be modified for the infinite-horizon discounted or ergodic settings. In these settings, we only need to slightly modify Eq.(3.2) due to having different forms of Bellman equation. With the added bonus term, we can similarly establish the optimism principle and sample complexity upper bounds.

(Assumptions 4.2 and 4.5) The eigen-decay condition captures the intrinsic complexity of RKHS. As discussed in the first two points for **Reviewer 1**, our results can be extended to general RKHS by bounding the the covering number N_∞ and the effective dimension γ_T under different eigen-decay conditions. Our assumption of exponential decay is common in nonparametric statistics, which leads to an infinite-dimensional RKHS. We will add results for general RKHS in revision.

{1} Mercer's Theorem, Feature Maps, and Smoothing, Minh et al. ICML, 2006.

{2} Random Features for Large-Scale Kernel Machines, Rahimi and Recht. NeurIPS, 2007.