We thank all the reviewers for their time and valuable feedback. We will improve the draft carefully based on your comments.

Response to Reviewer 1

Intuitive interpretation of the dual kernel norm (Thm 3.2)? The purpose of Thm 3.2 is to clarify how the kernel Bellman loss is related to the error $\delta = V - V_\pi$. This is achieved by "shifting" the Bellman operator on $V(s)$ to the kernel, yielding an *adjoint operator* applied on the kernel (and hence the dual kernel). That said, the definition of the dual kernel is rather technically motivated. We tend to think it as *"the kernel whose kernel norm equals our kernel Bellman loss.*

The (dual) kernel norm is not directly related to L_p norm. But if the maximum (resp. minimum) eigenvalue of the kernel is positive, then the kernel norm can be upper (resp. lower) bounded by L_2 norm. We will include more intuitive interpretation in the revision.

Computation time While the exact gradient involves $O(n^2)$ computation, we can obtain stochastic gradients using mini-batches, whose computation does not scale with n. This process is straightforward for U-statistics. For V-statistics, we need to estimate its diagonal and off-diagonal parts separately to obtain an unbiased stochastic gradient. Details will be added in the final version.

Response to Reviewer 2

The connection to the nested optimization formulation of LSTD. We thank the reviewer for insightfully relating this work to nested optimization formulations of LSTD and BRM algorithms. We agree that they are closely related and will give a thorough discussion in the revision. But our approach is derived from a very different perspective and is both simpler and more practical for general nonlinear function classes, and differs in other ways including (1) the primal and dual spaces can be specified independently; (2) we specifically project the Bellman residual onto RKHS that results in a closed form solution; (3) algorithmically our method solves a single optimization problem without the two-step procedure in previous work; among others.

Meanwhile, we believe that we can develop results similar to our Corollary 3.3 to explicitly clarify the concrete relation between LSTD and kernel loss in some special cases. We will discuss this extensively in the revision.

The properties of the empirical loss are not shown. We agree with the reviewer's comments on the issue of biasedness (see also our remark in L116–123). We do want to emphasize the difference between unbiasedness and consistency here. Even though our estimator is biased with non-iid data (even by using U-statistics), we still yield a consistent estimator under standard regularity conditions of the Markov chain; this is much better than the squared TD error used by residual gradient, which is inconsistent.

Establishing statistical properties of the estimator is an interesting problem, and can be achieved with the now-standard methods. But the anslysis for the non-IID case is quite technical, and can be a distraction of this work's main focus. Therefore, we prefer to study it in a separate work that focuses on statistical guarantees and uncertainty estimation.

Interplay of the function space $V(s)$ belongs and the kernel corresponded RKHS. The role of the value function space (denote it by $\mathcal{V}$) and kernel space $\mathcal{H}$ are similar to the generator and discriminator spaces in GAN (Goodfellow et al. 14). They plays orthogonal roles, so it is not easy to say which is more important. But $\mathcal{V}$ and $\mathcal{H}$ do need to be compatible with each other in that sense that $\mathcal{H}$ should be chosen to include $\{V - \mathcal{B}_\pi V \colon \forall V \in \mathcal{V}\}$, which holds when $\mathcal{H}$ is universal.

Clarification of the paper by Ormoneit & Sen, 2002. We agree with the reviewer that the kernel smoothing used in "Kernel-Based Reinforcement Learning" is not the same as the more general kernel methods used in the paper and other works. It is mentioned in the paper as a related work, and we will make the distinction explicit.

The purpose of Thm 3.2. Thm 3.2 is meant to clarify how the kernel Bellman loss is related to the error $\delta = V - V_\pi$ of the values functions (see also our response to Reviewer 1). We will consider to reform Theorem 3.2 into a "Dual kernel interpretation" inside Section 3.2, so that it does not interrupt the presentation in Section 3.1.

Response to Reviewer 3

The meaning of plots in Figure 2(c) & (d). The red dots in Fig2(c) are the (*MSE, K-loss*) pairs obtained in the trajectory of our algorithm. The yellow dots in Fig2(c) are the (*MSE, L2-loss*) pairs obtained by in the trajectory of residual gradient. Fig2(d) is similar, but plot the (*Bellman-error, K-loss*) and (*Bellman-error, L2-loss*) pairs. As mentioned in L264–268, these scatterplots show good correlation between our K-loss and MSE/BE, suggesting it is a good proxy for learning the value function.