Many thanks for your valuable comments and suggestions. We will address the minor comments (such as to fix the typos, add the discussion about the parameters, and broaden the related works section) in the next version of the paper. We now respond to the major comments are as follows.

To Reviewer 2: We believe that the proposed techniques could be helpful in other RL settings such as continuous state spaces. Take RL with the linear model as an example. In this case, to learn Q^*, it suffices to learn the d-dimensional vector w^* (which we name as pseudo value function) because $Q_h^*(s,a) = \phi(s,a)^\top w_h^*$, where $\phi(s,a)$ is the d-dimensional feature vector of (s,a). To reduce the estimation variance via reference-advantage decomposition, we first aim at learning a rough estimation of w^* as the reference pseudo value function, using time steps that are only polynomially depending on A, H, d, $1/\epsilon$, $\ln(1/p)$, $\ln T$, where p is the failure probability, and ϵ is the accuracy parameter. More formally, we believe that the key is to prove an analogue of Lemma 5 for the linear model. We will add this discussion in the next version of the paper.

We will also discuss the work on policy certificates (Dann et al., 2019) in related work section.

To Reviewer 3: At line 397, we mentioned that $[c_1, c_2, c_3] = [2, 2, 5]$ is sufficient. The safe choice in the current version of the paper for c_4 is $c_4 = 57603$, which seems quite large because we bound $(1 + \frac{1}{H})$ by 2 in many places for convenience. When H is sufficiently large (say greater or equal to 200), we can bring the choice of c_4 to around 200. We will add this discussion ot the next version of the paper.

To Reviewer 4: Thank you for your suggestion. We will include the table for all notations in the next version of the paper.