We thank the reviewers for constructive comments and questions.

One common major concern of the reviewers is whether EBF can be efficiently implemented. We have been thinking about this problem since the submission deadline and have made substantial progress in designing a practical version of EBF algorithm. To solve the optimization problem in line5 Algorithm1 efficiently, we relax some constraints (e.g. π_k is deterministic), and apply EVI alike techniques to compute the constrainted optimal policy in the extended MDP. However, we have not finished the details completely before the rebuttal deadline. As a result, we have to give up adding the implementation of EBF algorithm to this submission.

Other concerns are answered as below. The typos, wrong references, missing references and unclear expressions like Def. 4 would be fixed carefully in the final version (if possible):

To reviewer1:

1. About intuition: we can catch more information about h^* from the history trajectory (See line1 Algorithm2). One important difference to previous methods is that, the order of samples in the trajectory matters in EBF algorithm, while previous methods only use $N_{s,a}$ and $N_{s,a,s'}$ to build confidence set. As a result, $\mathcal{H}$ is a tighter confidence set for h^*, which enables us to prove ④ and part of ③ are lower order terms. (See Lemma5 and Appendix.C.5)
2. We do not bound $|h_k - h^*|_\infty$. Indeed, it suffices to bound $N_{s,a,s'}^{(t_k)}|\delta^*_{s,s'} - \delta_{k,s,s'}|$ up to $\tilde{O}(\sqrt{T})$ and this is exactly what we do.
3. We will mention the literature which first uses Bernstein's inequality to bound the uncertainty.
4. An MDP is flat iff $r_{s,a} + P_{s,a}^T h^* = h^*_s + \rho^*$ for any s, a. We use the notation $reg_{s,a}$ because we regard it as a generalization of $\Delta_a = \mu^* - \mu_a$ in multi-armed bandit (MAB) problem. We will explain these expressions more clearly.
5. As for the usage of Lemma1 in the proof of Theorem1, we apply Lemma1 to the virtual MDP with increased reward function. Although the original MDP M might not be flat, the new MDP is flat after increasing $r_{s,a}$ by $reg_{s,a}$. (See Appendix.C.5)
6. You mean REGAL.C rather than REGAL.D? We can search among the MDPs with constant gains so that the problem is well-posed, although it is intractable in practice.
7. The modified version of (2) refers to line1 in Algorithm2.
8. We are sorry that we are unaware of the state of the art. Consequently, we only improve an $\sqrt{S}$ (or $\sqrt{\Gamma}$) factor compared to the work you mentioned. Nevertheless, to out best of knowledge, this is the first upper bound which matches the lower bound with logarithm factors ignored. We will also mention [Talebi and Maillard, 2018]. In our analysis, the dominant term in regret is $\sum_{s,a}\sqrt{N_{s,a}^{(T)}V(P_{s,a},h^*)} \leq \sqrt{\sum_{s,a}N_{s,a}^{(T)}\sum_{s,a}V(P_{s,a},h^*)} = \sqrt{T\sum_{s,a}V(P_{s,a},h^*)}$, which outperforms the result in [Talebi and Maillard, 2018] by at least an $\sqrt{S}$ factor.

To reviewer3:

1. We will add a reference of upper bound of $\tilde{O}(\sqrt{N})$ in line 142.
2. We will check the related works section carefully. We will mention the works in the comments of reviewer1. The analysis about REGAL.C [Bartlett and Tewari 09] is correct, although that paper contains some other mistakes.

To reviewer4:

1. You mean a problem-dependent regret bound of $O(poly(S,A,H)\sum_{s,a}\log(T)/reg_{s,a})$ like the regret bound of $O(\sum_a \log(T)/\Delta_a)$ in the MAB problem? We can prove this regret bound is unreachable in the worst case where some state s has $o(T)$ visit count. Under the assumption the visit count (in expectation) of each state s is at least CT for some conditional number C, our method for estimating the bias function works, and thus it is hopeful get a regret bound of $O(poly(S,A,H,\frac{1}{C})\sum_{s,a}\log(T)/reg_{s,a})$. However, this assumption seems too strong for undiscountted RL.
2. In the case H is known, the example for lower bound of $\Omega(\sqrt{SATH})$ was proposed in [Bartlett and Tewari, 2009], although the authors claimed a wrong lower bound.
3. Yes. There is a similar mistake in the proof of Lemma3 [Osband and Van Roy 16].
4. Yes. But regret bound of $\tilde{O}(\sqrt{SATH})$ has been proved for finite horizon MDP with efficient algorithms (e.g. [Azar et al.2017].
5. We have sent an email to the authors, but did not get replies before the rebuttal deadline. We ensure that this issue will be carefully dealt with according to their suggestion.