The authors thank the reviewers for their helpful comments.

=== R1 ===

1) We have indeed thought about a lower bound, but do not yet have a full result yet. The contribution for the suboptimal arms is essentially tight (excluding logs and constants) from the bandit literature. However, for the sample complexity along the optimal state-actions we are not sure: the extra $1/(1-\gamma)$ factor in the upper bound, which is in some way unavoidable due to the worst case lower bound, stems from the discounted sum of visit probabilities, which might interact in a non-trivial way with the variances in constructing a lower bound. Finally, the constant term $S/(1-\gamma)^2$ is likely to be avoidable: a paper titled "On the Optimality of Sparse Model-Based Planning for Markov Decision Processes" that just appeared on Arxiv shows how to reduce that dependence for small ϵ in model-based approaches like ours, and their technique seems applicable to our case. **3)** We thank the reviewer for the careful reading of the proofs. The reviewer helpfully identified two small errors, that only impact the numerical constants. The first is in Equation 60 where there is indeed a lower order term that increases the numerical constant. By definition of B_{ksa} in Appendix A we have:

$$\sum_{(s,a)} \left(\overline{w}^{\pi,\rho}_{sa} - \widehat{\overline{w}}^{\pi,k,\rho}_{sa}\right)\left(2B^k_{sa}\right) = 2\sum_{(s,a)} \left(\overline{w}^{\pi,\rho}_{sa} - \widehat{\overline{w}}^{\pi,k,\rho}_{sa}\right)\left(\frac{2c_n}{(1-\gamma)(n_{sa}-1)} + \gamma\sqrt{\frac{2c_n}{n_{sa}}}\epsilon_k\right)$$

We examine the two different terms. For the first term using the definition of CI^k_{sa} and lemma 10:

$$2\sum_{(s,a)} \left(\overline{w}^{\pi,\rho}_{sa} - \widehat{\overline{w}}^{\pi,k,\rho}_{sa}\right)\frac{2c_n}{(1-\gamma)(n_{sa}-1)} = 2\sum_{(s,a)} \left(\overline{w}^{\pi,\rho}_{sa} - \widehat{\overline{w}}^{\pi,k,\rho}_{sa}\right)\Bigg(6 \times \underbrace{\frac{c_n}{3(1-\gamma)(n_{sa}-1)}}_{\leq CI^k_{sa}}\Bigg) \leq 12\sum_{(s,a)} \left(\overline{w}^{\pi,\rho}_{sa} - \widehat{\overline{w}}^{\pi,k,\rho}_{sa}\right)CI^k_{sa} \leq 12Cp(n_{min})\epsilon^\pi_k$$

and for the second term using the definition of $Cp(n_{min})$ in Appendix A and $\sum_{(s,a)}\left(\overline{w}^{\pi,\rho}_{sa} + \widehat{\overline{w}}^{\pi,k,\rho}_{sa}\right) = 2/(1-\gamma)$:

$$2\sum_{(s,a)} \left(\overline{w}^{\pi,\rho}_{sa} - \widehat{\overline{w}}^{\pi,k,\rho}_{sa}\right)\Bigg(\underbrace{\gamma\sqrt{\frac{2c_n}{n_{sa}}}}_{\leq(1-\gamma)Cp(n_{min})}\epsilon_k\Bigg) \leq 4\sum_{(s,a)} \left(\overline{w}^{\pi,\rho}_{sa} + \widehat{\overline{w}}^{\pi,k,\rho}_{sa}\right)(1-\gamma)Cp(n_{min})\epsilon_k = 8Cp(n_{min})\epsilon_k$$

In summary, equation (60) in the paper would become:

$$\sum_{(s,a)} \overline{w}^{\pi,\rho}_{sa}(CI_{sa}(n^{k+1}_{sa}) + 2B_{ksa}) + 15Cp(n_{min})\epsilon^\pi_k + 8Cp(n_{min})\epsilon_k$$

This way the rest of the proofs remain unchanged in the appendix (since we show $\epsilon^\pi_k \leq \epsilon_k$ for the good policies), and the numerical constants can be incorporated into the $\tilde{O}$ notation. There are ways that avoid increasing the constants as we showed above, but the argument was not self-contained enough to be explained in the rebuttal. **4)** The reviewer is again correct in obtaining $\|V^\star - V^{\pi_k}\|_\infty \leq 2\epsilon_k/(1 - Cp(n_{min}))$ as a final bound. Lemma 18 gives a value for $Cp(n_{min}) \leq 1/100$ and hence $\|V^\star - V^{\pi_k}\|_\infty \leq 2.03\epsilon_k$, instead of the reported $\|V^\star - V^{\pi_k}\|_\infty \leq 2\epsilon_k$.

===R3=== The notation n^k_{sa} represents the number of samples allocated in the kth phase of the algorithm to the (s,a) pair. For "why problem dependent structure can remove the dependence of horizon for suboptimal action", as the reviewer notes, the approach is quite technical and we will strive to better convey the intuition for why suboptimal actions do not require as many samples. The key ideas are in Fact 1 and Lemma 1, where we highlight how suboptimality depends on the distribution of visited state-action pairs, and how by adaptively allocating samples, in a way that depends on the gaps, we can avoid a horizon dependence for suboptimal actions. Finally, some authors do 'heuristic' translations of sample complexity between finite horizon and infinite horizon, where the number of steps for the finite case is roughly translated into the $1/(1-\gamma)$ factor, but as the reviewer points out, we need to define the word 'horizon' for our submission.

===R4=== The authors understand that the main suggestion for improvement are a method for computing the maximum likelihood MDP. We can report the maximum likelihood formulas for the rewards and transition probabilities in the appendix; after this, an algorithm like policy iteration (applied to the MDP with the computed maximum likelihood rewards and transition probabilities) can give the empirically optimal policy and value function.

=== Numerical Experiments R1, R3, R4 === We have worked towards an implementation to answer the reviewers' request of providing experiments, but unfortunately we did not complete it in time for the rebuttal. In addition, obtaining an implementation that takes full advantage of our problem dependent analysis involves a more careful computation of the numerical constants (which matter in practice) and using the law of iterated logarithms to lower the log dependence to log-log (this is standard practice to improve the practical performance of algorithms based on concentration inequalities).