We thank all reviewers for their careful reading of our paper and their constructive feedback. We respond to each reviewer separately below.

Reviewer #1 Thank you for expressing your appreciation of our results and your thoughtful comments! Regarding your remarks:

- While the paper is indeed somewhat lengthy, its length is far from being unusual for theoretical papers at NeurIPS, especially in this topic. For instance, we refer to some relevant papers from NeurIPS 2019: "Non-Asymptotic Gap-Dependent Regret Bounds for Tabular MDPs" (54 pages), "Exploration Bonus for Regret Minimization in Discrete and Continuous Average Reward MDPs" (34 pages), "Tight Regret Bounds for Model-Based Reinforcement Learning with Greedy Policies" (54 pages).
- We agree that theoretically sound optimistic RL algorithms are still not quite mature enough for real-world applications, but we would like to point out that there is steady progress in this field of research, with several new results being published at NeurIPS, ICML and COLT each year. Our work will be of interest to members of the community working on theoretical RL along with those working to bridge this gap between theory and practice, and so we believe the audience of our work will extend beyond "a handful of authors".
- We absolutely agree that our assertion about the model-optimistic framework yielding a simpler analysis is subjective, and we hope that our writing didn't suggest otherwise. We will make this clearer for the final version. In any case, we believe that many researchers may share our views about the simplicity of the analysis of model-based approaches and will find our results insightful.

Reviewer #2 We absolutely agree that addressing the gap between the best available bounds for model-optimistic and value-optimistic algorithms is an important open challenge, and that our paper did not manage to close this gap. In the present paper, we dispel the commonly held belief that model-optimistic methods would be difficult to implement, which at the very least suggests that such algorithms may be more powerful than they are commonly thought to be. In fact, we think that there is still no sufficient evidence for the claim that "value-optimisic algorithms [would be] more statistically efficient" than model-optimistic ones[1], and believe that the primal-dual view we introduce in this paper may prove beneficial for making progress towards addressing this important question. For instance, our primal view could enable constructing more sophisticated (non-local) confidence sets for the transition functions that could lead to tighter performance guarantees. In Section 5 of our paper, we demonstrate that this is indeed the case for linear MDPs where a model-based perspective leads to state of the art algorithms. In the tabular setting, as the reviewer points out, closing the gap is more difficult, and beyond the scope of the present (already lengthy) paper. However, we believe that the results presented in our work will serve as an important stepping stone towards developing such extensions. We will add further discussion of this to the final version of the paper.

Reviewer #3 Thank you for your positive evaluation of our paper! Regarding your questions:

- Finding an "optimal D" is indeed a very interesting question, but also a rather complex one. For a statistically valid analysis, one has to jointly pick a divergence D and a confidence width ϵ, in a way that the primal confidence intervals remain valid and the exploration bonuses (defined in terms of the conjugate D_*) are small. The interdependence of these factors makes it difficult to reverse-engineer a divergence from confidence bounds used by existing value-optimistic algorithms, and so far, we have not been able to derive tighter bounds following this (otherwise very tempting) approach. We believe that achieving minimax-optimal regret bounds may necessitate using more sophisticated confidence bounds in place of the local ones we analyze in this paper, but we leave this to future work.
- It is definitely possible to derive (uniform) PAC bounds using the techniques developed in our paper, since PAC-MDP algorithms also rely on the same notion of optimism as used for proving regret bounds. The key technical step in both PAC and regret analyses is bounding the instantaneous regret $\Delta_{1,t}$ in each episode, and then bounding $\sum_t \Delta_{1,t}$ to obtain a bound on the regret, or $\sum_t \mathbb{I}_{\{\Delta_{1,t} > \varepsilon\}}$ to obtain a PAC bound. Thus, one can also derive PAC bounds by following the steps in Theorem 4. We focused on regret bounds in the present paper since this framework is the most well-studied, but we will point out the possibility of providing PAC bounds in the final version of the paper—thank you for suggesting this extension!

Reviewer #4 Thank you for your review. In particular, thanks for the suggestion of adding more discussion about the relationship between this work and empirically successful work. We will do this in the final version.

[1] This view is currently only supported by comparing *upper bounds* without a separating lower bound.