We thank all reviewers for their detailed feedback.

1 Reviewer 1

- About the solution for under-exploration: We agree that uniform exploration may not be the best possible approach. However, we think it makes the important point that adding some exploration in response to approximate inference can guarantee sub-linear regret, even though this particular form of exploration might not be optimal.
- About stochastic optimism: Thank you for the reference, we agree that stochastic optimism in RL might be related. At first glance, stochastic optimism looks quite different from a bound on the error of alpha-divergence, but it would definitely be interesting to see if there is a deeper connection.
- About experiments: We agree that Q_t and Z_t (which serve to illustrate the intuition about over-exploration and under-exploration) might not be natural approximations. However Figure 3 shows variational inference and ensemble sampling on a 50 armed bandit instance, which is more realistic.

2 Reviewer 2

- About Garbage-in Reward-out paper [Kveton, 2019]: Thank you for the reference. This paper is related but does not study the same problem – they are not concerned with the case where the Thompson sampling's inference oracle is approximate. We would cite and discuss it more thoroughly in a revision.
- About presentation issues: Thank you very much for the comments. We will revise accordingly.
- About the implication of Theorem 1: Typically, approximate inference methods minimize divergences. Broadly speaking, we show that making a divergence a small constant, alone, is not enough to guarantee sub-linear regret. We do not mean to imply that low regret is *impossible* but simply that making an alpha-divergence a small constant alone is not sufficient. We will clarify this point.
- About the implications of all theorem statements:

  Broadly speaking, Theorem 1 implies that when the approximation scheme over-explores, even though the posterior may concentrate, we suffer regret because the approximation chooses the sub-optimal arm with higher probability than the posterior at every time-step due to over-exploration.

  On the other hand, Theorem 2 implies that when the approximation scheme under-explores, the posterior may not concentrate and therefore chooses the sub-optimal arm most of the times, leading the approximation to do the same. Theorem 3 strengthens Theorem 2's observation by showing that adding forced exploration will help the posterior to concentrate and choose the optimal arm most of the times, leading the under-explored approximation to do the same.
- About the correctness of the theorem statements: We believe the results are correct. It would be great if the reviewer could point out the points of concern.

3 Reviewer 3

- We apologize for the grammatical mistakes and thank you so much for listing them.
- Thank you also for the suggested edits about Bayesian agent in line 52 and reverse KL divergence in line 76. We will revise accordingly.