Thank you to all reviewers! The response to reviewer 2 may be most important to how the paper is interpreted:

Reviewer 2: *while the principle is certainly common in the literature, this is the first paper to demonstrate frequentist regret guarantees for perturbation induced exploration. Proof techniques do not appear to be terribly different than the prior arm, with the key differences appearing in Lemmas 4-5.. . . . . . However, I do not know if the fact that randomized value functions do in fact ensure frequentist regret guarantees is terribly surprising.*

Managing to give such a clean, seemingly straightforward, frequentist analysis of RLSVI seems to be a major contribution over prior art. The journal paper Osband et al. [2017] develops a theory of recursive stochastic-dominance relations to study the algorithm, hence requiring very different techniques than the rest of the RL literature. The paper on frequentist analysis of posterior sampling by Agrawal and Jia [2017] builds on those stochastic-dominance techniques, is immensely technical, requires modifying Thompson sampling to get the proof to work, and contains a critical flaw in the proof currently posted online. I've worked very hard to uncover a new proof that hopefully makes it easy for future researchers to transfer results known for optimistic algorithms over to randomized value function approaches.

On whether the results are surprising: Strong theory sometimes takes years to develop and in the meantime people can start to get used to the main the ideas. This paper tries to provide some backing for the claim that "Training a value function estimation scheme on noise-perturbed data generates a highly sophisticated form of exploration that is fundamentally quite different from what is generated by employing stochastic policies." It has been just three years since the first paper making such a claim was published [Osband et al., 2016]. To my understanding, this claim was often met with skepticism, especially because it lacked a frequentist regret bound to back it up. Such a bound has been elusive since then. Things can seem quite clear in hindsight, but I think the claims in this paper would have been shocking 7 years ago (before any analysis of Thompson sampling in bandits even exited.) That's noteworthy, since we're studying extremely old questions in sequential decision making.

Responses to Reviewer 1: [Paraphrasing] *(1) Would a similar analysis yield a high-probability regret bound? . . . (2) While the proof for Lemma 4 & 5 is described very well in the main text, it would be helpful to have a short explanation how this is used to achieve Lemma 6.. . . (3) How does this setting for β affect the empirical performance of the algorithm. . . . (4) The authors chose the setting with time-dependent dynamics which is a little less common than the default setting where dynamics and rewards are identically distributed across time steps within the episode. . .*

(1) Thanks so much! I now believe the high probability bounds work out. Effectively, the techniques in this paper bound, with high probability, the conditional expected regret $\mathbb{E}[V(M,\pi^*) - V(M,\pi^k) \mid \mathcal{H}^k] \leq B_k$ by some simple terms B_k whose sum we know how to control. Rather than take an expectation, the Azuma-Hoeffdin'g inequality should bound the sum of martingale differences: $\sum_{k=1}^{K}\left(\left(V(M,\pi^*) - V(M,\pi^k)\right) - \mathbb{E}[V(M,\pi^*) - V(M,\pi^k) \mid \mathcal{H}_{k-1}]\right)$ The main challenge seems to be that B_k involves terms that are not uniformly bounded, since we add Gaussian noise. The standard ways in which researchers bounds things like $B_1 + \cdots + B_K$ could get unusually messy as a result.

(2-4) In the revision, I'll explain more about Lemma 6 and clearly highlight open questions regrading points (3) and (4). I think setting β too large is similar to setting overly large optimism bonuses for optimistic algorithms. For practical applications, I suspect the noise variances one adds should also be adaptive to the data. We're essentially (recursively) applying linear-regression to minimize the Bellman residuals. One can write down variants of RLSVI that calibrate the noise they add to the variance of observed residuals, but I'd like to do careful empirical evaluation. I'll try to work through the time-homogeneous analysis. The challenge is that mis-estimation of a single state could lead to error in every Bellman update. One needs to be careful to avoid introducing even more factors of H.

Responses to Reviewer 3: *In response to the comments on bootstrapping and the correct form of prior randomness .*

Thanks, this really gets to the crux of what makes boostrap-like methods for exploration so different from the treatment statistics books. You're right that in initial periods, RLSVI is really no better than uniform exploration, though perhaps the same comment applies to optimistic algorithms. But my understanding is that, no matter what, any algorithm will visit certain states/actions many times. These become well understood and the variance of the injected noise begins to vanish at those states/actions. Some other parts of the state/action space are still poorly understood. Because we add lots of noise to the estimated rewards at those states, there is a significant chance they appear to be even better than they really are in any given episode, in which case the algorithm will deftly navigate through the well understood part of the state space trying to reach them. Once the algorithm starts actively trying to reach poorly understood states, it's performing the kind of multi-period exploration that's essential for efficient RL.

To inject prior randomness, there is a natural counterpart (to sampling Q) for linear models, where you regularize the parameter vector to a prior sample. This proecdure actually corresponds to Algorithm 1 if specialize it to the tabular case (think $\theta = Q$). Things are more subtle with neural networks but [Osband et al., 2018] offers one approach. You're right, that algorithm otherwise applies directly to settings with function approximation.