We thank the reviewers for their valuable feedback. A common point raised by the reviewers is that the clarity of the paper could be improved. We agree and, as suggested by the reviewers, will move some proofs to the appendix and replace them with more intuitive explanations. We address reviewer specific comments below:

R3: We thank R3 for recognising the value of decoupling reward and transition uncertainty, and of the intuition provided by our theory about the failure modes of existing methods. R3 says: "[SU] lacks adequate motivation for why the approach would not suffer from the drawbacks highlighted in part 1". SU is motivated in Section 3 (please see lines 81-85 and the ensuing discussion). We satisfy our Definition 2 which allows us to avoid the pitfalls highlighted in Section 3 (please see lines 146-148). The mechanism through which SU avoids these issues is illustrated in the sketch proof of proposition 3 on page 6. We will reemphasise these points in the next revision.

Regarding the numbered list of questions: (1) *Exploration deterministic or stochatic?*: Stochastic—we use posterior sampling (please see line 127 and Algorithm 1); (2) *Disconnect between pseudocode and text*: We have not found any. Can you please provide more detail?; (3) *"[F]ailure of BDQN and UBE is surprising . . . I do not see how they fail so much"*: Failure of UBE is predicted in Proposition 1 (please see line 100). For BDQN, we hypothesise that before finding the reward signal, P_Q depends more on the random initialisation of the NN then on the actual MDP. Please note that the code and instructions to reproduce all experiments can be found in the supplementary material; (4) *Section 5.3 is unnecessary*: Section 5.3 shows we reliably outperform BootDQN+Prior (the strongest competitor of SU) on the benchmark proposed in the BootDQN+Prior paper (Osband et al., 2018)—we believe this to be one of the strongest empirical results in our paper; (5) *"y-axis in Figure 4—clipped or is-between"*: clipped (please see caption of Figure 4); (6) *Explanation of assumptions about state-action embeddings:* Most assumptions follow from the definition of successor features (Dayan, 1993; Barreto et al., 2017). For the rest, please see lines 133 (unit norm), and 142–143 (non-negativity). (8) *Are "tied actions" equivalent to "stochastic transtions"?*: No, tied actions means that a_1 is always mapped to UP and a_2 to DOWN *in each state*, as opposed to this mapping being *different between states*. This mapping is randomised at the beginning but then kept fix, leading to deterministic transitions (please see lines 359-360); (9) *Are action just another input to the network:* No, the state is fed in and all actions are considered to determine highest Q value (please see Section 4.1, and lines 592–605 in the appendix for more detail).

R2: We thank R2 for recognising the empirical strength of SU and the promising nature of our work in regard to future research. R2 observes that BootDQN does not satisfy the definition of *Randomised Policy Iteration* (RPI): This is intentional as BootDQN does not suffer from the issues discussed in Section 3 (the reasons to prefer SU over BootDQN are so far purely practical: a significantly lower computational cost and better empirical performance). It may be more natural to define an RPI method as an algorithm iterating *policy improvement* and *value prediction* steps while maintaining a distribution over the values and/or policies. We will distinguish between this more general definition and the "single policy" RPI methods in the next revision; please note that this will not affect our theoretical claims.

R2's Proposition 1 comments: (i) It is *not* the case that "the analysis fixes the policy" which, as R2 points out, would be quite limiting. The result holds for any algorithm which employs a factorised Q function distribution with symmetric marginals. The confusion perhaps comes from the π superscript used in statement of Proposition 1; we will adjust the notation in the next revision. (ii) It indeed may seem that function approximation will lead to high correlation between Q values of nearby states. However, our experiments in Section 5.1 show that BDQN, which uses neural network function approximation, fails to outperform the uniform exploration policy (this phenomenon was present in all architectures we tested). As mentioned in response to R3's question (3), we hypothesise this is due to P_Q's dependence on initialisation before finding the reward signal. SU can be seen as a simple fix which can leverage information about transitions even without observing any rewards. We agree that gaining thorough theoretical understanding of why BDQN fails in Section 5.1 is an interesting direction of future research.

R2's Proposition 2 comments: The purpose of this proposition is to prove that "propagation of uncertainty" is not *necessary* to satisfy our Definition 2. That propagation of uncertainty is not *sufficient* for effective exploration is shown by Proposition 1 and experimentally in Section 5.1, meaning that SU's and BootDQN's success cannot be ascribed to propagation of uncertainty when posterior sampling is used. However, we do agree with R2 that matching PSRL's distribution over policies directly would be preferable to satisfying Definition 2. Doing so in a computationally tractable way in large scale settings remains a challenge though which is why all contemporary algorithms (including SU) employ approximations. We will clarify these points in the next version of our manuscript.

R1: We thank R1 for recognising SU's strong empirical performance, and our contributions to the ongoing theoretical exploration of PSRL. We are glad R1 highlighted relative simplicity of SU which may lead to its wider adoption, and are thankful for the suggestions regarding writing which we will implement in the next revision of our paper.