We thank all reviewers and we will modify the paper to clarify each of the points raised, as discussed/clarified below.

Similar points across reviewers: **(1)** Re statistical significance of empirical results, our presentation was misleading and instead we will provide confidence intervals (CIs) that clarify/quantify the statistical significance of our results. I.e., the 95% CIs of all mean scores are relatively small for all operators; e.g., such CIs for cartpole are $< \pm 0.3$ for each operator, much smaller than the differences in their mean scores. **(2)** Re results for constant β, we will expand the discussion in Sec 4.5 (L338-341) noting constant β performs worse, and will provide the corresponding numerical results; e.g., mean cartpole scores for $\beta = 1$ and $\beta \sim U[0, 1)$ are 187 compared with 191 for $\beta \sim U[0, 2)$. **(3)** Re defns of terms, instead of providing a reference as in L92-L94, we will add such defns. E.g.: stochastic ordering (s.o.): r.v. X is stochastically $\leq$ to r.v. Y if $\mathbb{P}[X > z] \leq \mathbb{P}[Y > z], \forall z$; convex ordering (c.o.): r.v. X is $<$ r.v. Y under c.o. iff $\mathbb{E}[f(X)] \leq \mathbb{E}[f(Y)]$, $\forall$ convex functions f. **(4)** Re theoretical results establishing benefits of our approach, we will provide some more intuition. For any (x, a) in eq (4) where the action a is not the optimal action, there will very often (i.e. for many k) be instances where $V_k(x) > Q_k(x, a)$, so eq (4) will make $Q(x, a)$ (eventually) deviate more from $V(x)$; OTOH, for action a s.t. $Q(x, a) = V(x)$, then $V_k(x) > Q_k(x, a)$ will only happen rarely, and thus eq (4) will not affect the end value of $V(x)$. These both reflect the concepts of optimality preserving and action gap increasing. Moreover, we observe that the multiplier in front of $V_k(x) - Q_k(x, a)$ (i.e. β_k) is desired to be large individually, but its overall efforts should not be so large as to affect $V(x)$. We therefore introduce a family of RSOs, where β_k is allowed to take on any value, but its average remains < 1. Furthermore, we establish that greater variability in β_k will lead to larger action gaps and that s.o. (c.o.) in β_k will lead to s.o. (c.o.) in the action gaps. **(5)** Re theoretical proofs, because our RSOs introduce probabilistic elements on top of the original MDP, it is natural for us to employ probabilistic arguments in our analysis/proofs. E.g., in L421-424, we use the $\limsup$ and $\liminf$ for set sequences for the ease of derivations. The probabilistic nature of the problem also affords us the liberty to exploit weak convergence limits (convergence in probability) to identify the limit of $V_k(x)$ after establishing a stronger a.s. convergence for $V_k(x)$. More importantly, the stochastic nature of the problem allows us to consider s.o. (c.o.), which are common machinery in probability theory and which we exploit to establish important orderings of performance among different RSOs. While some of this may not be very familiar within the AI community, we believe these additions broaden the spectrum of ideas and methodologies that can be exploited to help improve solutions of fundamental problems in RL and beyond.

R1. **(1)** Benefit of stochastic β_k is addressed by our theoretical results (Thms 3.2-3.4) and by our empirical results demonstrating significant improvements under stochastic β_k. **(2)** L260 ff. were intended to note that distributions with lower means and variances performed worse than $U[0, 2)$ in our experiments, which we will clarify/expand. Further Thms 3.2-3.4 are intended to help find the best β_k sequence. **(3)** Indeed, the submission contained a typo in eq (9): π^{b_k} in both inequalities should be a, which is the focus of the arguments that follow. **(4)** We should have explicitly stated that $V_k(x) + f_k$ is uniformly upper bounded from the facts that the rewards are bounded functions and $\gamma \in (0, 1)$. **(5)** In L421-L424, we establish the (right) inequality in eq. (9) by considering the limit superior and limit inferior of the sequence of sets on which the probabilities are calculated. Therefore, we examine events that happen infinitely often for the $\limsup$ and all but finite exceptions for $\liminf$. **(6)** We view the problem of finding the maximally efficient operator as one of finding a sequence of β_k that produces dominating performance. We believe that the statement is correct from this viewpoint. But the statement does not discuss how the optimization should be conducted, where different methods could have different implications. We will clarify these points. **(7)** The value function $V(x)$ indeed does not change, but $Q(x, a)$ changes and this leads to a larger action gap. This should then lead to more efficient ways of ultimately finding $V(x)$ via updating $Q(x, a)$, as indicated in refs [5] and [12].

R2. To address concerns of the role of randomness, we will expand L263-265 which notes the same trends were observed when we varied exploration of the ϵ-greedy algorithm over a wide range of ϵ (even for deterministic operators).

R3. **(1)** We appreciate the comment on "robust" in other contexts, and will be more careful/clearer in our usage. **(2)** Re proof of Thm 3.1, the inequality in L408 follows from $V_k(x) \geq Q_k(x, a), \forall a$, by defn of $V_k(x)$. The 3^{rd} relation below L411 follows directly from the defn of $\mathcal{T}_B$ and the 4^{th} relation below L411 is directly due to the order relation of $\mathcal{T}_B$ and $\mathcal{T}_{\beta_k}$. Lastly, regarding the a.s. convergence and its weak form, convergence in probability: By L417, we already established that $V_k(x)$ converges a.s., so the purpose of the next paragraph is to identify the limit, as stated; For that purpose, we only need to identify the limit of the weak convergence, since a.s. convergence naturally implies that $V_k(x)$ also weakly converges to the limit. **(3)** Re proof of Thm 3.2, it is quite standard to prove results by introducing a general function with minimally restrictive properties (increasing) and then appropriately using this function and its properties; the ordering $\mathbb{E}[f(\hat{Q}_{k+1})] \geq \mathbb{E}[f(\bar{Q}_{k+1})]$ leads to $\hat{Q}_k \geq_{st} \bar{Q}_k$ because of the properties of $f(\cdot)$ and the defn of s.o. ($\geq_{st}$). Similarly, in the last part of the proof, $f(\cdot)$ is used to establish the desired s.o. of the action gap. **(4)** Re proof of Thm 3.3, with the addition of the defn of c.o., the proof follows along similar lines to that of Thm 3.2 as stated. **(5)** The proof of Thm 3.4 is not based on induction and is proven using a basic relation between variance and conditional expectation, with the direct derivation establishing the general result for each k. **(6)** Q-learning provides convergence to the values of all $Q(x, a), \forall x, a$, asymptotically over time. Our theoretical results study the behavior of $Q(x, a)$ under different operators, establishing the benefits of our RSOs over other (deterministic) operators such as in [5].