First of all, we would like to thank the reviewers for their valuable comments, which will help us improve the revised manuscript. We endeavour to address all the major remarks below.

Second, we would like to notify all the reviewers about a major theoretical improvement allowing us to transform the conjecture at line 107 into a theorem. At submission time, the proof was missing two steps (lines 443–444 and 465), and we decided to present it as a conjecture for the sake of caution. In the meantime, we have proved them and we propose to make it a theorem.

Third, a remark common to all reviewers is that Algorithm 3 (convex hull policy) lacks a detailed explanation, which we acknowledge: it was a questionable decision due to space constraints. We will include a descriptive paragraph.

Reviewer #1

In principle it could be framed [...] why this is not a good idea. Swapping the min and max would lead to a different set of policies. Indeed, the BMDP solution is generally not a saddle point: max-min $<$ min-max.

When creating batch samples what happens when the budget is zero? The episodes ends with a negative reward? If the agent explores/exploits with zero budget, it will always select the estimated safest action, with minimal estimated expected cost Q_c. If costs are still incurred, the model Q_θ will be updated accordingly.

As it is not a contraction in general, in which cases should we not use the proposed approach? The Remark 1 and the counter-example in Theorem 2 both indicate that the proposed approach is more likely to diverge when Q^* is steep, i.e. in problems where a small increase in budget β leads to substantial gains in rewards. As in most theoretical works, the required assumptions may be transgressed in real world. But, it does not necessarily imply that the algorithm would not work in practice. For a sensitive real world application, we would not refrain from using the algorithm, but would strongly advise to supervise it.

How can the budget lie between two Q_c, is it not the case that all of them should respect the budget constraint? $Q_c^*(\overline{s}, \overline{a})$ is the cost induced by *first* executing action $\overline{a}$ and only *then* following the optimal budgeted policy π^*. Among all these actions, some of them may not be feasible and exceed the budget β, which is why an additional optimisation step is required in the definition of the greedy policy.

Reviewer #2

(A.1) The cost constraints only apply to the final trained budgeted policy, we indeed do not care about the costs incurred during training. To enforce a specific β at test time, set the initial augmented state to $\overline{s}_0 = (s_0, \beta)$.

(A.2) If we reach Line 9 in Algorithm 3, it must be that the condition $\beta < q_c^2$ in Line 5 never holds, which means that every action verifies $q_c \leq \beta$: it is a case where the budget is always respected.

(A.3) We will follow your suggestion: indeed Algorithm 4 is more specific to our work than Algorithms 1/2.

(B) We will make the introduction clearer and properly introduce the "set of solutions" techniques like FTQ(λ) along with a proper comparison to the BMDP approach. Namely, they do not recover optimal budgeted policies, in addition to being inefficient.

Minor comment 2 Indeed, but one of them is just a toy example to illustrate the risk-sensitive exploration only.

Minor comment 3 This is absolutely correct. The problem is that since we framed the problem in an augmented space where the budget β is part of the state $\overline{s}$, its evolution can only be described within the dynamics of $\overline{s}$ (which we specify). We agree that this makes the notations quite obscure for a very simple idea, but on the other hand casting the problem in such a way drastically simplifies the next definitions and proofs, e.g. that of Proposition 1.

Reviewer #3

Presentation of the result on the optimality operator. The main criticism concerned the presentation of the contractivity of $\mathcal{T}$ over smooth Q-functions as a conjecture rather than a proven result. We hope that these reservations have been lifted by the upgrade of the conjecture to a theorem.

The exploration strategy from Section 3.2 is merely described: This strategy is motivated by the observation that a wide variety of risk levels needs to be experienced during training, which can be achieved by enforcing the risk constraints during data collection. This intuition was meant to be conveyed by the corridor example where a conventional greedy exploration procedure fails to visit the safe region. We will add an additional discussion to clarify this point.

The implementation proposed in Section 4 seems to rely heavily on results from Boutilier and Lu, 2016, but a lot of the necessary context is missing. Several differences exist between the approach from Boutilier and Lu (2016) - the greedy budget allocation (GBA) algorithm - and ours: they rely on a finite set of non-dominated budget points B which grows exponentially with the horizon and becomes uncountably infinite in continuous state spaces. They also compute and sort a matrix of bang-per-buck ratios of size $|S| \times B$, which is again infeasible when S is continuous. In contrast, we instead rely on estimating the optimal cost-to-go Q_c^*, which requires an additional min constraint in its definition that does not appear in Boutilier and Lu (2016) (they only estimate V_r^*).