Dear all reviewers, we highly appreciate your valuable comments and will reflect your comments in the revision.

» Revewer 1

» *"1. I think more could be done on the experimental front. ... in better RL in a number of settings."*

In all the environments except for Walker2D-disc, OC3 and SoftRobust successfully acquired options corresponding to decomposed skills required for solving the tasks. Especially, OC3 produces robust options. In HopperIceBlock-disc, for example, OC3 produces options corresponding to walking on slippery grounds and jumping onto a box (Figures 1 and 2). In HalfCheetah-disc, OC3 produces an option for running (highlighted in green in Figure 3) and an option for stabilizing the cheetah-bot's body (highlighted in red in Figure 3), which is used mainly in the rare-case model parameter setups. Acquisition of such decomposed robust skills is useful when transferring to a different domain, and they can be also used for post-hoc human analysis and maintenance, which is an important advantage of option-based RL over flat-policy RL.

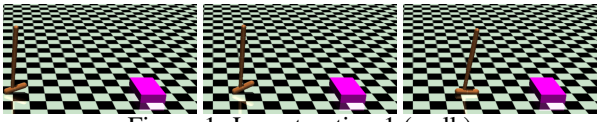

Figure 1: Learnt option 1 (walk)

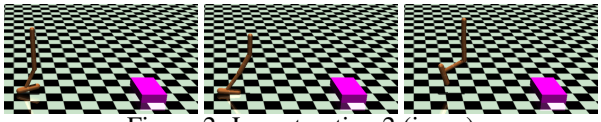

Figure 2: Learnt option 2 (jump)

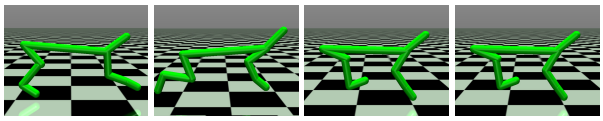

Figure 3: Learnt options (shown in green when option 1 is taking the control and in red otherwise) in Halfcheetah-cont. Left: the case with a normative model parameter setup. Right: the case with an abnormal model parameters setup.

» *"2. The soft robust loss baseline is not that bad ... to demonstrate the benefits of the proposed framework."*

We have conducted additional experiments with the condition that the model parameter distribution and the parameter value ranges in the test phase are more uncertain (in terms of of entropy) than those in the training phase. The results confirmed that OC3 outperforms SoftRobust with respect to multiple performance measures (average return/loss and CVaR) (the results will be reported in the revised version). The sim2real experiment couldn't be done in this time frame.

» Reviewer 2

» *" However, correct me if I misunderstood, ... it is unclear to me what the originality is of the proposed algorithm. "*

Let us clarify the difference between our algorithm and Algorithm 2 in [6][1]. In addition to the type of parameters you pointed out, Algorithm 2 in [6] and our algorithm are primarily different in the form of augmented MDPs. In Algorithm 2 in [6], the augmented MDPs take the form of $\langle S', A, R', \gamma, T', \mathbb{P}'_0 \rangle$, while, in our algorithm, the augmented MDPs take the form of $\left\langle S', A, R', \gamma, \mathbb{E}_p\left[T'_p\right], \mathbb{P}'_0 \right\rangle$. Here T' is the transition function without model parameters, and the other elements are the same as those in our paper. In sum, the types of transition function are different.

This difference comes from the difference of optimization objectives. In [6], the optimization objective is the expected loss without a model parameter uncertainty ($\mathbb{E}_{\mathcal{R}'}\left[\mathcal{R}'\right]$), while in our optimization objective, the model parameter uncertainty is introduced by taking the expectation of the model parameter distribution ($\mathbb{E}_{\mathcal{R}',p}\left[\mathcal{R}'\right] = \sum_p \mathbb{P}(p)\mathbb{E}_{\mathcal{R}'}\left[\mathcal{R}' \mid p\right]$).

You may be concerned that modifying the augmented MDPs in [6] to use $\mathbb{E}_p\left[T'_p\right]$ for dealing with the parameter uncertainty is straightfoward. It turned out, however, that this seemingly simple modification required five pages of proofs for justification, and any theoretical discussion for this has not been provided in previous work (e.g., in [6], [32], and [1]). This is the reason why we described theoretical discussions (i.e., Eq. 11 $\sim$ 14 based on our option policy gradient theorems in Appendix) for this modification, which should help many researchers in the field.

» Reviewer 3 : Thank you very much for your comments. We will modify our paper as you pointed out.

» Reviewer 4

» *"1) While I understand the paper is proposing ... is only to extend this CVaR MDP framework to include options."*

In the derivation of our CVaR option critic, the optimization of the soft robust loss $\mathbb{E}_{\mathcal{R}',p}\left[\mathcal{R}'\right]$ in augmented MDPs is necessary (please see our answer to Reviewer 2's Q). So, as a part of the derivation, we also propose a new option critic architecture to deal with this loss. Note that, due to the model parameter uncertainty (i.e., soft robust loss $\mathbb{E}_{\mathcal{R}',p}\left[\mathcal{R}'\right]$), the standard option-critic framework (the Bacon's one) w. Chow's CVaR PG cannot be applied to our setting. Both of the previous frameworks (and Tamer's framework) cannot deal with the case with model parameter uncertainty.

» *"3) While the experimental ... the vanilla CVaR PG algorithm compare with the option-critic counterpart?"*

OC3 produces some useful portable options even in Halfcheetah and Walker2D-cont (see our answer to Reviewer 1's Q). Note that, since Chow's framework [6] (and Tamer's framework) cannot deal with model parameter uncertainty, these cannot be applied to our setting. Finally, we will modify the notations as you suggested.

[1] We assume that you mean Algorithm "2" (Actor-Critic Algorithms for CVaR Optimization), not Algorithm "1", in [6].