Thank you for the constructive comments and suggestions. We will incorporate all presentation improvements suggested.

Theoretical results (Reviewer 1): [$\boldsymbol{\omega}^\intercal$ and $\arg_Q$ cancel each other] By definition (line 143), $\arg_Q \sup_{a\in\mathcal{A},\boldsymbol{\omega}'\in\Omega} \boldsymbol{\omega}^\intercal \boldsymbol{Q}(s,a,\boldsymbol{\omega}') := \boldsymbol{Q}(s,a',\boldsymbol{\omega}'')$, i.e., the $\arg_Q$ operator extracts the $\boldsymbol{Q}$ value that results in the largest utility using the preferences $\boldsymbol{\omega}$. Therefore, linearizing this $\boldsymbol{Q}$ with the same $\boldsymbol{\omega}$ results in exactly the same supremum, i.e., $\boldsymbol{\omega}^\intercal \arg_Q \sup_{a\in\mathcal{A},\boldsymbol{\omega}'\in\Omega} \boldsymbol{\omega}^\intercal \boldsymbol{Q}(s,a,\boldsymbol{\omega}') = \boldsymbol{\omega}^\intercal \boldsymbol{Q}(s,a',\boldsymbol{\omega}'') = \sup_{a\in\mathcal{A},\boldsymbol{\omega}'\in\Omega} \boldsymbol{\omega}^\intercal \boldsymbol{Q}(s,a',\boldsymbol{\omega}')$. **Note the supremum is over $\boldsymbol{\omega}'$, not $\boldsymbol{\omega}$.**

[Theorem 1] Thanks for catching the typo - $\boldsymbol{Q}$ should be $\boldsymbol{Q}^*$. We realize that the proofs are a bit compressed - we will update the paper with more detailed derivations for all proofs. Here is Thm. 1 in detail (starting step 2 under line 601):

$$
\begin{aligned}
\boldsymbol{\omega}^\intercal \mathcal{T}\boldsymbol{Q}^*(s,a,\boldsymbol{\omega}) &= \boldsymbol{\omega}^\intercal \boldsymbol{r}(s,a) + \gamma \cdot \boldsymbol{\omega}^\intercal \mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)} \arg_Q \sup_{a'\in\mathcal{A},\boldsymbol{\omega}'\in\Omega} \boldsymbol{\omega}^\intercal \boldsymbol{Q}^*(s',a',\boldsymbol{\omega}') \\
\text{(linearity of exp. \& cancel } \boldsymbol{\omega}^\intercal \text{ and } \arg_Q) &= \boldsymbol{\omega}^\intercal \boldsymbol{r}(s,a) + \gamma \cdot \mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)} \sup_{a'\in\mathcal{A},\boldsymbol{\omega}'\in\Omega} \boldsymbol{\omega}^\intercal \boldsymbol{Q}^*(s',a',\boldsymbol{\omega}') \\
\text{(insert eq. (20), def. of } \boldsymbol{Q}^*) &= \boldsymbol{\omega}^\intercal \boldsymbol{r}(s,a) + \gamma \cdot \mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)} \sup_{a'\in\mathcal{A},\boldsymbol{\omega}'\in\Omega} \boldsymbol{\omega}^\intercal \left\{ \arg_Q \sup_{\pi\in\Pi} \boldsymbol{\omega}'^\intercal \mathbb{E}_{\substack{\tau\sim(\mathcal{P},\pi) \\ |s_0=s',a_0=a'}} \left[\sum_{t=0}^{\infty} \gamma^t \boldsymbol{r}(s_t,a_t) \right] \right\} \\
\text{(use def. of } \arg_Q \text{, explained below)} &= \boldsymbol{\omega}^\intercal \boldsymbol{r}(s,a) + \gamma \cdot \mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)} \sup_{a'\in\mathcal{A}} \boldsymbol{\omega}^\intercal \left\{ \arg_Q \sup_{\pi\in\Pi} \boldsymbol{\omega}^\intercal \mathbb{E}_{\substack{\tau\sim(\mathcal{P},\pi) \\ |s_0=s',a_0=a'}} \left[\sum_{t=0}^{\infty} \gamma^t \boldsymbol{r}(s_t,a_t) \right] \right\} \\
\text{(rearrange expectation and sup)} &= \boldsymbol{\omega}^\intercal \boldsymbol{r}(s,a) + \gamma \cdot \boldsymbol{\omega}^\intercal \arg_Q \sup_{\pi\in\Pi} \boldsymbol{\omega}^\intercal \mathbb{E}_{\substack{\tau\sim(\mathcal{P},\pi) \\ s_0\sim\mathcal{P}(\cdot|s,a)}} \left[\sum_{t=0}^{\infty} \gamma^t \boldsymbol{r}(s_t,a_t) \right] \\
\text{(merge 1st term to sum \& use def. of } \boldsymbol{Q}^* \text{ again)} &= \boldsymbol{\omega}^\intercal \left\{ \arg_Q \sup_{\pi\in\Pi} \boldsymbol{\omega}^\intercal \mathbb{E}_{\substack{\tau\sim(\mathcal{P},\pi) \\ |s_0=s,a_0=a}} \left[\sum_{t=0}^{\infty} \gamma^t \boldsymbol{r}(s_t,a_t) \right] \right\} = \boldsymbol{\omega}^\intercal \boldsymbol{Q}^*(s,a,\boldsymbol{\omega})
\end{aligned}
$$

The fourth equation is due to a sandwich inequality, $\boldsymbol{\omega}^\intercal \arg_Q \sup_{\pi\in\Pi} \boldsymbol{\omega}^\intercal \boldsymbol{Q}^\pi \le \sup_{\boldsymbol{\omega}'\in\Omega} \boldsymbol{\omega}^\intercal \arg_Q \sup_{\pi\in\Pi} \boldsymbol{\omega}'^\intercal \boldsymbol{Q}^\pi = \boldsymbol{\omega}^\intercal \arg_Q \sup_{\pi\in\Pi} \boldsymbol{\omega}_*'^\intercal \boldsymbol{Q}^\pi = \boldsymbol{\omega}^\intercal \boldsymbol{Q}^{\pi'_{\boldsymbol{\omega}'_*}} \le \boldsymbol{\omega}^\intercal \arg_Q \sup_{\pi\in\Pi} \boldsymbol{\omega}^\intercal \boldsymbol{Q}^\pi$, where $\boldsymbol{\omega}'_*$ and $\pi'_{\boldsymbol{\omega}'_*}$ are preference and policy corresponding to the supremums.

[Theorem 2] Step 2 to 3 (line 614) is because $|\mathbb{E}[\cdot]| \le \mathbb{E}[|\cdot|] \le \sup|\cdot|$, and step 3 to 4 results from the cancellation between $\boldsymbol{\omega}^\intercal$ and $\arg_Q$ (as justified above). After line 616, step 2 to 3 arises from the w.l.o.g. assumption that $\boldsymbol{\omega}^\intercal \boldsymbol{Q}(s',a',\boldsymbol{\omega}') - \sup_{a'',\boldsymbol{\omega}''} \boldsymbol{\omega}^\intercal \boldsymbol{Q}'(s',a'',\boldsymbol{\omega}'') \ge 0$, as stated in lines 612 and 615. Thus, the whole expression in $|\cdot|$ is nonnegative and $\boldsymbol{\omega}^\intercal \boldsymbol{Q}(s',a',\boldsymbol{\omega}') - \boldsymbol{\omega}^\intercal \boldsymbol{Q}'(s',a',\boldsymbol{\omega}') \ge 0$. We can discard the last two terms since $\boldsymbol{\omega}^\intercal \boldsymbol{Q}'(s',a',\boldsymbol{\omega}') \le \sup_{a''\in\mathcal{A},\boldsymbol{\omega}''\in\Omega} \boldsymbol{\omega}^\intercal \boldsymbol{Q}'(s',a'',\boldsymbol{\omega}'')$. Step 3 to 4 is because $\sup_{s',\boldsymbol{\omega}'} f(s',a',\boldsymbol{\omega}') \le \sup_{s',a'',\boldsymbol{\omega}'} f(s',a'',\boldsymbol{\omega}')$ holds for any a' and $f(\cdot)$.

Empirical results (Reviewers 1 and 3): [Multiple runs and error bars] Each data point in Table 1 indicates the **mean and standard deviation** over **5 independent** training and test runs, for all methods in all four domains. The error bars in Figure 4 are standard deviations of CR and AE estimated from 5 independent runs under each configuration. This is mentioned in lines 228, 860-862, 877, 881, but we will consolidate and make this clearer for the reader.

[Statistical tests] We performed the unpaired t-test between our envelope model and the baselines and achieved significance scores of $p < 0.05$ vs MOFQI on all domains, $p < 0.01$ vs CN+OLS on DST and $p < 0.05$ vs Scalarized on FTN, Dialog and SuperMario. We will add this information to the results table.

Comparison with Abels, et al. (Reviewers 2 and 3): There are 3 key contributions that distinguish our work from Abels et. al., 2019. We will add a better description of these to the paper as well as better explain figures 2 & 3.

[Algorithmic] Our algorithm (envelope Q-learning), utilizes the convex envelope of the solution frontier to update parameters of the policy network, using an optimality filter $\mathcal{H}$ (line 142) which maintains $\sup_{\boldsymbol{\omega}'} \boldsymbol{\omega}^\intercal Q(\cdot,\cdot,\boldsymbol{\omega}')$. This allows our method to quickly align one preference with optimal rewards and trajectories that may have been explored under other preferences. Abels et al. on the other hand, use scalarized updates that optimizes the scalar utility and hence cannot use the information of $\max_a Q(s,a,\boldsymbol{\omega}')$ to update the optimal solution aligned with a different $\boldsymbol{\omega}$. As illustrated in Figure 2 (c), assuming we have found two optimal solutions D and F in the CCS, misaligned with preferences $\boldsymbol{\omega}_2$ and $\boldsymbol{\omega}_1$. The scalarized update cannot use the information of $\max_a Q(s,a,\boldsymbol{\omega}_1)$ (corresponding to F) to update the optimal solution aligned with $\boldsymbol{\omega}_2$ or vice versa. It only searches along $\boldsymbol{\omega}_1$ direction leading to non-optimal L, even if solution D has been seen under $\boldsymbol{\omega}_2$. Hence, our algorithm has better sample efficiency, as is also seen from the empirical results.

[Theoretical] Further, we introduce a **theoretical framework** for designing and analyzing value-based MORL algorithms, and **convergence proofs** for our envelope Q-learning algorithm. Abels et al., whose method can also be analyzed under our framework, do not provide theoretical analyses of the correctness or convergence of their algorithm.

[Empirical] We also provide **new evaluation metrics and benchmark environments for MORL** – CR and AE. In terms of experiments, Abels et al. only evaluate on two synthetic domains – DST and Minecart. We apply our algorithm to a wider variety of domains including DST, FTN and **two complex larger scale domains** – task-oriented dialog and supermario. Our FTN domain (128 solutions) is a scaled up, more complex version of Minecart (< 10 solutions).