



# Eliciting User Preferences for Personalized Multi-Objective Decision Making through Comparative Feedback [1]

Empirical Evaluation for Theorem 1

---

**Presenter:** Yuwei Cheng, Department of Statistics, University of Chicago

# Presentation Outline

1. Problem Setup
2. Methodology
3. Results
4. Conclusion
5. Limitations
6. Appendix

1. **Motivation:** Many real world problems require people to make sequential decisions that balance multiple but sometimes conflicting objectives. For example, in the area of autonomous driving, safety, speed, and comfort are all desired objectives, while speed could negatively impact safety.
2. **Goal:** Design an optimal personalized policy for a given user through lowest possible user pairwise feedback
3. **Gap:** [1] provides a provably efficient algorithm to estimate user's personalized policy however with no simulation study

# Notations in Multi-Objective Reinforcement Learning

1. If the decision problem has  $K$  objectives to consider, then a user is characterized by its preference vector  $\omega^* \in \mathcal{R}^K$ ,  $\|\omega^*\|_1 = 1$
2. The policy  $\pi : |S| \rightarrow \Delta_{|A|}$
3. The state value function of policy  $\pi$  starting from initial state  $s_0$  is  $V^\pi(s_0) \in \mathcal{R}^K$ .  
 $V^\pi = E_{s_0 \sim \rho}[V^\pi(s_0)] \in \mathcal{R}^K$ ,  $\rho$  is the initial state distribution
4. The personalized value of policy  $\pi$  is  $\langle \omega^*, V^\pi \rangle \in \mathcal{R}^+$
5. The optimal policy  $\pi^* := \arg \max_{\pi \in \Pi} \langle \omega^*, V^\pi \rangle$  and its corresponding optimal personalized value  $\nu^* := \langle \omega^*, V^{\pi^*} \rangle$

# Methodology – Preference Estimation

1. Identification of basis policy (see 1)
2. Computation of basis ratios (see 2)
3. Solve a linear system  $\hat{A}\hat{\omega} = e_1$

$$\hat{A} = \begin{bmatrix} V^{\pi_1 T} \\ (\hat{\alpha}_1 V^{\pi_1} - V^{\pi_2})^T \\ \dots \\ (\hat{\alpha}_{d-1} V^{\pi_1} - V^{\pi_d})^T \end{bmatrix} \quad (1)$$

## Theorem

Consider the algorithm of computation  $\hat{A}$  defined in Eq(1) and any solution  $\hat{\omega}$  to  $\hat{A}x = e_1$  and outputting the policy  $\pi^{\hat{\omega}} = \arg \max_{\pi \in \Pi} \langle \hat{\omega}, V^{\pi} \rangle$ , which is the optimal personalized policy for preference vector  $\hat{\omega}$ . Then the output policy  $\pi^{\hat{\omega}}$  satisfying that  $\nu^* - \langle \omega^*, V^{\pi^{\hat{\omega}}} \rangle \leq O((\sqrt{K} + 1)^{d + \frac{14}{3}} \epsilon^{\frac{1}{3}})$  by using  $O(K \log(\frac{K}{\epsilon}))$  comparison queries.

# Methodology – MO-GridWorld

1. Remove the red-colored self-absorbing states with rewards -10
2. Randomly sample K cells from the grids except the obstacle states with reward 1
3. The reward function in MORL setting is  $K \times |S| \times |S| \times |A|$  matrix
4. Add "STAY" action in addition to the previous four actions. Taking "STAY" action will always receive reward 0
5. Same noise structure in state transition

Parameter	$\theta$	$\gamma$	noise	$\epsilon$	K	rep
Description	tolerance threshold	discounting factor	-	indistinguishability	no.of objectives	-
Value	0.01	0.99	0.1	0.01	[3, 10]	5

**Table 1:** Experiment Parameters

# Results – Sanity Check 1

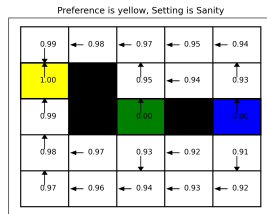
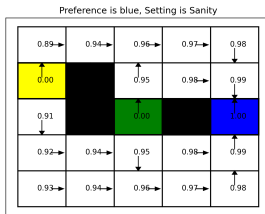
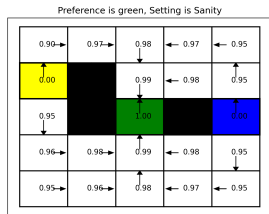
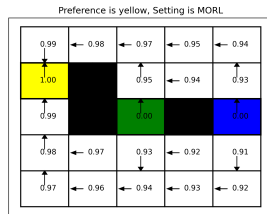
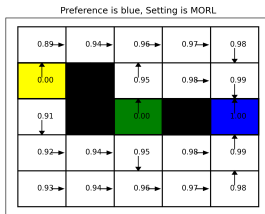
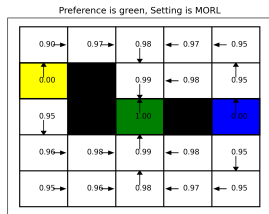


Figure 1: Sanity Check of Value Iteration for MORL

## Results – Sanity Check 2

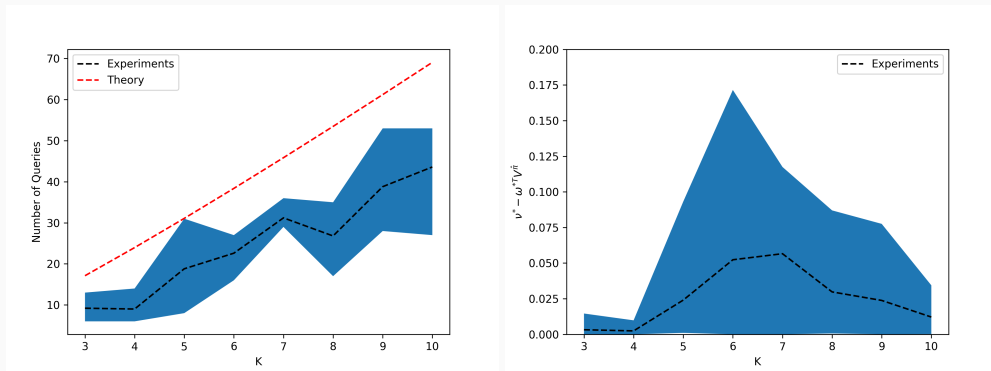
Preference is Doing Nothing, Setting is MORL

0.00	0.00	0.00	0.00	0.00
0.00		0.00	0.00	0.00
0.00		0.00		0.00
0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00

Figure 2: Sanity Check of Policy Evaluation for MORL



# Results – Empirical Evaluation of Theorem 1



**Figure 3:** Simulation Results of Number of Comparison Queries Required in Practice to Estimate  $\omega^*$  and Absolute Performance Gap in Practice

## Results – Relative Performance Gap

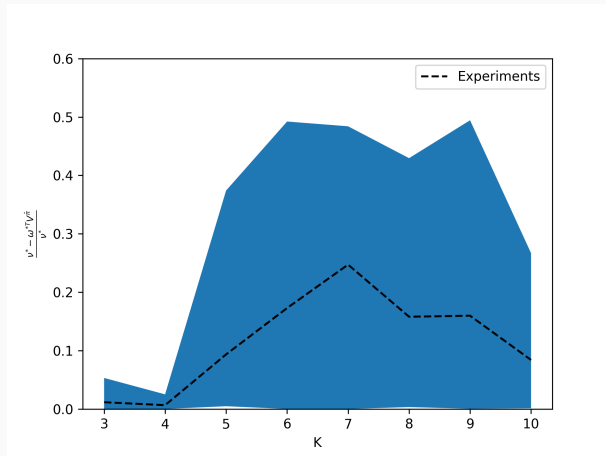


Figure 4: Simulation Results of Relative Performance Gap  $\frac{\nu^* - \langle \omega^*, V^{\pi^{\hat{\omega}}} \rangle}{\nu^*}$

# Results – Poor Preference Estimation

Trial	True Preference			Estimated Preference			$\frac{\nu^* - \langle \omega^*_{\nu} \pi^{\omega} \rangle}{\nu^*}$
0	0.33	0.34	0.33	0.05	0.05	-0.9	0.05
1	0.45	0.06	0.49	0.47	0.03	0.50	2e-08
2	0.39	0.40	0.21	0.47	0.50	0.03	2e-03
3	0.22	0.39	0.39	0.27	0.29	0.45	2e-03
4	0.41	0.30	0.29	0.42	0.31	0.27	4e-08

Table 2: Compare True Preference and Estimated Preference

K	3	4	5	6	7	8	9	10
Mean	2.8	3.8	4.2	5.4	5.8	7.6	7.4	9.6
Min	2	3	3	4	5	7	4	9
Max	3	4	5	6	7	8	9	10
Max $\frac{\nu^* - \langle \omega^*_{\nu} \pi^{\omega} \rangle}{\nu^*}$	2	3	3	6	5	7	4	9

Table 3: Counts of Aligned Sign between True Preference & Estimated Preference

1. Empirical assessment does not show any violation of *Theorem 1*
2. The relative performance optimality gap could be as large as 50%
3. This poor performance might be caused by the poor preference estimation assessed by the proportion of flipped signs
4. I suggest to provide worst case analysis for the relative performance gap and bound the difference between  $\hat{\omega}$  and  $\omega^*$  by putting necessary restrictions on  $\hat{\omega}$

1. The implemented MORL environment is almost the simplest tabular MDP. Given its simple structure, it might not be able to test the robustness of the  $O((\sqrt{K} + 1)^{d+\frac{14}{3}} \epsilon^{\frac{1}{3}})$  in the worse case, especially when  $H$  is large
2. Randomly sampling  $K$  cells from the grid and setting their rewards as 1 is not a good way to model  $K$  different objectives.

# Identification of basis policy

---

## Algorithm 1: Identification of Basis Policies

---

Initialize  $\pi^{e^*} \leftarrow \pi^{e_1}$

**for**  $j = 2, \dots, k$  **do**

    | compare  $\pi^{e^*}$  and  $\pi^{e_j}$

    | if  $\pi^{e_j} > \pi^{e^*}$  then  $\pi^{e^*} \leftarrow \pi^{e_j}$

**end**

$\pi_1 \leftarrow \pi^{e^*}$  and  $u_1 \leftarrow \frac{V^{\pi^{e^*}}}{\|V^{\pi^{e^*}}\|_2}$

**for**  $i = 2, \dots, k$  **do**

    | Arbitrarily pick an orthonormal basis  $\rho_1, \dots, \rho_{k+1-i}$  of  $\text{span}(V^{\pi_1}, \dots, V^{\pi_{i-1}})^\perp$

    |  $j_{\max} \leftarrow \arg \max_{j \in [k+1-i]} \max(|\nu^{\rho_j}|, |\nu^{-\rho_j}|)$

    | **if**  $\max(|\nu^{\rho_{j_{\max}}}|, |\nu^{-\rho_{j_{\max}}}|)$  **then**

        |  $\pi_i \leftarrow \pi^{\rho_{j_{\max}}}$  if  $|\nu^{\rho_{j_{\max}}}| > |\nu^{-\rho_{j_{\max}}}|$  else  $\pi_i \leftarrow \pi^{-\rho_{j_{\max}}}$ .  $u_i \leftarrow \rho_{j_{\max}}$

    | **end**

    | **else** output  $(\pi_1, \pi_2, \dots), (u_1, u_2, \dots)$  and **stop**

**end**

---

# Computations of Basis Ratios

---

## Algorithm 2: Computations of Basis Ratios

---

**Input:**  $(V^{\pi_1}, \dots, V^{\pi_d})$  and  $C_\alpha = 2K$  (see Lemma 1 [1]),  $\hat{\alpha}_i \in [0, C_\alpha], \forall i$


**for**  $i = 1, \dots, d - 1$  **do**

- let  $l = 0, h = 2C_\alpha, \hat{\alpha}_i = C_\alpha$
- while** *True* **do**
  - if**  $\hat{\alpha}_i > 1$  **then**
    - compare  $\pi_1$  and  $\frac{1}{\hat{\alpha}_i}\pi_{i+1} + (1 - \frac{1}{\hat{\alpha}_i})\pi_0$ . **If**  $\pi_1 > \frac{1}{\hat{\alpha}_i}\pi_{i+1} + (1 - \frac{1}{\hat{\alpha}_i})\pi_0$ , **then**  
 $h \leftarrow \hat{\alpha}_i, \hat{\alpha}_i \leftarrow \frac{l+h}{2}$ . **If**  $\pi_1 < \frac{1}{\hat{\alpha}_i}\pi_{i+1} + (1 - \frac{1}{\hat{\alpha}_i})\pi_0$ , **then**  $l \leftarrow \hat{\alpha}_i, \hat{\alpha}_i \leftarrow \frac{l+h}{2}$
  - end**
  - else** compare  $\pi_{i+1}$  and  $\hat{\alpha}_1\pi_1 + (1 - \hat{\alpha}_i)\pi_0$ . **If**  $\hat{\alpha}_1\pi_1 + (1 - \hat{\alpha}_i)\pi_0 > \pi_{i+1}$ , **then**  
 $h \leftarrow \hat{\alpha}_i, \hat{\alpha}_i \leftarrow \frac{l+h}{2}$ . **If**  $\hat{\alpha}_1\pi_1 + (1 - \hat{\alpha}_i)\pi_0 < \pi_{i+1}$ , **then**  $l \leftarrow \hat{\alpha}_i, \hat{\alpha}_i \leftarrow \frac{l+h}{2}$ .
  - if** Indistinguishable **then** break
- end**

**end**

**Output**  $\hat{\alpha}_1, \dots, \hat{\alpha}_{d-1}$

---

-  H. Shao, L. Cohen, A. Blum, Y. Mansour, A. Saha, and M. R. Walter.  
Eliciting user preferences for personalized multi-objective decision making through comparative feedback, 2023.



QUESTIONS