Multi-Objective Reinforcement Learning Using Routed Ensembles and Trajectory Attribution

---

Hongyang R. Zhang

Assistant Professor, Computer Science
Northeastern University, Boston, MA





Based on joint work with Zhenshuo Zhang (PhD student, Northeastern) and Ruoxuan Xiong (Assistant Professor, Emory)

INFORMS 2026, San Francisco, November 3, 2026

Motivation

Trajectory-level data arise from a variety of applications in robotic control, reasoning models and post-training



Multi-Objective Reinforcement Learning for Large-Scale Tote Allocation in Human-Robot Collaborative Fulfillment Centers

Sikata Sengupta [1 *] Guangyi Liu [2 *] Omer Gottesman [2]
Joseph W Durham [2] Michael Kearns [1 2] Aaron Roth [1 2] Michael Caldara [2]

arXiv:2602.24182v1 [cs.LG] 27 Feb 2026

Abstract

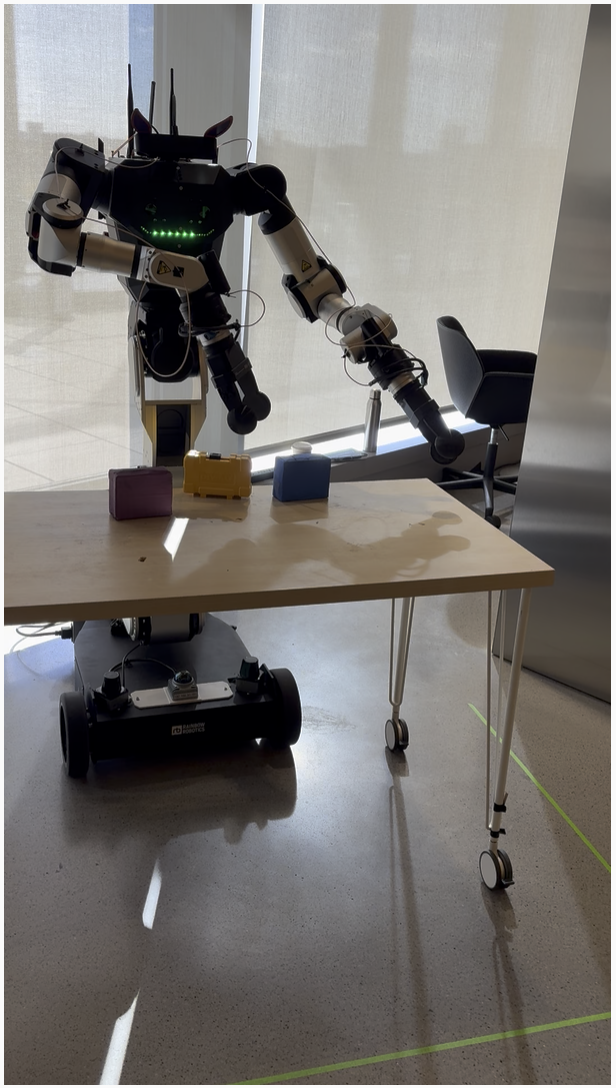
Optimizing the consolidation process in container-based fulfillment centers requires trading off competing objectives such as processing speed, resource usage, and space utilization while adhering to a range of real-world operational constraints. This process involves moving items between containers via a combination of human and robotic workstations to free up space for inbound inventory and increase container utilization. We formulate this problem as a large-scale Multi-Objective Reinforcement Learning (MORL) task with high-dimensional state spaces and dynamic system behavior. Our method builds on recent theoretical advances in solving constrained RL problems via best-response and no-regret dynamics in zero-sum games, enabling principled minimax policy learning. Policy evaluation on realistic warehouse simulations shows that our approach effectively trades off objectives, and we empirically observe that it learns a single policy that simultaneously satisfies all constraints, even if this is not theoretically guaranteed. We further introduce a theoretical framework to handle the problem of error cancellation, where time-averaged solutions display oscillatory behavior. This method returns a single iterate whose Lagrangian value is close to the minimax value of the game. These results demonstrate the promise of MORL in solving complex, high-impact decision-making problems in large-scale industrial systems.

operations. For example, Amazon's robotic fulfillment system, "Sequoia," leverages these collaborations to store inventory 75% faster and process customer orders 25% faster than its prior generation fulfillment system through the integration of multiple robotic products used to containerize inventory and present it at an ergonomic height for processing at an employee workstation. Key to unlocking these speed-ups is efficient resource management within the storage system. In this work, we focus on a particular space management process called "consolidation," wherein items are moved between containers to improve utilization (Figure 1, right) and free up space for new inbound items. Items are stored in containers, known as "totes", which are cyclically inbounded into storage shelves, consolidated to reclaim space, and picked from to fulfill customer orders. Efficient tote selection for consolidation is thus central to optimizing both storage utilization and throughput in this environment, but requires careful management of multiple operational objectives.

For illustration purposes, we assume the following stylized consolidation process: (1) totes can be processed by both humans and robots, each with different manipulation capabilities. We assume that humans are capable of safely manipulating any item, whereas robots are only capable of safely manipulating some items. The task of both operators is to transfer items from a "source" tote to one or more "destination" totes. Source totes are ejected when emptied, and destination totes are ejected when full. A range of factors influence whether a tote is better suited for a human or robot to consolidate including the particular items in the source tote, the destination tote fullness, and the tote properties

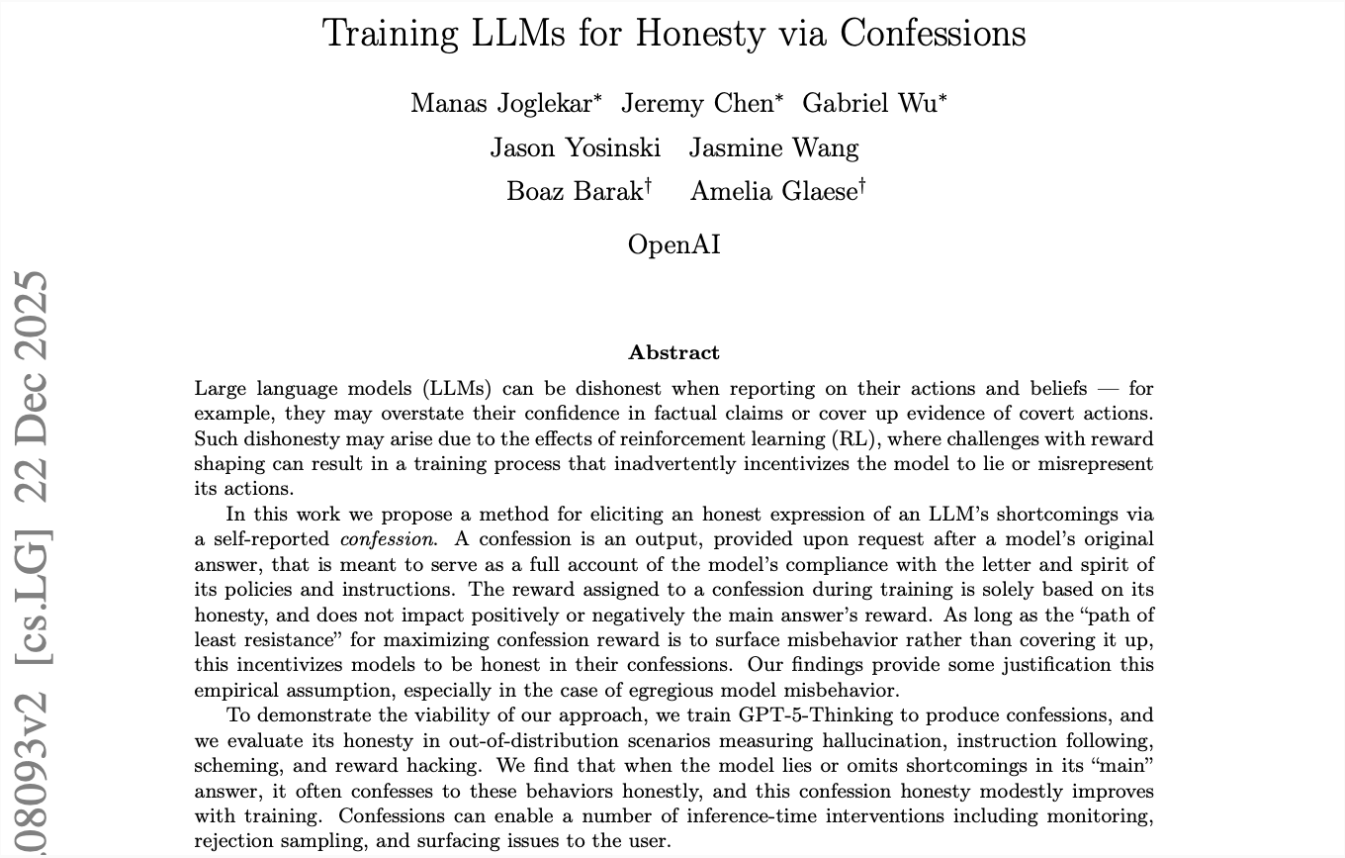
Motivation

Trajectory-level data arise from a variety of applications in robotic control, reasoning models and post-training



Motivation

Trajectory-level data arise from a variety of applications in robotic control, reasoning models and post-training



Training LLMs for Honesty via Confessions

Manas Joglekar* Jeremy Chen* Gabriel Wu*
Jason Yosinski Jasmine Wang
Boaz Barak† Amelia Glaese†
OpenAI

08093v2 [cs.LG] 22 Dec 2025

Abstract

Large language models (LLMs) can be dishonest when reporting on their actions and beliefs — for example, they may overstate their confidence in factual claims or cover up evidence of covert actions. Such dishonesty may arise due to the effects of reinforcement learning (RL), where challenges with reward shaping can result in a training process that inadvertently incentivizes the model to lie or misrepresent its actions.

In this work we propose a method for eliciting an honest expression of an LLM's shortcomings via a self-reported *confession*. A confession is an output, provided upon request after a model's original answer, that is meant to serve as a full account of the model's compliance with the letter and spirit of its policies and instructions. The reward assigned to a confession during training is solely based on its honesty, and does not impact positively or negatively the main answer's reward. As long as the "path of least resistance" for maximizing confession reward is to surface misbehavior rather than covering it up, this incentivizes models to be honest in their confessions. Our findings provide some justification this empirical assumption, especially in the case of egregious model misbehavior.

To demonstrate the viability of our approach, we train GPT-5-Thinking to produce confessions, and we evaluate its honesty in out-of-distribution scenarios measuring hallucination, instruction following, scheming, and reward hacking. We find that when the model lies or omits shortcomings in its "main" answer, it often confesses to these behaviors honestly, and this confession honesty modestly improves with training. Confessions can enable a number of inference-time interventions including monitoring, rejection sampling, and surfacing issues to the user.

RL is widely used for post-training reasoning models (RLHF, RLVR), where one needs to balance between multiple evaluation criteria (honesty, safety, etc)

Talk outline

Central question: How should we balance competing objectives when we design policies in *multi-objective* reinforcement learning?

Motivation

Problem setup and examples

Theoretical analysis

Algorithms

Conclusion

Problem setup

Consider a Markov decision process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \Pr[\cdot \mid \cdot, \cdot], \mathbf{r}, \Omega, f_\Omega, H)$, with state space $\mathcal{S}$, action space $\mathcal{A}$, transition model $\Pr : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, reward vector $r \in \mathbb{R}^m$, preference space Ω and aggregator f_Ω, and number of horizons H

Problem setup

Consider a Markov decision process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \Pr[\cdot \mid \cdot, \cdot], \mathbf{r}, \Omega, f_\Omega, H)$, with state space $\mathcal{S}$, action space $\mathcal{A}$, transition model $\Pr : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, reward vector $r \in \mathbb{R}^m$, preference space Ω and aggregator f_Ω, and number of horizons H

In an RL environment, a trajectory at time $t = 0, 1, \ldots, T-1$ consists of a sequence of state-action pairs

$$\tau^{(t)} = (s_0^{(t)}, a_0^{(t)}, \mathbf{r}_0^{(t)}, \ldots, s_H^{(t)}, a_H^{(t)}, \mathbf{r}_H^{(t)})$$

Problem setup

Consider a Markov decision process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \Pr[\cdot \mid \cdot, \cdot], \mathbf{r}, \Omega, f_\Omega, H)$, with state space $\mathcal{S}$, action space $\mathcal{A}$, transition model $\Pr : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, reward vector $r \in \mathbb{R}^m$, preference space Ω and aggregator f_Ω, and number of horizons H

In an RL environment, a trajectory at time $t = 0, 1, \ldots, T-1$ consists of a sequence of state-action pairs

$$\tau^{(t)} = (s_0^{(t)}, a_0^{(t)}, \mathbf{r}_0^{(t)}, \ldots, s_H^{(t)}, a_H^{(t)}, \mathbf{r}_H^{(t)})$$

Given the reward function $\mathbf{r}$, we can aggregate the cumulative return from the trajectory as:

$$\mathbf{G}(\tau) = \sum_{h=0}^{H} \mathbf{r}(s_h, a_h) \in \mathbb{R}^m$$

The expected return of policy π from an initial distribution $s_0 \sim \rho$:

$$\mathbf{J}(\pi) = \mathop{\mathbb{E}}_{s_0 \sim \rho, \tau \sim \pi} [\mathbf{G}(\tau)] \in \mathbb{R}^m$$

Examples

Consider a simple MDP with two objectives $m = 2$. We illustrate the Pareto frontier in terms of the achieved rewards below, in three regimes:

i) aligned objectives ($(1, 1)$ vs. $(1.2, 1.2)$)

ii) conflicting objectives ($(2, 0)$ vs. $(-2, 0)$) with an approximately linear Pareto frontier: two corner policies whose mixtures span the frontier

iii) conflicting objectives with a strictly concave frontier

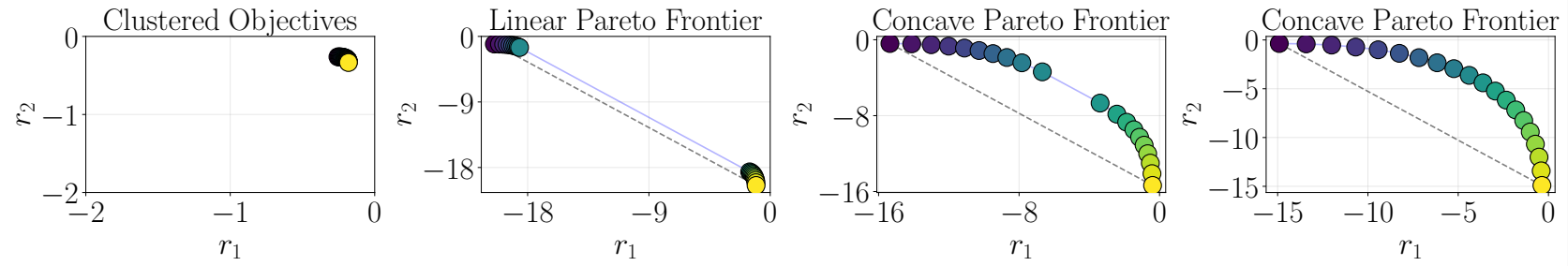


Figure 1: Visualize the Pareto frontier of two objectives (r_1, r_2) for policies optimized under varying preference weights $w \in [0, 1]$ in three regimes

Examples

Consider a simple MDP with two objectives $m = 2$. We illustrate the Pareto frontier in terms of the achieved rewards below, in three regimes:

i) aligned objectives ($(1, 1)$ vs. $(1.2, 1.2)$)

ii) conflicting objectives ($(2, 0)$ vs. $(-2, 0)$) with an approximately linear Pareto frontier: two corner policies whose mixtures span the frontier

iii) conflicting objectives with a strictly concave frontier

Key questions:

- How shall we tackle the problem of objective conflicts in multi-objective reinforcement learning?
- From a computational perspective, is there a procedure to efficiently approximate the Pareto frontier of optimal policies in a multi-objective Markov decision process?

Our focus: we will show how to learn ensemble policies corresponding to a small set of π

Examples

Consider a simple MDP with two objectives $m = 2$. We illustrate the Pareto frontier in terms of the achieved rewards below, in three regimes:

i) aligned objectives ($(1, 1)$ vs. $(1.2, 1.2)$)

ii) conflicting objectives ($(2, 0)$ vs. $(-2, 0)$) with an approximately linear Pareto frontier: two corner policies whose mixtures span the frontier

iii) conflicting objectives with a strictly concave frontier

Key questions:

- How shall we tackle the problem of objective conflicts in multi-objective reinforcement learning?
- From a computational perspective, is there a procedure to efficiently approximate the Pareto frontier of optimal policies in a multi-objective Markov decision process?

Our focus: we will show how to learn ensemble policies corresponding to a small set of π

Outline

Motivation

Problem setup and examples

Theoretical analysis

Algorithms

Conclusion

Theoretical analysis

We focus on linear aggregation operators:

$$\mathcal{J}(\pi; w) = \langle w, \mathbf{J}(\pi) \rangle$$

Definition: Given K basis policies $\pi_1^*, \ldots, \pi_K^*$, let

$$\overline{d}_h(s,a) := \frac{1}{K} \sum_{k=1}^{K} d_h^{\pi_k^*}(s,a)$$

We introduce a measure called *objective conflict coefficient*

$$\Gamma := \max_{h \in [H]} \max_{k \in [K]} \chi^2 \left(d_h^{\pi_k^*} \| \overline{d}_h \right) \tag{1}$$

Regret bounds

Assume there exists a basis of size K such that the optimal policy can always be expressed as mixture of the basis for any $w \in \Omega$. There exists an algorithm that outputs K policies such that the cumulative regret at time T satisfies (with probability at least $1 - \delta$)

$$\mathrm{Reg}(T) \leq \mathcal{O}_T \left(H \sqrt{C_{\mathrm{cov}} (1 + \Gamma) \log(KT\delta^{-1}) T} \right) \tag{2}$$

Theoretical analysis

Proof technique: Builds on *coverability*, a complexity measure defined on a fixed policy class Π [XFB+23]:

$$C_{\mathrm{cov}} := \inf_{\mu \in \Delta(\mathcal{S}\times\mathcal{A})} \sup_{\pi\in\Pi} \max_{h\in[H]} \sup_{(s,a)} \frac{d_h^\pi(s,a)}{\mu_h(s,a)},$$

and earlier work on transfer learning theory [BBC+10]

Extensions and relaxations:

- **Nearly-tight regret lower bound:** We complement the regret bound with *a matching lower bound* in the order of $\Omega_T\left(H\sqrt{(1+\Gamma)T}\right)$, establishing the tightness of the dependence on Γ and T up to C_{cov} and log factors
- **Low-rank basis assumption:** The proof can be extended to account for the existence of an approximate low-rank basis (resulting regret bound preserves the same qualitative dependence on K and Γ)

Theoretical analysis

Proof technique: Builds on *coverability*, a complexity measure defined on a fixed policy class Π [XFB+23]:

$$C_{\mathrm{cov}} := \inf_{\mu \in \Delta(\mathcal{S}\times\mathcal{A})} \sup_{\pi\in\Pi} \max_{h\in[H]} \sup_{(s,a)} \frac{d_h^\pi(s,a)}{\mu_h(s,a)},$$

and earlier work on transfer learning theory [BBC+10]

Extensions and relaxations:

- **Nearly-tight regret lower bound:** We complement the regret bound with *a matching lower bound* in the order of $\Omega_T\left(H\sqrt{(1+\Gamma)T}\right)$, establishing the tightness of the dependence on Γ and T up to C_{cov} and log factors
- **Low-rank basis assumption:** The proof can be extended to account for the existence of an approximate low-rank basis (resulting regret bound preserves the same qualitative dependence on K and Γ)

Theoretical analysis

Proof technique: Builds on *coverability*, a complexity measure defined on a fixed policy class Π [XFB+23]:

$$C_{\mathrm{cov}} := \inf_{\mu \in \Delta(\mathcal{S} \times \mathcal{A})} \sup_{\pi \in \Pi} \max_{h \in [H]} \sup_{(s,a)} \frac{d_h^\pi(s,a)}{\mu_h(s,a)},$$

and earlier work on transfer learning theory [BBC+10]

Extensions and relaxations:

- **Nearly-tight regret lower bound:** We complement the regret bound with *a matching lower bound* in the order of $\Omega_T\left(H\sqrt{(1+\Gamma)T}\right)$, establishing the tightness of the dependence on Γ and T up to C_{cov} and log factors
- **Low-rank basis assumption:** The proof can be extended to account for the existence of an approximate low-rank basis (resulting regret bound preserves the same qualitative dependence on K and Γ)

Outline

Motivation

Problem setup and examples

Theoretical analysis

Algorithms

Conclusion

Learning ensemble policies

Goal: Design an *ensemble policy learning* algorithm that [ZDY+26]

i) explicitly learn a set of N basis policies, with each trajectory routed to the highest-reward basis, and

ii) learn a preference-conditioned transfer operator $f_w : \Pi^N \to \Pi$ that maps these basis policies to a composite policy for any given preference w.

Learning objective:

$$\max_{\pi_1,\pi_2,\ldots,\pi_N} \mathop{\mathbb{E}}_{s_0\sim\rho, w\sim\mathcal{P}} \left[\left\langle w, \sum_{h=0}^{H} \mathbf{r}(s_h, f_w) \right\rangle \right] \tag{3}$$

Optimal ensemble policy follows a routing mechanism

Fix N policies $\{\pi_1, \ldots, \pi_N\}$. Suppose $f_w \in \mathrm{conv}\{\pi_1, \ldots, \pi_N\}$ and $d_h^{f_w}(s,a) = \sum_{i=1}^{N} \alpha_i(w) d_h^{\pi_i}(s,a)$ for $\alpha(w) \in \Delta^{N-1}$. There exists an optimal routing function $\widehat{f}$ for equation (3) such that

$$\text{For every } w \in \Omega,\ \widehat{f}_w = \pi_j, \text{ where } j \in \arg\max_{i\in\{1,2,\ldots,N\}} \langle w, J(\pi_i)\rangle$$

Learning ensemble policies

Goal: Design an *ensemble policy learning* algorithm that [ZDY+26]

i) explicitly learn a set of N basis policies, with each trajectory routed to the highest-reward basis, and

ii) learn a preference-conditioned transfer operator $f_w : \Pi^N \to \Pi$ that maps these basis policies to a composite policy for any given preference w.

Learning objective:

$$\max_{\pi_1,\pi_2,\ldots,\pi_N} \mathop{\mathbb{E}}_{s_0\sim\rho, w\sim\mathcal{P}} \left[\left\langle w, \sum_{h=0}^{H} \mathbf{r}(s_h, f_w) \right\rangle \right] \tag{3}$$

Optimal ensemble policy follows a routing mechanism

Fix N policies $\{\pi_1, \ldots, \pi_N\}$. Suppose $f_w \in \mathrm{conv}\{\pi_1, \ldots, \pi_N\}$ and $d_h^{f_w}(s,a) = \sum_{i=1}^{N} \alpha_i(w) d_h^{\pi_i}(s,a)$ for $\alpha(w) \in \Delta^{N-1}$. There exists an optimal routing function $\widehat{f}$ for equation (3) such that

$$\text{For every } w \in \Omega,\ \widehat{f}_w = \pi_j, \text{ where } j \in \arg\max_{i\in\{1,2,\ldots,N\}} \langle w, J(\pi_i)\rangle$$

Learning ensemble policies

Goal: Design an *ensemble policy learning* algorithm that [ZDY+26]

i) explicitly learn a set of N basis policies, with each trajectory routed to the highest-reward basis, and

ii) learn a preference-conditioned transfer operator $f_w : \Pi^N \to \Pi$ that maps these basis policies to a composite policy for any given preference w.

Learning objective:

$$\max_{\pi_1,\pi_2,\ldots,\pi_N} \mathop{\mathbb{E}}_{s_0\sim\rho, w\sim\mathcal{P}} \left[\left\langle w, \sum_{h=0}^{H} \mathbf{r}(s_h, f_w) \right\rangle \right] \tag{3}$$

Optimal ensemble policy follows a routing mechanism

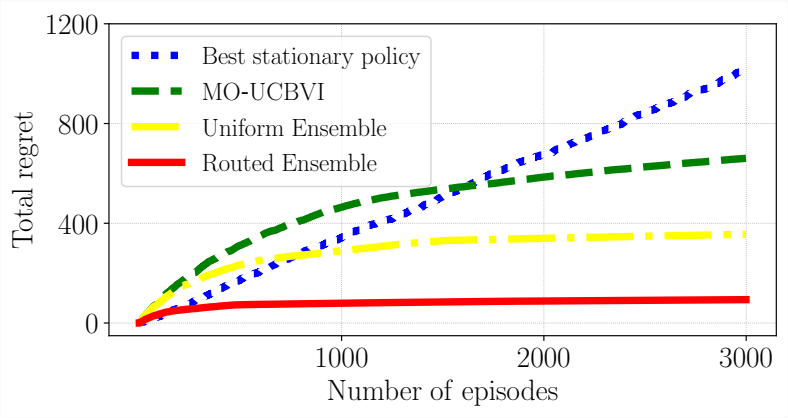
Fix N policies $\{\pi_1, \ldots, \pi_N\}$. Suppose $f_w \in \mathrm{conv}\{\pi_1, \ldots, \pi_N\}$ and $d_h^{f_w}(s,a) = \sum_{i=1}^{N} \alpha_i(w) d_h^{\pi_i}(s,a)$ for $\alpha(w) \in \Delta^{N-1}$. There exists an optimal routing function $\widehat{f}$ for equation (3) such that

$$\text{For every } w \in \Omega,\ \widehat{f}_w = \pi_j, \text{ where } j \in \arg\max_{i\in\{1,2,\ldots,N\}} \langle w, J(\pi_i)\rangle$$

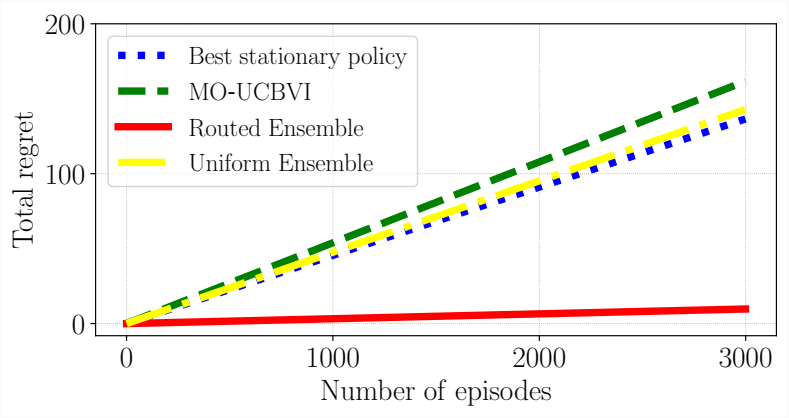
Illustration in synthetic environments

In this experiment, we compare ensemble policy learning with:

1. learning an optimal (stationary) policy that maximizes the average reward
2. learning one policy for each objective (using UCB-VI)
3. uniformly assigning each trajectory to one group and learn an ensemble based on the partition



(a) Learning Pareto frontier



(b) Learning under corrupted examples

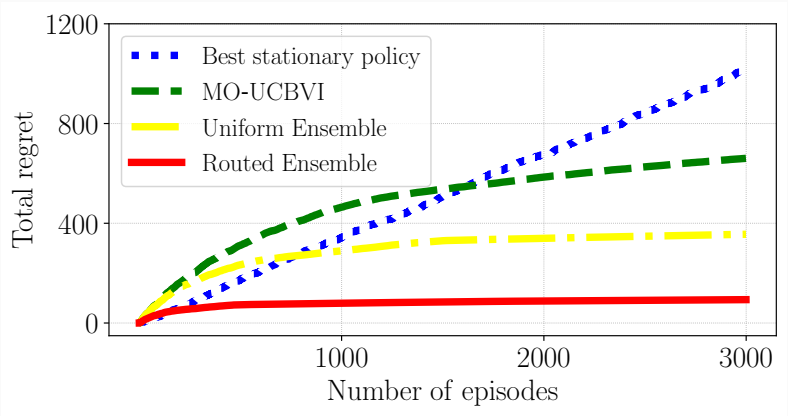
Figure 1: Two synthetic environments: Compare the cumulative regret of routed ensembles against MO-UCBVI, uniform ensembles, and the best stationary policy

More evaluation in real-world RL environments can be found in our paper

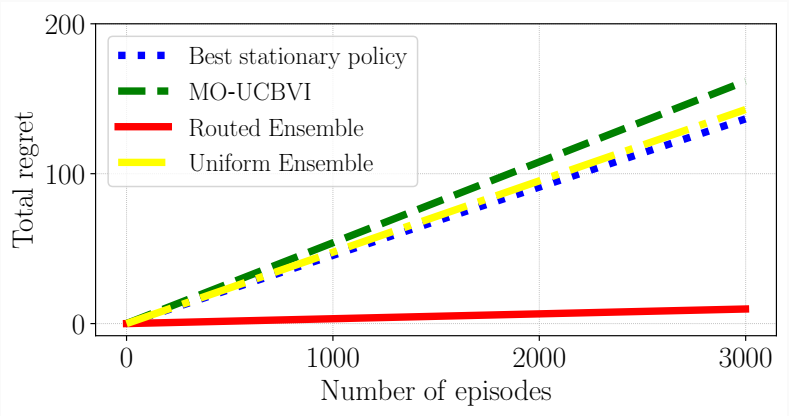
Illustration in synthetic environments

In this experiment, we compare ensemble policy learning with:

1. learning an optimal (stationary) policy that maximizes the average reward
2. learning one policy for each objective (using UCB-VI)
3. uniformly assigning each trajectory to one group and learn an ensemble based on the partition



(a) Learning Pareto frontier



(b) Learning under corrupted examples

Figure 1: Two synthetic environments: Compare the cumulative regret of routed ensembles against MO-UCBVI, uniform ensembles, and the best stationary policy

More evaluation in real-world RL environments can be found in our paper

Trajectory attribution

Rather than partitioning the trajectories into disjoint sets, what if some of the trajectories are corrupted or completely irrelevant?

Trajectory attribution: how could we quantify the influence of individual trajectories upon the learned optimal policy?

1. **Kernel surrogate models:** Estimate a kernel for the input-output mapping of a policy learning algorithm over subsets of objectives
2. **Policy gradient estimation:** An efficient implementation of the kernel surrogate model (based on gradients at a meta-initialization policy)
3. **Clustering and ensemble:** Cluster similar objectives into N groups

Trajectory attribution

Rather than partitioning the trajectories into disjoint sets, what if some of the trajectories are corrupted or completely irrelevant?

Trajectory attribution: how could we quantify the influence of individual trajectories upon the learned optimal policy?

1. **Kernel surrogate models:** Estimate a kernel for the input-output mapping of a policy learning algorithm over subsets of objectives
2. **Policy gradient estimation:** An efficient implementation of the kernel surrogate model (based on gradients at a meta-initialization policy)
3. **Clustering and ensemble:** Cluster similar objectives into N groups

Evaluations on Meta-World (multi-task robotic manipulation), MO-Gymnasium (multi-objective robotic control), and RLHF (HelpSteer2 and UltraFeedback) settings demonstrate both the accuracy and validity of the above trajectory attribution algorithm [ZDY+26]

Summary

- **New theoretical framework** on multi-objective reinforcement learning: We introduce a new measure called the objective conflict coefficients, and derive nearly-optimal regret bounds [ZXZ26]
- **New algorithms** for this problem: learning ensemble policies using routing, and trajectory selection using kernel surrogate models [ZDY+26][1]

Open questions

- What would be a good mechanism to align strategic or game-theoretic incentives of individuals with the overall objective of platform designer (e.g., social welfare, honesty, truthfulness)?
- Is there an efficient algorithm co compute the Pareto frontier of the optimal policies in multi-objective RL?

[1] Link to code repository:
https://github.com/VirtuosoResearch/Multi-objective-reinforcement-learning-code

Summary

- **New theoretical framework** on multi-objective reinforcement learning: We introduce a new measure called the objective conflict coefficients, and derive nearly-optimal regret bounds [ZXZ26]
- **New algorithms** for this problem: learning ensemble policies using routing, and trajectory selection using kernel surrogate models [ZDY+26][1]

Open questions

- What would be a good mechanism to align strategic or game-theoretic incentives of individuals with the overall objective of platform designer (e.g., social welfare, honesty, truthfulness)?
- Is there an efficient algorithm co compute the Pareto frontier of the optimal policies in multi-objective RL?

[1] Link to code repository:
https://github.com/VirtuosoResearch/Multi-objective-reinforcement-learning-code

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). **"A theory of learning from different domains".** In: *Machine learning* 79, pp. 151–175 (14–16).

Xie, T., Foster, D. J., Bai, Y., Jiang, N., and Kakade, S. M. (2023). **"The Role of Coverage in Online Reinforcement Learning".** In: *The Eleventh International Conference on Learning Representations* (14–16).

Zhang, Z., Duan, M., Ye, Y., and Zhang, H. R. (2026a). **"Scalable Multi-Objective and Meta Reinforcement Learning via Gradient Estimation".** In: *AAAI Conference on Artificial Intelligence* (18–20, 24–26).

Zhang, Z., Xiong, R., and Zhang, H. R. (2026b). **"Multi-Objective Reinforcement Learning Using Routed Ensembles and Trajectory Attribution".** In: *Advances in Neural Information Processing Systems* (25, 26).