Adaptive Policy Backbone via Shared Network

Bumgeun Park[1] and Donghwan Lee[1]

[1]Korea Advanced Institute of Science & Technology (KAIST)

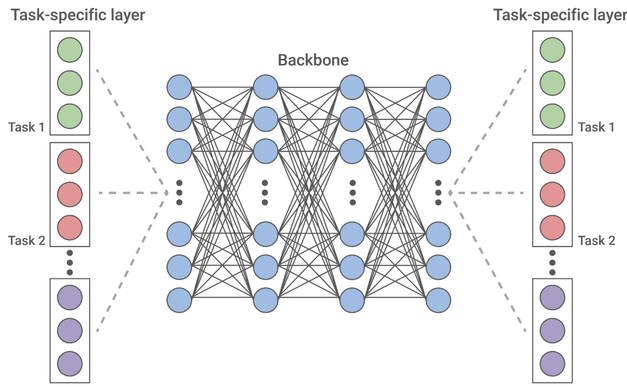
ICML International Conference On Machine Learning

TL;DR

- ❏ We argue that **updating only the first and last layers of the policy network** suffices for effective adaptation to new tasks.
- ❏ To support our main claim, we provide a **theoretical analysis** of the policy structures in MDPs.
- ❏ Interestingly, we observe that even **randomly initialized backbone can yield competitive policies** when paired with appropriate linear pre- and post- mappings.
- ❏ Empirical results demonstrate that this structural constraint can **enhance the generalization capability** of resulting policies.

Introduction

- Although each optimal policy is task-specific, they share a **common component** induced by structural similarities in the underlying MDPs.
- We propose **A**daptive **P**olicy **B**ackbone (APB), which consists of a frozen backbone paired with lightweight, task-specific pre- and post-backbone linear layers.
- We argue that APB can improve **generalization capability**, with its adaptability stemming from a unique synergy between its structural design and meta-learned representations.



Lemma & Theorem

Lemma. *Let $\Pi^{\pi_1}, \Pi^{\pi_2} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}||\mathcal{A}|}$ be the policy matrices for any π_1 and π_2, respectively, and let matrix $A \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ satisfies $AV^{\pi_1} = V^{\pi_2}$. Then there exists another matrix $B \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ such that*

$$A\Pi^{\pi_1} B = \Pi^{\pi_2}$$

- The target policy and source policy matrices are related through linear transformation.
- This serves as the key intermediate result for deriving the unified policy representation theorem.

Assumption. *(Isomorphism) A is a permutation matrix (i.e., the two MDPs are isomorphic up to a permutation of the state space).*

Theorem. *Under the **Isomorphism Assumption**, let $e_s \in \mathbb{R}^{|S|}$ denote the one-hot vector whose s-th entry is 1 and all other entries are 0. Then π_2 can be expressed with any policy π_1 as*

$$\pi_2(\cdot \mid s) = h\Big(\pi_1\big(\cdot \mid g(e_s)\big)\Big)$$

where $g : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ is a linear map (e.g., $g(e_s) = e_s^\top A$) and $h : \mathbb{R}^{|\mathcal{A}|} \to \mathbb{R}^{|\mathcal{A}|}$ is a (possibly state-dependent) linear map.

- The target policy can be fully expressed by the source policy through **linear transformations** applied before and after the policy map.
- This justifies why a frozen backbone only needs linear adaptation layers.
- Although the isomorphism assumption is restrictive and the analysis does not fully capture the complexity of deep RL, the result still provides useful theoretical intuition.

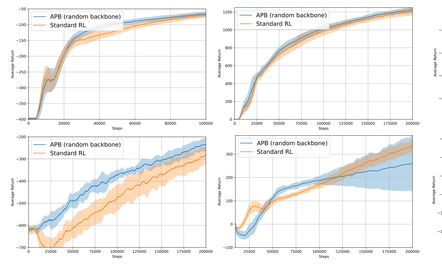
***Remarks**

Notably, the Theorem imposes no specific conditions on the policy backbone, implying that even a **randomly initialized backbone** can yield competitive policies when paired with appropriate linear pre- and post-mappings. This suggests that the APB's structural constraints—high-dimensional projection via a fixed backbone followed by task-specific adaptation—act as a **powerful inductive bias**.

Empirical Summary

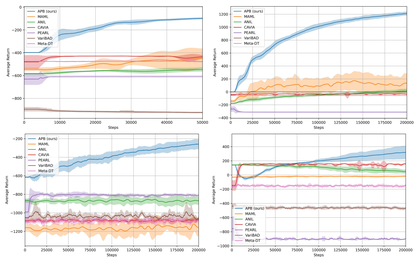
Q) Is training only the pre- and post-backbone linear layers sufficient for effective task adaptation, even when the backbone parameters are randomly initialized?

A)



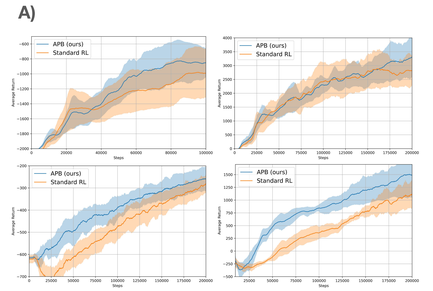
Q) Can the proposed method adapt OOD tasks where existing meta-RL baselines typically fail?

A)



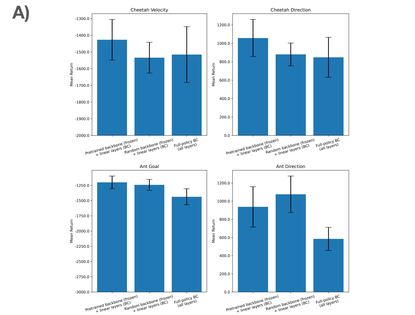
Q) Can the proposed method improve generalization capability of policy on OOD tasks?

A)



Q) How does the meta-trained backbone differ from the randomly initialized one?

A)



- Task adaptation can be effectively achieved by training only the linear pre- and post-mappings, a capability that remarkably persists even with a **randomly initialized backbone**.
- The APB demonstrates robust adaptation performance in OOD scenarios.
- The proposed structural constraints inherently improve **generalization capability**, which is further amplified when paired with **meta-learned representations**.

Contact with me on LinkedIn
Email : j4t123@kaist.ac.kr