

Constrained GPI for Zero-Shot Transfer in Reinforcement Learning

Jaekyeom Kim¹, Seohong Park² & Gunhee Kim¹

1



SEOUL NATIONAL UNIV.
VISION & LEARNING

2

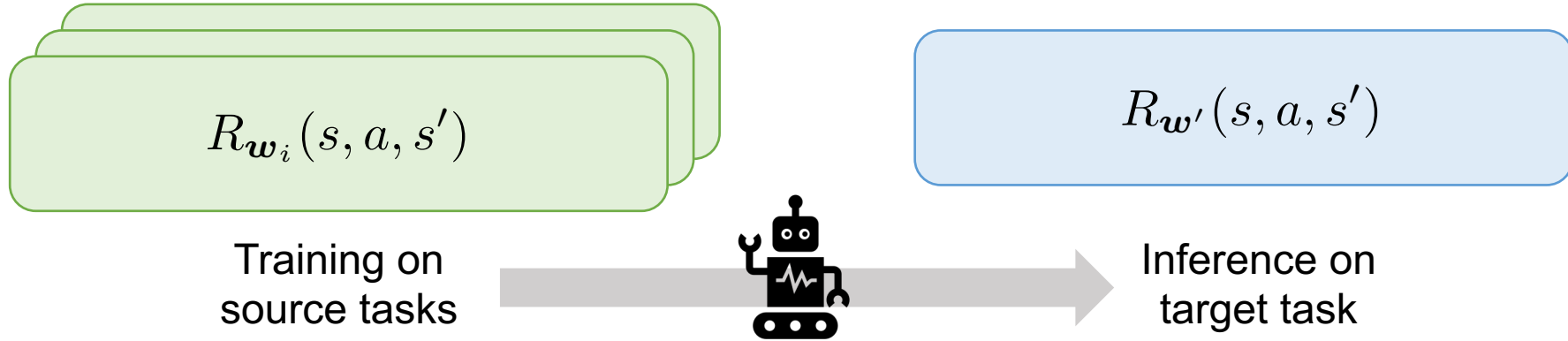


BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH



NEURAL INFORMATION
PROCESSING SYSTEMS

Zero-Shot Transfer in RL

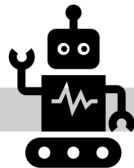


- Target tasks may be unknown during training
- Impractical or impossible to train on all possible tasks
- Re-using trained policies for new tasks could be useful

Zero-Shot Transfer in RL

$$R_{w_i}(s, a, s') = \phi(s, a, s')^\top w_i$$

Training on
source tasks



Inference on
target task

$$R_{w'}(s, a, s') = \phi(s, a, s')^\top w'$$

- Successor features (SFs) [1]

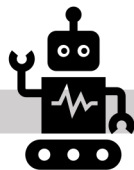
$$Q_w^\pi(s, a) = \psi^\pi(s, a)^\top w$$

- Linear decomposition of value functions
- Allows fast policy *evaluation* on new tasks

Zero-Shot Transfer in RL

$$R_{w_i}(s, a, s') = \phi(s, a, s')^\top w_i$$

Training on
source tasks



Inference on
target task

$$R_{w'}(s, a, s') = \phi(s, a, s')^\top w'$$

- Successor features (SFs) + Generalized policy improvement (GPI) [1]

$$\pi_{\text{GPI}}(s) \in \operatorname{argmax}_a \max_i \tilde{Q}_{w'}^{\pi_i}(s, a)$$

- Combining multiple value functions for policy *improvement*

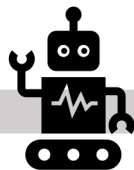
Only fixed sets of SFs can be used for GPI!

No exploitation of “smoothness” of optimal values w.r.t. tasks!

Zero-Shot Transfer in RL

$$R_{w_i}(s, a, s') = \phi(s, a, s')^\top w_i$$

Training on
source tasks



Inference on
target task

$$R_{w'}(s, a, s') = \phi(s, a, s')^\top w'$$

- Universal successor features approximators (USFAs) [2]

$$Q_w^{\pi_z}(s, a) = \psi(s, a, z)^\top w$$

- Performing function approximation even w.r.t. tasks

Function approximators may not generalize well especially to novel, distant tasks!

Key Idea 1



Can we exploit the linearly decomposed reward structure to improve the universal function approximation?

Bounding Optimal Values for Novel Tasks

Theorem 1 (simplified). Given $w' = \sum_{w \in \mathcal{T}} \alpha_w w$ for $\alpha_w \in \mathbb{R}, \forall w \in \mathcal{T}$ the optimal value is lower- and upper-bounded as

$$L_{w', \mathcal{T}}(s, a) \leq Q_{w'}^{\pi_{w'}}(s, a) \leq U_{w', \mathcal{T}, \alpha}(s, a),$$

$$L_{w', \mathcal{T}}(s, a) := \max_{w \in \mathcal{T}} \left[\tilde{Q}_{w'}^{\pi_w}(s, a) - \epsilon_{w'}^{\pi_w}(s, a) \right],$$

$$U_{w', \mathcal{T}, \alpha}(s, a) := \sum_{w \in \mathcal{T}} \max \left\{ \alpha_w \left(\tilde{Q}_w^{\pi_w}(s, a) + \epsilon_w^{\pi_w}(s, a) \right), \alpha_w \frac{1}{1 - \gamma} r_w^{\min} \right\},$$

for maximum approximation error ϵ and minimum reward r_w^{\min} .

- Relaxes prior bounds [3] to a wider range of tasks
- Tightest upper bounds can be computed with an LP solver

Bounding Optimal Values for Novel Tasks

Theorem 1 (simplified). Given $w' = \sum_{w \in \mathcal{T}} \alpha_w w$ for $\alpha_w \in \mathbb{R}, \forall w \in \mathcal{T}$ the optimal value is lower- and upper-bounded as

$$L_{w', \mathcal{T}}(s, a) \leq Q_{w'}^{\pi_{w'}}(s, a) \leq U_{w', \mathcal{T}, \alpha}(s, a),$$

$$L_{w', \mathcal{T}}(s, a) := \max_{w \in \mathcal{T}} \left[\tilde{Q}_{w'}^{\pi_w}(s, a) - \epsilon_{w'}^{\pi_w}(s, a) \right],$$

$$U_{w', \mathcal{T}, \alpha}(s, a) := \sum_{w \in \mathcal{T}} \max \left\{ \alpha_w \left(\tilde{Q}_w^{\pi_w}(s, a) + \epsilon_w^{\pi_w}(s, a) \right), \alpha_w \frac{1}{1 - \gamma} r_w^{\min} \right\},$$

for maximum approximation error ϵ and minimum reward r_w^{\min} .

Can bound optimal values for new tasks using
source SFs with small approximation errors

Constrained Training

- To use the lower and upper bounds as constraints for training
- Training of the universal approximators can be equipped with the constraints

$$L_{\mathbf{w}', \mathcal{T}}(s, a) \leq \tilde{\psi}(s, a, \mathbf{w}')^\top \mathbf{w}' \leq U_{\mathbf{w}', \mathcal{T}, \xi(\mathbf{w}', \mathcal{T}, s, a)}(s, a) \quad \text{for } \mathbf{w}' \in \mathcal{W}$$

- Needs sampling of tasks in source tasks' linear span for the constraints

Key Idea 2



Only source successor features
are needed and considered trustworthy

Constrained GPI

- To apply the constraints at test time right before using GPI

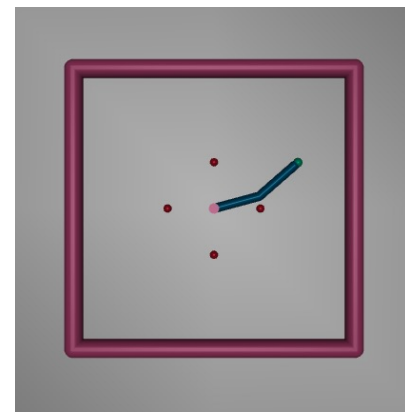
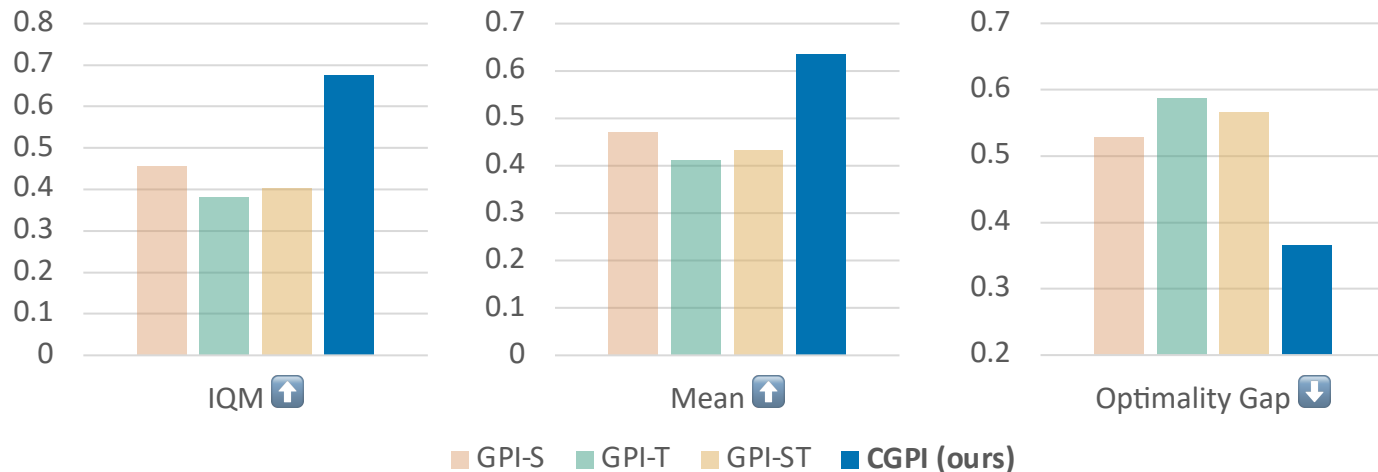
$$\pi_{\text{CGPI}}(s) \in \operatorname{argmax}_a \max_{z \in \mathcal{C}} \left[\min \left\{ \max \left\{ \tilde{Q}_{\mathbf{w}'}^{\pi_z}(s, a), L_{\mathbf{w}', \mathcal{T}}(s, a) \right\}, U_{\mathbf{w}', \mathcal{T}, \xi(\mathbf{w}', \mathcal{T}, s, a)}(s, a) \right\} \right]$$

- Provides an analogous effect to the constrained training
- Do not affect (or hinder) the accuracy of source successor features
- No need for any modification to the training and allows re-using existing models

Experiments

- Robot arm manipulation environment [4]
 - Four goals with different rewards
 - Agent is trained only with each goal with a fixed positive reward and tested on mixtures of positive and negative rewards

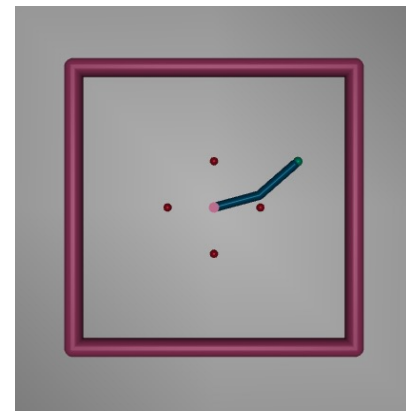
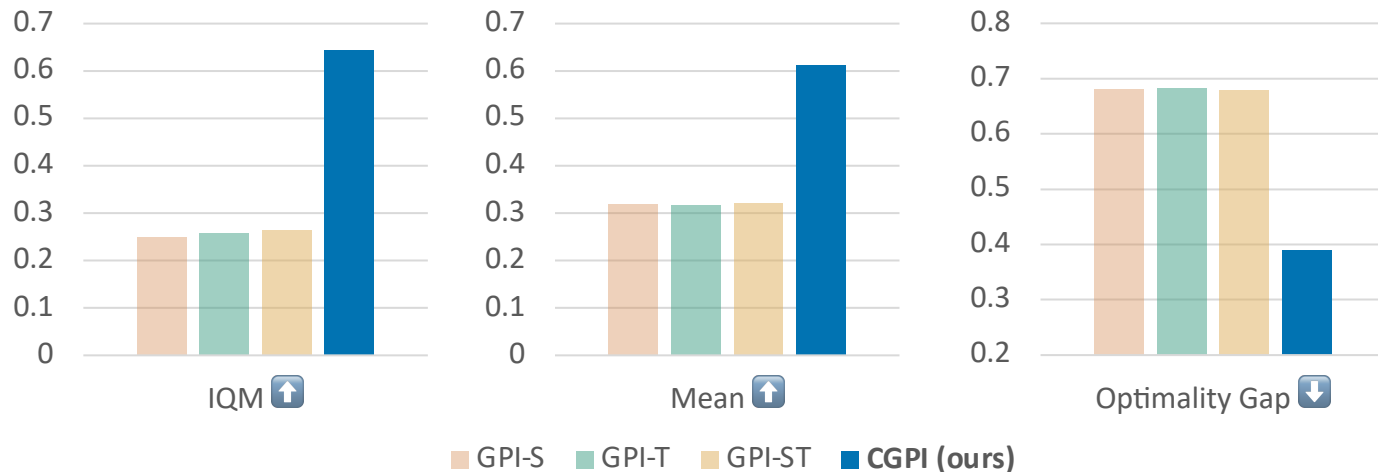
All



Experiments

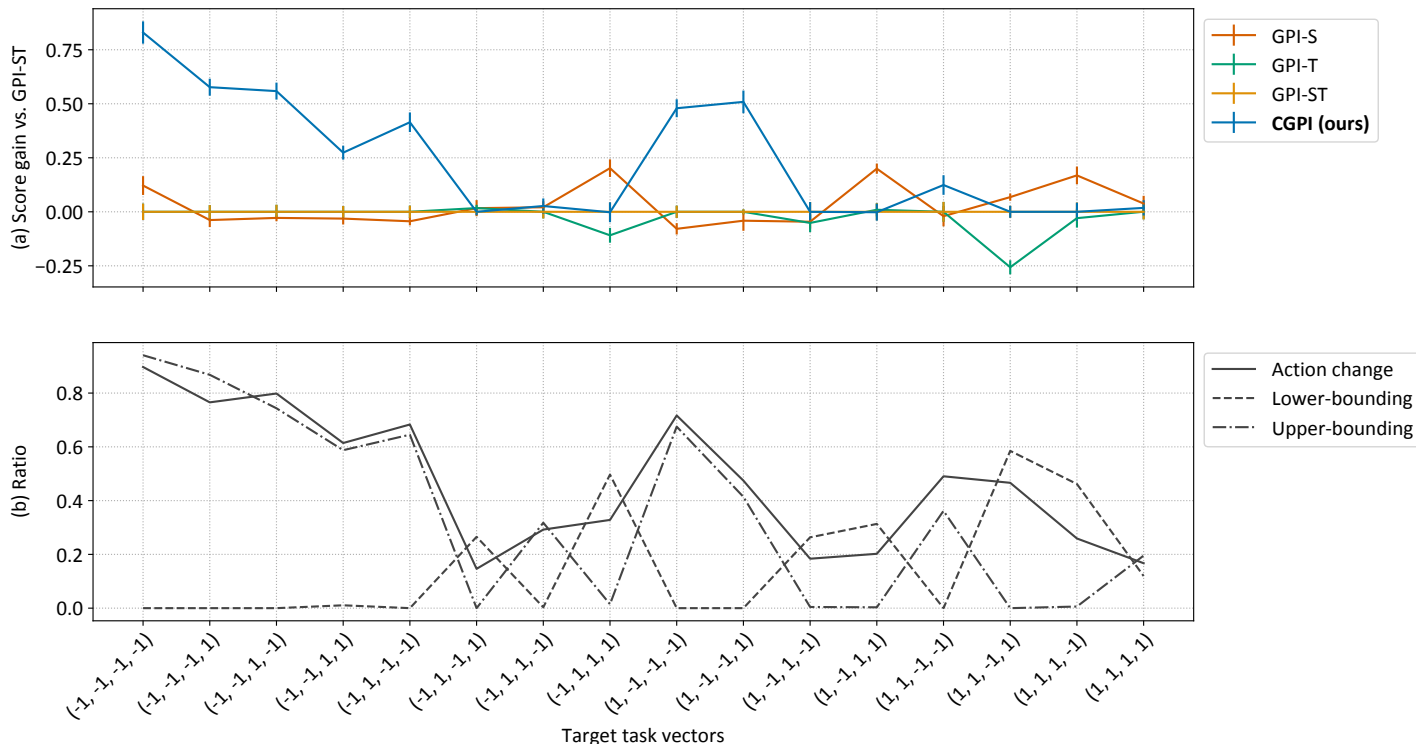
- Robot arm manipulation environment [4]
 - Four goals with different rewards
 - Agent is trained only with each goal with a fixed positive reward and tested on mixtures of positive and negative rewards

Harsh



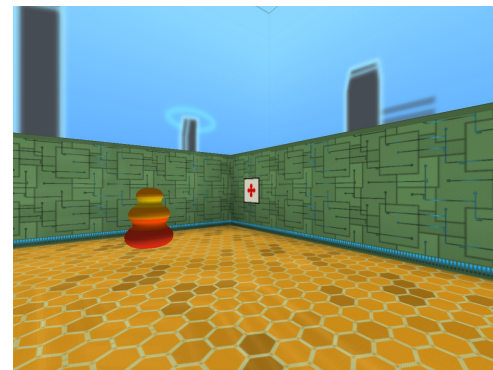
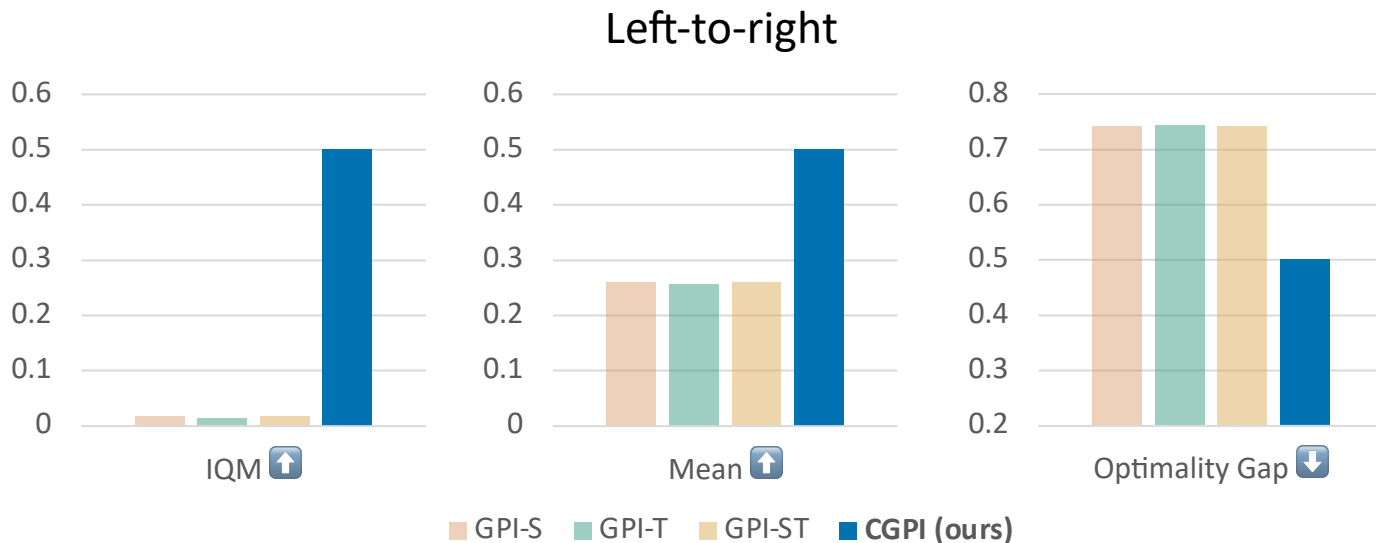
Experiments

- Robot arm manipulation environment [4]



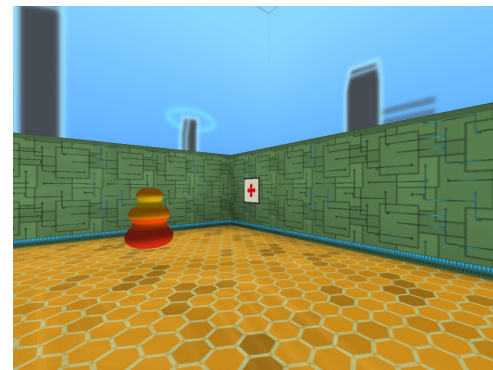
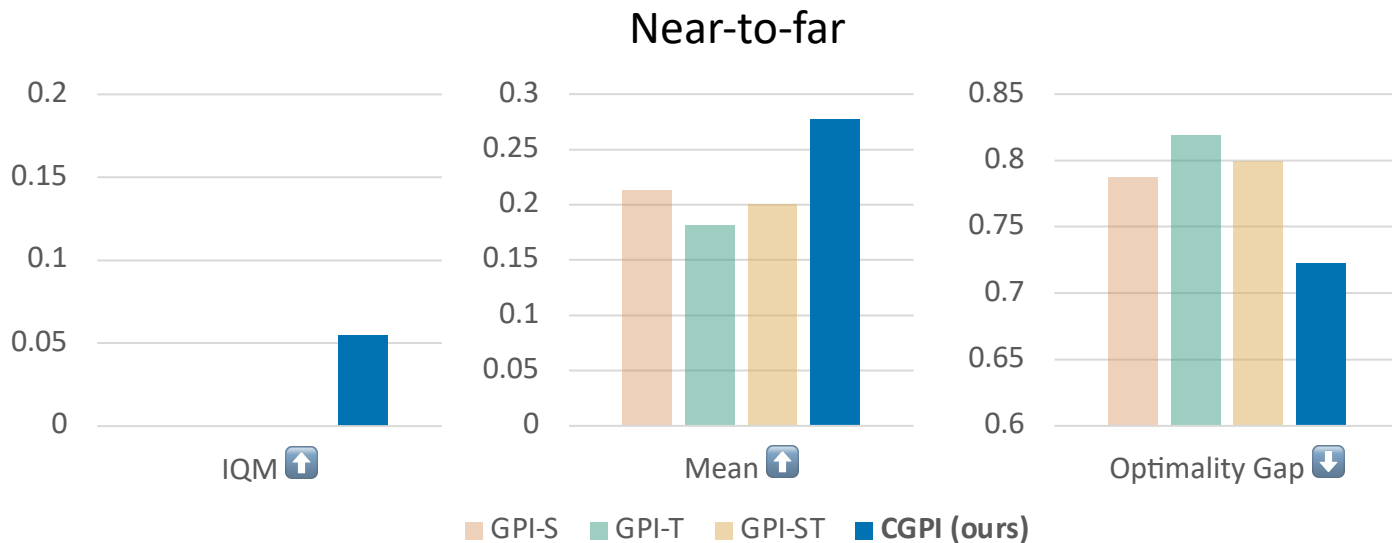
Experiments

- DeepMind Lab environment [5] (with learned ϕ)
 - 3D first-person view, partially observable visual environment
 - The goal object with sparse reward functions
 - Test on goals from a completely disjoint area



Experiments

- DeepMind Lab environment [5] (with learned ϕ)
 - 3D first-person view, partially observable visual environment
 - The goal object with sparse reward functions
 - Test on goals from a completely disjoint area




Conclusion

- Presented lower and upper bounds on optimal values for novel tasks using source successor features
- Proposed constrained GPI, a simple test-time approach to bounding approximation errors and improving performance
- Showed notable performance improvements in robot arm manipulation and 3D first-person view environments

Thank you!

 <https://jaekyeom.github.io/projects/cgpi/>

 jaekyeom@snu.ac.kr