Multi-Step Generalized Policy Improvement by Leveraging Approximate Models

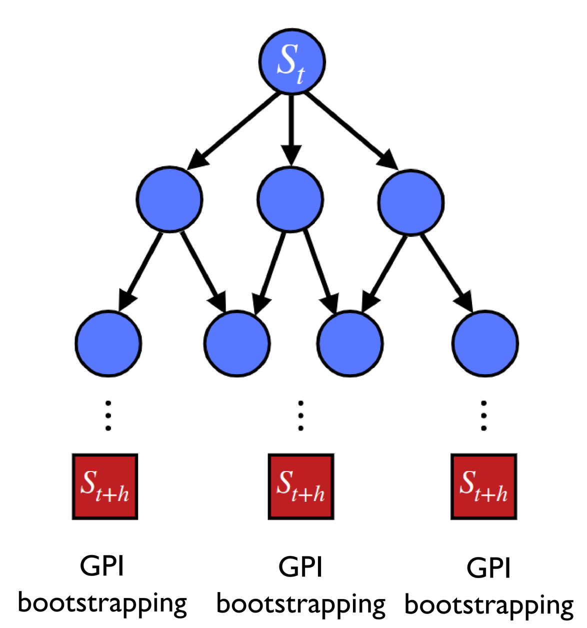
Lucas N. Alegre Ana L. C. Bazzan Ann Nowé Bruno C. da Silva

lnalegre@inf.ufrgs.br @lnalegre github.com/LucasAlegre

Contribution

***h*-GPI**: Multi-Step Generalized Policy Improvement

- Interpolates between model-free GPI and fully model-based planning as a function of the planning horizon h
- Zero-shot policy transfer with performance guarantees by exploiting approximate, imperfect models



h planning steps with approximate model

zero-shot GPI in states reachable in h steps

Successor Features (SFs)

Linear reward: $r_{\mathbf{w}}(s,a,s') = \phi(s,a,s') \cdot \mathbf{w}$

SFs: $\psi^\pi(s,a) \triangleq \mathbb{E}_\pi\left[\sum_{i=0}^{\infty} \gamma^i \phi_{t+i} \mid S_t = s, A_t = a\right]$

Generalized Policy Evaluation (GPE): $q^\pi_{\mathbf{w}}(s,a) = \psi^\pi(s,a) \cdot \mathbf{w}$

Generalized Policy Improvement (GPI)

Computation of a policy π' that improves over a **set** *of policies* $\Pi = \{\pi_i\}_{i=1}^n$

$$\boxed{\pi^{GPI}(s;\mathbf{w}) = \arg\max_{a\in\mathcal{A}} \max_{\pi\in\Pi} q^\pi_{\mathbf{w}}(s,a)}$$

GPI Theorem

$$q^{GPI}_{\mathbf{w}}(s,a) \geq \max_{\pi\in\Pi} q^\pi_{\mathbf{w}}(s,a) \quad \textit{for any } \mathbf{w} \in \mathcal{W}$$

h-GPI: Multi-Step Generalized Policy Improvement

- The ***h*-GPI** policy with planning horizon $h \geq 0$ is defined as:

$$\pi^{h\text{-}GPI}(s) \in \arg\max_{a\in\mathcal{A}} (\mathcal{T}^*_m)^h \max_{\pi\in\Pi} q^\pi(s,a)$$

$$=$$

$$\arg\max_{a\in\mathcal{A}} \max_{\mu_1\ldots\mu_{h-1}} \mathbb{E}_m\left[\underbrace{\sum_{k=0}^{h-1}\gamma^k r(S_{t+k},\mu_k(S_{t+k}))}_{\text{online planning}} + \underbrace{\gamma^h \max_{a'\in\mathcal{A}}\max_{\pi\in\Pi} q^\pi(S_{t+h},a')}_{\text{GPI}} \mid \mu_0(S_t) = a\right]$$

where μ_k is any policy the agent could choose to deploy at time k.

$\Pi = \{\pi_i\}_{i=1}^n$: set of policies $\quad m = (p, r)$: model

We characterize h-GPI's performance lower bound and optimality gap as a function of:

- h (planning horizon)
- $\{w_i\}_{i=1}^n$ (reward weights for which policies in Π are optimal)
- ϵ (action-value function error)
- $\epsilon_p, \epsilon_r,$ (model errors w.r.t. transition function p and reward r)

Theorem 1 (lower bound):

$$q^{h-GPI}(s,a) \geq \max_{\pi\in\Pi} q^\pi(s,a) - \frac{2}{1-\gamma}\left(\gamma^h\epsilon + c(\epsilon_p,\epsilon_r,h)\right)$$

Theorem 2 (optimality gap):

$$q^*_{\mathbf{w}} - q^{h-GPI}_{\mathbf{w}}(s,a) \geq \frac{2}{1-\gamma}\left(\phi_{max}\min_i ||\mathbf{w}-\mathbf{w}_i|| + \gamma^h\epsilon + c(\epsilon_p,\epsilon_r,h)\right)$$

where $c(\epsilon_p,\epsilon_r,h)) = \frac{1-\gamma^h}{1-\gamma}(\epsilon_r + \gamma\epsilon_p v^*_{max})$

Zero-Shot Transfer with h-GPI and SFs

Goal: Solve any task in $\mathcal{M}^\phi \triangleq \{\, M = (\mathcal{S},\mathcal{A},p,r_{\mathbf{w}},\gamma) \mid r_{\mathbf{w}} = \phi(s,a,s')\cdot\mathbf{w} \,\}$

Algorithm 1: h-GPI with Successor Features

Input : Model $\hat{m} = (\hat{p},\hat{\phi})$, SFs $\{\hat{\psi}^{\pi_i}\}_{i=1}^n$, planning horizon $h \geq 0$, state s, reward weights $\mathbf{w}$

for *action* $a \in \mathcal{A}$ **do**

  Let $S_t = s$, $\mu_0(s) = a$

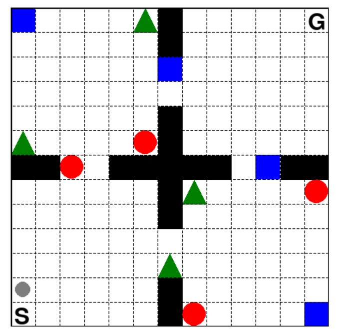
  Compute $(\mathcal{T}^*_{\hat{m}})^h \max_{\pi\in\Pi} \hat{q}^\pi_{\mathbf{w}}(s,a) \leftarrow$

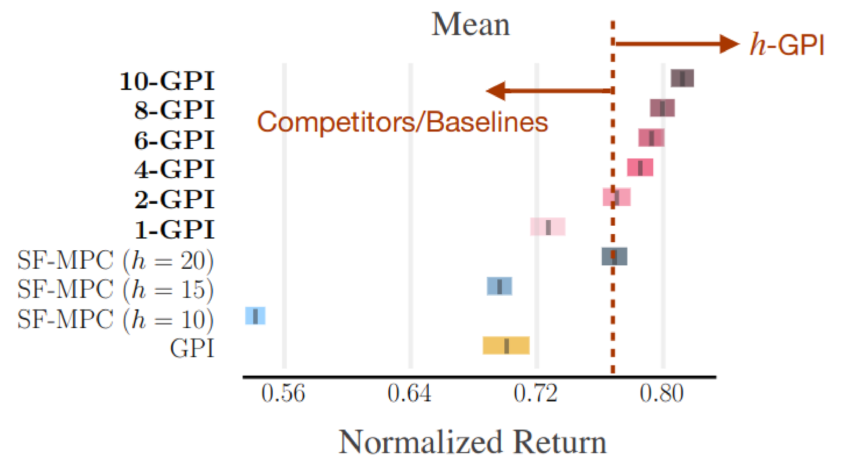
  $$\max_{\mu_1\ldots\mu_{h-1}} \mathbb{E}_{\hat{m}}\left[\sum_{k=0}^{h-1}\gamma^k\hat{\phi}_{t+k}(\hat{S}_{t+k},\mu_k(\hat{S}_{t+k}))\cdot\mathbf{w} + \gamma^h \max_{a'\in\mathcal{A}}\max_{\pi\in\Pi}\hat{\psi}^\pi(\hat{S}_{t+h},a')\cdot\mathbf{w}\right]$$

Return: $\pi^{h\text{-GPI}}(s;\mathbf{w}) \in \arg\max_{a\in\mathcal{A}}(\mathcal{T}^*_{\hat{m}})^h \max_{\pi_i\in\Pi}\hat{q}^{\pi_i}_{\mathbf{w}}(s,a)$

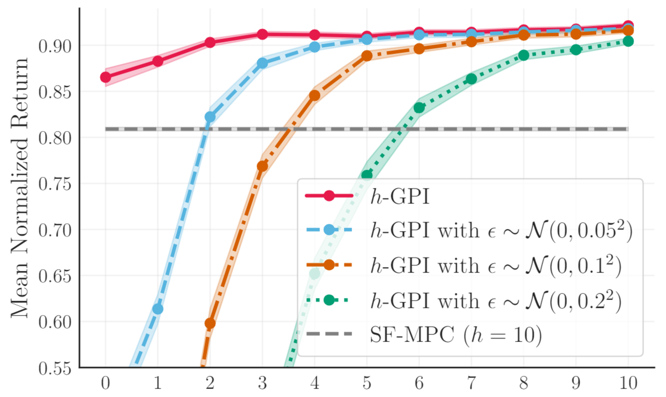
Experiments & Results

Four Room



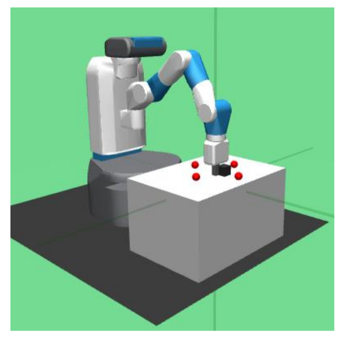


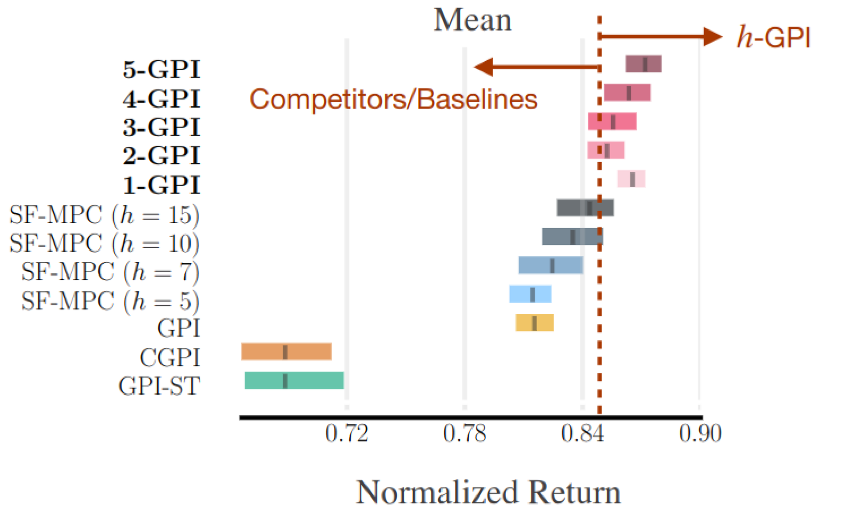
h-GPI outperforms SF-MPC baseline using *ten times* fewer planning steps



h-GPI is less susceptible to value function approximation errors as h increases

Fetch Push





h-GPI outperforms competitors under all values of h using a learned model

Discussion & Conclusion

- h-GPI: multi-step extension of GPI
  - Interpolates between model-free GPI ($h = 0$) and fully model-based planning ($h \to \infty$)
  - Exploits approximate models
  - Solves tasks in a zero-shot manner
- h *trades-off* approximation errors in the agent's:
  - Learned model
  - Action-value functions