DeepMind

Generalised Policy Improvement with Geometric Policy Composition

Shantanu Thakoor*, **Mark Rowland***, Diana Borsa,
Will Dabney, Rémi Munos, André Barreto

ICML 2022

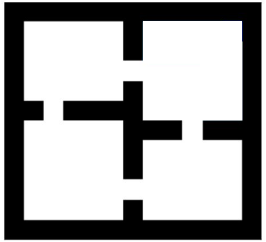
Policy improvement

Policy improvement

A motivating problem: Transfer

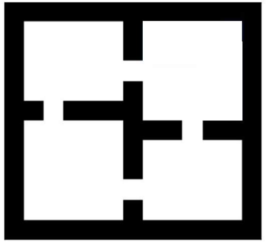
Policy improvement

A motivating problem: Transfer



Policy improvement

A motivating problem: Transfer

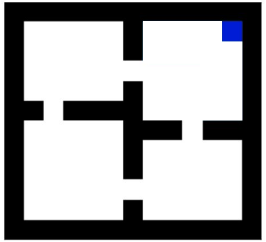


Known policies

$\pi_{\mathrm{U}}, \pi_{\mathrm{L}}, \pi_{\mathrm{R}}, \pi_{\mathrm{D}}$

Policy improvement

A motivating problem: Transfer

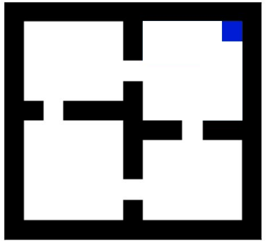


Known policies
$\pi_{\mathrm{U}}, \pi_{\mathrm{L}}, \pi_{\mathrm{R}}, \pi_{\mathrm{D}}$

New goal location indicated.

Policy improvement

A motivating problem: Transfer



Known policies

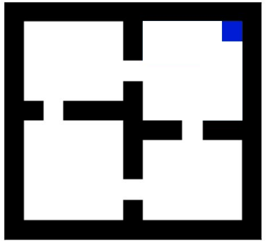
$\pi_{\mathrm{U}}, \pi_{\mathrm{L}}, \pi_{\mathrm{R}}, \pi_{\mathrm{D}}$

New goal location indicated.

Quickly derive improved policy for new task.

Policy improvement

A motivating problem: Transfer



Known policies
$\pi_{\mathrm{U}}, \pi_{\mathrm{L}}, \pi_{\mathrm{R}}, \pi_{\mathrm{D}}$

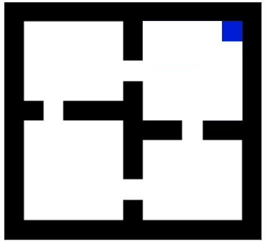
New goal location indicated.

Quickly derive improved policy for new task.

Generalised policy improvement (GPI)
(Barreto et al., 2017)

Policy improvement

A motivating problem: Transfer



Known policies
$\pi_{\mathrm{U}}, \pi_{\mathrm{L}}, \pi_{\mathrm{R}}, \pi_{\mathrm{D}}$

New goal location indicated.

Quickly derive improved policy for new task.

Generalised policy improvement (GPI)
(Barreto et al., 2017)

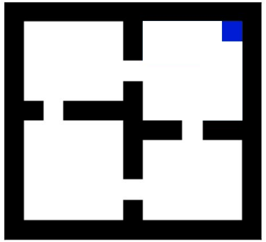
π_1

$\vdots$

π_k

Policy improvement

A motivating problem: Transfer



Known policies
$\pi_{\mathrm{U}}, \pi_{\mathrm{L}}, \pi_{\mathrm{R}}, \pi_{\mathrm{D}}$

New goal location indicated.

Quickly derive improved policy for new task.

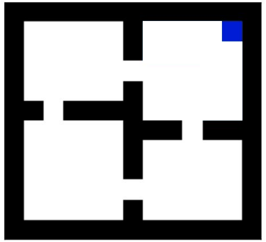
Generalised policy improvement (GPI)
(Barreto et al., 2017)

π_1
$\vdots$
π_k

$$\arg\max_{a} \max_{i=1,\ldots,k} Q^{\pi_i}(x, a)$$

Policy improvement

A motivating problem: Transfer



Known policies
$\pi_{\mathrm{U}}, \pi_{\mathrm{L}}, \pi_{\mathrm{R}}, \pi_{\mathrm{D}}$

New goal location indicated.

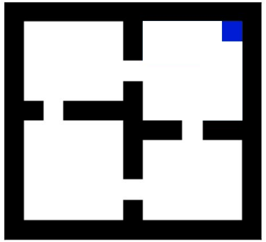
Quickly derive improved policy for new task.

Generalised policy improvement (GPI)
(Barreto et al., 2017)

$$\pi_1, \ldots, \pi_k \rightarrow \boxed{\arg\max_a \max_{i=1,\ldots,k} Q^{\pi_i}(x,a)} \rightarrow \pi'$$

Policy improvement

A motivating problem: Transfer



Known policies
$\pi_{\mathrm{U}}, \pi_{\mathrm{L}}, \pi_{\mathrm{R}}, \pi_{\mathrm{D}}$

New goal location indicated.

Quickly derive improved policy for new task.

Generalised policy improvement (GPI)
(Barreto et al., 2017)

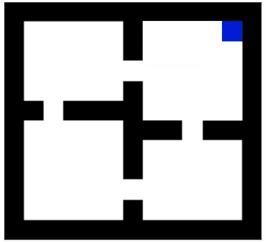
$\pi_1, \ldots, \pi_k$

$$\arg\max_a \max_{i=1,\ldots,k} Q^{\pi_i}(x,a) \rightarrow \pi'$$

Guarantee: $Q^{\pi'} \geq \max_{i=1,\ldots,k} Q^{\pi_i}$

Policy improvement

A motivating problem: Transfer



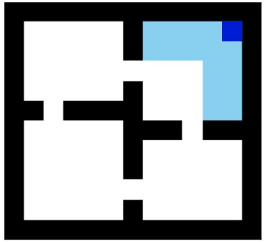
Known policies

$\pi_U, \pi_L, \pi_R, \pi_D$

New goal location indicated.

Quickly derive improved policy for new task.

In this case, GPI produces optimal behaviour only at nearby states.



Generalised policy improvement (GPI)

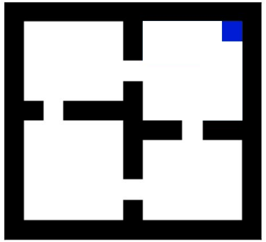
(Barreto et al., 2017)

$$\pi_1, \ldots, \pi_k \rightarrow \boxed{\arg\max_a \max_{i=1,\ldots,k} Q^{\pi_i}(x,a)} \rightarrow \pi'$$

Guarantee: $Q^{\pi'} \geq \max_{i=1,\ldots,k} Q^{\pi_i}$

Policy improvement

A motivating problem: Transfer

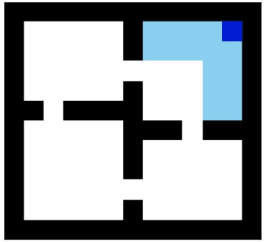


Known policies
$\pi_{\mathrm{U}}, \pi_{\mathrm{L}}, \pi_{\mathrm{R}}, \pi_{\mathrm{D}}$

New goal location indicated.

Quickly derive improved policy for new task.

In this case, GPI produces optimal behaviour only at nearby states.



Central question:
Can we use more information about $\pi_1, \ldots, \pi_k$ to get an even stronger improvement than GPI?

Generalised policy improvement (GPI)

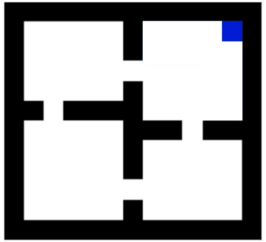
(Barreto et al., 2017)

$$\pi_1, \ldots, \pi_k \rightarrow \arg\max_a \max_{i=1,\ldots,k} Q^{\pi_i}(x,a) \rightarrow \pi'$$

Guarantee: $Q^{\pi'} \geq \max_{i=1,\ldots,k} Q^{\pi_i}$

Policy improvement

A motivating problem: Transfer



Known policies
$\pi_{\mathrm{U}}, \pi_{\mathrm{L}}, \pi_{\mathrm{R}}, \pi_{\mathrm{D}}$

New goal location indicated.

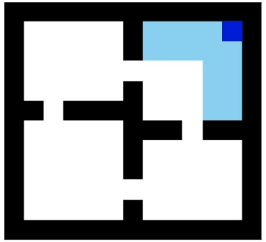
Quickly derive improved policy for new task.

Generalised policy improvement (GPI)
(Barreto et al., 2017)

$$\pi_1, \ldots, \pi_k \rightarrow \boxed{\arg\max_a \max_{i=1,\ldots,k} Q^{\pi_i}(x,a)} \rightarrow \pi'$$

Guarantee: $Q^{\pi'} \geq \max_{i=1,\ldots,k} Q^{\pi_i}$

In this case, GPI produces optimal behaviour only at nearby states.

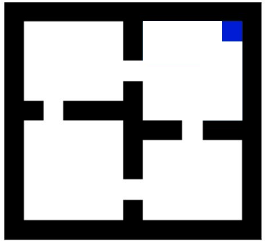


Central question:
Can we use more information about $\pi_1, \ldots, \pi_k$ to get an even stronger improvement than GPI?

Core ideas:

Policy improvement

A motivating problem: Transfer



Known policies
$\pi_{\mathrm{U}}, \pi_{\mathrm{L}}, \pi_{\mathrm{R}}, \pi_{\mathrm{D}}$

New goal location indicated.

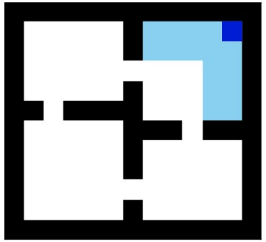
Quickly derive improved policy for new task.

Generalised policy improvement (GPI)
(Barreto et al., 2017)

$\pi_1, \ldots, \pi_k \rightarrow \boxed{\arg\max_a \max_{i=1,\ldots,k} Q^{\pi_i}(x,a)} \rightarrow \pi'$

Guarantee: $Q^{\pi'} \geq \max_{i=1,\ldots,k} Q^{\pi_i}$

In this case, GPI produces optimal behaviour only at nearby states.



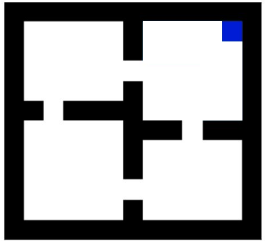
Central question:
Can we use more information about $\pi_1, \ldots, \pi_k$ to get an even stronger improvement than GPI?

Core ideas:

Evaluate certain **non-Markov behaviours** that switch amongst $\pi_1, \ldots, \pi_k$ within episodes, without any additional learning.

Policy improvement

A motivating problem: Transfer



Known policies
$\pi_{\mathrm{U}}, \pi_{\mathrm{L}}, \pi_{\mathrm{R}}, \pi_{\mathrm{D}}$

New goal location indicated.

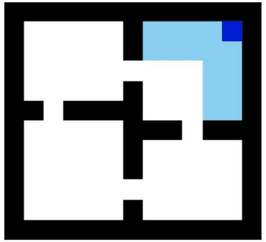
Quickly derive improved policy for new task.

Generalised policy improvement (GPI)
(Barreto et al., 2017)

$$\pi_1, \ldots, \pi_k \rightarrow \boxed{\arg\max_a \max_{i=1,\ldots,k} Q^{\pi_i}(x,a)} \rightarrow \pi'$$

Guarantee: $Q^{\pi'} \geq \max_{i=1,\ldots,k} Q^{\pi_i}$

In this case, GPI produces optimal behaviour only at nearby states.



Central question:
Can we use more information about $\pi_1, \ldots, \pi_k$ to get an even stronger improvement than GPI?

Core ideas:

Evaluate certain **non-Markov behaviours** that switch amongst $\pi_1, \ldots, \pi_k$ within episodes, without any additional learning.

Strengthen GPI to improve over these non-Markov behaviours too.

Improving over non-Markov geometric switching policies (GSPs)

Evaluating GSPs with geometric horizon models (GHMs)

Learning GHMs with cross-entropy TD (CETD)

Improving over non-Markov geometric switching policies (GSPs)

Evaluating GSPs with geometric horizon models (GHMs)

Learning GHMs with cross-entropy TD (CETD)

Improving over non-Markov policies

Improving over non-Markov policies

A **geometric switching policy (GSP)**

$$\nu = \pi^{(1)} \xrightarrow{\alpha} \cdots \xrightarrow{\alpha} \pi^{(n)}$$

is a non-Markov behaviour that:

- Begins the episode using $\pi^{(1)}$.
- At each time step, switches to
  the next policy in the list with probability α.

Switching times are **geometrically** distributed.

Improving over non-Markov policies

A **geometric switching policy (GSP)**

$$\nu = \pi^{(1)} \xrightarrow{\alpha} \cdots \xrightarrow{\alpha} \pi^{(n)}$$

is a non-Markov behaviour that:

- Begins the episode using $\pi^{(1)}$.
- At each time step, switches to the next policy in the list with probability α.

Switching times are **geometrically** distributed.

<u>Geometric generalised policy improvement</u> (GGPI)

Improving over non-Markov policies

A **geometric switching policy (GSP)**

$$\nu = \pi^{(1)} \overset{\alpha}{\to} \cdots \overset{\alpha}{\to} \pi^{(n)}$$

is a non-Markov behaviour that:

- Begins the episode using $\pi^{(1)}$.
- At each time step, switches to the next policy in the list with probability α.

Switching times are **geometrically** distributed.

Geometric generalised policy improvement (GGPI)

π_1

$\vdots$

π_k

Base policies

Improving over non-Markov policies

A **geometric switching policy (GSP)**

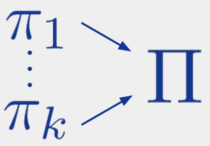
$$\nu = \pi^{(1)} \xrightarrow{\alpha} \cdots \xrightarrow{\alpha} \pi^{(n)}$$

is a non-Markov behaviour that:

- Begins the episode using $\pi^{(1)}$.
- At each time step, switches to the next policy in the list with probability α.

Switching times are **geometrically** distributed.

Geometric generalised policy improvement (GGPI)



Base policies

Set of GSPs

Improving over non-Markov policies

A **geometric switching policy (GSP)**

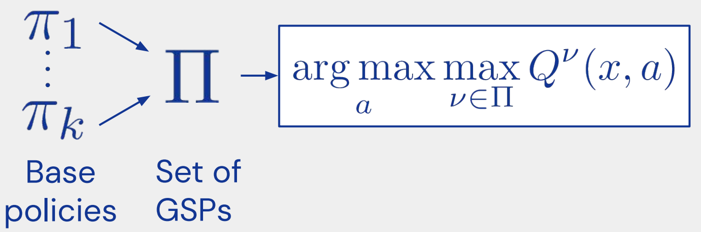
$$\nu = \pi^{(1)} \xrightarrow{\alpha} \cdots \xrightarrow{\alpha} \pi^{(n)}$$

is a non-Markov behaviour that:

- Begins the episode using $\pi^{(1)}$.
- At each time step, switches to the next policy in the list with probability α.

Switching times are **geometrically** distributed.

Geometric generalised policy improvement (GGPI)



Improving over non-Markov policies

A **geometric switching policy (GSP)**

$$\nu = \pi^{(1)} \xrightarrow{\alpha} \cdots \xrightarrow{\alpha} \pi^{(n)}$$

is a non-Markov behaviour that:

- Begins the episode using $\pi^{(1)}$.
- At each time step, switches to the next policy in the list with probability α.

Switching times are **geometrically** distributed.

Geometric generalised policy improvement (GGPI)

$$\pi_1, \ldots, \pi_k \rightarrow \Pi \rightarrow \boxed{\arg\max_a \max_{\nu \in \Pi} Q^{\nu}(x, a)} \rightarrow \pi'$$

Base policies ($\pi_1, \ldots, \pi_k$) — Set of GSPs (Π)

Improving over non-Markov policies

A **geometric switching policy (GSP)**

$$\nu = \pi^{(1)} \xrightarrow{\alpha} \cdots \xrightarrow{\alpha} \pi^{(n)}$$

is a non-Markov behaviour that:

- Begins the episode using $\pi^{(1)}$.
- At each time step, switches to the next policy in the list with probability α.

Switching times are **geometrically** distributed.

Geometric generalised policy improvement (GGPI)

$\pi_1, \ldots, \pi_k$ → Π → $\arg\max_a \max_{\nu \in \Pi} Q^{\nu}(x, a)$ → π'

Base policies · Set of GSPs

Guarantee: $Q^{\pi'} \geq \max_{\nu \in \Pi} Q^{\nu}$

as long as closure condition on Π holds (see paper)

Improving over non-Markov policies

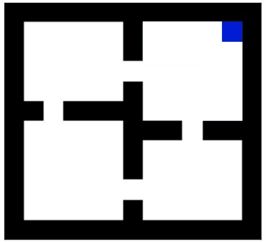
A **geometric switching policy (GSP)**

$$\nu = \pi^{(1)} \xrightarrow{\alpha} \cdots \xrightarrow{\alpha} \pi^{(n)}$$

is a non-Markov behaviour that:

- Begins the episode using $\pi^{(1)}$.
- At each time step, switches to the next policy in the list with probability α.

Switching times are **geometrically** distributed.



Known policies $\pi_U, \pi_L, \pi_R, \pi_D$

New goal location indicated.

Quickly derive improved policy for new task.

Geometric generalised policy improvement (GGPI)

$$\pi_1, \ldots, \pi_k \rightarrow \Pi \rightarrow \boxed{\arg\max_a \max_{\nu \in \Pi} Q^\nu(x,a)} \rightarrow \pi'$$

Base policies — Set of GSPs

Guarantee: $Q^{\pi'} \geq \max_{\nu \in \Pi} Q^\nu$

as long as closure condition on Π holds (see paper)

Improving over non-Markov policies

A **geometric switching policy (GSP)**

$$\nu = \pi^{(1)} \xrightarrow{\alpha} \cdots \xrightarrow{\alpha} \pi^{(n)}$$

is a non-Markov behaviour that:

- Begins the episode using $\pi^{(1)}$.
- At each time step, switches to the next policy in the list with probability α.

Switching times are **geometrically** distributed.

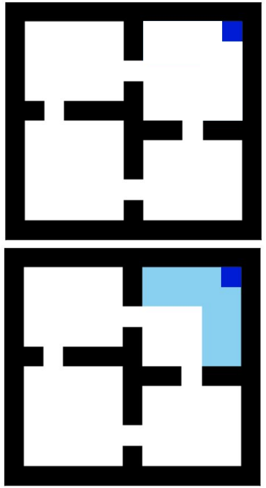
Geometric generalised policy improvement (GGPI)

$$\pi_1, \ldots, \pi_k \rightarrow \Pi \rightarrow \boxed{\arg\max_a \max_{\nu \in \Pi} Q^{\nu}(x, a)} \rightarrow \pi'$$

Base policies — Set of GSPs

Guarantee: $Q^{\pi'} \geq \max_{\nu \in \Pi} Q^{\nu}$

as long as closure condition on Π holds (see paper)



Known policies

$\pi_{\mathrm{U}}, \pi_{\mathrm{L}}, \pi_{\mathrm{R}}, \pi_{\mathrm{D}}$

New goal location indicated.

Quickly derive improved policy for new task.

GPI produces optimal behaviour only at nearby states.

Improving over non-Markov policies

A **geometric switching policy (GSP)**

$$\nu = \pi^{(1)} \overset{\alpha}{\rightarrow} \cdots \overset{\alpha}{\rightarrow} \pi^{(n)}$$

is a non-Markov behaviour that:

- Begins the episode using $\pi^{(1)}$.
- At each time step, switches to the next policy in the list with probability α.

Switching times are **geometrically** distributed.

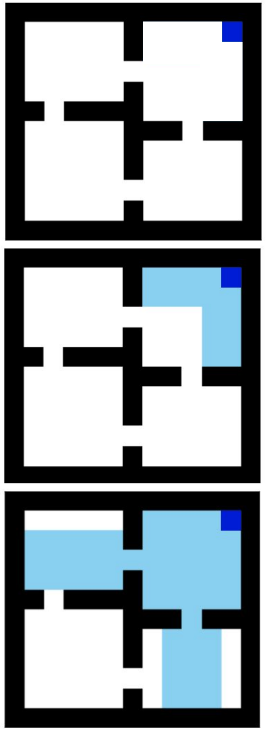
Geometric generalised policy improvement (GGPI)

$$\pi_1, \ldots, \pi_k \rightarrow \Pi \rightarrow \boxed{\arg\max_a \max_{\nu \in \Pi} Q^{\nu}(x, a)} \rightarrow \pi'$$

Base policies — Set of GSPs

Guarantee: $Q^{\pi'} \geq \max_{\nu \in \Pi} Q^{\nu}$

as long as closure condition on Π holds (see paper)



Known policies $\pi_U, \pi_L, \pi_R, \pi_D$

New goal location indicated.

Quickly derive improved policy for new task.

GPI produces optimal behaviour only at nearby states.

GGPI with depth-2 GSPs $\Pi = \{\pi_U \rightarrow \pi_R, \ldots\}$ obtains optimal behaviour in many more states.

Improving over non-Markov policies

A **geometric switching policy (GSP)**

$$\nu = \pi^{(1)} \overset{\alpha}{\to} \cdots \overset{\alpha}{\to} \pi^{(n)}$$

is a non-Markov behaviour that:

- Begins the episode using $\pi^{(1)}$.
- At each time step, switches to the next policy in the list with probability α.

Switching times are **geometrically** distributed.

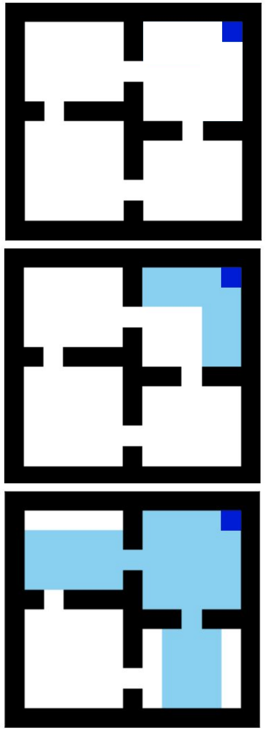
Geometric generalised policy improvement (GGPI)

$\pi_1, \ldots, \pi_k \to \Pi \to \boxed{\arg\max_a \max_{\nu \in \Pi} Q^{\nu}(x,a)} \to \pi'$

Base policies — Set of GSPs

Guarantee: $Q^{\pi'} \geq \max_{\nu \in \Pi} Q^{\nu}$

as long as closure condition on Π holds (see paper)



Known policies $\pi_U, \pi_L, \pi_R, \pi_D$

New goal location indicated.

Quickly derive improved policy for new task.

GPI produces optimal behaviour only at nearby states.

GGPI with depth-2 GSPs $\Pi = \{\pi_U \to \pi_R, \ldots\}$ obtains optimal behaviour in many more states.

In order to implement, need a way of estimating GSP values $Q^{\nu}(x,a)$ for new reward functions, without requiring further learning.

Improving over non-Markov geometric switching policies (GSPs)

Evaluating GSPs with geometric horizon models (GHMs)

Learning GHMs with cross-entropy TD (CETD)

Improving over non-Markov geometric switching policies (GSPs)

Evaluating GSPs with geometric horizon models (GHMs)

Learning GHMs with cross-entropy TD (CETD)

Policy evaluation with geometric horizon models

Policy evaluation with geometric horizon models

Geometric horizon models
(also γ-models (Janner et al., 2020), β-models (Sutton, 1995))

Policy evaluation with geometric horizon models

Geometric horizon models
(also γ-models (Janner et al., 2020), β-models (Sutton, 1995))

For policy π and discount β, a **geometric horizon model (GHM)** μ_β^π is a generative model for the corresponding discounted visitation distributions.

Policy evaluation with geometric horizon models

Geometric horizon models
(also γ-models (Janner et al., 2020), β-models (Sutton, 1995))

For policy π and discount β, a **geometric horizon model (GHM)** μ_β^π is a generative model for the corresponding discounted visitation distributions.

Environment rollout

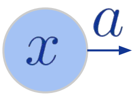
x

Policy evaluation with geometric horizon models

Geometric horizon models
(also γ-models (Janner et al., 2020), β-models (Sutton, 1995))

For policy π and discount β, a **geometric horizon model (GHM)** μ_β^π is a generative model for the corresponding discounted visitation distributions.

Environment rollout



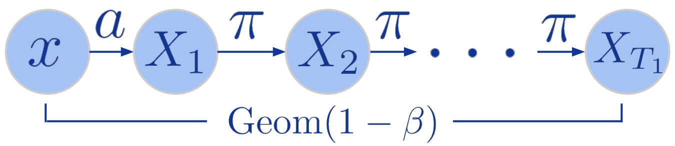
Policy evaluation with geometric horizon models

Geometric horizon models

(also γ-models (Janner et al., 2020), β-models (Sutton, 1995))

For policy π and discount β, a **geometric horizon model (GHM)** μ_β^π is a generative model for the corresponding discounted visitation distributions.

Environment rollout

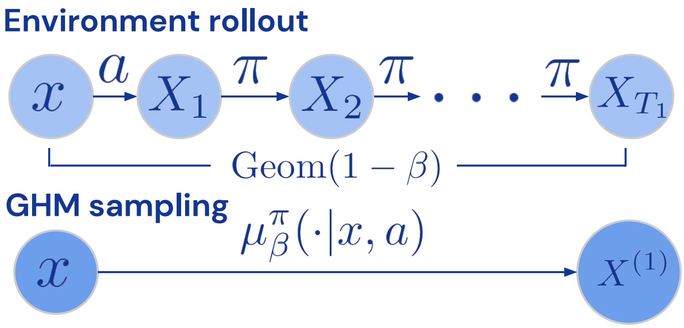


Policy evaluation with geometric horizon models

Geometric horizon models

(also γ-models (Janner et al., 2020), β-models (Sutton, 1995))

For policy π and discount β, a **geometric horizon model (GHM)** μ_β^π is a generative model for the corresponding discounted visitation distributions.



Policy evaluation with geometric horizon models

Policy evaluation with geometric horizon models

Compose these models to evaluate GSPs (extending Markov results from Janner et al. (2020))

Policy evaluation with geometric horizon models

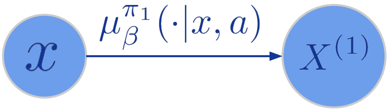
Compose these models to evaluate GSPs (extending Markov results from Janner et al. (2020))

Aim: Evaluate $\nu = \pi_1 \overset{\alpha}{\rightarrow} \cdots \overset{\alpha}{\rightarrow} \pi_n$

Policy evaluation with geometric horizon models

Compose these models to evaluate GSPs (extending Markov results from Janner et al. (2020))

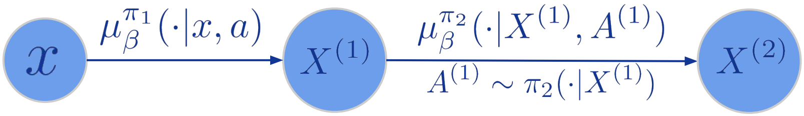
Aim: Evaluate $\nu = \pi_1 \stackrel{\alpha}{\to} \cdots \stackrel{\alpha}{\to} \pi_n$



Policy evaluation with geometric horizon models

Compose these models to evaluate GSPs (extending Markov results from Janner et al. (2020))

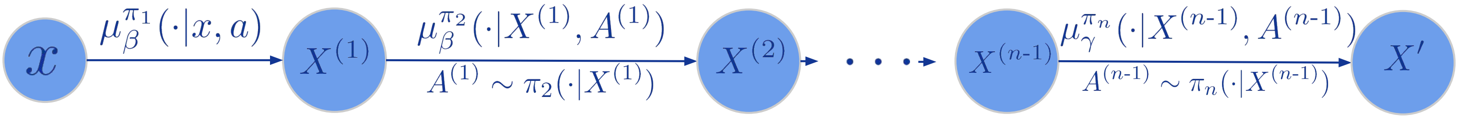
Aim: Evaluate $\nu = \pi_1 \overset{\alpha}{\to} \cdots \overset{\alpha}{\to} \pi_n$



Policy evaluation with geometric horizon models

Compose these models to evaluate GSPs (extending Markov results from Janner et al. (2020))

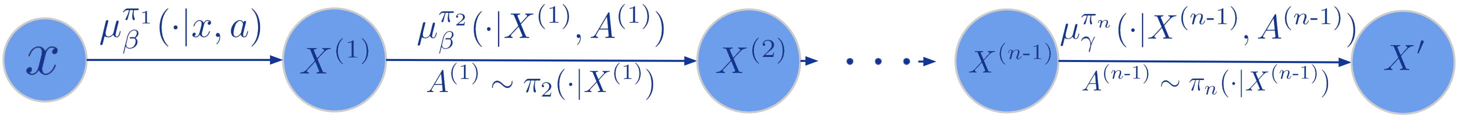
Aim: Evaluate $\nu = \pi_1 \overset{\alpha}{\rightarrow} \cdots \overset{\alpha}{\rightarrow} \pi_n$



Policy evaluation with geometric horizon models

Compose these models to evaluate GSPs (extending Markov results from Janner et al. (2020))

Aim: Evaluate $\nu = \pi_1 \xrightarrow{\alpha} \cdots \xrightarrow{\alpha} \pi_n$

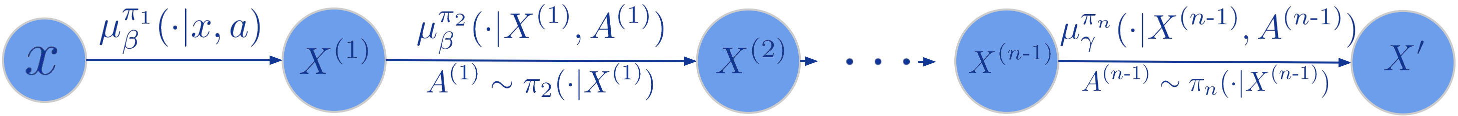


$$r(x) + \frac{\gamma}{1-\gamma}\left[\sum_{m=1}^{n-1} \frac{1-\gamma}{1-\beta}\left(\frac{\gamma-\beta}{1-\beta}\right)^{m-1} r(X^{(m)}) + \left(\frac{\gamma-\beta}{1-\beta}\right)^{n-1} r(X')\right]$$

Policy evaluation with geometric horizon models

Compose these models to evaluate GSPs (extending Markov results from Janner et al. (2020))

Aim: Evaluate $\nu = \pi_1 \overset{\alpha}{\to} \cdots \overset{\alpha}{\to} \pi_n$



$$r(x) + \frac{\gamma}{1-\gamma}\left[\sum_{m=1}^{n-1} \frac{1-\gamma}{1-\beta}\left(\frac{\gamma-\beta}{1-\beta}\right)^{m-1} r(X^{(m)}) + \left(\frac{\gamma-\beta}{1-\beta}\right)^{n-1} r(X')\right]$$

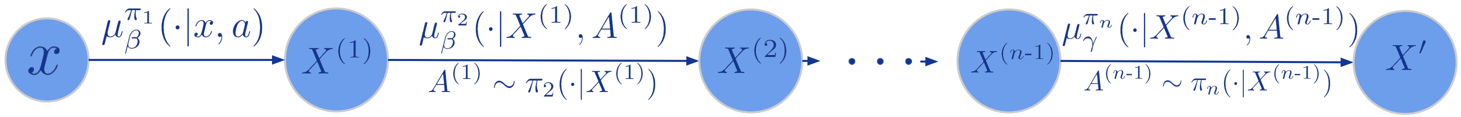
Proposition

This is an unbiased estimate of the value $Q_\gamma^\nu(x, a)$ of the GSP $\nu = \pi_1 \overset{\alpha}{\to} \cdots \overset{\alpha}{\to} \pi_n$, where $\alpha = \frac{\gamma - \beta}{\gamma}$.

Policy evaluation with geometric horizon models

Compose these models to evaluate GSPs (extending Markov results from Janner et al. (2020))

Aim: Evaluate $\nu = \pi_1 \overset{\alpha}{\to} \cdots \overset{\alpha}{\to} \pi_n$



$$r(x) + \frac{\gamma}{1-\gamma}\left[\sum_{m=1}^{n-1} \frac{1-\gamma}{1-\beta}\left(\frac{\gamma-\beta}{1-\beta}\right)^{m-1} r(X^{(m)}) + \left(\frac{\gamma-\beta}{1-\beta}\right)^{n-1} r(X')\right]$$

Proposition

This is an unbiased estimate of the value $Q_\gamma^\nu(x, a)$ of the GSP $\nu = \pi_1 \overset{\alpha}{\to} \cdots \overset{\alpha}{\to} \pi_n$, where $\alpha = \dfrac{\gamma - \beta}{\gamma}$.

Takeaway: Composing GHMs allows us to evaluate GSPs without further learning.

Improving over non-Markov geometric switching policies (GSPs)

Evaluating GSPs with geometric horizon models (GHMs)

Learning GHMs with cross-entropy TD (CETD)

Improving over non-Markov geometric switching policies (GSPs)

Evaluating GSPs with geometric horizon models (GHMs)

Learning GHMs with cross-entropy TD (CETD)

Learning geometric horizon models

Learning geometric horizon models

Cross-entropy temporal-difference learning (CETD)

Learning geometric horizon models

Cross-entropy temporal-difference learning (CETD)

MLE with a bootstrapped target distribution.

Learning geometric horizon models

Cross-entropy temporal-difference learning (CETD)

MLE with a bootstrapped target distribution.

Aim: Learn μ_β^π with a parameterised approximator μ_θ .

Learning geometric horizon models

Cross-entropy temporal-difference learning (CETD)
MLE with a bootstrapped target distribution.

Aim: Learn μ_β^π with a parameterised approximator μ_θ .

Algorithm:

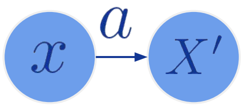
Learning geometric horizon models

<u>Cross-entropy temporal-difference learning</u> (CETD)
MLE with a bootstrapped target distribution.

Aim: Learn μ_β^π with a parameterised approximator μ_θ .

Algorithm:

Observe transition (x, a, X')



Learning geometric horizon models

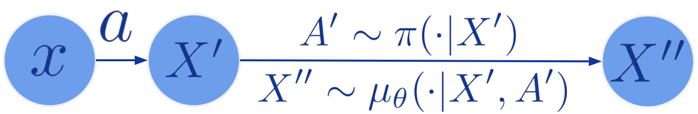
Cross-entropy temporal-difference learning (CETD)
MLE with a bootstrapped target distribution.

Aim: Learn μ_β^π with a parameterised approximator μ_θ .

Algorithm:

Observe transition (x, a, X')

Sample action $A' \sim \pi(\cdot|X')$
and bootstrap state $X'' \sim \mu_\theta(\cdot|X', A')$



Learning geometric horizon models

Cross-entropy temporal-difference learning (CETD)

MLE with a bootstrapped target distribution.

Aim: Learn μ_β^π with a parameterised approximator μ_θ .

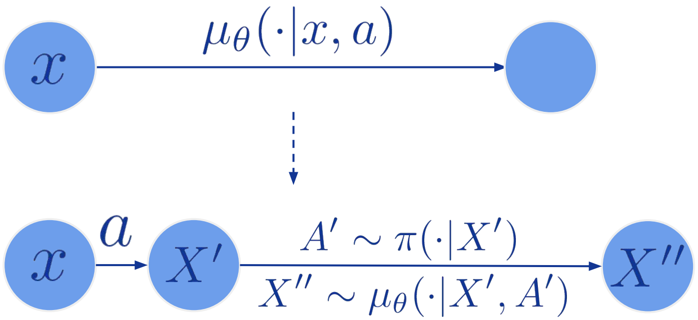
Algorithm:

Observe transition (x, a, X')

Sample action $A' \sim \pi(\cdot|X')$
and bootstrap state $X'' \sim \mu_\theta(\cdot|X', A')$

Gradient descent on

$$-(1-\beta)\log \mu_\theta(X'|x,a) - \beta \log \mu_\theta(X''|x,a)$$



Learning geometric horizon models

Cross-entropy temporal-difference learning (CETD)
MLE with a bootstrapped target distribution.

Aim: Learn μ_β^π with a parameterised approximator μ_θ .

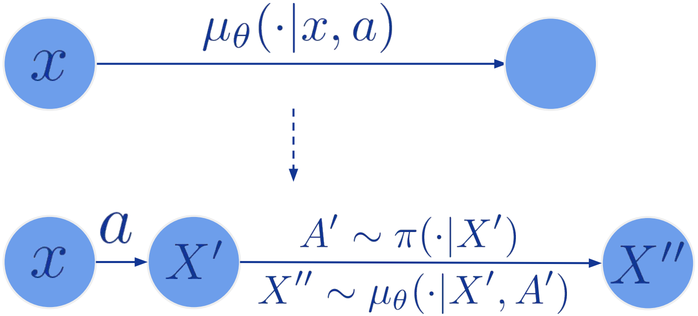
Algorithm:

Observe transition (x, a, X')

Sample action $A' \sim \pi(\cdot|X')$
and bootstrap state $X'' \sim \mu_\theta(\cdot|X', A')$

Gradient descent on

$$-(1-\beta)\log\mu_\theta(X'|x,a) - \beta\log\mu_\theta(X''|x,a)$$



Theorem: Almost-sure convergence to μ_β^π in tabular setting (under appropriate conditions).

Improving over non-Markov geometric switching policies (GSPs)

Evaluating GSPs with geometric horizon models (GHMs)

Learning GHMs with cross-entropy TD (CETD)

Implementation with deep reinforcement learning

Implementation with deep reinforcement learning

MuJoCo (Todorov, 2012) Ant, with pre-trained policies to move up/right/down/left.

(Todorov, 2012)

Implementation with deep reinforcement learning

MuJoCo (Todorov, 2012) Ant, with pre-trained policies to move up/right/down/left.

Test time: Each episode, new target location revealed via reward function.
Goal: Reach target without any additional learning.

(Todorov, 2012)

Implementation with deep reinforcement learning

MuJoCo (Todorov, 2012) Ant, with pre-trained policies to move up/right/down/left.

Test time: Each episode, new target location revealed via reward function.
Goal: Reach target without any additional learning.

GHMs implemented as conditional VAEs (Sohn et al., 2015; Kingma & Welling, 2014; Rezende et al., 2014) and trained on the ELBO of the CETD objective.

(Todorov, 2012)

Implementation with deep reinforcement learning

MuJoCo (Todorov, 2012) Ant, with pre-trained policies to move up/right/down/left.

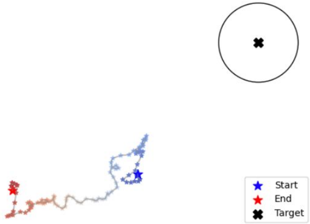

(Todorov, 2012)

Test time: Each episode, new target location revealed via reward function.
Goal: Reach target without any additional learning.

GHMs implemented as conditional VAEs (Sohn et al., 2015; Kingma & Welling, 2014; Rezende et al., 2014) and trained on the ELBO of the CETD objective.

GPI (baseline)

Representative episodes



Implementation with deep reinforcement learning

(Todorov, 2012)

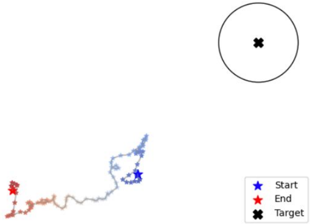
MuJoCo (Todorov, 2012) Ant, with pre-trained policies to move up/right/down/left.

Test time: Each episode, new target location revealed via reward function.
Goal: Reach target without any additional learning.

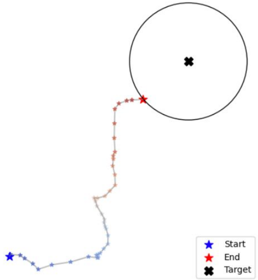
GHMs implemented as conditional VAEs (Sohn et al., 2015; Kingma & Welling, 2014; Rezende et al., 2014) and trained on the ELBO of the CETD objective.

Representative episodes

GPI (baseline)



Depth-2 GGPI



Implementation with deep reinforcement learning

MuJoCo (Todorov, 2012) Ant, with pre-trained policies to move up/right/down/left.

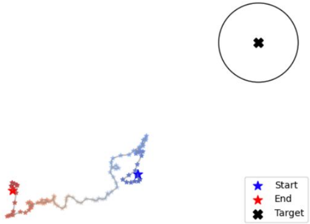
Test time: Each episode, new target location revealed via reward function.
Goal: Reach target without any additional learning.

GHMs implemented as conditional VAEs (Sohn et al., 2015; Kingma & Welling, 2014; Rezende et al., 2014) and trained on the ELBO of the CETD objective.

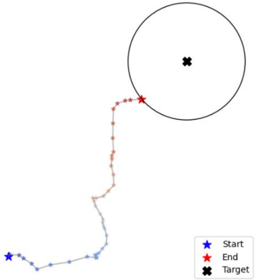
(Todorov, 2012)

Representative episodes

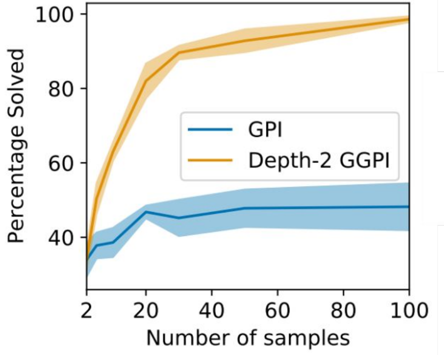
GPI (baseline)



Depth-2 GGPI



Overall results



Related work

Generative/density modelling for discounted visitation distributions

- *Gamma-models* (Janner et al., 2020)
- *Successor states* (Blier et al., 2021, Touati & Ollivier, 2021)
- *Contrastive density modelling* (Eysenbach et al., 2020)

Successor representation, successor features, and generalised policy improvement for transfer

- *Successor representation* (Dayan, 1993; Kulkarni et al., 2016)
- *Beta-models, multi-time models* (Sutton, 1995; Precup et al., 1998)
- *Successor features and generalised policy improvement* (Barreto et al.; 2017, 2020)

Option modelling

- *Compositional option models* (Silver & Ciosek, 2012)
- *Universal option models* (Yao et al., 2014)

Policy improvement

- *Multi-step improvement* (Efroni et al., 2018; 2019; 2020)

And many more: see paper.

Further work and limitations

Further work and limitations

Geometric switching times are key to the theory in this work.

Further work and limitations

Geometric switching times are key to the theory in this work.

Extensions of GHMs that do not need to model full agent state.

Further work and limitations

Geometric switching times are key to the theory in this work.

Extensions of GHMs that do not need to model full agent state.

Potentially exponential number of GSPs to consider.

Conclusion

Conclusion

Framework for stronger policy improvement:

- Compose **geometric horizon models** to evaluate non-Markov **geometric switching policies**.
- Use **geometric generalised policy improvement** to improve over collections of GSPs.

Conclusion

Framework for stronger policy improvement:

- Compose **geometric horizon models** to evaluate non-Markov **geometric switching policies**.
- Use **geometric generalised policy improvement** to improve over collections of GSPs.

Theory for policy evaluation, improvement guarantee, and GHM convergence.

Conclusion

Framework for stronger policy improvement:

- Compose **geometric horizon models** to evaluate non-Markov **geometric switching policies**.
- Use **geometric generalised policy improvement** to improve over collections of GSPs.

Theory for policy evaluation, improvement guarantee, and GHM convergence.

Applications to policy iteration and transfer learning.

Conclusion

Framework for stronger policy improvement:

- Compose **geometric horizon models** to evaluate non-Markov **geometric switching policies**.
- Use **geometric generalised policy improvement** to improve over collections of GSPs.

Theory for policy evaluation, improvement guarantee, and GHM convergence.

Applications to policy iteration and transfer learning.

Thank you!
Poster: Hall E #932