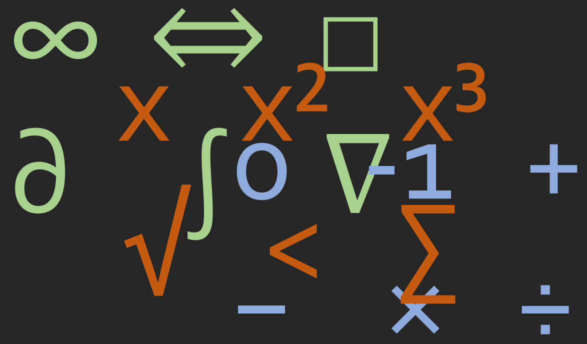
Seminar in Deep Reinforcement Learning

The Path to Continual Learning
Curriculum Learning

Ramon Witschi, ETH Computer Science MSc, 19.05.2020

What is a Curriculum?



$0 \quad -1 \quad +$

$- \quad \times \quad \div$

$\infty \quad \Longleftrightarrow \quad \square$

$\partial \quad x \quad \int \quad x^2 \quad \nabla \quad x^3$

$\sqrt{} \quad < \quad \sum$

$0 \quad -1 \quad +$

$- \quad \times \quad \div$

$x \quad x^2 \quad x^3$

$\surd \quad < \quad \sum$

$\infty \quad \Leftrightarrow \quad \square$

$\partial \quad \int \quad \nabla$

$0 \quad -1 \quad +$

$- \quad \times \quad \div$

$\rightarrow$

$x \quad x^2 \quad x^3$

$\surd \quad < \quad \sum$

$\infty \quad \Leftrightarrow \quad \square$

$\partial \quad \int \quad \nabla$

$0 \quad -1 \quad +$

$- \quad \times \quad \div$

→

$x \quad x^2 \quad x^3$

$\sqrt{} \quad < \quad \sum$

→

$\infty \quad \Leftrightarrow \quad \square$

$\partial \quad \int \quad \nabla$

Curriculum over Training Data!

[Curriculum Learning (ICML, 2009) – Bengio et al.]

1 Start with simple examples

2 Gradually add more difficult ones

3 Arrive at target training distribution

Curriculum Learning (ICML, 2009) – Bengio et al.

1 Start with simple examples

2 Gradually add more difficult ones

3 Arrive at target training distribution

[Curriculum Learning (ICML, 2009) – Bengio et al.]

1 Start with simple examples

2 Gradually add more difficult ones

3 Arrive at target training distribution

[Curriculum Learning (ICML, 2009) – Bengio et al.](#)

Empirical Results
Faster Training & sometimes higher Test Scores

Curriculum Learning (ICML, 2009) – Bengio et al.

Faster Training proven on Linear Regression (Convex Optimization) ☺

Curriculum Learning by Transfer Learning: Theory and Experiments with Deep Networks – Weinshall, Cohen & Amir

Curriculum Learning meets Reinforcement Learning

Intrinsic Motivation and Automatic Curricula via Asymmetric Self-Play

Sukhbaatar et al.

Reverse Curriculum Generation for Reinforcement Learning

Florensa et al.

Mix & Match – Agent Curricula for Reinforcement Learning

Czarnecki et al.

Curriculum Learning meets Reinforcement Learning

Intrinsic Motivation and Automatic Curricula via Asymmetric Self-Play

Sukhbaatar et al.

Reverse Curriculum Generation for Reinforcement Learning

Florensa et al.

Mix & Match – Agent Curricula for Reinforcement Learning

Czarnecki et al.

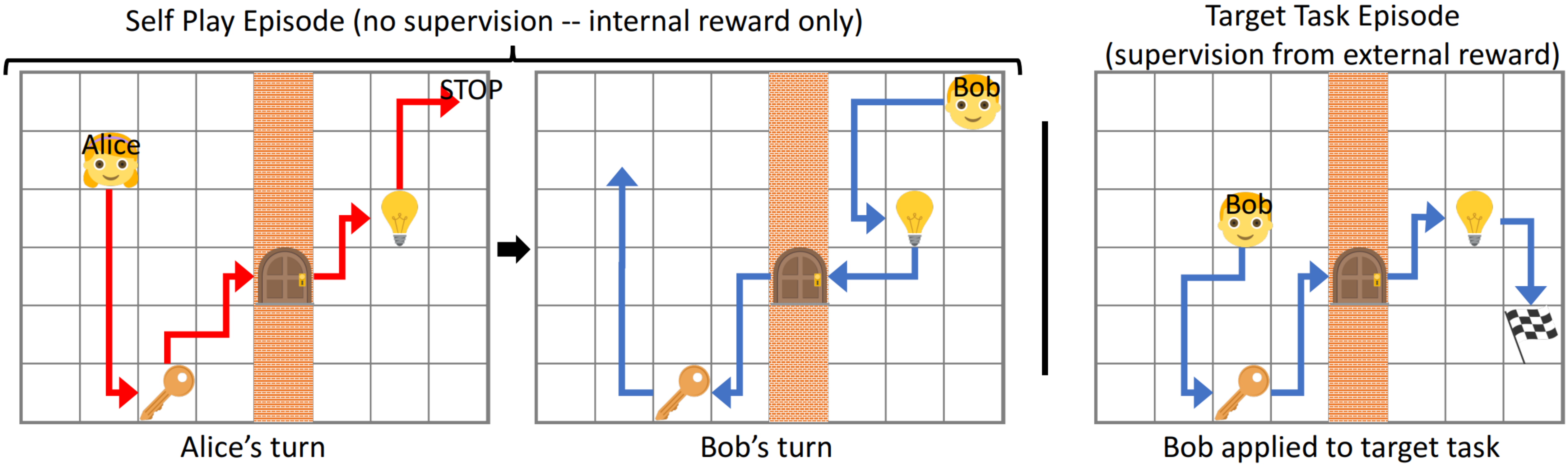
Model-Free Reinforcement Learning

Sample Inefficient ☹

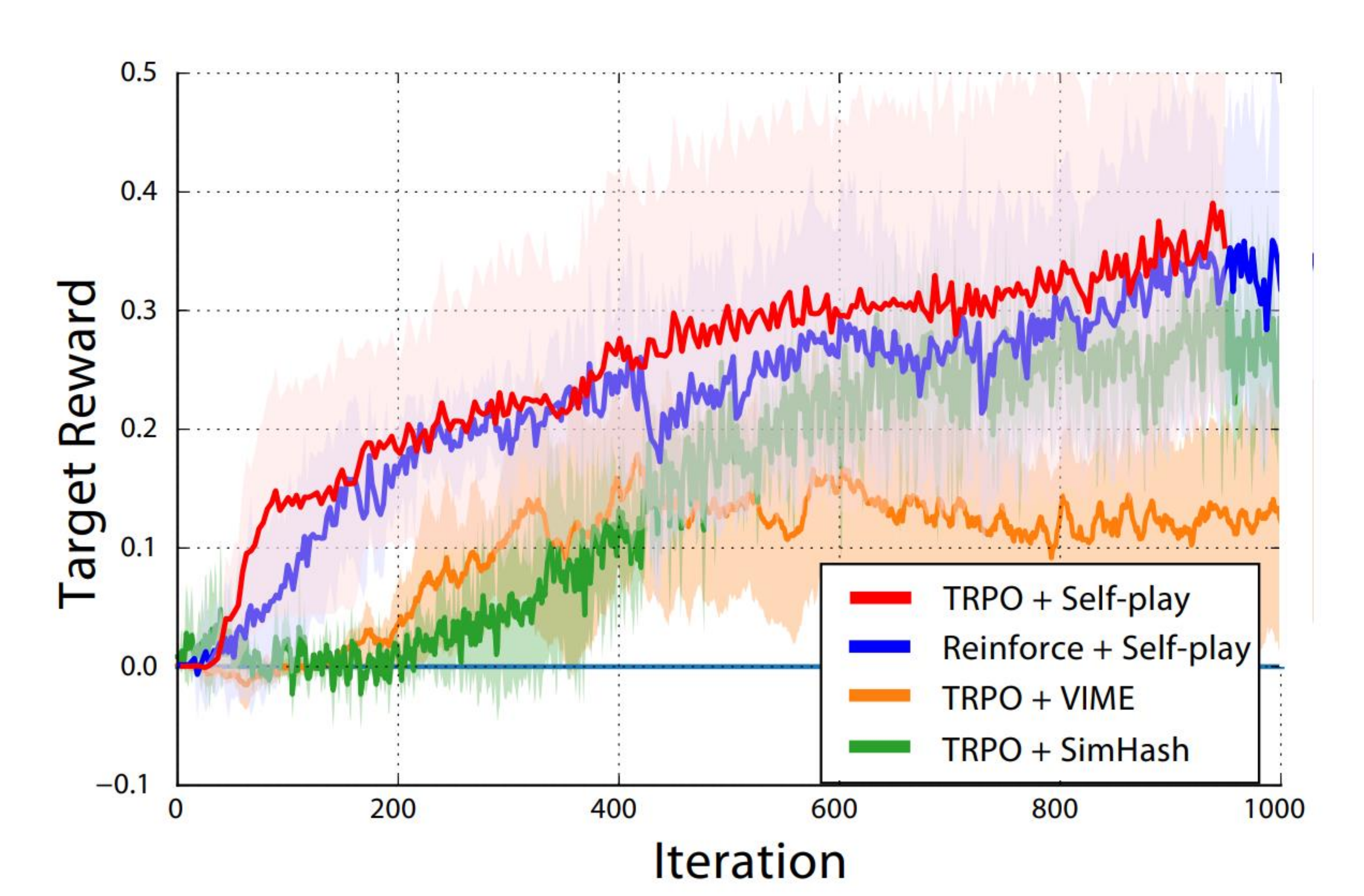
Jointly learn Environment and optimize for Reward

“Unsupervised” Exploration!

Framework



Self Play Episode (no supervision -- internal reward only)
Target Task Episode
(supervision from external reward)
STOP
Alice
Bob
Bob
Alice's turn
Bob's turn
Bob applied to target task



Target Reward
Iteration
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
0
200
400
600
800
1000
TRPO + Self-play
Reinforce + Self-play
TRPO + VIME
TRPO + SimHash

Internal Reward Structure

$$R_A = \max(0, t_B - t_A) \xRightarrow{\substack{Bob\ fast \\ or \\ too\ slow}} 0$$

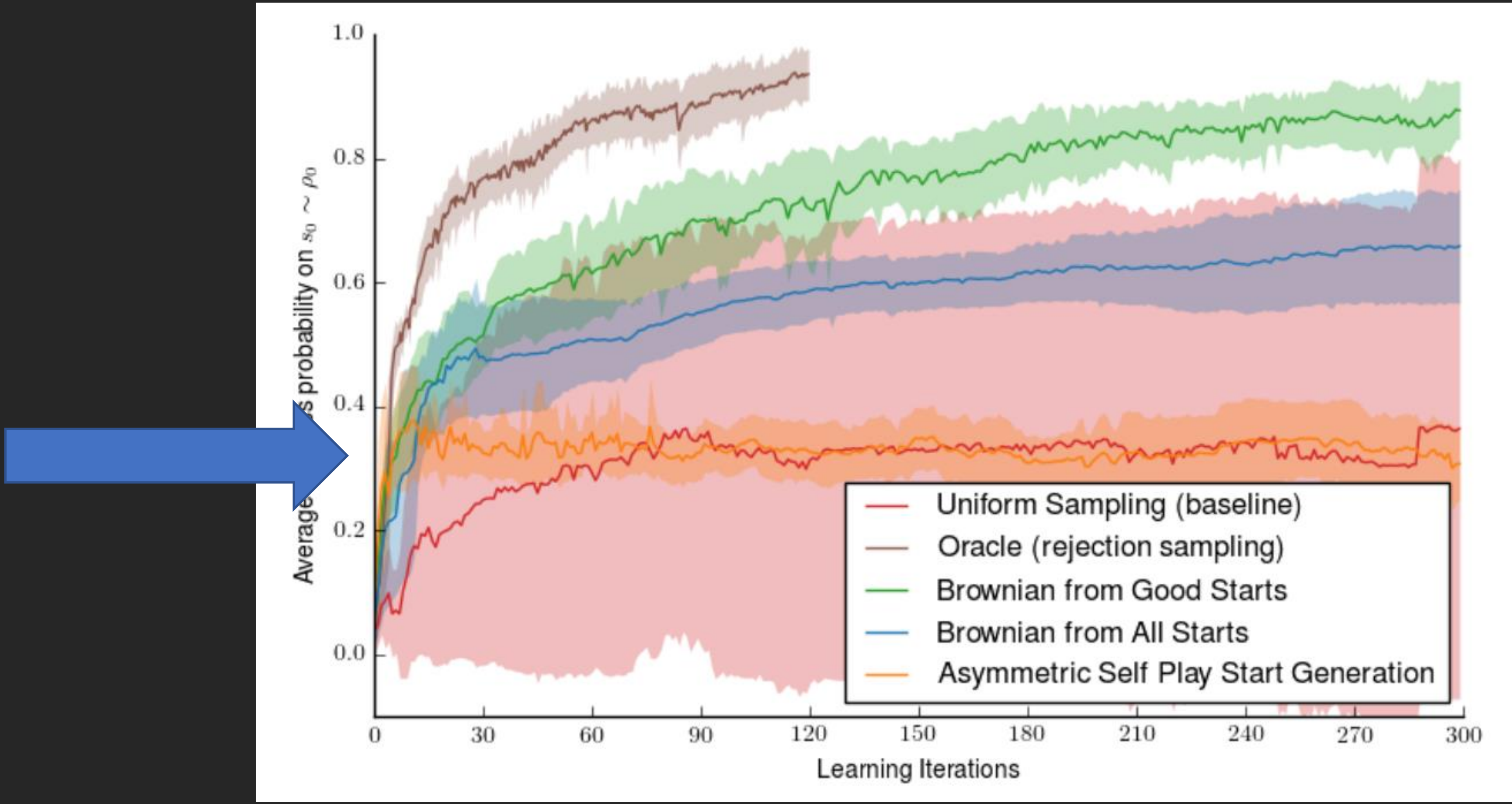
☹

Automatically creates a Curriculum over Exploration Tasks!

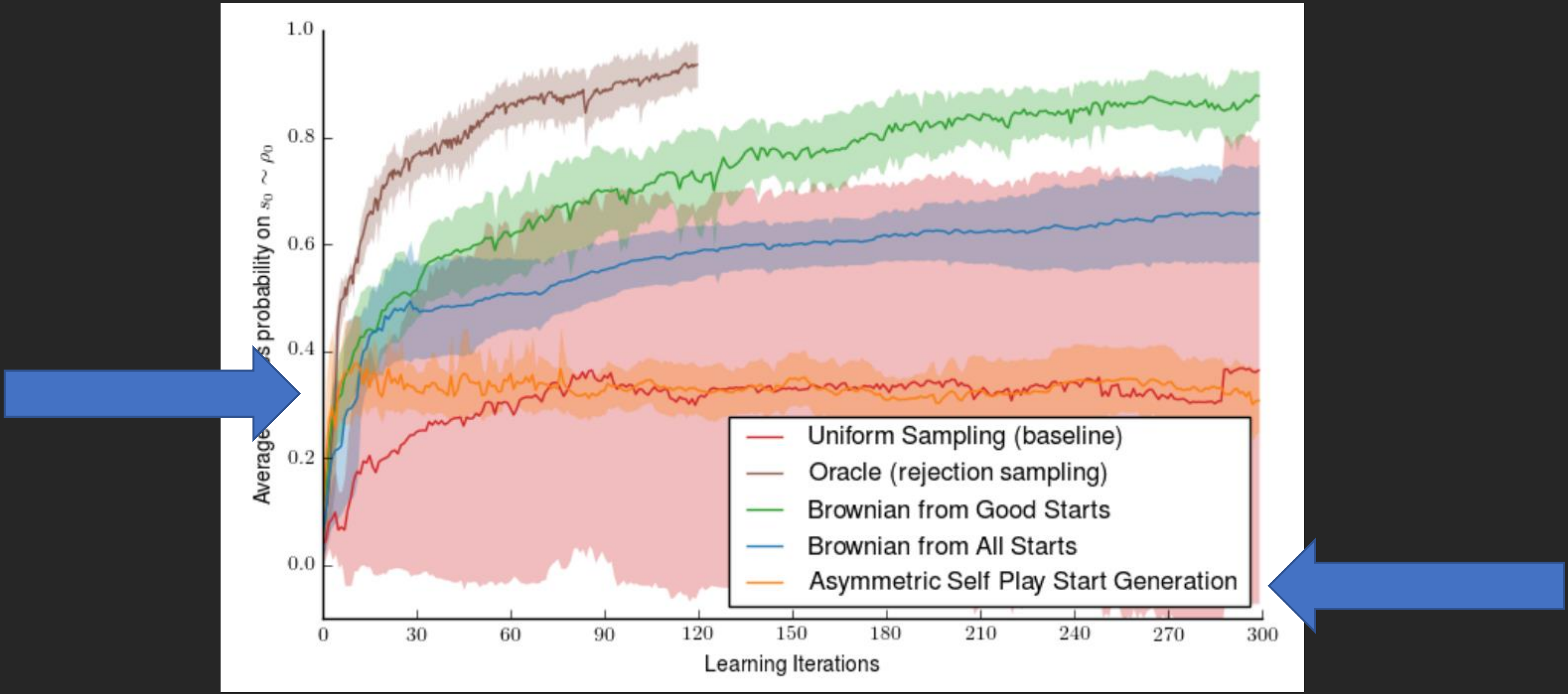
Internal Reward Structure

$$R_A = \max(0, t_B - t_A) \xRightarrow[\text{too slow}]{\text{job fast or}} 0$$

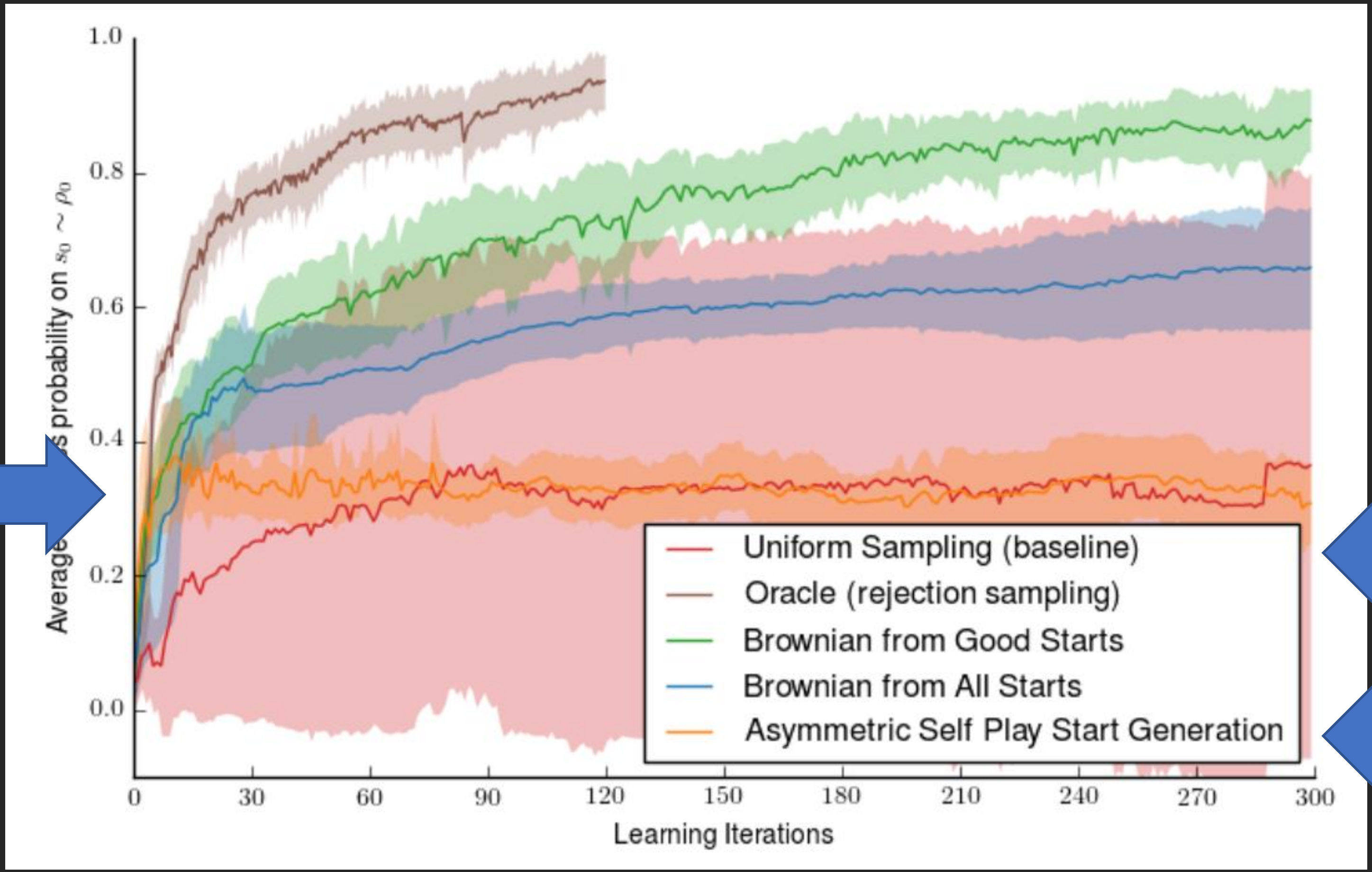
☹



Average success probability on $s_0 \sim \rho_0$
Learning Iterations
Uniform Sampling (baseline)
Oracle (rejection sampling)
Brownian from Good Starts
Brownian from All Starts
Asymmetric Self Play Start Generation



Average success probability on $s_0 \sim \rho_0$
Learning Iterations
Uniform Sampling (baseline)
Oracle (rejection sampling)
Brownian from Good Starts
Brownian from All Starts
Asymmetric Self Play Start Generation



Average success probability on $s_0 \sim \rho_0$
Learning Iterations
Uniform Sampling (baseline)
Oracle (rejection sampling)
Brownian from Good Starts
Brownian from All Starts
Asymmetric Self Play Start Generation

Curriculum Learning meets Reinforcement Learning

Intrinsic Motivation and Automatic Curricula via Asymmetric Self-Play

Sukhbaatar et al.

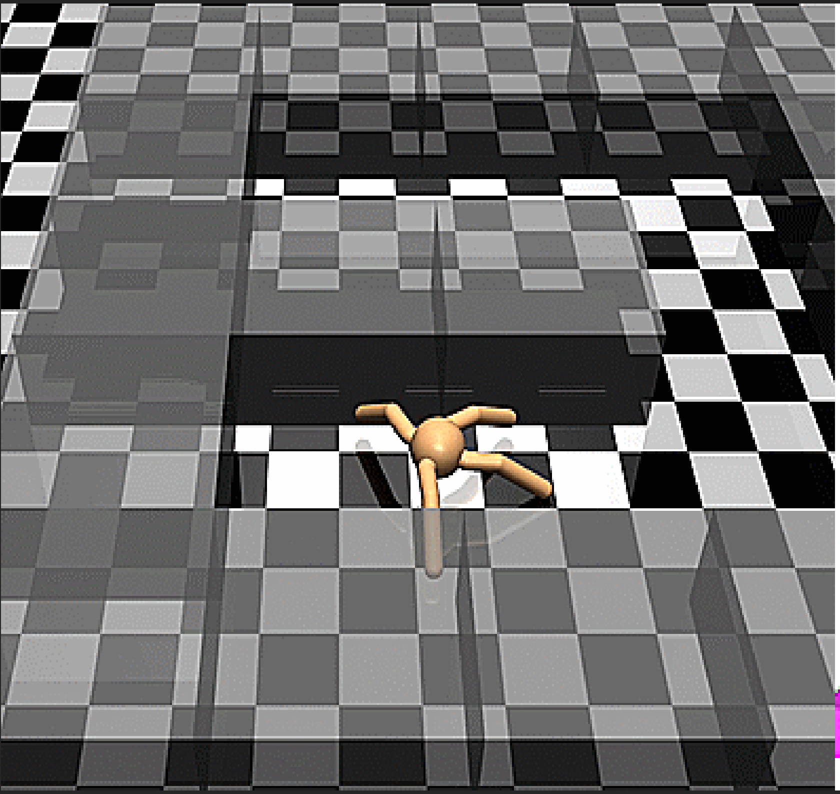
Reverse Curriculum Generation for Reinforcement Learning

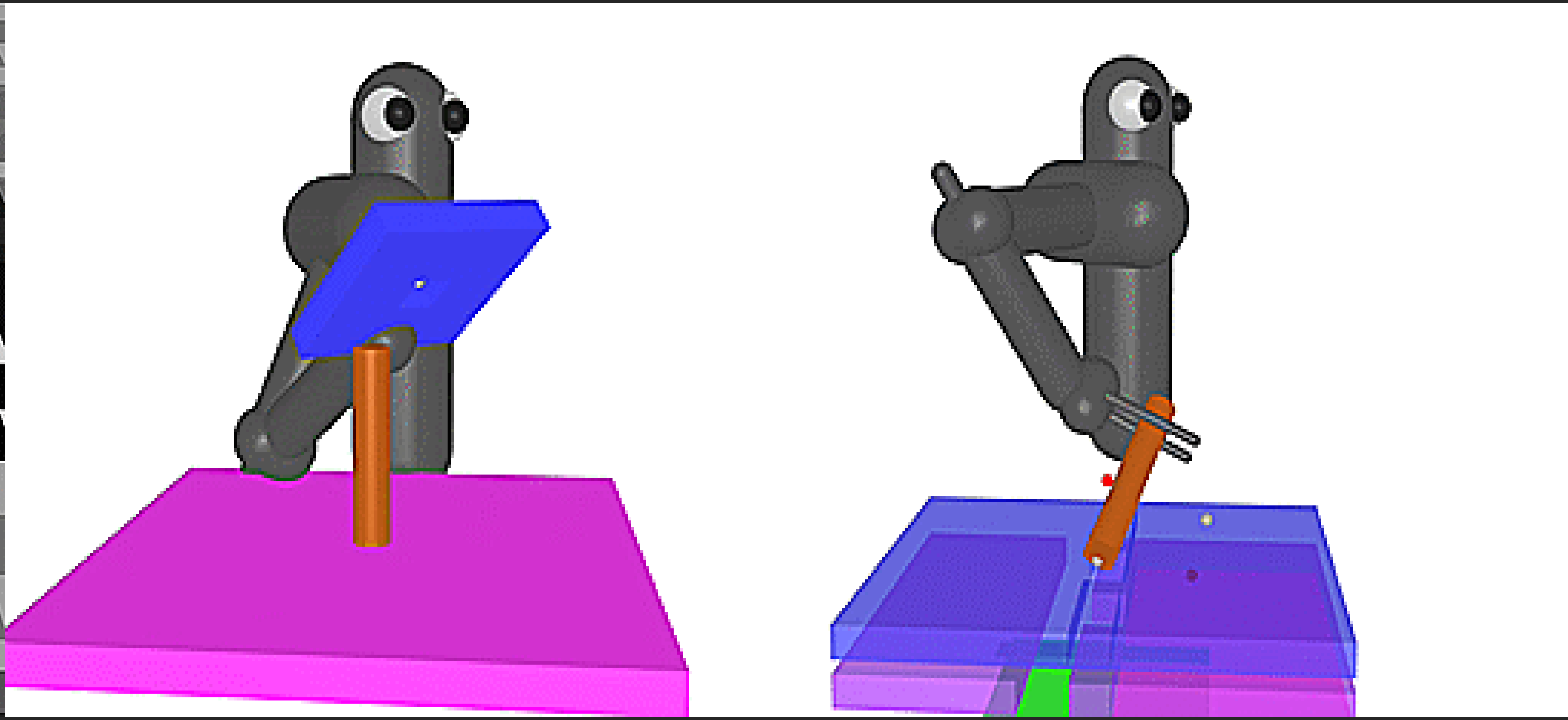
Florensa et al.

Mix & Match – Agent Curricula for Reinforcement Learning

Czarnecki et al.

Goal-Oriented Target Tasks





Goal-Oriented Target Tasks

Binary Reward Signal ☹

Goal-Oriented Target Tasks

Binary Reward Signal ☹

\+ Model-Free Reinforcement Learning ☹

Goal-Oriented Target Tasks

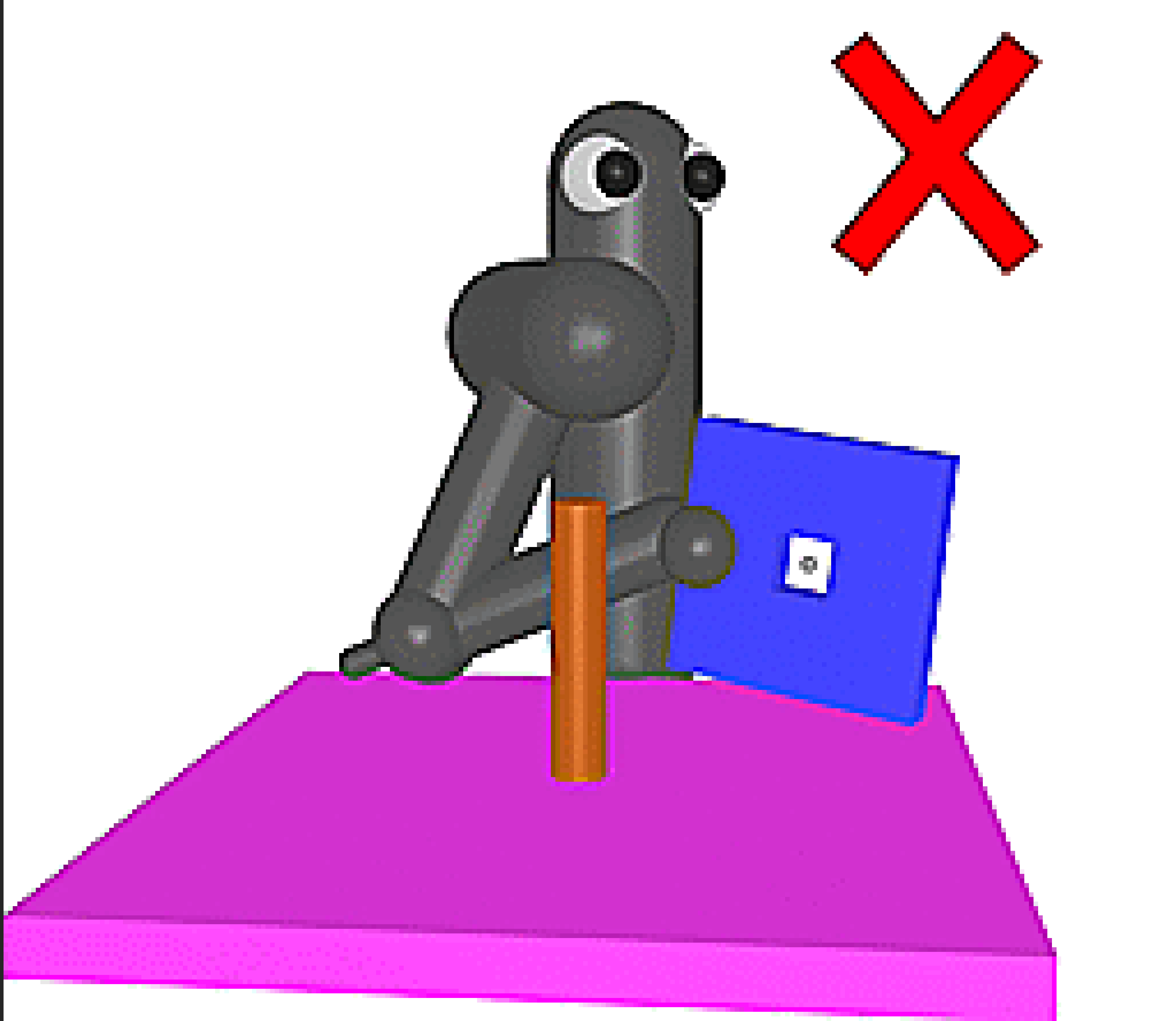
Binary Reward Signal ☹

\+ Model-Free Reinforcement Learning ☹

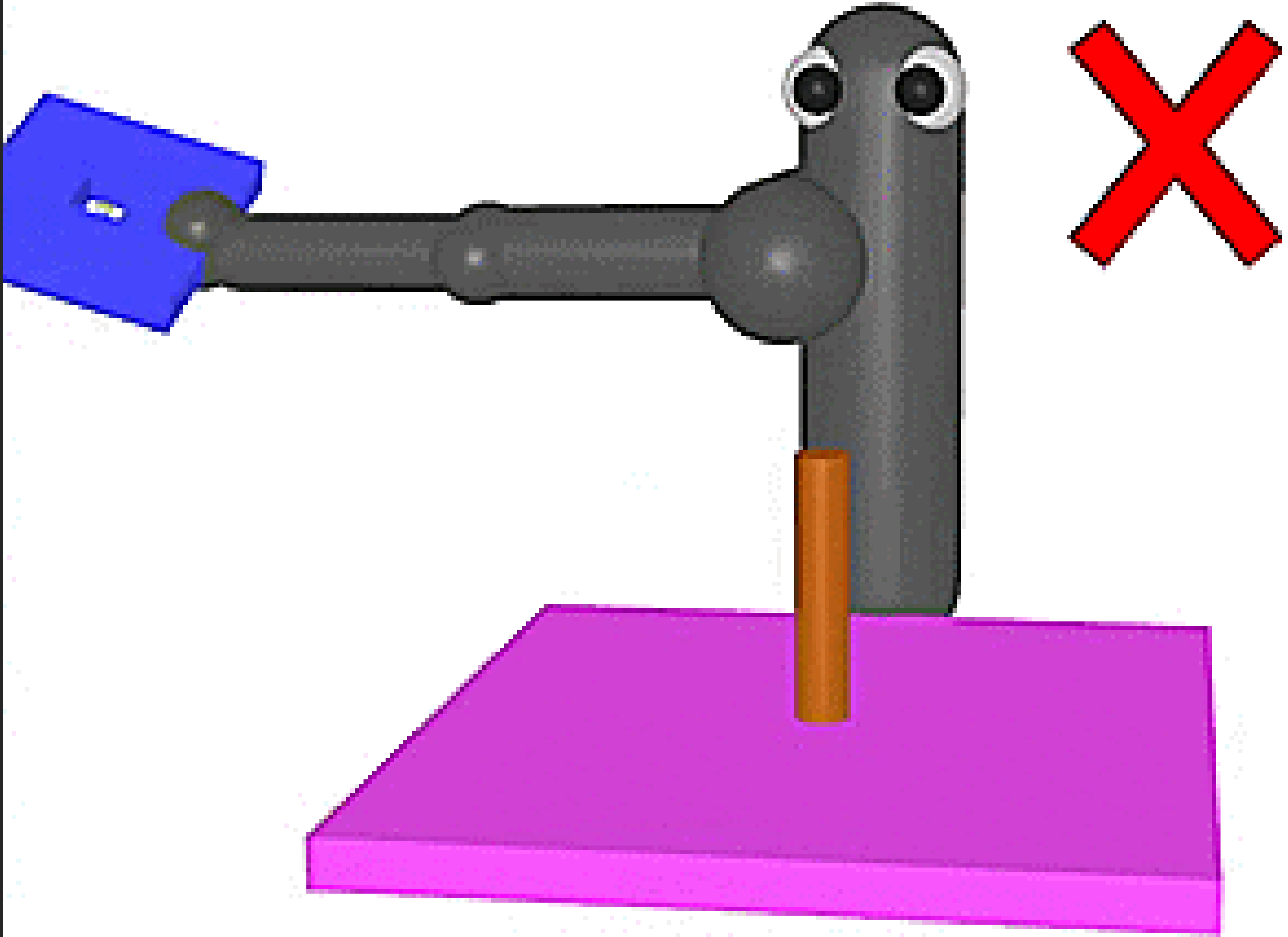
= (╯°□°）╯︵ ┻━┻

How do We Train the Agent?

Random Sampling of Starting States?



Add Regularization Term?



What’s the Trick?

Easy to Win, if you Start at the Goal!

Reverse Curriculum

1. Start almost there
2. Start increasingly further away
3. Profit from work already done

Reverse Curriculum

1. Start almost there
2. Start increasingly further away
3. Profit from work already done

Reverse Curriculum

1. Start almost there
2. Start increasingly further away
3. Profit from work already done

Automatically creates a Curriculum over Start States!

States of Intermediate Difficulty (SoIDs)

1. States Close to s^g may be good Start States
2. Random Walk in State-Space ☹
3. Brownian Motion in Action-Space ☺

States of Intermediate Difficulty (SoIDs)

1. States Close to s^g may be good Start States
2. Random Walk in State-Space
3. Brownian Motion in Action-Space

States of Intermediate Difficulty (SoIDs)

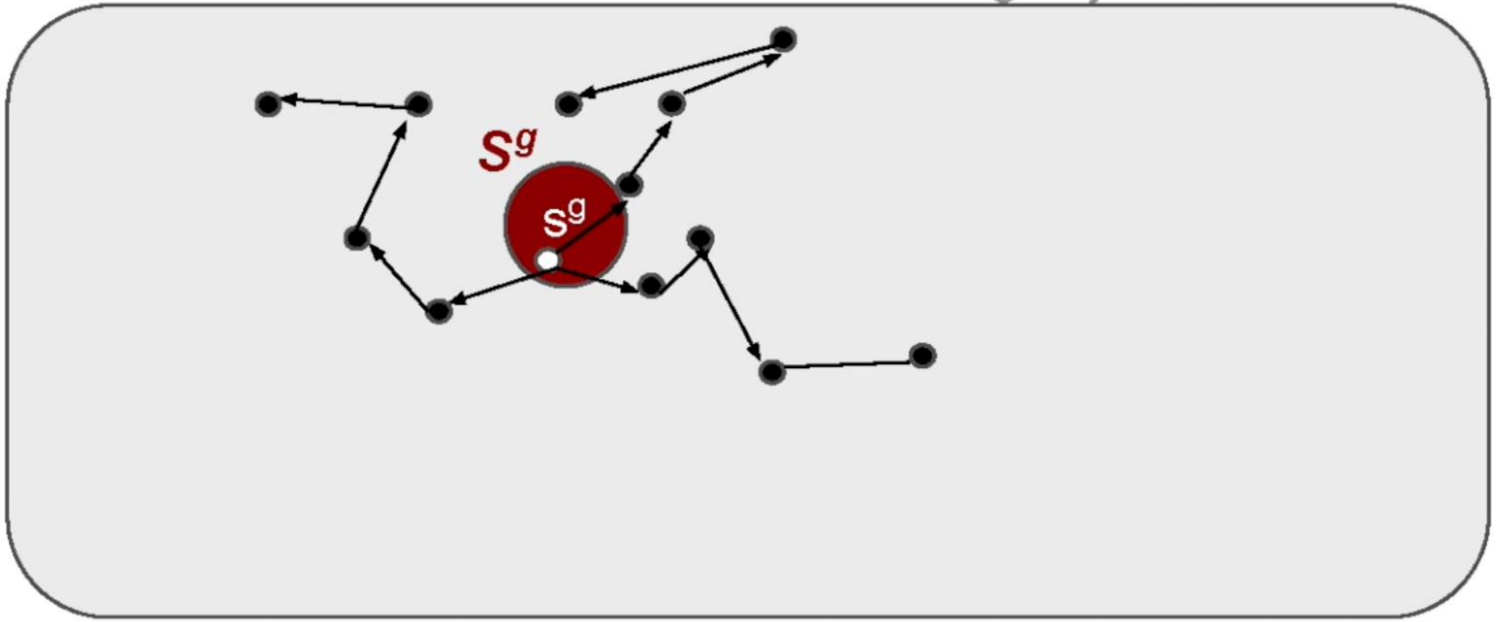
1. States Close to s^g may be good Start States
2. Random Walk in State-Space
3. Brownian Motion in Action-Space

S^g: goal states we want to reach from everywhere.

s^g: one goal state is provided

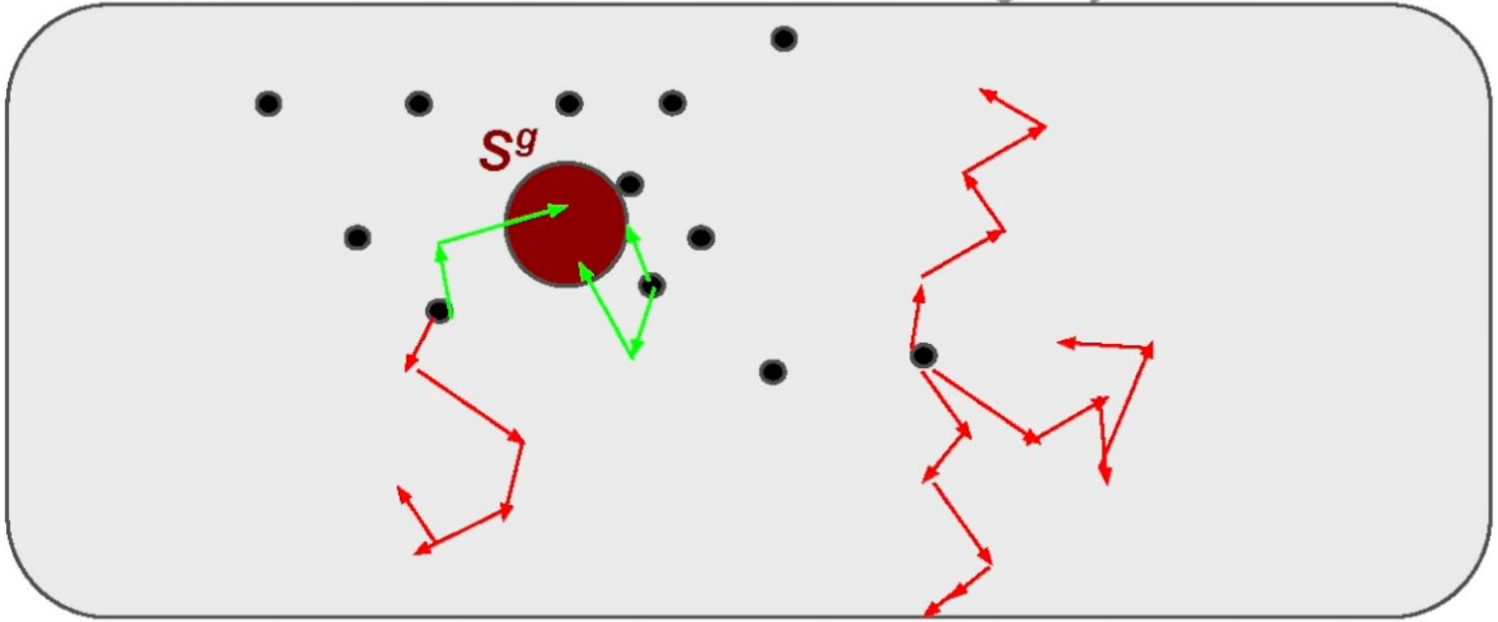
Iteration 1:

- Run Brownian motion
- Obtain trajectories from collected starts to train policy
- Label and filter starts based on training trajectories



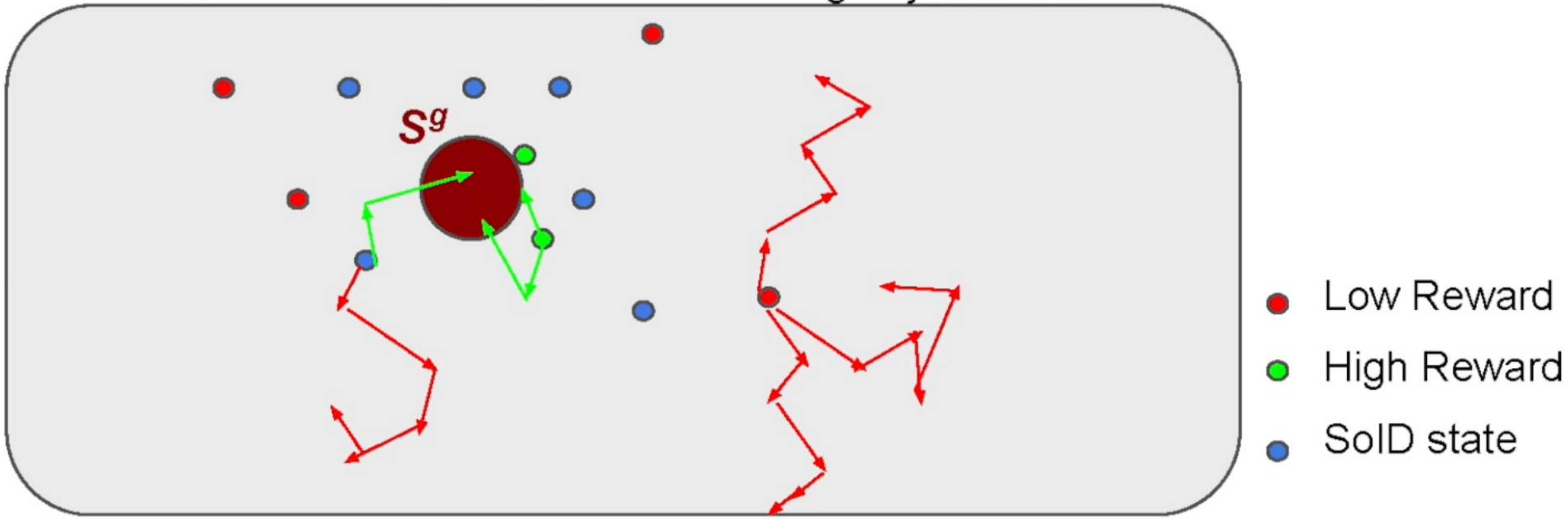
Iteration 1:

- Run Brownian motion
- Obtain trajectories from collected starts to train policy
- Label and filter starts based on training trajectories



Iteration 1:

- Run Brownian motion
- Obtain trajectories from collected starts to train policy
- Label and filter starts based on training trajectories

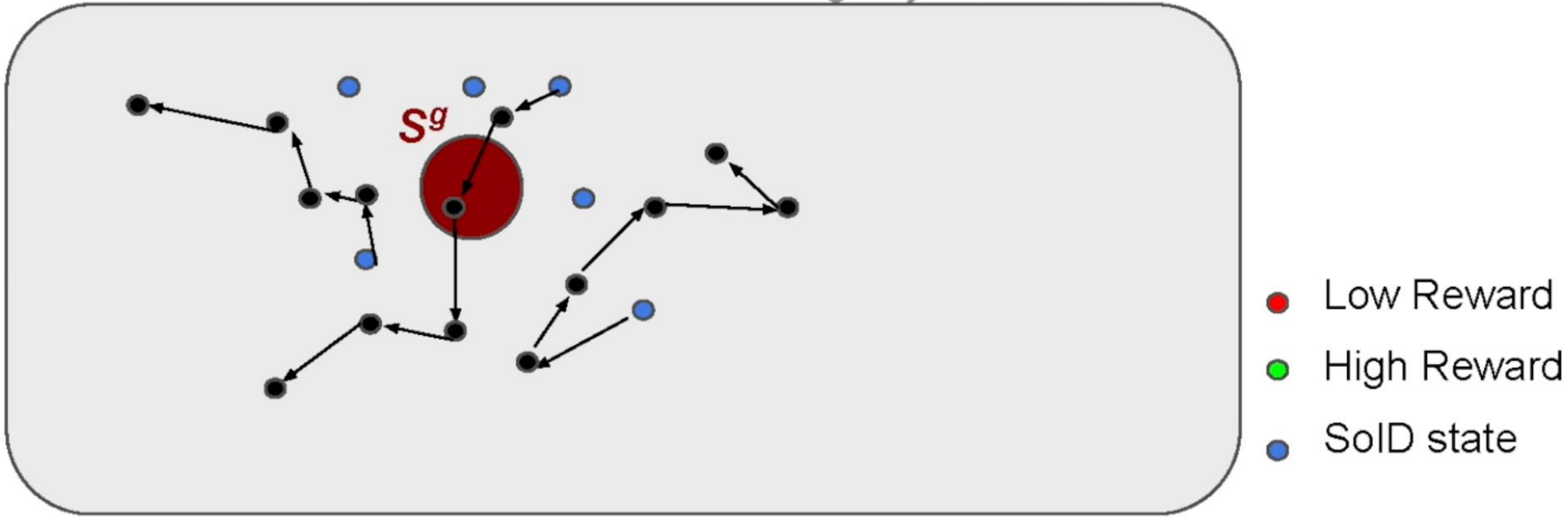


Iteration 1:

- Run Brownian motion
- Obtain trajectories from collected starts to train policy
- Label and filter starts based on training trajectories

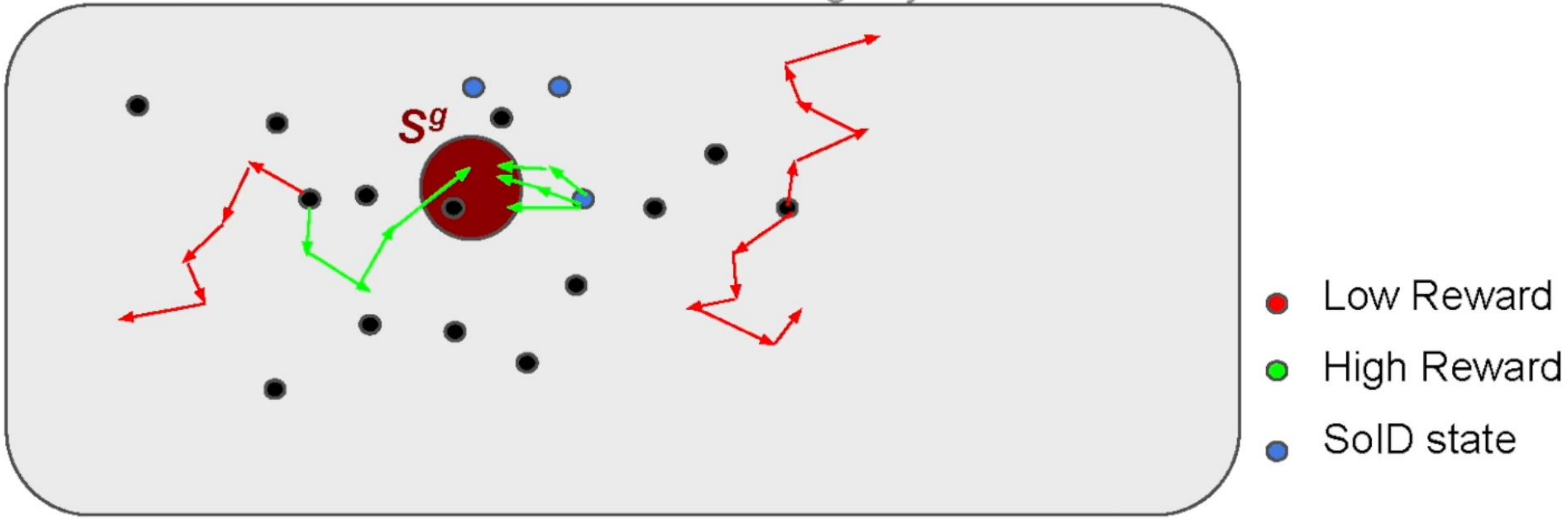
Iteration 2:

- Run Brownian motion
- Obtain trajectories from collected starts to train policy
- Label and filter starts based on training trajectories



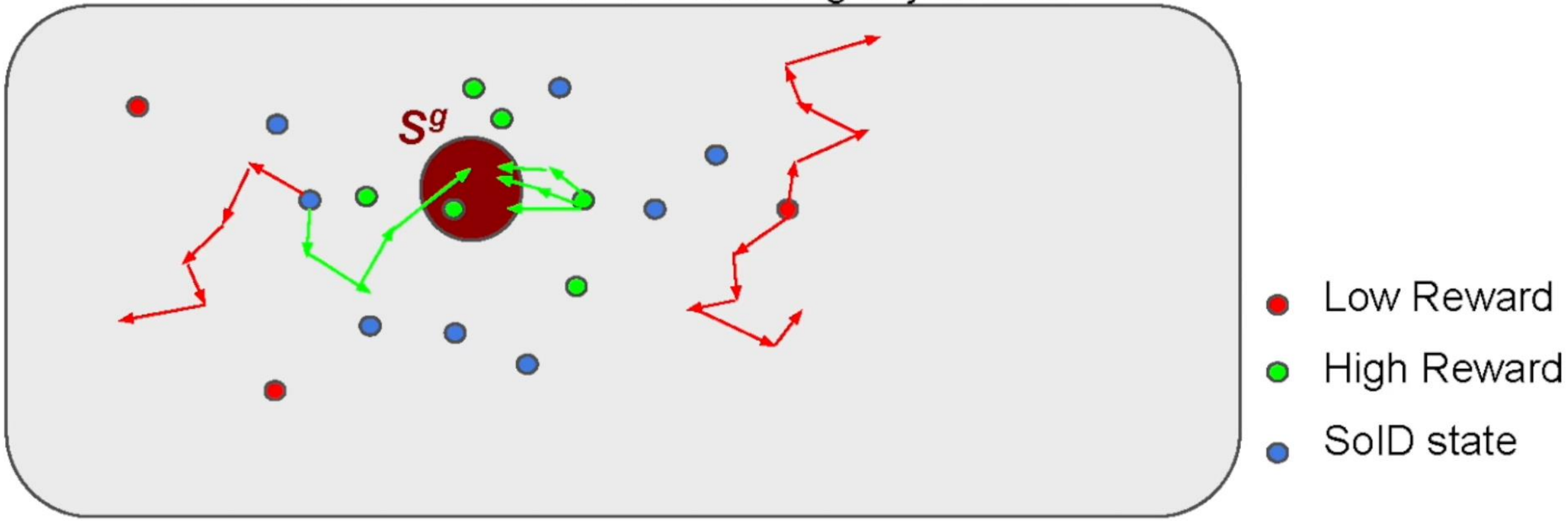
Iteration 2:

- Run Brownian motion
- Obtain trajectories from collected starts to train policy
- Label and filter starts based on training trajectories



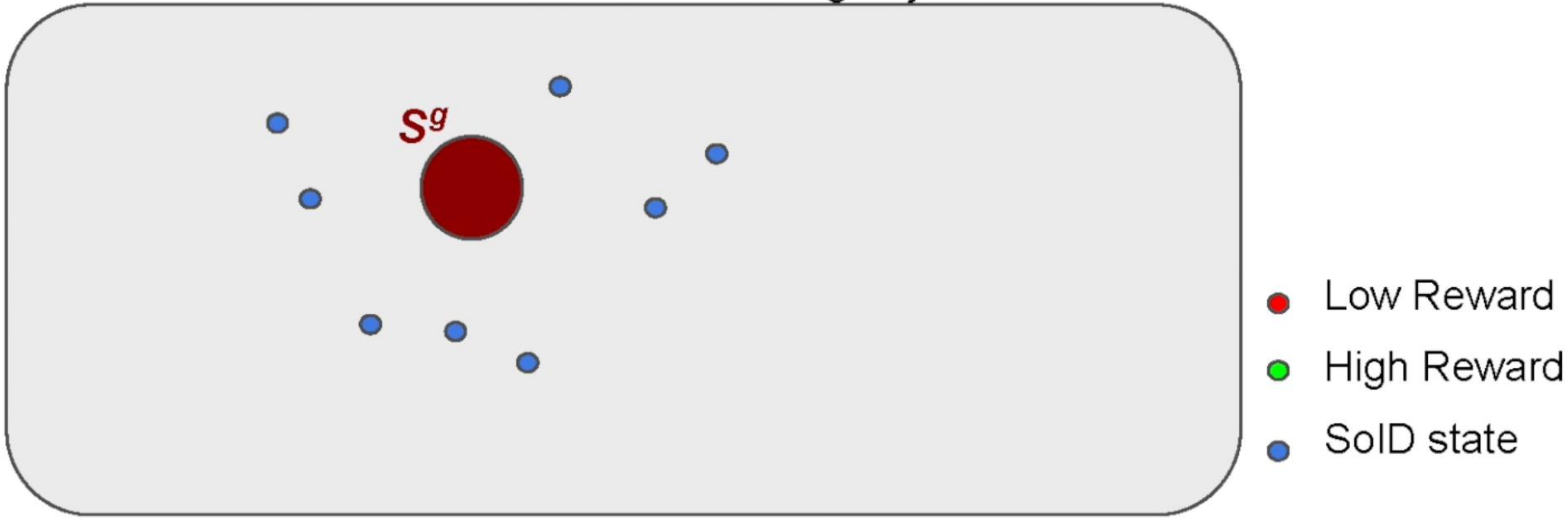
Iteration 2:

- Run Brownian motion
- Obtain trajectories from collected starts to train policy
- Label and filter starts based on training trajectories



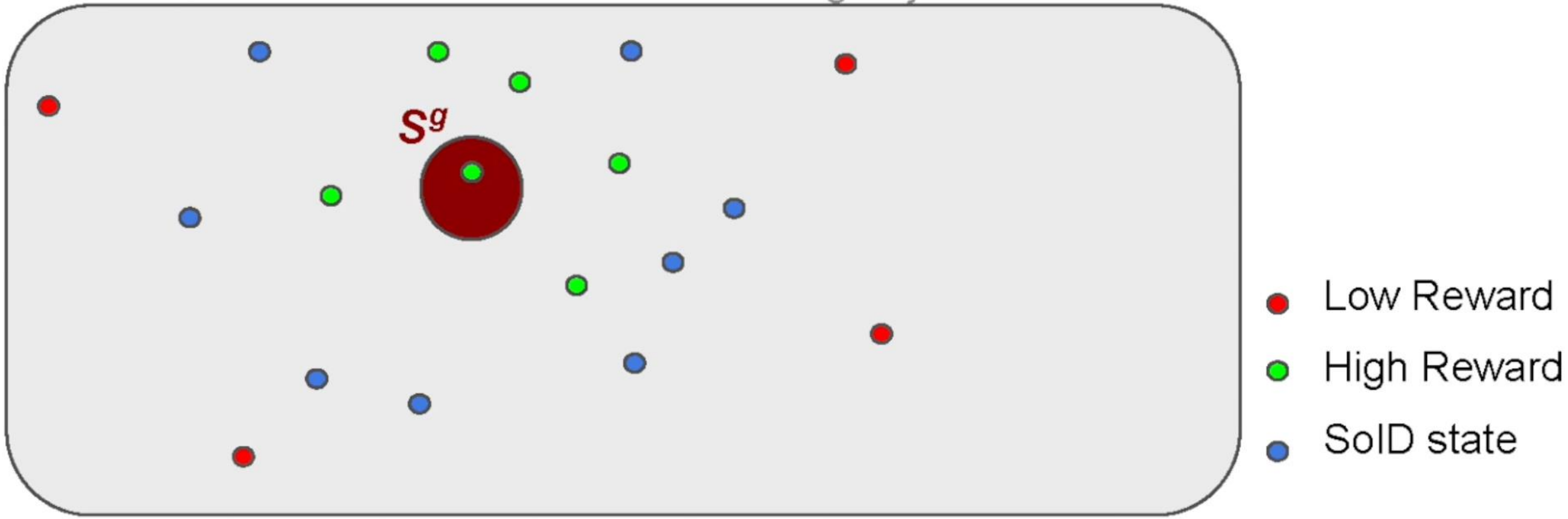
Iteration 2:

- Run Brownian motion
- Obtain trajectories from collected starts to train policy
- Label and filter starts based on training trajectories



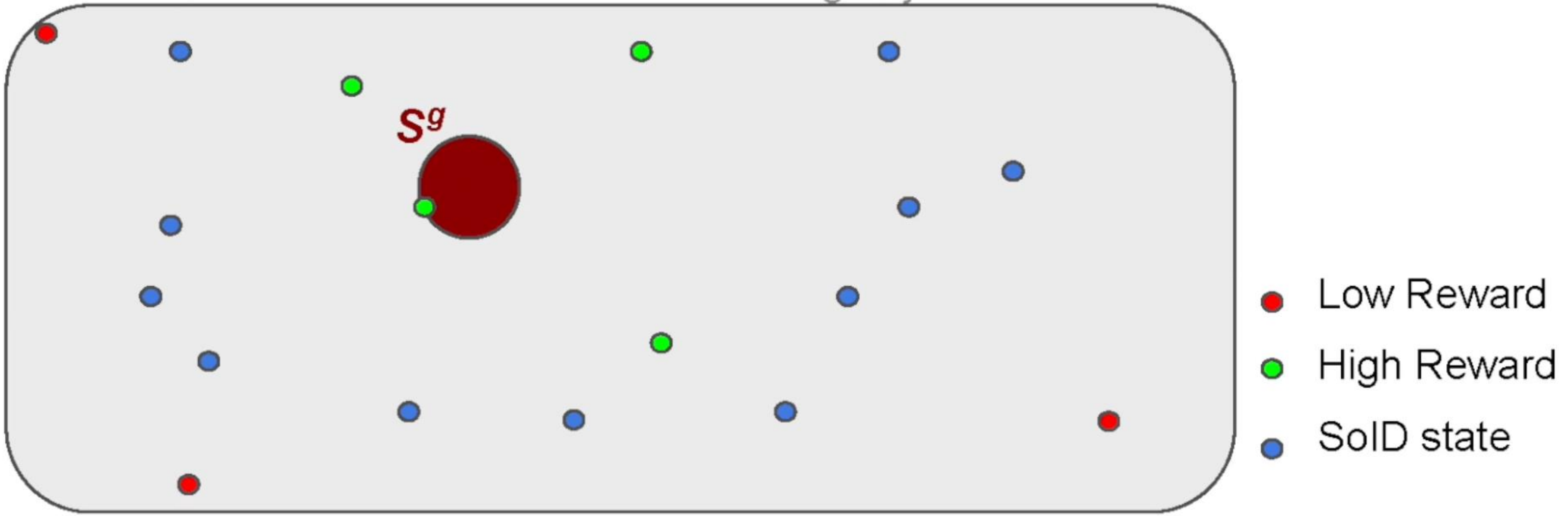
Iteration 3:

- Run Brownian motion
- Obtain trajectories from collected starts to train policy
- Label and filter starts based on training trajectories



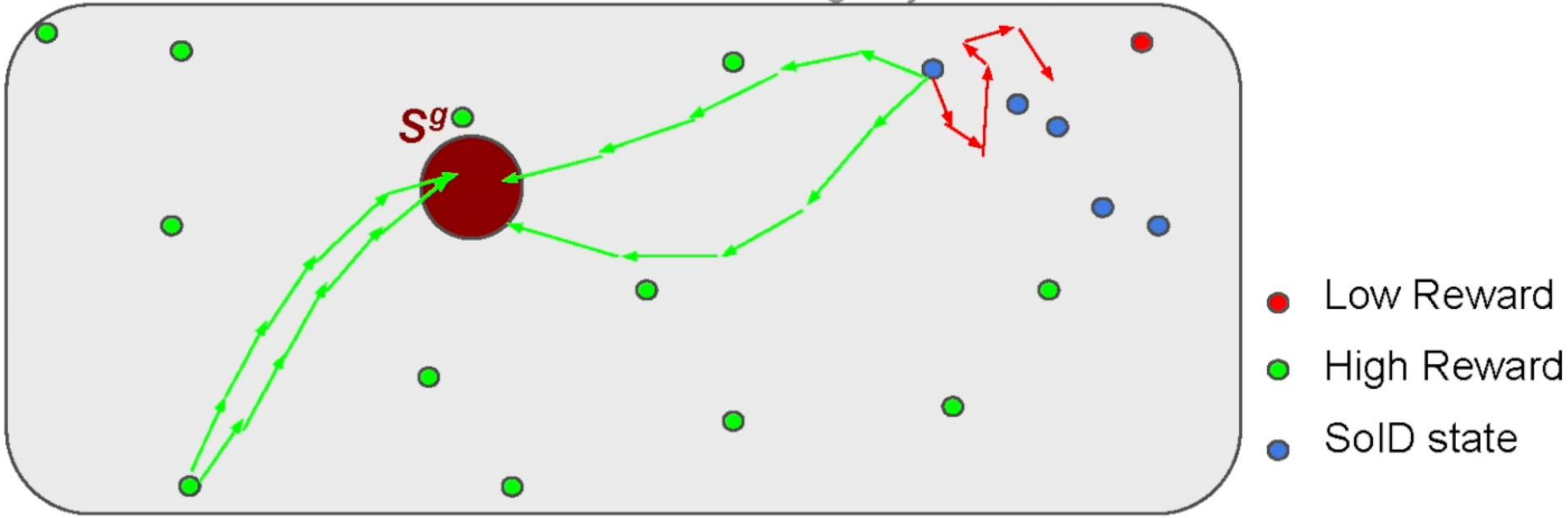
Iteration 4:

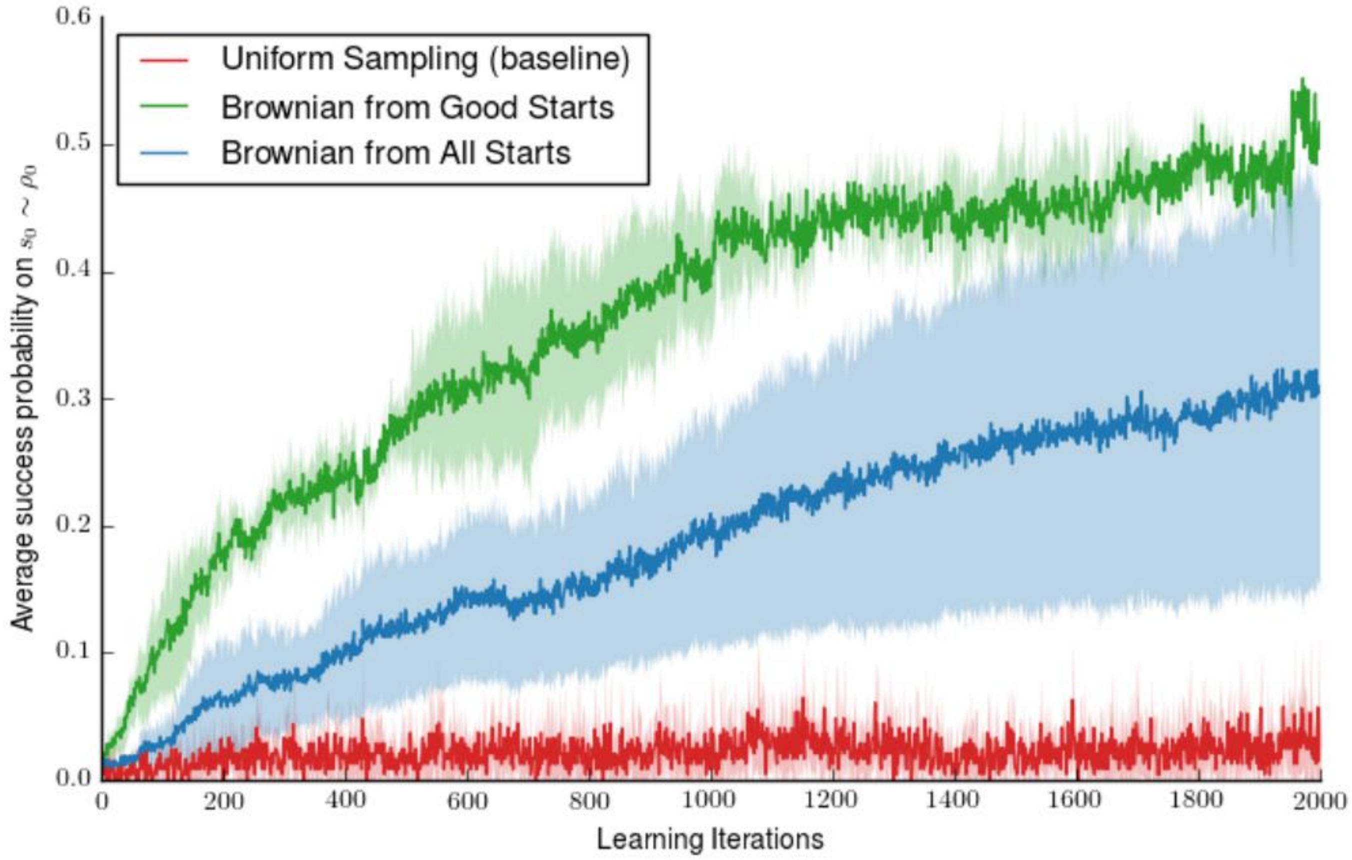
- Run Brownian motion
- Obtain trajectories from collected starts to train policy
- Label and filter starts based on training trajectories



Iteration 5:

- Run Brownian motion
- Obtain trajectories from collected starts to train policy
- Label and filter starts based on training trajectories





(d) Key insertion task

Curriculum Learning meets Reinforcement Learning

Intrinsic Motivation and Automatic Curricula via Asymmetric Self-Play

Sukhbaatar et al.

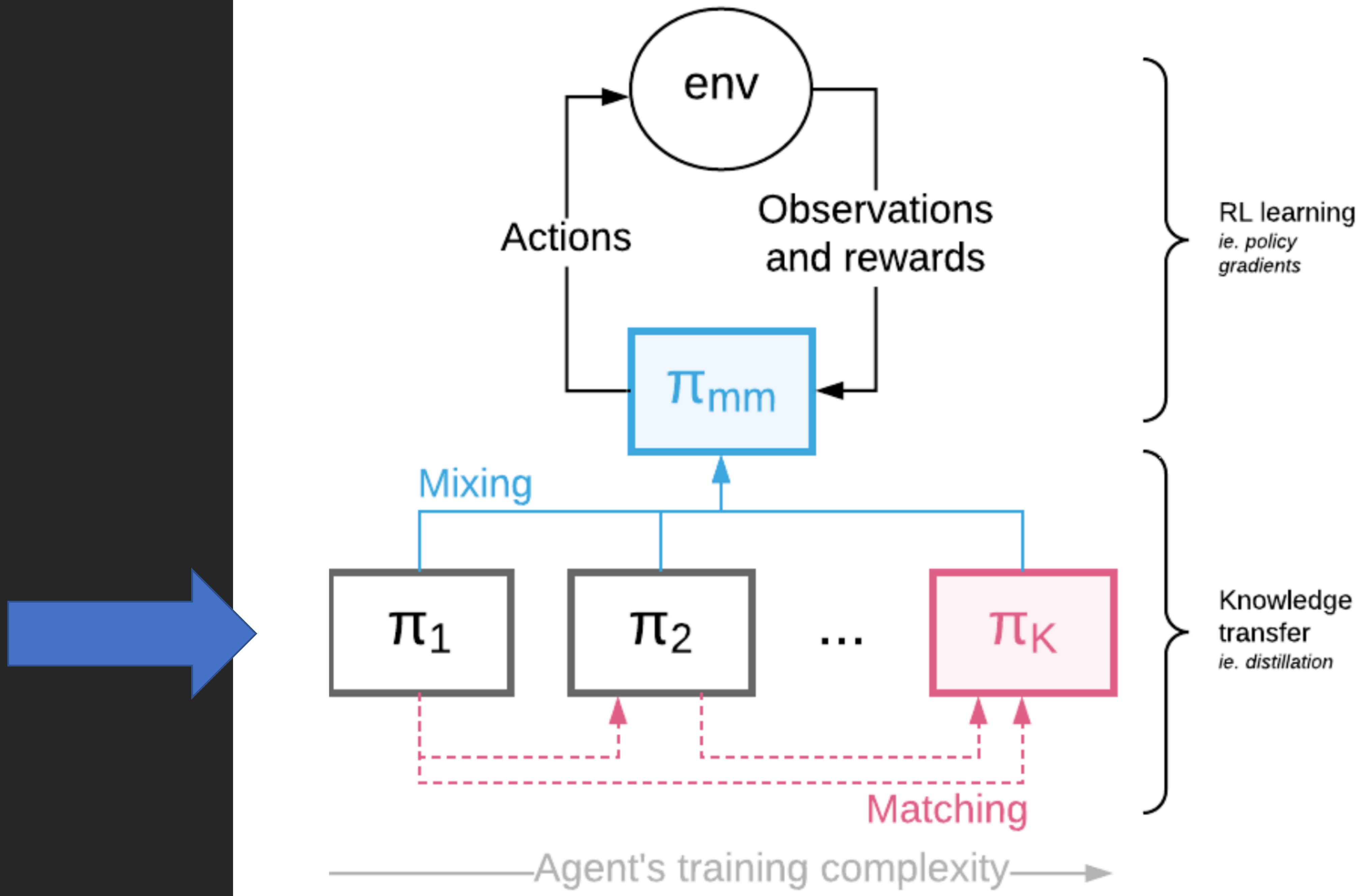
Reverse Curriculum Generation for Reinforcement Learning

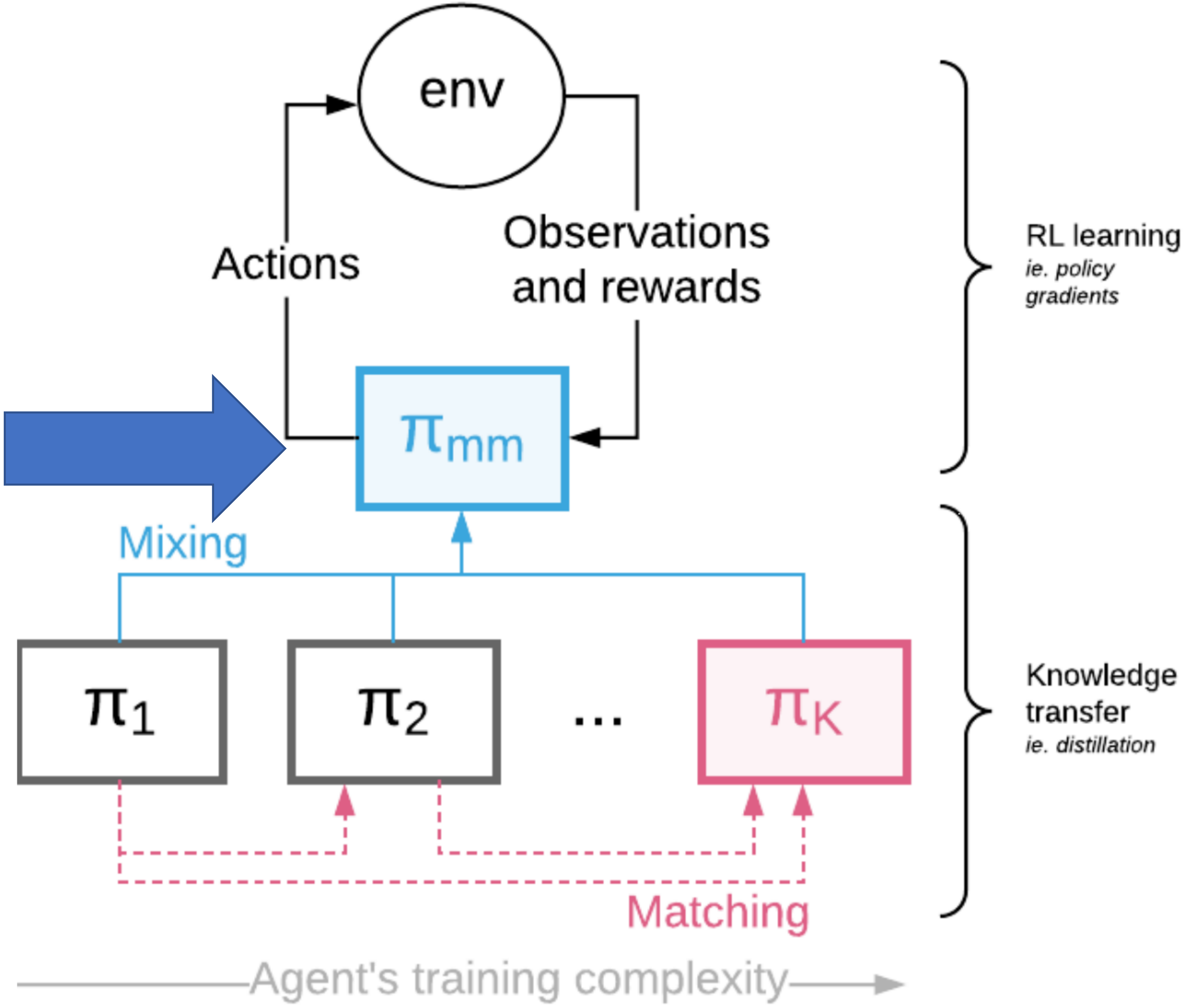
Florensa et al.

Mix & Match – Agent Curricula for Reinforcement Learning

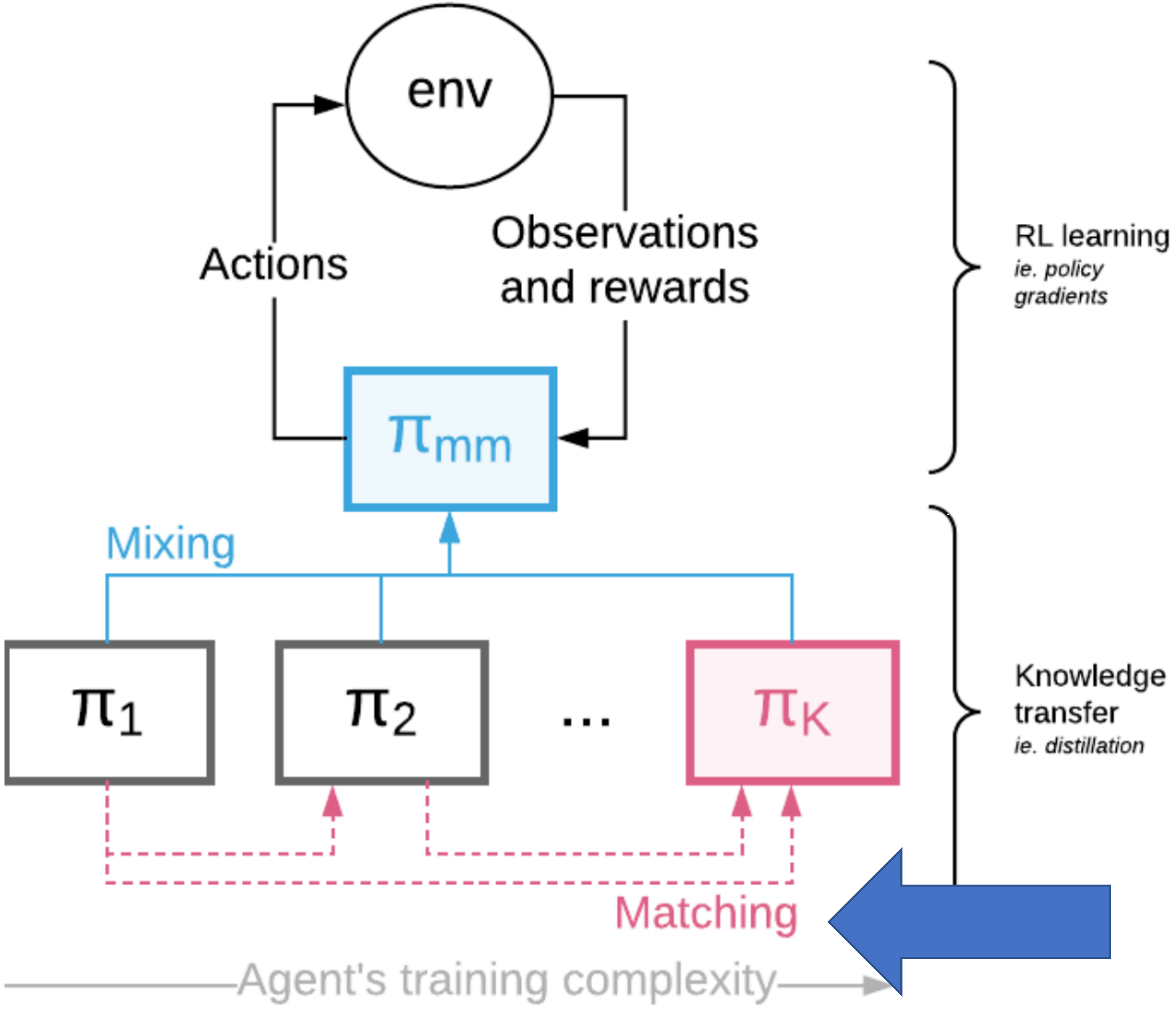
Czarnecki et al.

Rethinking the Notion of Curriculum

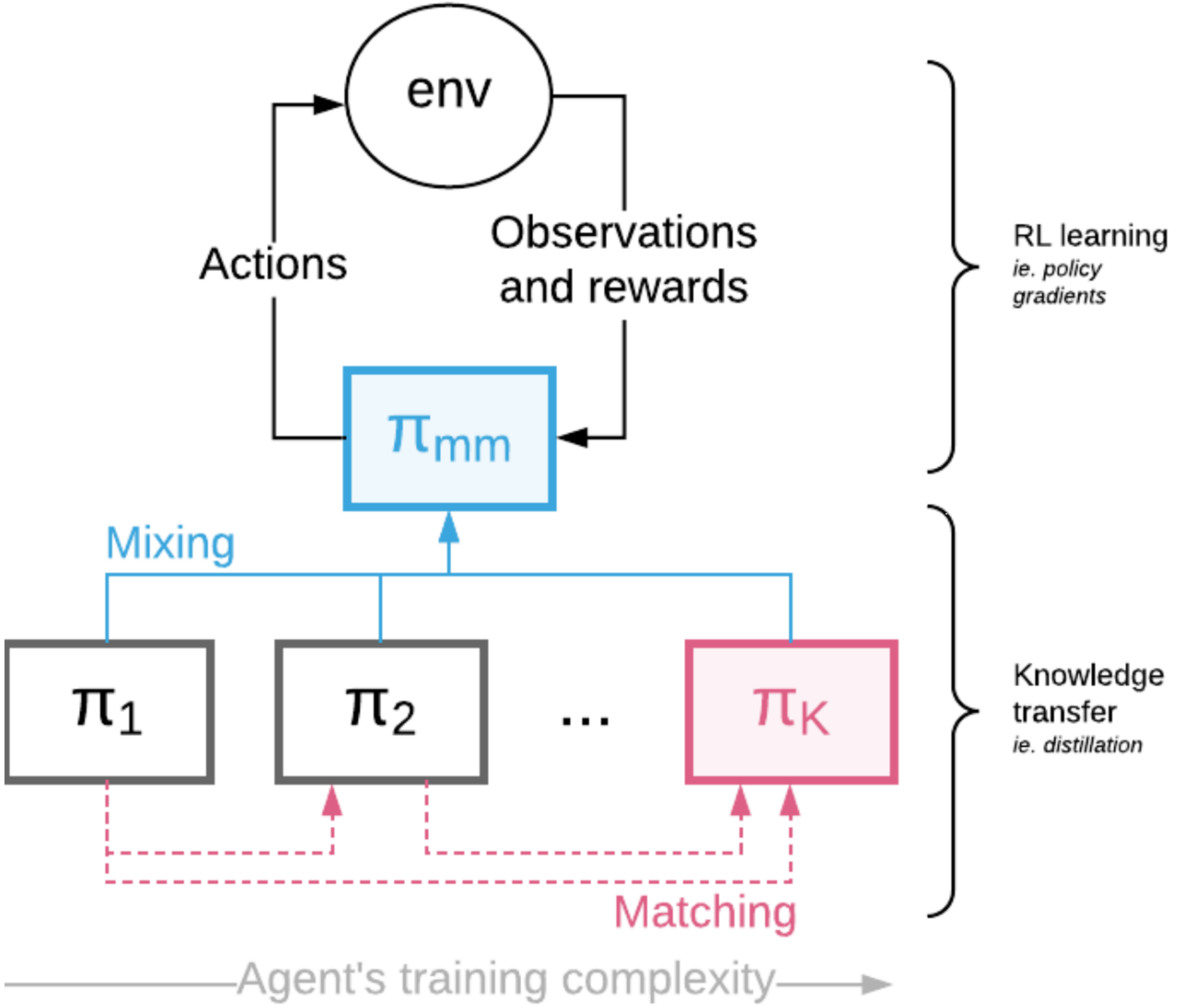

env
Actions
Observations and rewards
π_{mm}
Mixing
π_1
π_2
...
π_K
Matching
Agent's training complexity
RL learning
ie. policy gradients
Knowledge transfer
ie. distillation



env
Actions
Observations
and rewards
π_{mm}
Mixing
π_1
π_2
...
π_K
Matching
Agent's training complexity
RL learning
ie. policy
gradients
Knowledge
transfer
ie. distillation



env
Actions
Observations and rewards
π_{mm}
Mixing
π_1
π_2
...
π_K
Matching
Agent's training complexity
RL learning
ie. policy gradients
Knowledge transfer
ie. distillation


env
Actions
Observations and rewards
π_{mm}
Mixing
π_1
π_2
...
π_K
Matching
Agent's training complexity
RL learning
ie. policy gradients
Knowledge transfer
ie. distillation

Curriculum not Automatic!

What’s the Difficulty of an Agent?

Agents are Neural Networks!

*for all practical purposes

Architectural Components

or

Performable Actions

or

Jointly-Learnable Tasks

and

Training Iterations

Architectural Components

or

Performable Actions

or

Jointly-Learnable Tasks

and

Training Iterations

Architectural Components
or
Performable Actions
or
Jointly-Learnable Tasks
and
Training Iterations

Architectural Components

or

Performable Actions

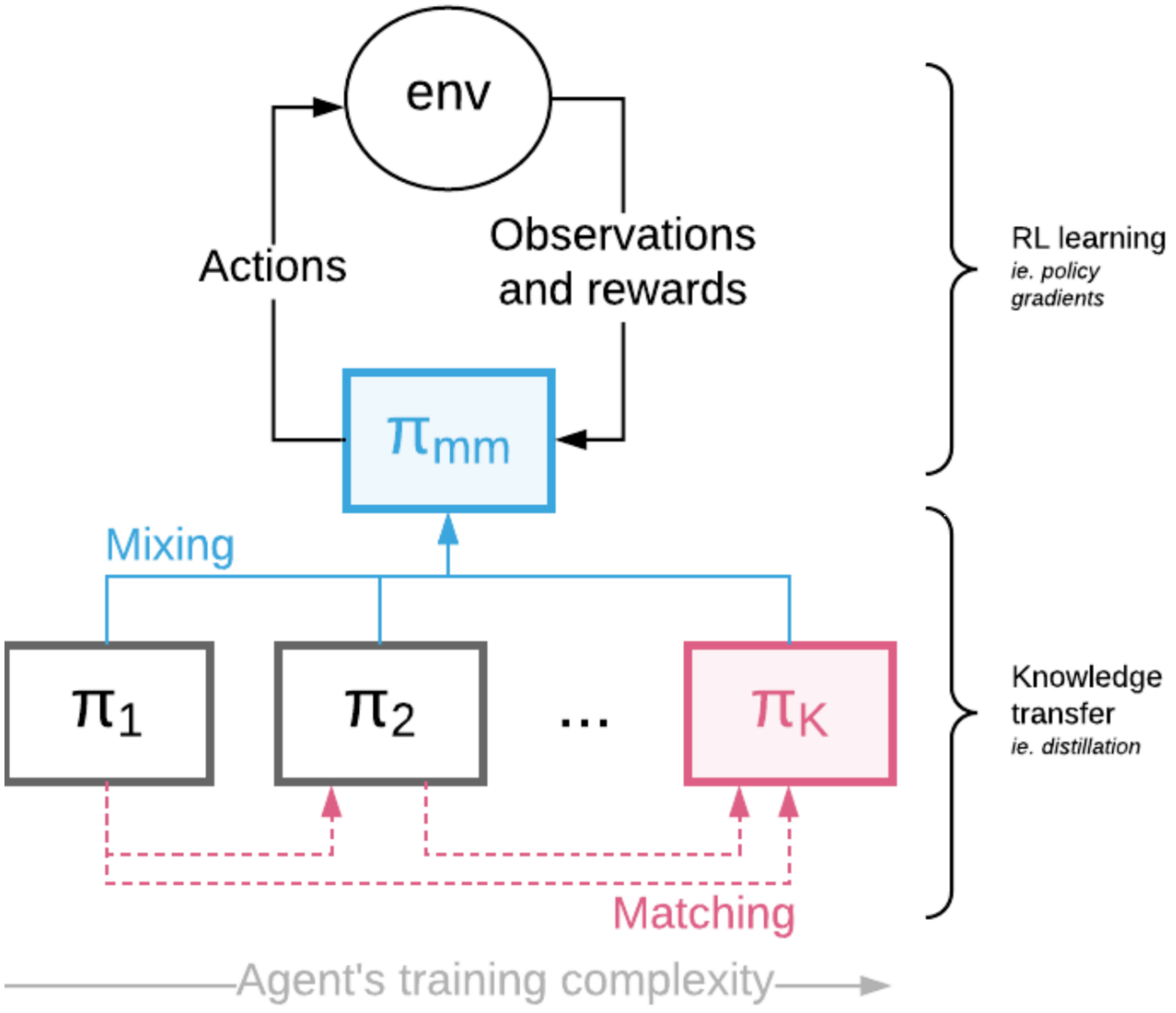
or

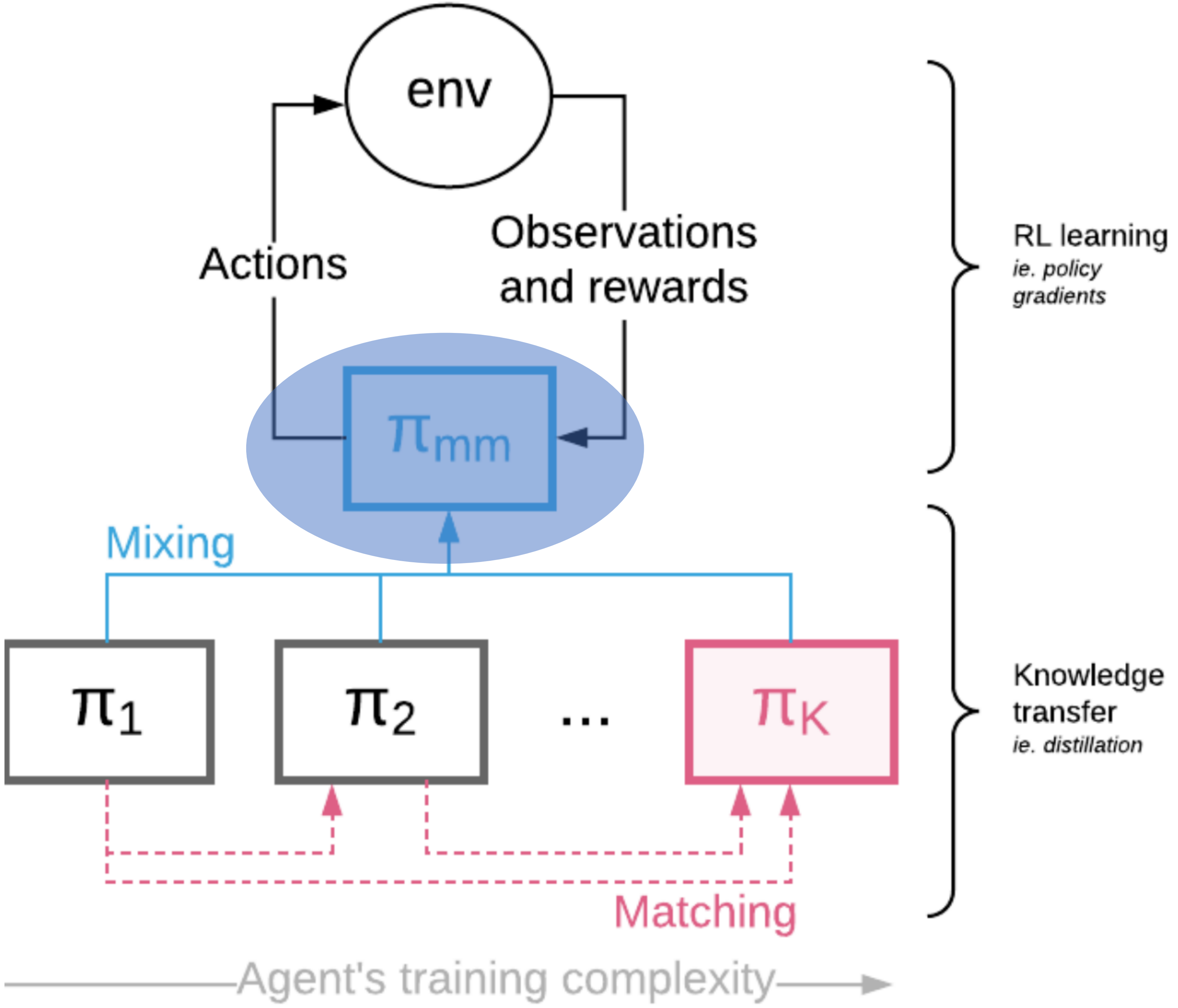
Jointly-Learnable Tasks

and

Training Iterations

Difficulty ✓


env
Actions
Observations and rewards
π_{mm}
Mixing
π_1
π_2
...
π_K
Matching
Agent's training complexity
RL learning
ie. policy gradients
Knowledge transfer
ie. distillation


env
Actions
Observations and rewards
π_{mm}
Mixing
π_1
π_2
...
π_K
Matching
Agent's training complexity
RL learning
ie. policy gradients
Knowledge transfer
ie. distillation

Scheduler: Tune Mixture Parameter $\boldsymbol{\alpha}$

Scheduler: Tune Mixture Parameter $\boldsymbol{\alpha}$

→ Could use hand crafted scheduler ☹

Scheduler: Tune Mixture Parameter α

Could use hand crafted scheduler ☹

Could use naive hyperparameter tuning ☹

Scheduler: Tune Mixture Parameter α

Could use hand crafted scheduler ☹

Could use naive hyperparameter tuning ☹

Population Based Training ☺

Population Based Training

Population Based Training

1 Tuning several mixture agents in parallel

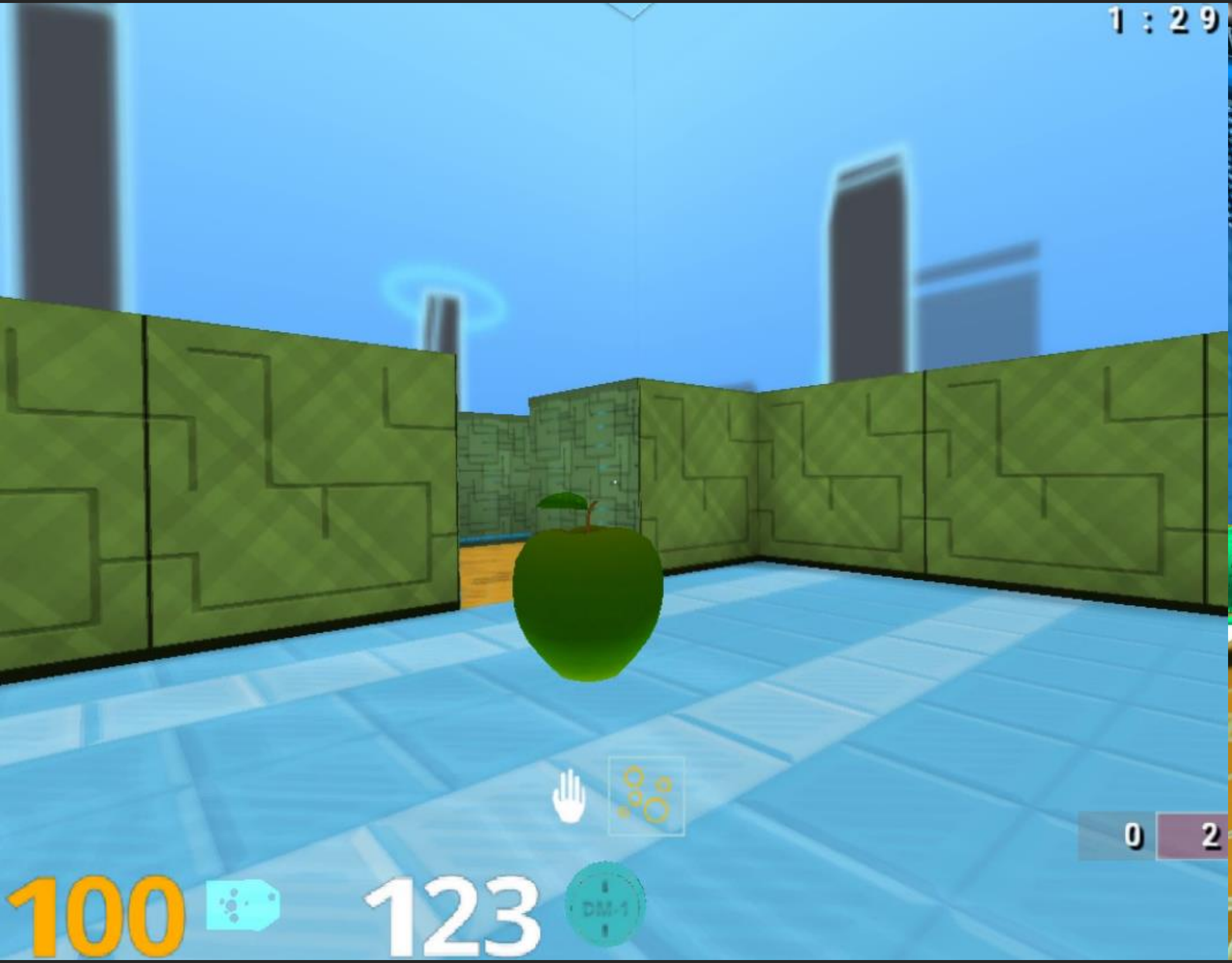
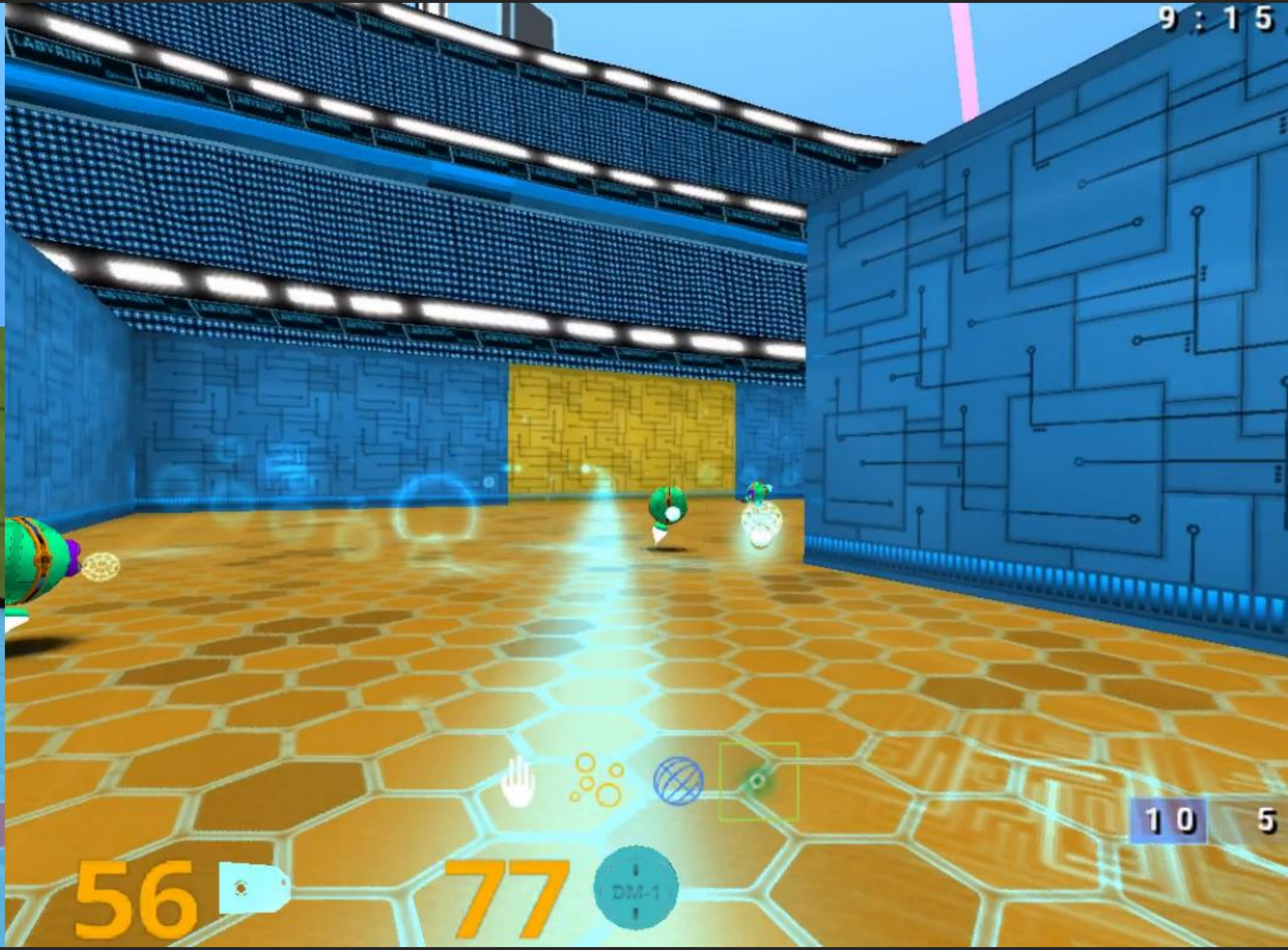
Population Based Training

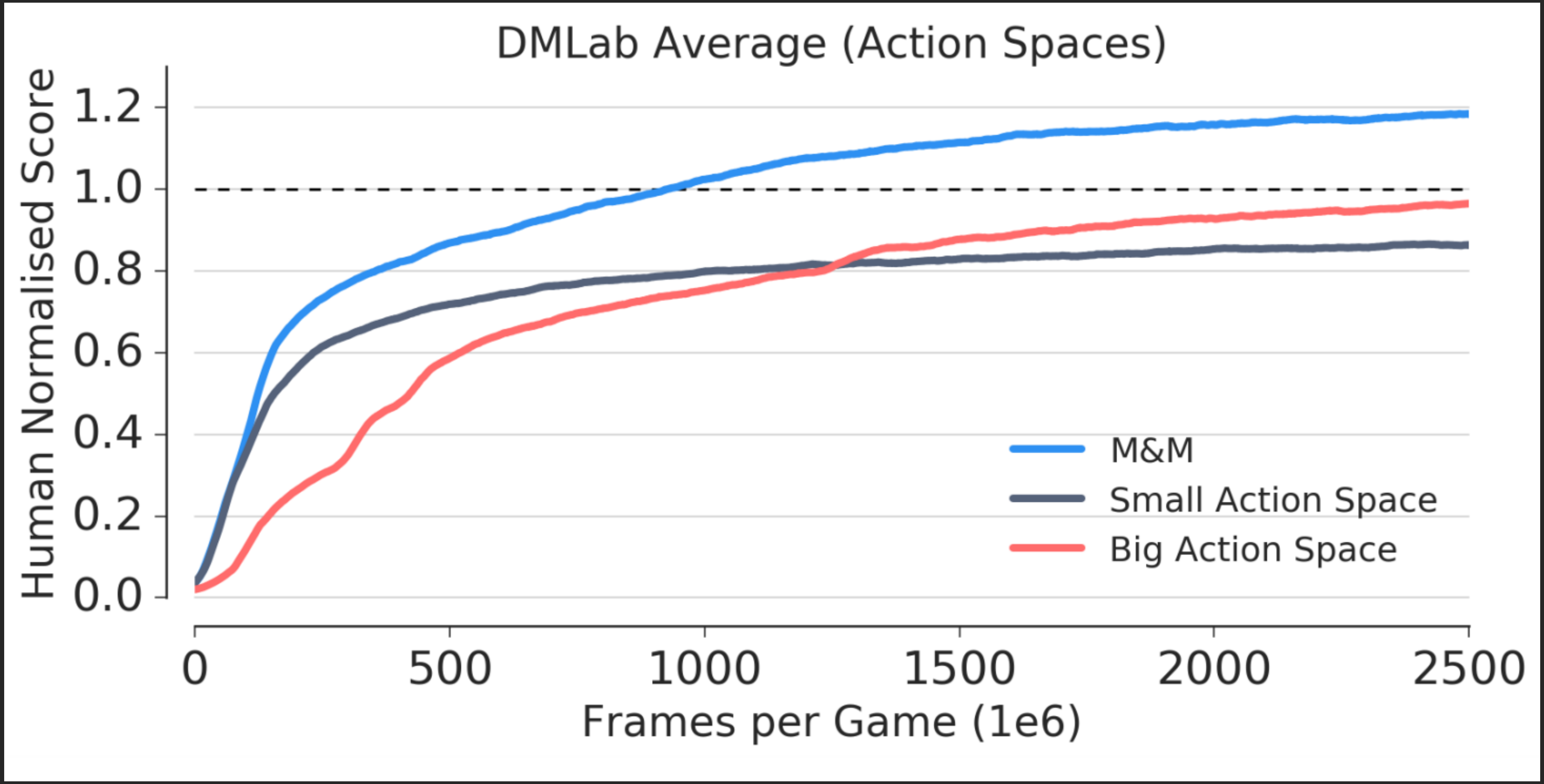
1. Tuning several mixture agents in parallel
2. Agent A periodically communicates with some B

Population Based Training

1. Tuning several mixture agents in parallel
2. Agent A periodically communicates with some B
3. Badly performing: Copy weights and hyperparameters (α)

Explore Search Space with badly performing Agents


1 : 29
0
2
100
123
DM-1

9 : 15
10
5
56
77
DM-1



DMLab Average (Action Spaces)
Human Normalised Score
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
2500
Frames per Game (1e6)
M&M
Small Action Space
Big Action Space

Curriculum Learning Is Here to Stay! ☺


Yes, Mr. Frodo. It's over now.

Take Care!