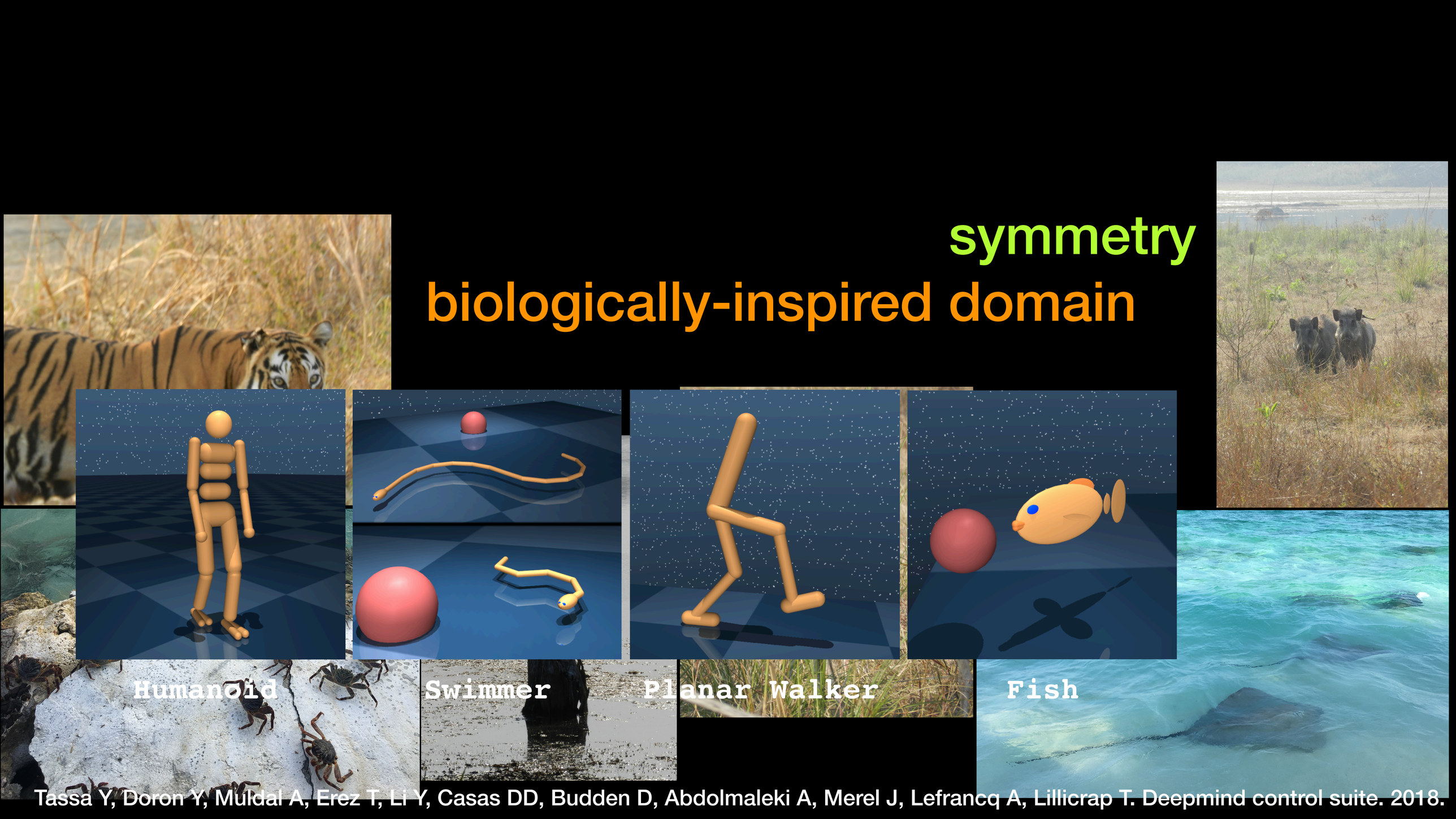
reinforcement learning for locomotion

Shruti Mishra, Abbas Abdolmaleki,
Arthur Guez, Piotr Trochim, Doina Precup

reinforcement learning

biologically-inspired domain

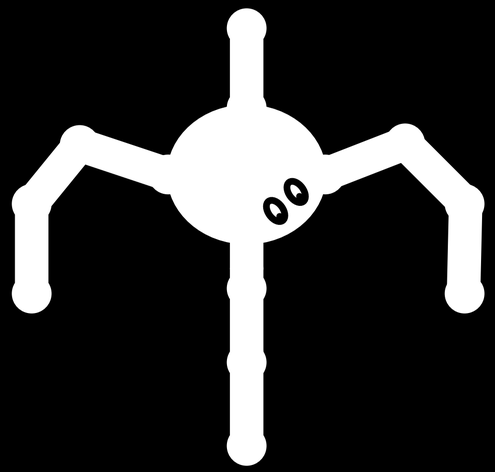


Tassa Y, Doron Y, Muldal A, Erez T, Li Y, Casas DD, Budden D, Abdolmaleki A, Merel J, Lefrancq A, Lillicrap T. Deepmind control suite. 2018.

Augmenting learning using symmetry
in a biologically-inspired domain

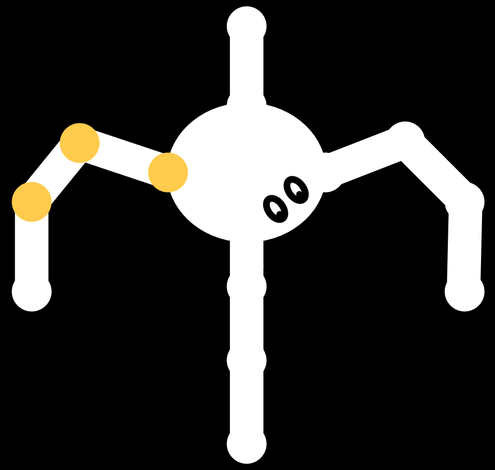
Augmenting learning using symmetry in a biologically-inspired domain

Domain: Quadruped



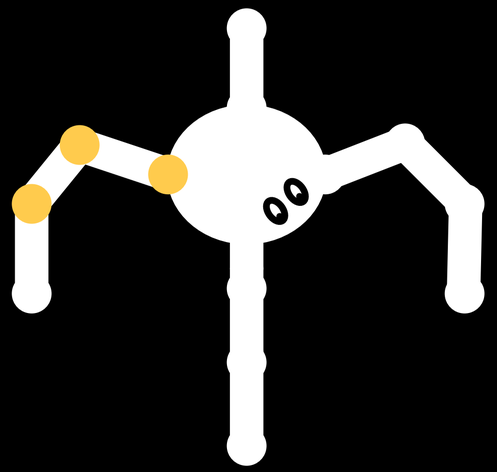
Tassa Y, Doron Y, Muldal A, Erez T, Li Y, Casas DD, Budden D, Abdolmaleki A, Merel J, Lefrancq A, Lillicrap T. DeepMind control suite. 2018.

Domain: Quadruped



Tassa Y, Doron Y, Muldal A, Erez T, Li Y, Casas DD, Budden D, Abdolmaleki A, Merel J, Lefrancq A, Lillicrap T. DeepMind control suite. 2018.

Domain: Quadruped



12 Actions

Each leg has 3 hinge joints → 12

78 Observations

Torso: Orientation relative to horizontal → 1

Torso: Forward velocity → 3

Torso: Accelerometer and gyroscope → 6

Each toe: Force and Torque → 12

Each joint: Instantaneous angle, angular velocity, activation → 44

Tassa Y, Doron Y, Muldal A, Erez T, Li Y, Casas DD, Budden D, Abdolmaleki A, Merel J, Lefrancq A, Lillicrap T. DeepMind control suite. 2018.

Task: `Walk` and `Run`

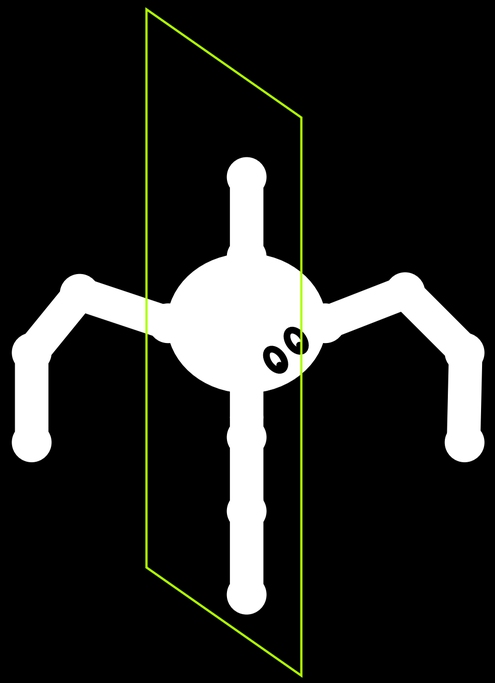

```
reward = move_reward * upright_reward
```

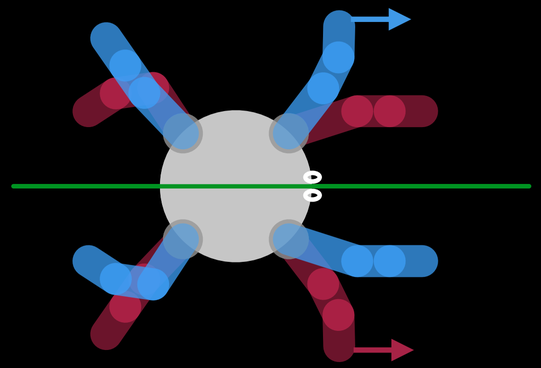
Tassa Y, Doron Y, Muldal A, Erez T, Li Y, Casas DD, Budden D, Abdolmaleki A, Merel J, Lefrancq A, Lillicrap T. DeepMind control suite. 2018.

Augmenting learning using symmetry in a biologically-inspired domain

Symmetry in the Quadruped Domain

Symmetry in the Quadruped Domain



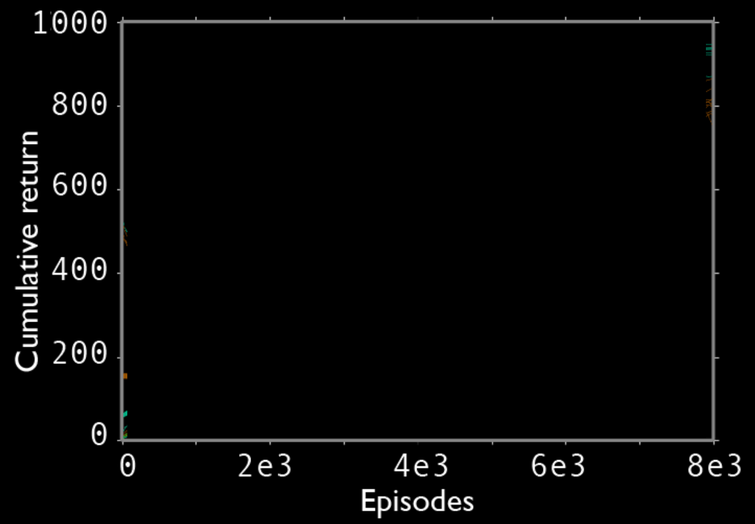


Augmenting learning using symmetry

`Critic` augments the `batches` of data from the `actor` with `batches` of mirrored data.

Preliminary results

Batch size = 256

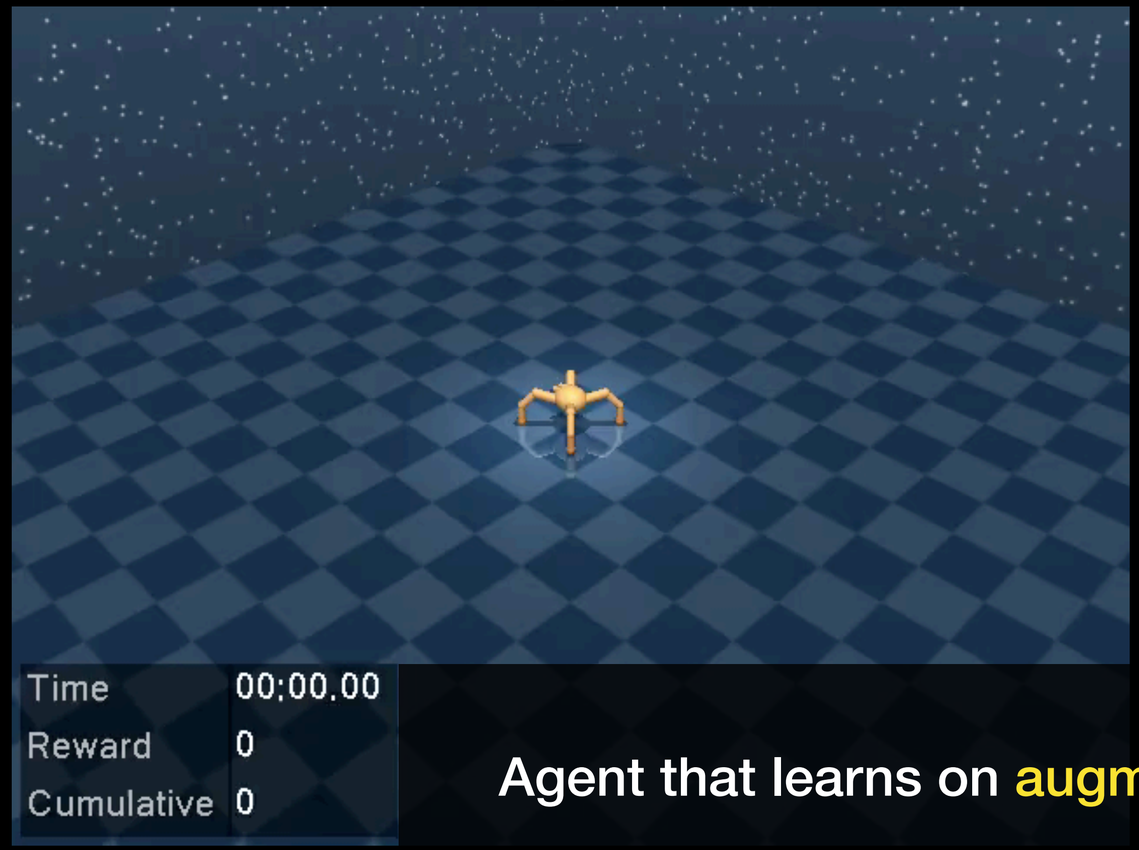


Walk task

Abdolmaleki A, Springenberg JT, Tassa Y, Munos R, Heess N, Riedmiller M. Maximum a posteriori policy optimisation. 2018.

Preliminary results

Learning on data



Learning on augmented data

Agent that learns on augmented data looks more natural.

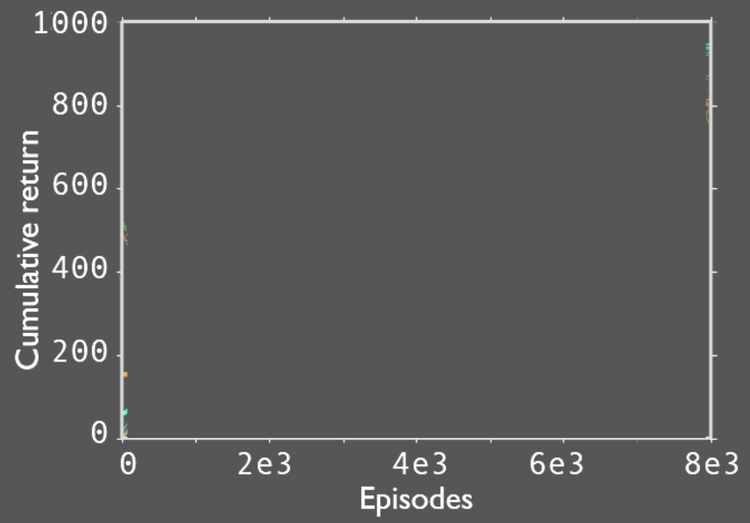
We need more quantitative analysis.

Algorithm

given batch-size (K), number of actions (N), Q-function $\hat{Q}$, old-policy π_k and replay-buffer
initialize π_θ **from the parameters of** $\pi^{(k)}$
repeat
    Sample states with batch of size N from replay buffer
    Step 1: sample based policy (weights)
    $q(a_i|s_j) = q_{ij}$, **computed as:**
    for j = 1,...,K **do**
        for i = 1,...,N **do**
            $a_i \sim \pi_k(a|s_j)$
            $Q_{ij} = \hat{Q}(s_j, a_i)$
            $q_{ij} \propto \exp(Q_{ij}/\text{temperature parameter}))$
        end for
    end for
    Calculate mirrored states and mirrored actions
    Step 2: update parametric policy
    Given the data-set $\{s_j, (a_i, q_{ij})_{i=1...N}\}_{j=1...K}$ and $\{s_j^{\text{mirrored}}, (a_i^{\text{mirrored}}, q_{ij})_{i=1...N}\}_{j=1...K}$
    Update the Policy by taking gradient of following weighted maximum likelihood objective
    $\max_\theta(\sum_j^K \sum_i^N q_{ij} \log \pi_\theta(a_i|s_j) + \sum_j^K \sum_i^N q_{ij} \log \pi_\theta(a_i^{\text{mirrored}}|s_j^{\text{mirrored}}))$
    (subject to additional (KL) regularization)
until Fixed number of steps
return $\pi_{(k+1)}$

Preliminary results

Batch size = 256



Walk task

Abdolmaleki A, Springenberg JT, Tassa Y, Munos R, Heess N, Riedmiller M. Maximum a posteriori policy optimisation. 2018.