Variational Information Maximization for Intrinsically Motivated Reinforcement Learning

Shakir Mohamed and Danilo J. Rezende

NIPS 2015

Presented by Sanghyeon Lee

Motivation

Introduction

Mutual Information is well used

1) Multi modal 2) Maximizing nosiy transmission channels 3) Learning behavior polices for exploration

Problem

Optimization Algorithm has large computational cost

Solution

Use Variational Inference for estimating mutual information

Contributions:

1. Propose Stochastic Variational Information Maximisation
2. Combine Variational information optimization and Deep learning for develop a algorithm for intrinsically-motivated RL
3. This methods has lower computational than previous algorithms

Motivation

Intrinsic Motivation

- Limitation of the standard RL approach ~ Agent is only able to learn using external rewards

What is intrinsic Motivation ?

- Each agent have internal desires (e.g hunger, boredom, curiosity)
- These desires allows the agent to continue to explore

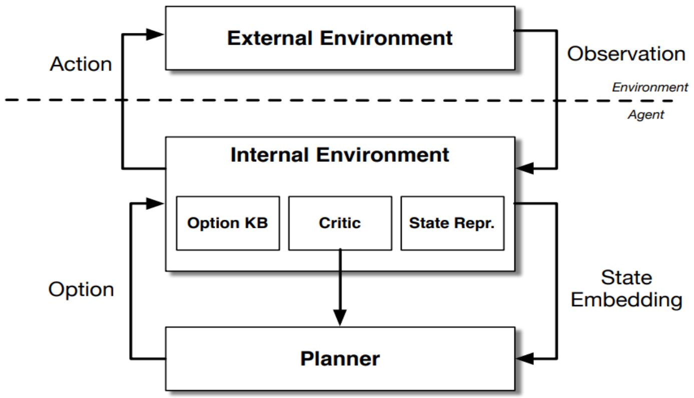
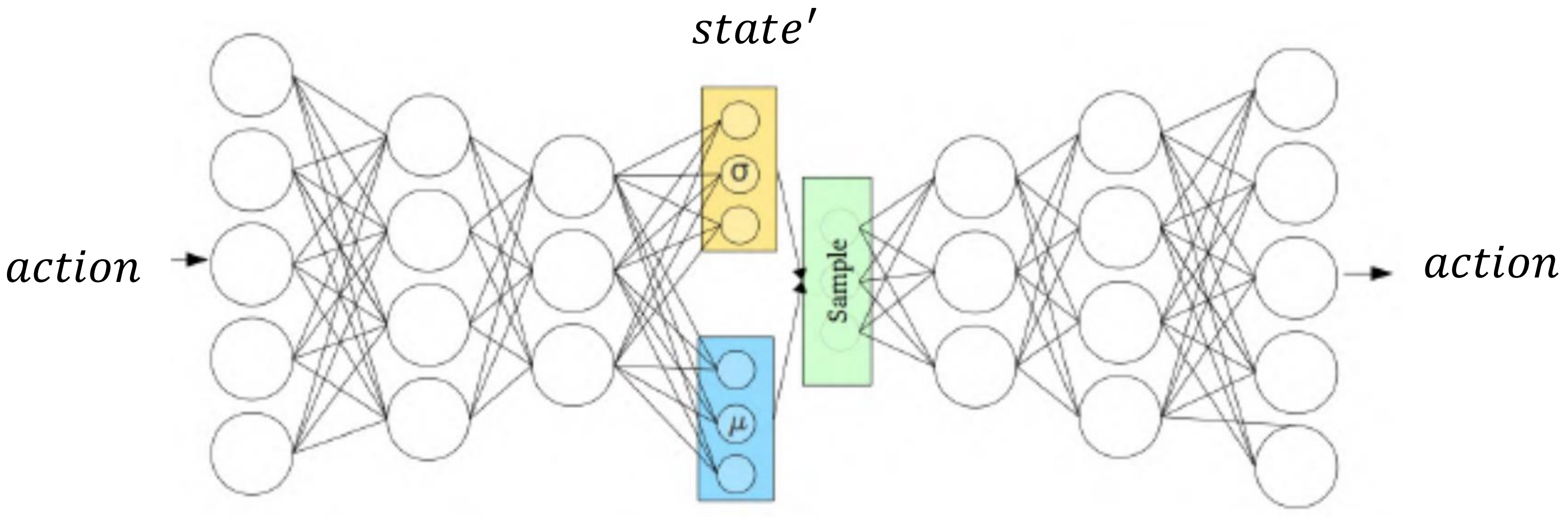


Figure 1: Perception-action loop separating environment into internal and external facets.



Method

Notation

$\mathbf{a} = \{a_1, \ldots., a_k\}$: Sequence of K primitive actions a_k leading to final state s'

$p(s'|a, s)$: K-step transition probability of the environment

$p(a, s'|s)$: Joint distribution of action sequences and the final state

$w(a|s)$: A distribution over K-step action sequences,

We want this policy be efficient exploration policy

& This policy is not used by the agent for acting

acting policy is determined by other learning algorithms (ex Q learning)

$p(s'|s)$: The joint probability marginalized over the action sequence

Method

Empowerment

- There are many formally define internal drives
- Commonly usages is Mutual Information
- Empowerment : Internal reward measure , Maximize Mutual Information

$$\mathrm{E(s)} = \max_{w} I^{w}(a, s'|s) = \max E_{p\left(s'\middle|a,s\right)w(a|s)}[\log\left(\frac{p(a, s'|s)}{w(a|s)p(s'|s)}\right)]$$

- $w(a|s)$: We want this policy be efficient exploration policy

  & This policy is not used by the agent for acting but internal

  acting policy is determined by other learning algorithms (ex Q learning)
- $p(s'|s)$: The joint probability marginalized over the action sequence
- This measure is of the amount of information contained in action sequences ***a*** about the future state s'

Method

Scalable Information Maximization

$H(a|s', s) = -E_{p(s'|a, s)w(a|s)}[\log p(a|s', s)]$, $H(a|s) = -E_{w(a|s)}[\log w(a|s)]$

$w(a|s)$: Internal policy

$p(a|s', s)$: True action posterior distribution ➔ intractable

$p(s'|a, s)$: we don't know transition dynamics of the environment $p(s'|a, s)$ - How to get this value?

➔ Sampling or Generate model of the Environment

& Previous algorithm – Blahut-Arimoto algorithm has inefficiency

➔ Variational Method

Method

Variational Informational Lower Bound

$$\mathrm{KL}[p(x|y)\|q(x|y)] \geq 0 \Rightarrow H(x|y) \leq -\mathbb{E}_{p(x|y)}\left[\log q_{\xi}(x|y)\right]$$

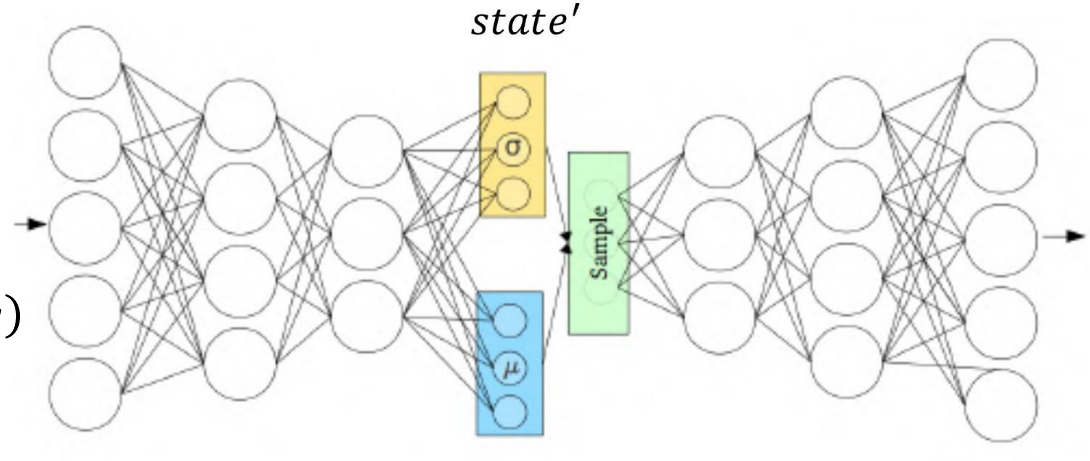
$$\mathcal{I}^{\omega}(\mathbf{s}) = H(\mathbf{a}|\mathbf{s}) - H(\mathbf{a}|\mathbf{s}', \mathbf{s}) \geq H(\mathbf{a}) + \mathbb{E}_{p(s'|a,s)\omega_{\theta}(a|s)}[\log q_{\xi}(\mathbf{a}|\mathbf{s}', \mathbf{s})] = \mathcal{I}^{\omega,q}(\mathbf{s})$$

$p(a|s',s)$ ➔ $q_{\xi}(a|s',s)$: We approximate intractable distribution from tractable distribution q

$w_{\theta}(a|,s)$ and $q_{\xi}(a|s',s)$ are the function of hyperparameters

$p(s'|\mathrm{a}, s)$

$state'$



$action$

$w_{\theta}(a|,s)$

$action$

$q_{\xi}(a|s',s)$

Method

Variational Informational Maximization

$$\hat{\mathcal{E}}(\mathbf{s}) = \max_{\omega,q} \mathcal{I}^{\omega,q}(\mathbf{s})\ s.t.\ H(\mathbf{a}|\mathbf{s}) < \epsilon,\ \ \hat{\mathcal{E}}(\mathbf{s}) = \max_{\omega,q} \mathbb{E}_{p(s'|a,s)\omega(a|s)}[-\frac{1}{\beta}\ln\omega(\mathbf{a}|\mathbf{s}) + \ln q_\xi(\mathbf{a}|\mathbf{s}',\mathbf{s})]$$

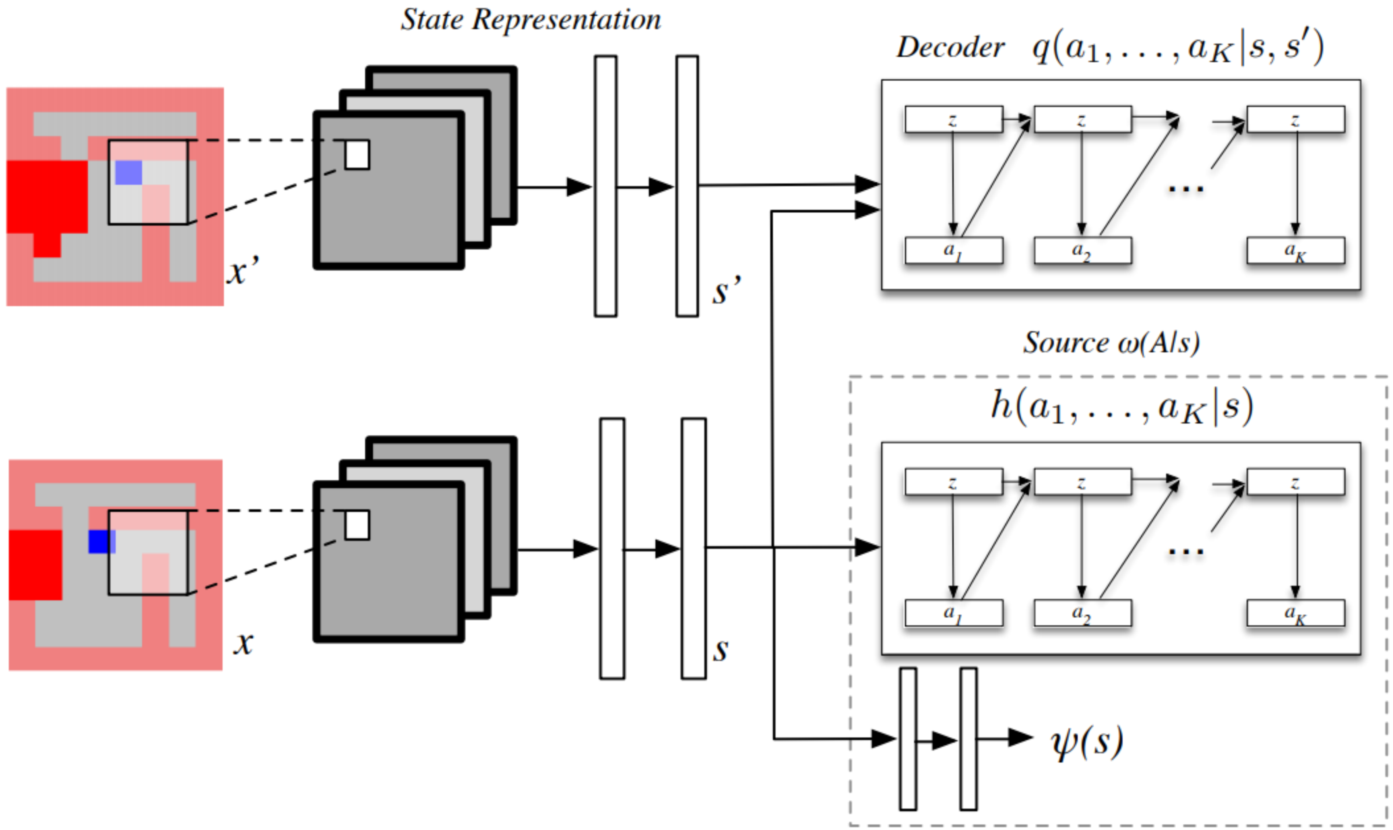
$\mathrm{H}(a|\mathrm{s}) < \epsilon$: Restriction

We ca
Evide
BOund

$$\hat{\phi} = \operatorname*{argmin}_{\phi} \mathrm{KL}[q(W;\phi) \,\|\, p(W|\mathcal{D})]$$

$$= \operatorname*{argmin}_{\phi} \int q(W;\phi)\log\frac{q(W;\phi)}{p(W)p(\mathcal{D}|W)}dW$$

$$= \operatorname*{argmin}_{\phi} \underbrace{\mathrm{KL}[q(W;\phi)\|p(W)]}_{\text{Regularization term not too far from prior}} - \underbrace{\mathbb{E}_{q(W;\phi)}[\log p(}_{\text{Expected likelih over dataset}}$$



Method

Encoder:

$$w^*(a|s) \approx h_\theta(a|s) \Rightarrow \hat{u}(s,a) \approx r_\theta(s,a)$$

$$w^*(a|s) = \frac{1}{Z(s)} \exp\big(\hat{u}(s,a)\big) \ \&$$

$$r_\theta(s,a) = ln h_\theta(a|s) + \psi_\theta(s) \ \&$$

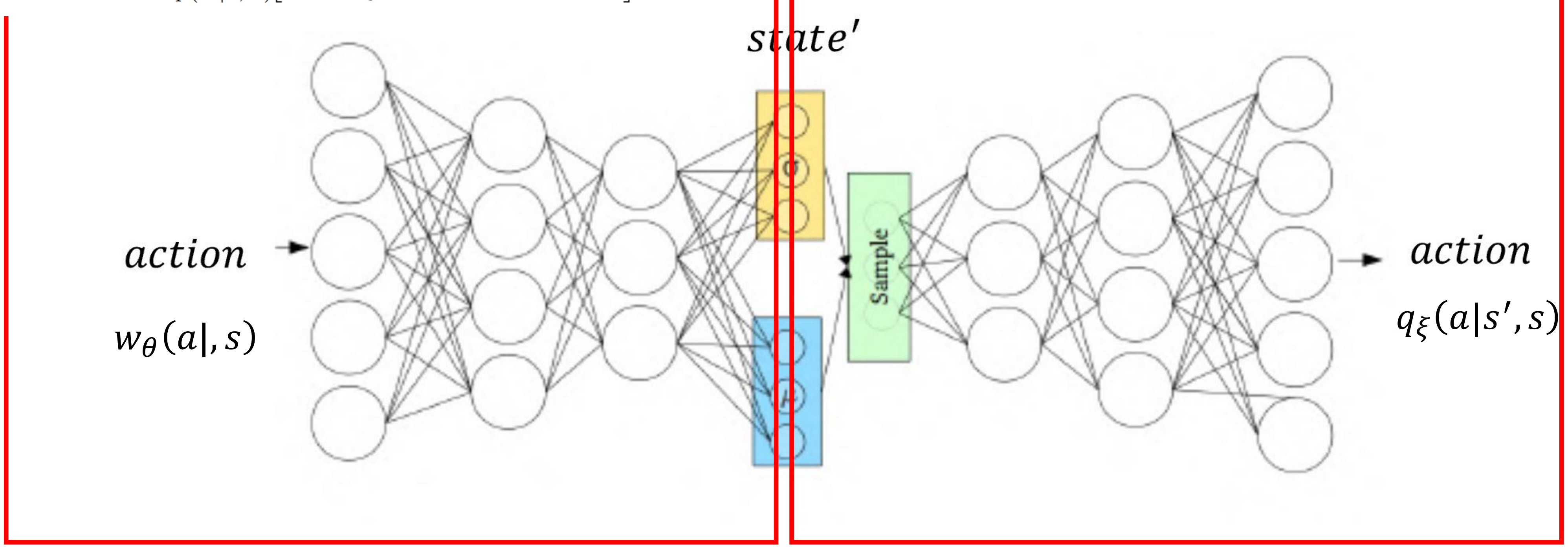
$$u(s,a) = E_{p(s^{'}|s,a)}[ln\ q_\xi(a|s',s)]$$

$$L(h_\theta,\psi_\theta) = \mathbb{E}_{p(s'|s,A)}\big[(\beta \ln q_\xi(\mathbf{a}|\mathbf{s},\mathbf{s}') - r_\theta(\mathbf{s},\mathbf{a}))^2\big]\ .$$

Decoder: MLE

$$q_\xi(\mathbf{a}|\mathbf{s}',\mathbf{s}) = q(a_1|\mathbf{s},\mathbf{s}') \prod_{k=2}^{K} q(a_k|f_\xi(a_{k-1},\mathbf{s},\mathbf{s}')),$$

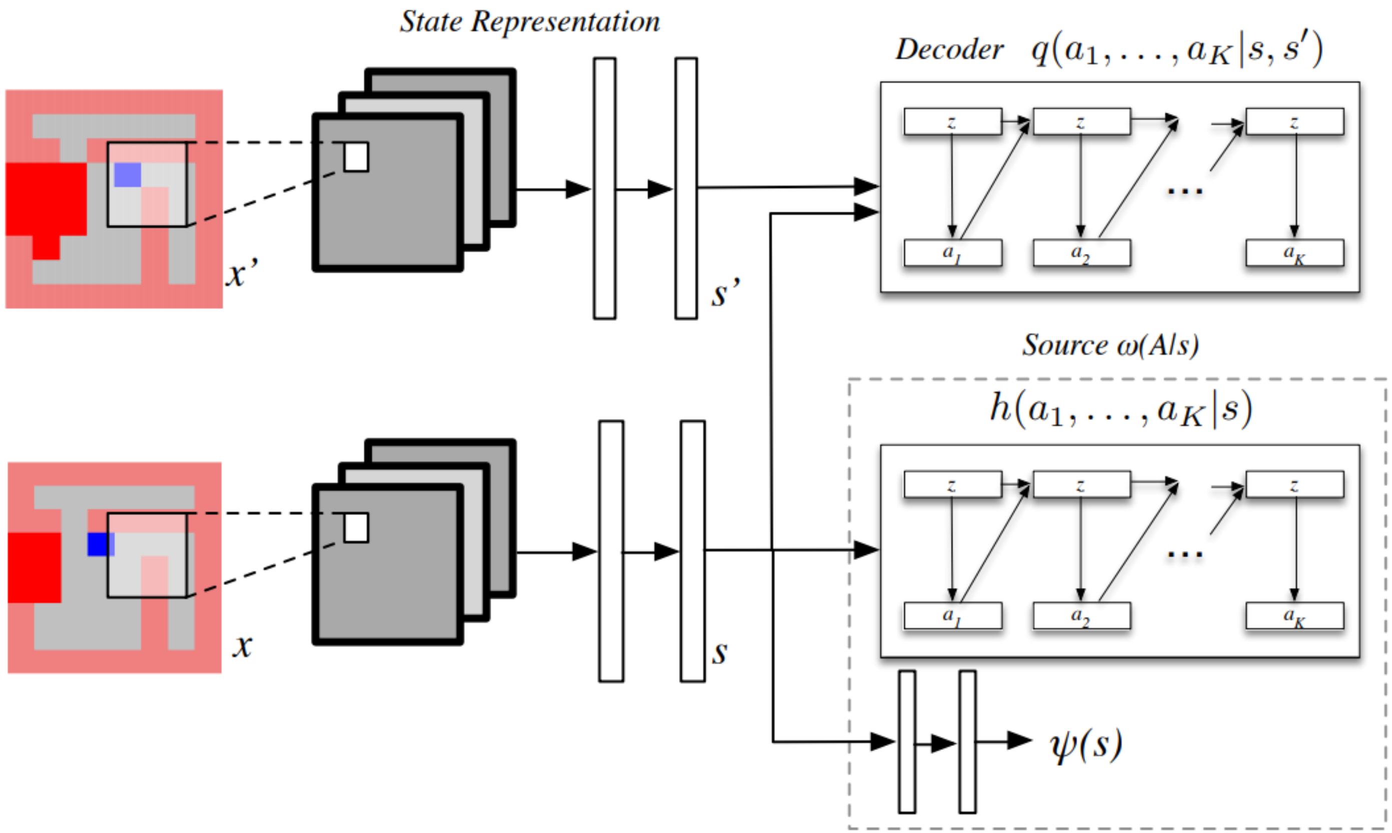
$p(s^{'}|a,s)$



Method

Algorithm 1: Stochastic Variational Information Maximisation for Empowerment

Parameters: ξ variational, λ convolutional, θ source
while not converged **do**
  $\mathbf{x} \leftarrow$ {Read current state}
  $\mathbf{s} = \text{ConvNet}_\lambda(\mathbf{x})$ {Compute state repr.}
  $A \sim \omega(\mathbf{a}|\mathbf{s})$ {Draw action sequence.}
  Obtain data $(\mathbf{x}, \mathbf{a}, \mathbf{x}')$ {Acting in env. }
  $\mathbf{s}' = \text{ConvNet}_\lambda(\mathbf{x}')$ {Compute state repr.}
  $\Delta\xi \propto \nabla_\xi \log q_\xi(\mathbf{a}|\mathbf{s}, \mathbf{s}')$ (18)
  $\Delta\theta \propto \nabla_\theta L(h_\theta, \psi_\theta)$ (8)
  $\Delta\lambda \propto \nabla_\lambda \log q_\xi(\mathbf{a}|\mathbf{s}, \mathbf{s}') + \nabla_\lambda L(h_\theta, \psi_\theta)$
end while
$\mathcal{E}(\mathbf{s}) = \frac{1}{\beta}\psi_\theta(\mathbf{s})$ {Empowerment}



Experiments

Effectiveness of the MI Bound

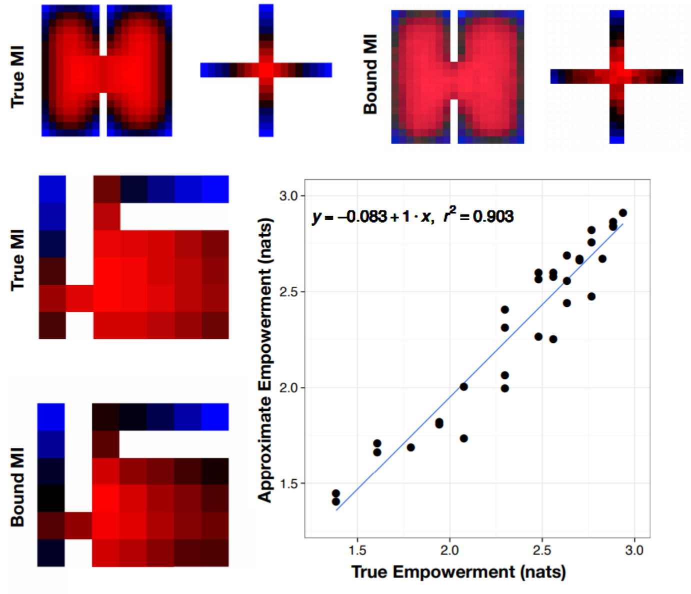


Figure 3: Comparing exact vs approximate empowerment. Heat maps: empowerment in 3 environments: two rooms, cross room, two-rooms; Scatter plot: agreement for two-rooms.

Dynamic Environments

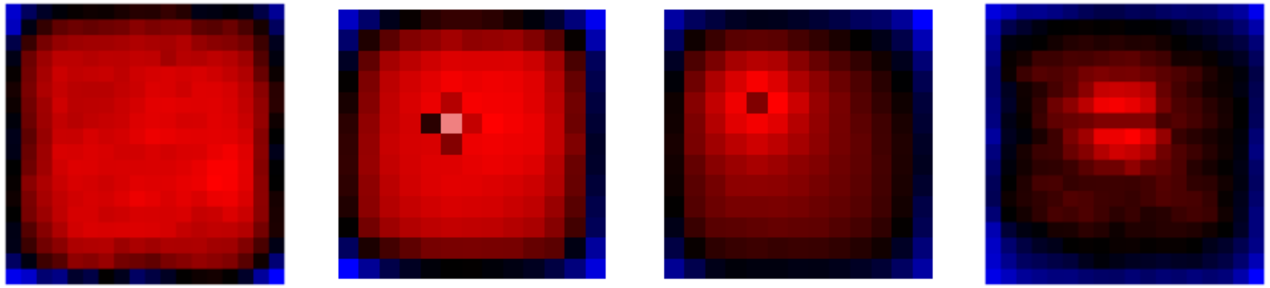


Figure 4: Empowerment for a room environment, showing a) an empty room, b) room with an obstacle c) room with a moveable box, d) room with row of moveable boxes.

Experiments

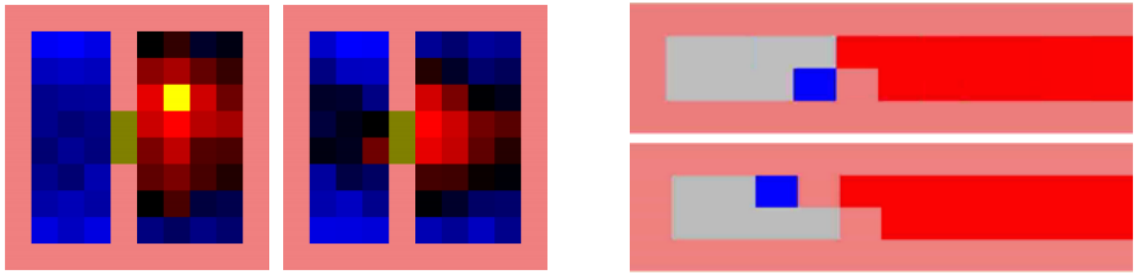


Figure 5: Left: empowerment landscape for agent and key scenario. Yellow is the key and green is the door. Right: Agent in a corridor with flowing lava. The agent places a bricks to stem the flow of lava.

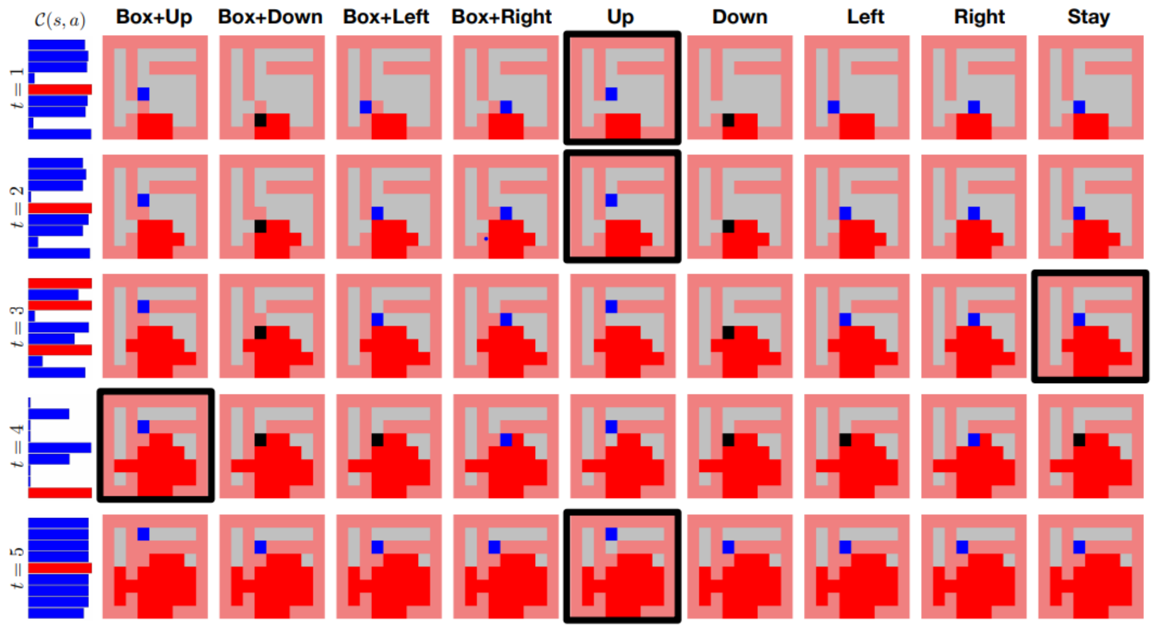


Figure 6: Empowerment planning in a lava-filled maze environment. Black panels show the path taken by the agent.