A Bandit Framework for Optimal Selection of Reinforcement Learning Agents

A. Merentitis[1(*)] K. Rasul[2], R. Vollgraf[2], A. S. Sheikh[2], U. Bergmann[2]

MOTIVATION

- The optimal inductive bias (architecture, hyperparameters, etc.) of a RL agent depends on the application.
- We propose a multi-arm bandit with the double objective of maximizing the reward while the agents are learning and selecting the best agent after a small number of learning steps.

MULTI-ARM BANDIT FRAMEWORK

Different types of bandit strategies are possible, from simple ϵ-greedy, to techniques like SoftMax (probability matching bandit), UCB1 (based on the optimism in the face of uncertainty principle) and EXP3 (adversarial bandit).

It has been shown in [Yi-Sun-2011] that it is beneficial for agents to take actions that maximize the reduction in uncertainty about the environment dynamics. This can be formalized as taking a sequence of actions a_t that maximize the sum of reductions in entropy. With the history of the agents up until time step t as $\xi_t = \{s_1, a_1, \ldots, s_t\}$, we can write the sum of entropy reductions as:

$$\sum_t \left(H\left(\Theta|\xi_t, a_t\right) - H\left(\Theta|S_{t+1}, \xi_t, a_t\right)\right). \tag{1}$$

As indicated in [Houthooft-2016], according to information theory, the individual terms express the mutual information between the next state distribution S_{t+1} and the model parameter distribution Θ, namely $I(S_{t+1}; \Theta|\xi_t, a_t)$. This mutual information can be written as:

$$I\left(S_{t+1}; \Theta|\xi_t, a_t\right) = \mathbb{E}_{s_{t+1}\sim P(\cdot|\xi_t, a_t)}[D_{KL}[p\left(\theta|\xi_t, a_t, s_{t+1}\right)||p\left(\theta|\xi_t\right)]], \tag{2}$$

The KL divergence term is expressing the difference between the new and the old beliefs of the agent regarding the environment dynamics, and the expectation is with respect. Under these assumptions the above formulation can also be interpreted as information gain [Houthooft-2016] according to the Variational Information Maximizing Exploration (VIME) exploration strategy.

CALCULATING THE INFORMATION GAIN

Formally, since calculating the posterior $p(\theta|D)$ for a dataset D is not feasible, we follow VIME and approximate it through an alternative distribution $q(\theta; \phi)$, parametrized by ϕ. In this setting we seek to minimize D_{KL} through maximization of the variational lower bound $L[q(\theta; \phi), D]$. The latter is formulated as:

$$L[q(\theta; \phi), D] = \mathbb{E}_{\theta\sim q(\cdot;\phi)}[\log p(D|\theta)] - D_{KL}[q(\theta; \phi)||p(\theta)]. \tag{3}$$

The information gain term can then can be expressed as:

$$I(S_{t+1}; \Theta|\xi_t, a_t) = D_{KL}[q(\theta; \phi_{t+1})||q(\theta; \phi)], \tag{4}$$

where ϕ_{t+1} represents the updated and ϕ_t the old parameters of the agent's belief regarding the environment dynamics.

RELATION BETWEEN INFORMATION GAIN AND TRUE ENVIRONMENT REWARDS

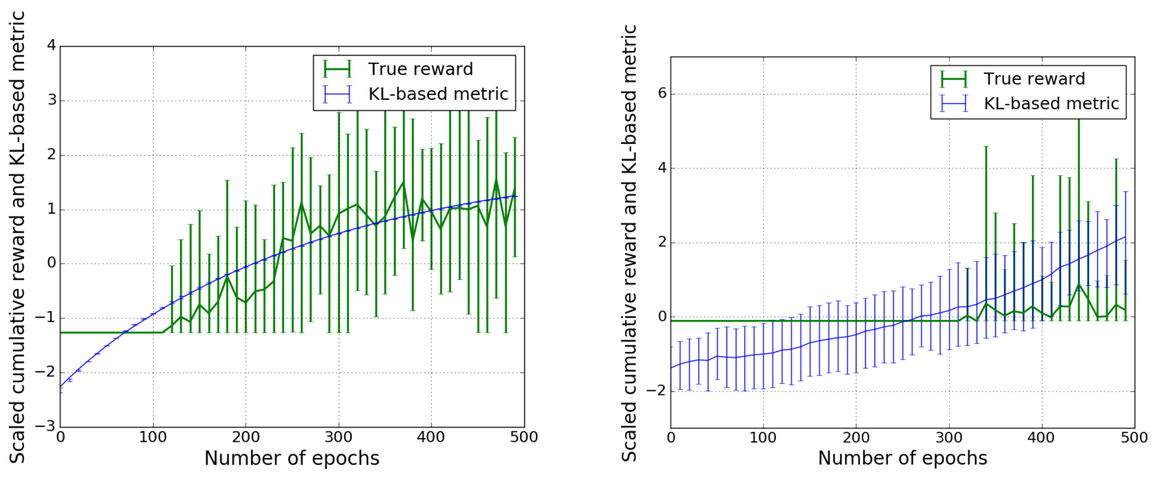


Figure 1: *Correlation between environment and information gain rewards for a good (left) and a suboptimal agent (right) for Mountain Car environment.*

WHAT WE GAIN WITH SURROGATE REWARDS

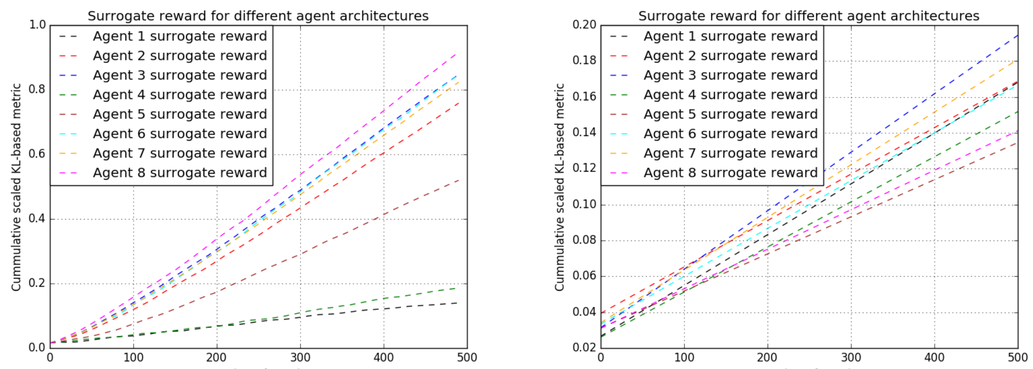
- To alleviate the problem of sparse rewards, the reinforcement learning agents are augmented with surrogate rewards.
- This helps the bandit framework to select the best agents early, since these rewards are smoother and less sparse than the environment reward.

- *1: OLX Berlin Hub*
- *2: Zalando Research*

(*) Most of this work was done while the author worked at Zalando Research

CUMULATIVE SURROGATE AND TRUE ENVIRONMENT REWARDS

Cumulative surrogate reward for Cartpole and Mountain Car envs.



Cumulative true reward for Cartpole and Mountain Car envs.

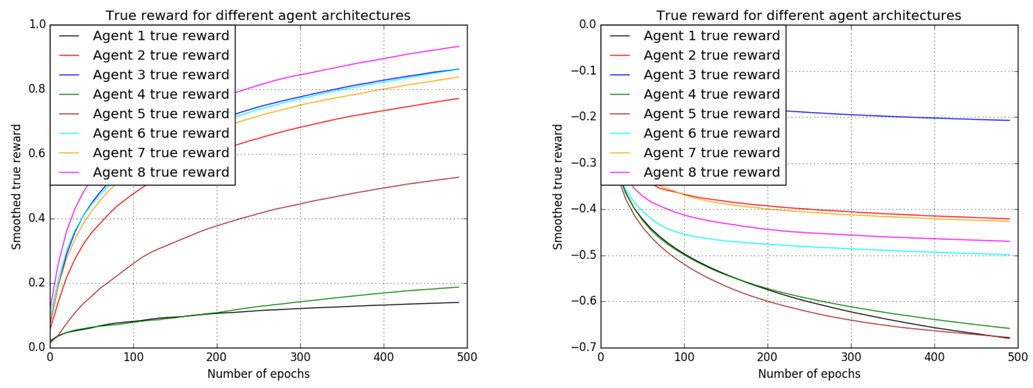


Figure 2: Comparison of the scaled cumulative surrogate and true rewards for the Cartpole and Mountain Car environments for different reinforcement learning agents. The relative order of the agents is similar between the first and second rows, indicating that the surrogate reward can be used to augment the true reward, especially early on when the latter is sparse and noisy.

FREQUENCY OF SELECTIONS OF THE DIFFERENT AGENTS

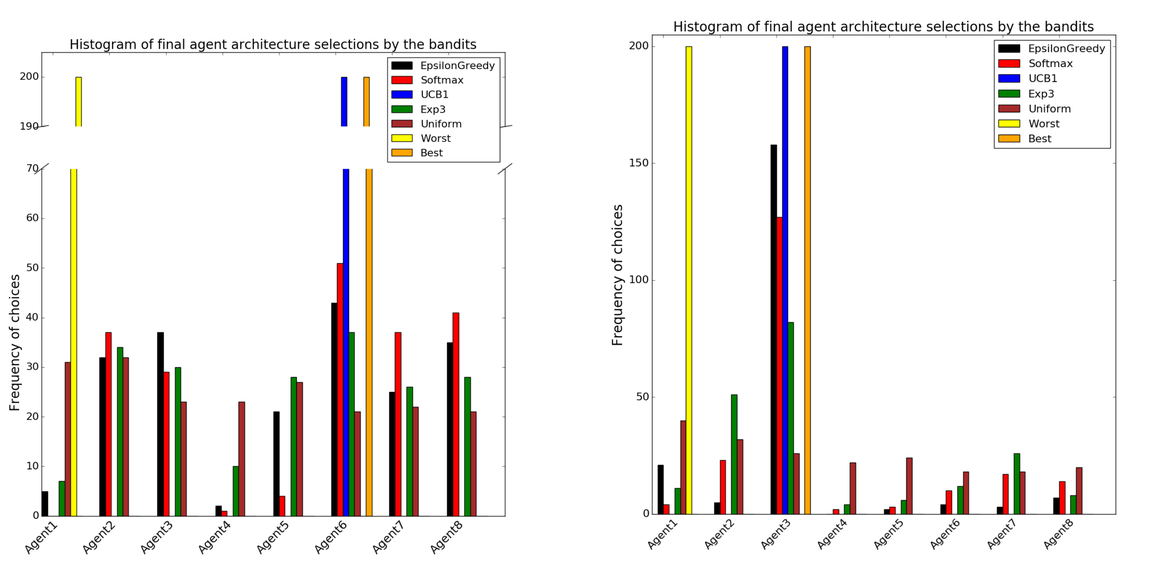


Figure 3: *Frequency of selections of the different agents for Lunar Lander (left) and Mountain Car (right) at the end of training for the different bandit algorithms. The UCB algorithm matches the oracle (Best) in selecting the best agent after training is complete.*

SCORE FOR DQN, GORILA, PRO HUMAN GAMER, AND THE AGENT SELECTED FROM THE BANDIT

Games	DQN Score	Gorila Score	Human Pro Score	Best Agent Score
Atlantis	85641 ± 17600	100069.16	29028	217810 ± 7256.4
Kung-Fu	23270 ± 5955	27543.33	22736	29860 ± 6793.1
Ms. Pacman	2311 ± 525	3233.50	15693	5708.0 ± 860.1
Seaquest	5286 ± 1310	13169.06	20182	17214 ± 2411.5
Sp. Invaders	1976 ± 893	1883.41	1652	3697.5 ± 2876.1
Zaxxon	4977 ± 1235	7129.33	9173	30610 ± 8169.0

CONCLUSIONS

- Introduced a bandit framework that offers a principled way of selecting between different RL agent architectures.
- Maximizing rewards during the learning process.
- Reliably select the best agent (in future expected rewards).
- Composite surrogate reward captures the certainty of the agents regarding the environment dynamics, for a given amount of environment interactions.
- Experimental results show that the bandit outperforms both a single non-optimal agent and uniform alternation between the agents.

- *"Planning to Be Surprised: Optimal Bayesian Exploration in Dynamic Environments", Yi-Sun et al. 2011*
- *"Curiosity-driven Exploration in Deep Reinforcement Learning via Bayesian Neural Networks", Houthooft et al. 2016*