

---

# Backpropagation-Free Parallel Deep Reinforcement Learning

---

**William H. Guss**

Machine Learning at Berkeley  
2650 Durant Ave, Berkeley CA, 94720  
wguss@ml.berkeley.edu

**Mike Zhong**

Machine Learning at Berkeley  
Berkeley CA, 94720  
lol@gmail.com

**Utkarsh S**

Machine Learning at Berkeley  
Berkeley CA, 94720  
philkuz@ml.berkeley.edu

**Max Johansen**

Machine Learning at Berkeley  
Berkeley CA, 94720  
max@ml.berkeley.edu

## Abstract

In this paper we conjecture that an agent, environment pair  $(\pi, E)$  trained using DDPG with an actor network  $\mu$  and critic network  $Q^\pi$  can be decomposed into a number of sub-agent, sub-environment pairs  $(\pi_n, E_n)$  ranging over every neuron in  $\mu$ ; that is, we show empirically that treating each neuron  $n$  as an agent  $\pi_n : \mathbb{R}^n \rightarrow \mathbb{R}$  of its inputs and optimizing a value function  $Q^{\pi_n}$  with respect to the weights of  $\pi_n$  is dual to optimizing  $Q^\pi$  with respect to the weights of  $\mu$ . Finally we propose a learning rule which simultaneously optimizes each  $\pi_n$  without error backpropagation achieving state of the art performance and speed across a variety of OpenAI Gym environments.

## Todo list

Introduction to DDPG and recent advances in deep RL. . . . .	2
Biological diffusion of dopamine in the brain $\implies$ error backpropagation is not biologically feasible. . . . .	2
Synthetic gradients are a step in the right direction, but still require eventual back propogation. . . . .	2
Therefore it is feasible that each neuron is maximizing the expectation on his future dopamine intake, and so we propose the following theorem. . . . .	2
A high level description of the section. . . . .	2
Insert reference and make this a footnote . . . . .	3
Cite lilicrap . . . . .	3
Make DDPG figure . . . . .	3
Cite deepmind . . . . .	3
Build up definitions required . . . . .	3
Write conjecture on decomposition which is free of neural configuration. Subject to change in later versions of ArXiv paper . . . . .	3
Emperical justification of the iff using the following experiment (s). . . . .	3

1. Training a network on Atari using DDPG and plotting average critic functions for neurons using window. . . . .	3
2. Possibly others. . . . .	3
Therefore we propose the following learning rule in aims to evidence the reverse, training $\mu$ using simultaneous optimization on all $Q_n$ w.r.t $\pi_n$ 's weights. . . . .	3
Proposal of the rule. Linear approximation of the $Q$ function for every neuron is good enough, (experimentally). . . . .	3
Implications of the rule to DDPG . . . . .	3
Implications of the rule to entirely recurrent networks (infinite time horizon and NO unrolling since the environment the local actions of the neuron which globally recur to that neuron again are <i>encoded</i> into $Q_n$ ; large time horizon probably implies that better regressor needed for $Q_n$ .) . . . . .	3
Parallelism, no error backprop, and only 2x operations, but no locking on GPU, so all can be run simultaneously if we cache! . . . . .	3
To validate the new learning rule we throw a fuck ton of experiments together on the following list (or better using OpenAI Gym). . . . .	4
1. Show that training decentralized policy gradient $\implies$ total policy optimization . . . . .	4
2. Show speed improvements on update step through parallelism (samples per second vs DDPG). . . . .	4
3. Show results are comparable with the state of the art. . . . .	4
We wrecked deep reinforcement learning using biological inspiration. . . . .	4
Would like to try the method with full recurrent networks and purely asynchronous implementation of leaky integration networks. . . . .	4
Would like to prove the conjecture. List possible methods of proof. . . . .	4

## 1 Introduction

Introduction to DDPG and recent advances in deep RL.

Biological diffusion of dopamine in the brain  $\implies$  error backpropagation is not biologically feasible.

Synthetic gradients are a step in the right direction, but still require eventual back propagation.

Therefore it is feasible that each neuron is maximizing the expectation on his future dopamine intake, and so we propose the following theorem.

## 2 Agent-Environment Value Decomposition

A high level description of the section.

### 2.1 Background

Recall the standard reinforcement learning setup. We say  $E$  is an *environment* if  $E \stackrel{\text{def}}{=} (\mathcal{S}, \mathcal{A}, \mathcal{R}, T, r)$  where  $T$  describes transition probability measure  $T(s_{t+1} | s_t, a_t)$  and  $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{R}$  is a reward function. Furthermore  $\mathcal{S}$ ,  $\mathcal{A}$ ,  $\mathcal{R}$  are the *state space*, *action space*, and *reward space* respectively. We restrict  $\mathcal{R}$  to a compact subset of  $\mathbb{R}$  and action space and state space to finite di-

mensional real vector spaces. As in DDPG we assume that the environment  $E$  is *fully observed*; that is, at any time step the state  $s_t$  is fully described by the observation presented,  $x_t$ , and not by the history  $(x_1, a_1, \dots, a_{t-1})$ .

We define the policy for an agent to be  $\pi : \mathcal{P}(\mathcal{A}) \times \mathcal{S} \rightarrow [0, 1]$ . In general the policy is a probability measure on some  $\sigma$ -algebra  $\mathcal{M} \subset \mathcal{P}(\mathcal{A})$  conditioned on  $\mathcal{S}$  so that  $\pi(\mathcal{A} \mid s \in \mathcal{S}) = 1$ . However, we will deal only with *deterministic* policies where for every  $s_t$  there is unique  $a_t$  so that  $\pi(\{a_t\} \mid s = s_t) = 1$  and the measure is 0 otherwise. Thus we will abuse notation and define a *deterministic agent* by a policy function  $\pi : \mathcal{S} \rightarrow \mathcal{A}$ .

For a policy  $\pi$  the action-value function is the expected future reward under  $\pi$  by performing  $a_t$  at state  $s_t$  using the Bellman equation

$$Q^\pi(s_t, a_t) = \mathbb{E}_{s_{t+1} \sim E} [r(s_t, a_t) + \gamma Q^\pi(s_{t+1}, \pi(s_{t+1}))] \quad (2.1.1)$$

with  $\gamma \in (0, 1)$  a discount factor, and the second expectation removed because  $\pi$  is deterministic. [Some survey] provides an extensive exposition into a justification of this equation and choice for the action-value of  $\pi$ , so we will assume such a choice is a valid measure of performance.

In deterministic policy gradient methods, we define an actor  $\mu : \mathcal{S} \rightarrow \mathcal{A}$  and a critic  $Q^\mu : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  and optimize  $Q^\mu(s_t, \mu(s_t))$  with respect to the parameters  $\theta^\mu$  of  $\mu$ . This method is provably the true policy gradient of  $\mu$  if  $Q^\mu$  is known. Recently (DDPG) utilizes the universality of DNNs in order to approximate both  $\mu$  and  $Q^\mu$  along with delayed weight-transfer networks to stabilize learning and prevent divergence as depicted in Figure 1. In order to decompose the action-value function we will make heavy use of this methodology at a scale local to each neuron in the flavor of (Synthetic gradients.)

## 2.2 Towards Neurocomputational Decomposition of $Q^\mu$

Build up definitions required

Write conjecture on decomposition which is free of neural configuration. Subject to change in later versions of ArXiv paper

Empirical justification of the iff using the following experiment (s).

1. Training a network on Atari using DDPG and plotting average critic functions for neurons using window.

2. Possibly others.

Therefore we propose the following learning rule in aims to evidence the reverse, training  $\mu$  using simultaneous optimization on all  $Q_n$  w.r.t  $\pi_n$ 's weights.

## 3 Decentralized Deep Deterministic Policy Gradient

Proposal of the rule. Linear approximation of the  $Q$  function for every neuron is good enough, (experimentally).

Implications of the rule to DDPG

Implications of the rule to entirely recurrent networks (infinite time horizon and NO unrolling since the environment the local actions of the neuron which globally recur to that neuron again are *encoded* into  $Q_n$ ; large time horizon probably implies that better regressor needed for  $Q_n$ .)

Insert reference and make this a footnote

Cite lil-i-crap

Make DDPG figure

Cite deep-mind

Parallelism, no error backprop, and only 2x operations, but no locking on GPU, so all can be run simultaneously if we cache!

## 4 Results

To validate the new learning rule we throw a fuck ton of experiments together on the following list (or better using OpenAI Gym).

```
blockworld1 1.156 1.511 0.466 1.299 -0.080 1.260
blockworld3da 0.340 0.705 0.889 2.225 -0.139 0.658
canada 0.303 1.735 0.176 0.688 0.125 1.157
canada2d 0.400 0.978 -0.285 0.119 -0.045 0.701
cart 0.938 1.336 1.096 1.258 0.343 1.216
cartpole 0.844 1.115 0.482 1.138 0.244 0.755
cartpoleBalance 0.951 1.000 0.335 0.996 -0.468 0.528
cartpoleParallelDouble 0.549 0.900 0.188 0.323 0.197 0.572
cartpoleSerialDouble 0.272 0.719 0.195 0.642 0.143 0.701
cartpoleSerialTriple 0.736 0.946 0.412 0.427 0.583 0.942
cheetah 0.903 1.206 0.457 0.792 -0.008 0.425
fixedReacher 0.849 1.021 0.693 0.981 0.259 0.927
fixedReacherDouble 0.924 0.996 0.872 0.943 0.290 0.995
fixedReacherSingle 0.954 1.000 0.827 0.995 0.620 0.999
gripper 0.655 0.972 0.406 0.790 0.461 0.816
gripperRandom 0.618 0.937 0.082 0.791 0.557 0.808
hardCheetah 1.311 1.990 1.204 1.431 -0.031 1.411
hopper 0.676 0.936 0.112 0.924 0.078 0.917
hyq 0.416 0.722 0.234 0.672 0.198 0.618
movingGripper 0.474 0.936 0.480 0.644 0.416 0.805
pendulum 0.946 1.021 0.663 1.055 0.099 0.951
reacher 0.720 0.987 0.194 0.878 0.231 0.953
reacher3daFixedTarget 0.585 0.943 0.453 0.922 0.204 0.631
reacher3daRandomTarget 0.467 0.739 0.374 0.735 -0.046 0.158
reacherSingle 0.981 1.102 1.000 1.083 1.010 1.083
walker2d 0.705 1.573 0.944 1.476 0.393 1.397
```

1. Show that training decentralized policy gradient  $\implies$  total policy optimization

2. Show speed improvements on update step through parallelism (samples per second vs DDPG).

3. Show results are comparable with the state of the art.

## 5 Conclusion

We wrecked deep reinforcement learning using biological inspiration.

### 5.1 Future Work

Would like to try the method with full recurrent networks and purely asynchronous implementation of leaky integration networks.

Would like to prove the conjecture. List possible methods of proof.