Learning algorithm

The training algorithm employed is similar to that used in the paper *Continuous Control with Deep Reinforcement Learning* by Lillicrap, et al. This approach, known as the Deep Deterministic Policy Gradient algorithm (DDPG), borrows ideas from policy gradient, actor-critic and Deep Q-Learning Network (DQN) methods.

Policy gradient methods use models like deep neural networks to directly find the optimal policy by maximizing the expected return given a state using gradient ascent. DPG differs in that actions resulting from such models are deterministic rather than stochastic – the network just outputs an action, not the probability of taking an action. This has two benefits. First, in theory, the resulting policy gradient should require less samples to estimate as one of the elements contributing to the variance, namely the action value, is thrown out. We are just left with the variance from the state. Second, a DPG network can output continuous values, unlike policy gradient methods which just output the probability of a discrete action. A potential downfall of course is that we don't explore the action space. To remedy this, the DPG is made off-policy (i.e. we select an action, but hold off on applying back-propagation to the parameters of the network) with random noise added to the action to allow for exploration. As pointed out in the forum for project 2, the addition of noise to actions had to vary by agent and perhaps more importantly for convergence, had to lessen over time to minimize exploration closer to convergence.

Deep DPG may be viewed as an actor-critic RL method in that the DPG network acts as an "actor" network to predict an action using the current policy. The gradient estimate used to update that network uses another "critic" network which is used to estimate expected return. In theory, doing this at every time step instead of using the actual return at the end of episode allows the algorithm to potentially learn if specific steps were good or bad for the (estimated) return.

Actor-critic methods use the advantage function rather than q-values because the latter lead to a lot of instability. To improve stability, I wait until one of the agents reaches the end of an episode before updating the networks. I also get better results by collecting the experience from both agents but use one replay buffer, and the same actor and critic networks as suggested in the benchmark implementation. Finally, like almost any network used in non-trivial deep learning problems, the stability of the critic network improved when gradient clipping was applied along with batch normalization.

Finally, DDPG borrows ideas from DQNs to stabilize approximating the above action and q-values using deep neural networks. In particular, DDPG uses a replay buffer to break the correlation between between time steps so networks can actually learn, and has separate target networks (one each for the actor and critic networks) to stabilize learning. The parameters of these separate target networks are updated using a soft update – i.e. a small percentage of these are updated every time the original local networks get updated rather than at set number of time steps as in the original DQN method.

The specific version of DDPG used in this project comprises three components: a set of 2 agents (the tennis "players") to explore the environment, a single model containing the actor and critic networks, and the DDPG training algorithm itself.

The full training algorithm is described in Algorithm 1 and 1a below.

Algorithm 1 Training procedure for Tennis DDPG

Require:
The environment *env*
The *agent*

- 2 sub-agents interacting with the environment
- Initialized with its local actor network $\mu(s|\theta^{\mu})$ and target network $\mu'(s|\theta^{\mu'})$ - initialize these to have the same weights
- Initialized with its local critic network $Q(s,a|\theta^{Q})$and target critic network $Q'(s,a|\theta^{Q'})$ - initialize these to have the same weights

update_size – number of updates to networks' parameters per agent
The maximum number of episodes *M*
The maximum number of time steps per episode *T*
Initialize:
Experience replay buffer *R* to capacity *n*
Steps:
for episode = 1, M do
  state $s_{t=1,1..2}$ ← reset env
  Initialize random process $\mathcal{N}$ for exploration
  for t = 1, T **do**
    Calculate dampening factor df ← square root of the episode inverted
    Select action for each sub-agent i such that $a_{t,i} \leftarrow \mu(s_{t,i}|\theta^{\mu}) + df * \mathcal{N}_{t,i}$
    For both sub-agents, execute actions; observe rewards r_t, new states s_{t+1} and done flags
    For each sub-agent i, store transition ($s_{t,i}$, $a_{t,i}$, $r_{t,i}$, $s_{t+1,i}$, $done\_flag_i$) in R
    break if either agent has the done flag set
  end for
  Update Actor-Critic Local and Target Networks (Algorithm 1a)
  end if
end for

Algorithm 1a Update Actor-Critic Local and Target Networks

Require:
Local actor network $\mu(s|\theta^{\mu})$ and target network $\mu'(s|\theta^{\mu'})$ – see Algorithm 1
Local critic network $Q(s,a|\theta^{Q})$and target critic network $Q'(s,a|\theta^{Q'})$ – see Algorithm 1

update_size – see Algorithm 1
R – replay buffer, see Algorithm 1
γ - discount factor
τ - for soft update of target parameters
Steps:

n_updates ← number of agents * update_size
for j = 1, n_updates **do**
Sample a random minibatch of N transitions (s_j, a_j, r_j, s_{j+1}) from R
Set $y_j = r_j + \gamma Q'(s_{j+1}, \mu'(s_{j+1}|\theta^{\mu'})|\theta^{Q'})$
Update critic by minimizing the loss: $L = \frac{1}{N}\sum_j (y_j - Q(s_j, a_j|\theta^Q)^2$
Update the actor policy by using the sampled gradient:

$$\nabla_{\theta^\mu} J \approx \frac{1}{N}\sum_j \nabla_a Q(s, a|\theta^Q)|_{s=s_j, a=\mu(s_j)} \nabla_{\theta^\mu}\mu(s|\theta^\mu)|_{s_j}$$

Update the target networks:

$$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}$$
$$\theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}$$

end for

Model

Actor: Each model (target and local network) is simply a 3-layer fully connected network. The input layer takes the state of the Tennis environment. The two hidden layers have ReLU activations. The output layer uses the tanh activation for the action value. Batch normalization is used in between layers.

Critic: Each model (target and local network) is also a 3-layer fully connected network. The input layer takes the state of the Tennis environment and uses a ReLU activation. The second layer concatenates the output from the first layer and the action from the actor network; it too uses a ReLU activation. The output layer only does a linear transformation of the second layer's output to generate the q-value.

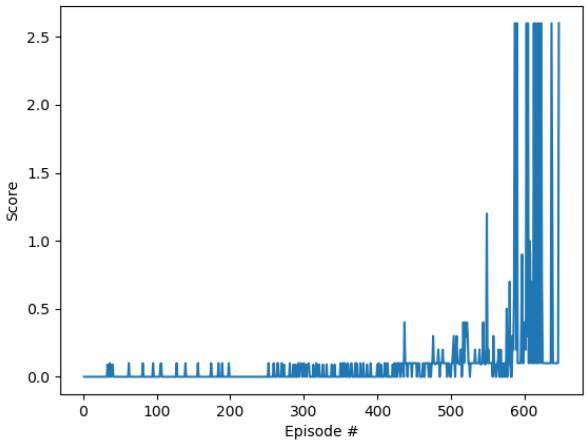
Hyperparameters

Name	Description	Value
n	experience replay buffer size	100000
N	SGD batch size	512
γ	discount factor	0.99
τ	for soft update of target parameters	0.001
lr_actor	actor network learning rate	0.001

lr_critic	critic network learning rate	0.001
M	maximum number of episodes	1000
T	maximum number of time steps per episode	1000
update_size	# of updates of networks per sub-agent	20
actor_fc1_units	# of units in actor's hidden layer 1	400
actor_fc2_units	# of units in actor's hidden layer 2	300
critic_fc1_units	# of units in critic's hidden layer 1	400
critic_fc2_units	# of units in critic's hidden layer 2	300
weight_decay	L2 weight decay	0
seed	seed to set random number generator for reproducible results	2

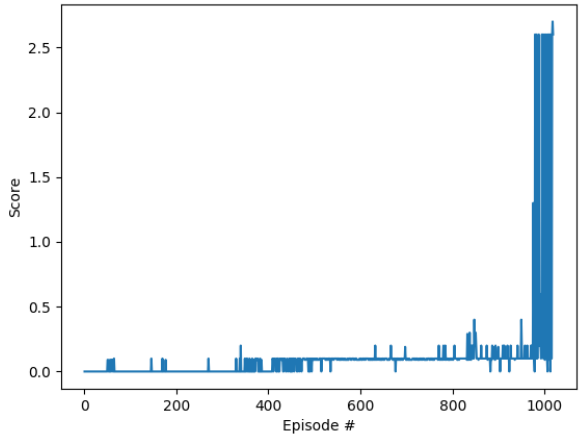
Results

Below is a plot of the average score (i.e. the sum of the rewards) per episode over a 100 episodes. Each individual score was the maximum of the two sub-agents.



The number of episodes needed to achieve a score greater than 0.5 (0.51) is 647 episodes using the hyper-parameters defined above.

The above results were obtained on MacOSX. Curiously, if I run the same algorithm on Ubuntu 16.04, the agents take longer to train. The number of episodes needed to achieve a score greater than 0.5 (0.51) becomes 1018 episodes using the hyper-parameters defined above.



I basically used the same algorithm for this project as I did for Project 2 (continuous control) and did not see this same behavior. That is, if I trained on one type of machine, I got the same results as I did on another.

Unlike the last project, however, batch normalization on the actor network was critical to stabilizing learning. Or at least I gave up after 3000 episodes without it.

Ideas for Future Work

Adding prioritized experience replay would likely help. And sampling for transitions just considers one transition, and not for N-step returns, which would also likely help.

Network stability was a serious problem for this project. Like project 2, I would also perhaps just switch to Trust Region Policy Optimization (TRPO). TRPO prevents large changes to the policy network's parameters by enforcing a KL divergence constraint on the size of the size of updates at each iteration. Another option is to use Truncated Natural Policy Gradient (TNPG) which effectively does the same thing (minimize gradient step size) using the Fisher Information Matrix.

I noticed that I got better results depending on *how often the initial Unity Tennis environment reset was called before training*. For example, if I called reset once before training, then I got the results above. If I called it twice, for example if I decided to print out statistics on the action and state sizes and then just call it again (inadvertently), training took twice as long! Clearly, the initial conditions are critical which suggests that I am not exploring the parameter space broadly enough to start. I would have to add a lot more noise to allow for more exploration, perhaps even switching to using parameter noise rather than action noise. Of course, this would impact the rate at which noise abates along perhaps with the learning rate.