Paper Review - *"Deterministic Policy Gradient Algorithms, Silver et al, 2014"*

Summary: The authors introduce a new algorithm DPG and argue that while the traditional RL algorithms like Q-learning and SARSA work well for discrete action spaces, they struggle to work for continuous action spaces which are more common in the real world. The main idea behind DPG is to learn a deterministic policy rather than a stochastic policy that maps states to the action. The authors show that it outperforms the stochastic counterparts in high-dimensional action spaces

Main Contributions: At a high level the idea proposed is based on the actor-critic method. This method has two components, the actor and the critic network. The actor network takes the current state as input and gives the deterministic actions as the output. This action is fed to the critic network along with the current state and outputs the expected rewards of the state-action pair. They used a deterministic policy gradient update rule where the actor-network is trained to maximize the expected rewards by the critic network. This means that the gradient of the policy is updated in the direction of the gradient of the expected reward.
The author tried both the on-policy deterministic and off-policy deterministic methods like COPDAC-B and COPDAC-Q to compare against their stochastic counterparts.

Strengths: The main advantage of DPG is their ability to work well in continuous and high-dimensional action spaces much better than the stochastic counterparts, which is also the main theme of the paper. By doing this, they can reduce computational costs as these methods converge better and faster. This also allows the gradients to be estimated more efficiently avoiding the problematic integral over the action spaces in the stochastic methods. The paper also shows that the optimal policy being learned is anyways deterministic under certain conditions.

Weakness: Due to the deterministic nature of the algorithm it may suffer from the lack of exploration due to the agent-only proceeding in the direction of the gradient of the expected reward and may get stuck in sub-optimal minima. Also, the authors only evaluated the work on MDPs and we don't know if they work for POMDPs. These algorithms are also more computationally expensive than simpler algorithms like Q-Learning.

Experiments: The authors evaluated their deterministic policy models against their stochastic counterparts to show that DPG algorithms are better and they estimate the gradients more efficiently. They show this in different problems namely Continuous Bandit with a high-dimensional quadratic cost function, second the continuous-action variants of the standard RL benchmark like Mountain Car, Pendulum, and 2D Puddle world. And at last, they evaluated on octopus arm. All these experiments showed that the DPG algorithms (COPDAC-B and COPDAC-Q) either performed better or at par with the stochastic algorithms. Another inference from the graphs which is not mentioned in the paper is that these algorithms converged faster than stochastic methods.

Extensions: The paper currently did not demonstrate scaling the approach to high-dimensional observation spaces so we can think of applying some techniques to explore that.
We can also analyze their work beyond the continuous control task evaluated in the paper such as robotics, autonomous driving, etc.