When multiple agents are learning the learning step of each of the agents depends on the policy of all other agents. “Learning with Opponent-Learning Awareness” (Foerster et al 2017) explicitly accounts for this dependency and differentiates through the learning step: the original paper relied on a Taylor expansion, second order gradients and a large batch size (4000). LOLA can be implemented exactly using DiCE:

**Motivation**

- The score function is commonly used in order to estimate gradients in Reinforcement Learning problems.
- This can be calculated via the Surrogate Loss approach.
- For order derivatives there is currently no satisfactory solution.
- Examples include Meta-Learning for RL and 2nd order optimisation.

**Setting and background**

- Stochastic Computation Graphs provide a framework for estimating gradients using Surrogate Losses (John Schulman et al 2015):

**Method: DiCE**

Where:

\[
L_{\text{DiCE}} = \sum_{c \in C} C(W_c) c
\]

1. \(C(W) \rightarrow 1,\)

2. \(\nabla \phi C(W) = \sum_{w \in W} \nabla \phi \log(p(w; \theta))\)

- Meets our requirements!

\[
\mathbb{E}[\nabla \phi L_{\text{DiCE}}] \rightarrow \nabla \phi L, \forall n \in \{0, 1, 2, \ldots \}
\]

**Implementation**

- DiCE can be implemented easily in standard deep learning libraries:

**Application Example: LOLA-DiCE**

- The original paper relied on a Taylor expansion, second order gradients and a large batch size (4000). LOLA can be implemented exactly using DiCE:

\[
L_{\text{LOLA}}^{(2)}(\theta_1, \theta_2) = \mathbb{E}_{\pi_{\theta_1}, \pi_{\theta_2} + \Delta \pi_{\theta_2}} \left[ L_{\text{LOLA}}^{(1)} \right]
\]

**Conclusions**

- DiCE offers a new approach to estimating gradients in stochastic computation graphs
- Arbitrary order derivatives of the DiCE objective correspond to gradient estimators
- DiCE can be implemented easily in standard learning frameworks!

**Acknowledgements and References**

Thanks to: Oxford-Google DeepMind Graduate Scholarship, the UK EPSRC CDT in Autonomous Intelligent Machines and Systems
