DiCE: The Infinitely Differentiable Monte Carlo Estimator

Jakob Foerster [1] **Gregory Farquhar** [*1] **Maruan Al-Shedivat** [*2]
Tim Rocktäschel [1] **Eric P. Xing** [2] **Shimon Whiteson** [1]

Abstract

The score function estimator is widely used for estimating gradients of stochastic objectives in *stochastic computation graphs* (SCG), *e.g.*, in reinforcement learning and meta-learning. While deriving the first order gradient estimators by differentiating a *surrogate loss* (SL) objective is computationally and conceptually simple, using the same approach for higher order derivatives is more challenging. Firstly, analytically deriving and implementing such estimators is laborious and not compliant with automatic differentiation. Secondly, repeatedly applying SL to construct new objectives for each order derivative involves increasingly cumbersome graph manipulations. Lastly, to match the first order gradient under differentiation, SL treats part of the cost as a fixed sample, which we show leads to missing and wrong terms for estimators of higher order derivatives. To address all these shortcomings in a unified way, we introduce DiCE, which provides a single objective that can be differentiated repeatedly, generating correct estimators of derivatives of any order in SCGs. Unlike SL, DiCE relies on automatic differentiation for performing the requisite graph manipulations. We verify the correctness of DiCE both through a proof and numerical evaluation of the DiCE derivative estimates. We also use DiCE to propose and evaluate a novel approach for multi-agent learning. Our code is available at github.com/alshedivat/lola.

1. Introduction

The score function trick is used to produce Monte Carlo estimates of gradients in settings with non-differentiable objectives, *e.g.*, in meta-learning and reinforcement learning. Estimating the first order gradients is computationally and conceptually simple. While the gradient estimators can be directly defined, it is often more convenient to define an objective whose derivative is the gradient estimator. Then the automatic-differentiation (auto-diff) toolbox, as implemented in deep learning libraries, can easily compute the gradient estimates with respect to all upstream parameters.

This is the method used by the *surrogate loss* (SL) approach (Schulman et al., 2015), which provides a recipe for building a surrogate objective from a *stochastic computation graph* (SCG). When differentiated, the SL yields an estimator for the first order gradient of the original objective.

However, estimating higher order derivatives is more challenging. Such estimators are useful for a number of optimization techniques, accelerating convergence in supervised settings (Dennis & Moré, 1977) and reinforcement learning (Furmston et al., 2016). Furthermore, they are vital for gradient-based meta-learning (Finn et al., 2017; Al-Shedivat et al., 2017; Li et al., 2017), which differentiates an objective after some number of first order learning steps. Estimators of higher order derivatives have also proven useful in multi-agent learning (Foerster et al., 2018), when one agent differentiates through the learning process of another agent.

Unfortunately, the first order gradient estimators mentioned above are fundamentally ill suited to calculating higher order derivatives via auto-diff. Due to the dependency on the sampling distribution, estimators of higher order derivatives require repeated application of the score function trick. Simply differentiating the first order estimator again, as was for example done by Finn et al. (2017), leads to missing terms.

To obtain higher order score function estimators, there are currently two unsatisfactory options. The first is to analytically derive and implement the estimators. However, this is laborious, error prone, and does not comply with the auto-diff paradigm. The second is to repeatedly apply the SL approach to construct new objectives for each further derivative estimate. However, each of these new objectives involves increasingly complex graph manipulations, defeating the purpose of a differentiable surrogate loss.

Moreover, to match the first order gradient after a single differentiation, the SL treats part of the cost as a fixed sam-

[*]Equal contribution [1]University of Oxford [2]Carnegie Mellon University. Correspondence to:
Jakob Foerster <jakob.foerster@cs.ox.ac.uk>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

ple, severing the dependency on the parameters. We show that this yields missing and incorrect terms in estimators of higher order derivatives. We believe that these difficulties have limited the usage and exploration of higher order methods in reinforcement learning tasks and other application areas that may be formulated as SCGs.

Therefore, we propose a novel technique, the *Infinitely **Di**fferentiable Monte-**C**arlo **E**stimator* (DiCE), to address all these shortcomings. DiCE constructs a single objective that evaluates to an estimate of the original objective, but can also be differentiated repeatedly to obtain correct estimators of derivatives of any order. Unlike the SL approach, DiCE relies on auto-diff as implemented for instance in TensorFlow (Abadi et al., 2016) or PyTorch (Paszke et al., 2017) to automatically perform the complex graph manipulations required for these estimators of higher order derivatives.

DiCE uses a novel operator, MagicBox ($\boxdot$), that acts on the set of those stochastic nodes $\mathcal{W}_c$ that influence each of the original losses in an SCG. Upon differentiation, this operator generates the correct derivatives associated with the sampling distribution:

$$\nabla_\theta \boxdot(\mathcal{W}_c) = \boxdot(\mathcal{W}_c) \nabla_\theta \sum_{w \in \mathcal{W}_c} \log(p(w;\theta)),$$

while returning 1 when evaluated: $\boxdot(\mathcal{W}) \rightarrowtail 1$. The MagicBox-operator can easily be implemented in standard deep learning libraries as follows:

$$\boxdot(\mathcal{W}) = \exp\big(\tau - \bot(\tau)\big),$$
$$\tau = \sum_{w \in \mathcal{W}} \log(p(w;\theta)),$$

where $\bot$ is an operator that sets the gradient of the operand to zero, so $\nabla_x \bot(x) = 0$. In addition, we show how to use a baseline for variance reduction in our formulation.

We verify the correctness of DiCE both through a proof and through numerical evaluation of the DiCE gradient estimates. To demonstrate the utility of DiCE, we also propose a novel approach for learning with opponent learning awareness (Foerster et al., 2018). We also open-source our code in TensorFlow. We hope this powerful and convenient novel objective will unlock further exploration and adoption of higher order learning methods in meta-learning, reinforcement learning, and other applications of SCGs. Already, DiCE is used to implement repeatedly differentiable gradient estimators with pyro.infer.util.Dice and tensorflow_probability.python.monte_carlo.expectation.

2. Background

Suppose x is a random variable, $x \sim p(x;\theta)$, f is a function of x and we want to compute $\nabla_\theta \mathbb{E}_x[f(x)]$. If the analytical gradients $\nabla_\theta f$ are unavailable or nonexistent, we can employ the *score function* (SF) estimator (Fu, 2006):

$$\nabla_\theta \mathbb{E}_x[f(x)] = \mathbb{E}_x[f(x) \nabla_\theta \log(p(x;\theta))] \tag{2.1}$$

If instead x is a deterministic function of θ and another random variable z, the operators ∇_θ and $\mathbb{E}_z$ commute, yielding the *pathwise derivative estimator* or *reparameterisation trick* (Kingma & Welling, 2013). In this work, we focus on the SF estimator, which can capture the interdependency of both the objective and the sampling distribution on the parameters θ, and therefore requires careful handling for estimators of higher order derivatives.[1]

2.1. Stochastic Computation Graphs

Gradient estimators for single random variables can be generalised using the formalism of a stochastic computation graph (SCG, Schulman et al., 2015). An SCG is a directed acyclic graph with four types of nodes: *input nodes*, Θ; *deterministic nodes*, $\mathcal{D}$; *cost nodes*, $\mathcal{C}$; and *stochastic nodes*, $\mathcal{S}$. Input nodes are set externally and can hold parameters we seek to optimise. Deterministic nodes are functions of their parent nodes, while stochastic nodes are distributions conditioned on their parent nodes. The set of cost nodes $\mathcal{C}$ are those associated with an objective $\mathcal{L} = \mathbb{E}[\sum_{c \in \mathcal{C}} c]$.

Let $v \prec w$ denote that node v *influences* node w, *i.e.*, there exists a path in the graph from v to w. If every node along the path is deterministic, v influences w deterministically which is denoted by $v \prec^D w$. See Figure 1 (top) for a simple SCG with an input node θ, a stochastic node x and a cost function f. Note that θ influences f deterministically ($\theta \prec^D f$) as well as stochastically via x ($\theta \prec f$).

2.2. Surrogate Losses

In order to estimate gradients of a sum of cost nodes, $\sum_{c \in \mathcal{C}} c$, in an arbitrary SCG, Schulman et al. (2015) introduce the notion of a *surrogate loss* (SL):

$$\mathrm{SL}(\Theta, \mathcal{S}) := \sum_{w \in \mathcal{S}} \log p(w \mid \mathrm{DEPS}_w) \hat{Q}_w + \sum_{c \in \mathcal{C}} c(\mathrm{DEPS}_c).$$

Here DEPS_w are the 'dependencies' of w: the set of stochastic or input nodes that deterministically influence the node w. Furthermore, $\hat{Q}_w$ is the sum of *sampled* costs $\hat{c}$ corresponding to the cost nodes influenced by w.

The hat notation on $\hat{Q}_w$ indicates that inside the SL, these costs are treated as fixed samples. This severs the functional dependency on θ that was present in the original stochastic computation graph.

The SL produces a gradient estimator when differentiated once (Schulman et al., 2015, Corollary 1):

$$\nabla_\theta \mathcal{L} = \mathbb{E}[\nabla_\theta \mathrm{SL}(\Theta, \mathcal{S})]. \tag{2.2}$$

[1] In the following, we use the terms 'gradient' and 'derivative' interchangeably.

Note that the use of sampled costs $\hat{Q}_w$ in the definition of the SL ensures that its first order gradients match the score function estimator, which does not contain a term of the form $\log(p)\nabla_\theta Q$.

Although Schulman et al. (2015) focus on first order gradients, they argue that the SL gradient estimates themselves can be treated as costs in an SCG and that the SL approach can be applied repeatedly to construct higher order gradient estimators. However, the use of sampled costs in the SL leads to missing dependencies and wrong estimates when calculating such higher order gradients, as we discuss in Section 3.2.

3. Higher Order Derivatives

In this section, we illustrate how to estimate higher order derivatives via repeated application of the score function (SF) trick and show that repeated application of the surrogate loss (SL) approach in stochastic computation graphs (SCGs) fails to capture all of the relevant terms for higher order gradient estimates.

3.1. Estimators of Higher Order Derivatives

We begin by revisiting the derivation of the score function estimator for the gradient of the expectation $\mathcal{L}$ of $f(x;\theta)$ over $x \sim p(x;\theta)$:

$$\begin{aligned}
\nabla_\theta \mathcal{L} &= \nabla_\theta \mathbb{E}_x\left[f(x;\theta)\right] \\
&= \nabla_\theta \sum_x p(x;\theta) f(x;\theta) \\
&= \sum_x \nabla_\theta \big(p(x;\theta) f(x;\theta)\big) \\
&= \sum_x \big(f(x;\theta)\nabla_\theta p(x;\theta) + p(x;\theta)\nabla_\theta f(x;\theta)\big) \\
&= \sum_x \big(f(x;\theta)p(x;\theta)\nabla_\theta \log(p(x;\theta)) \\
&\qquad + p(x;\theta)\nabla_\theta f(x;\theta)\big) \\
&= \mathbb{E}_x\left[f(x;\theta)\nabla_\theta \log(p(x;\theta)) + \nabla_\theta f(x;\theta)\right] \quad (3.1) \\
&= \mathbb{E}_x[g(x;\theta)].
\end{aligned}$$

The estimator $g(x;\theta)$ of the gradient of $\mathbb{E}_x\left[f(x;\theta)\right]$ consists of two distinct terms: (1) the term $f(x;\theta)\nabla_\theta \log(p(x;\theta))$ originating from $f(x;\theta)\nabla_\theta p(x;\theta)$ via the SF trick, and (2) the term $\nabla_\theta f(x;\theta)$, due to the direct dependence of f on θ. The second term is often ignored because f is often only a function of x but not of θ. However, even in that case, the gradient estimator g depends on both x and θ. We might be tempted to again apply the SL approach to $\nabla_\theta \mathbb{E}_x[g(x;\theta)]$ to produce estimates of higher order gradients of $\mathcal{L}$, but below we demonstrate that this fails. In Section 4, we introduce a practical algorithm for correctly producing such higher order gradient estimators in SCGs.

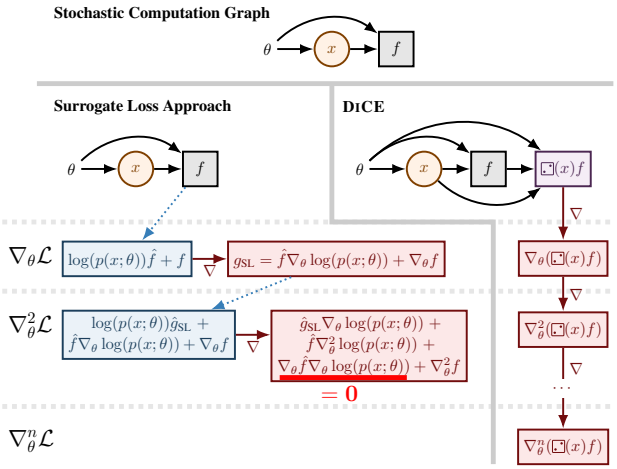

Figure 1. Simple example illustrating the difference of the Surrogate Loss (SL) approach to DiCE. Stochastic nodes are depicted in orange, costs in gray, surrogate losses in blue, DiCE in purple, and gradient estimators in red. Note that for second-order gradients, SL requires the construction of an intermediate stochastic computation graph and due to taking a sample of the cost $\hat{g}_{\text{SL}}$, the dependency on θ is lost, leading to an incorrect second-order gradient estimator. Arrows from θ, x and f to gradient estimators omitted for clarity.

3.2. Higher Order Surrogate Losses

While Schulman et al. (2015) focus on first order gradients, they state that a recursive application of SL can generate higher order gradient estimators. However, as we demonstrate in this section, because the SL approach treats part of the objective as a sampled cost, the corresponding terms lose a functional dependency on the sampling distribution. This leads to missing terms in the estimators of higher order gradients.

Consider the following example, where a single parameter θ defines a sampling distribution $p(x;\theta)$ and the objective is $f(x,\theta)$.

$$\begin{aligned}
\text{SL}(\mathcal{L}) &= \log p(x;\theta)\hat{f}(x) + f(x;\theta) \\
(\nabla_\theta \mathcal{L})_{\text{SL}} &= \mathbb{E}_x[\nabla_\theta \text{SL}(\mathcal{L})] \\
&= \mathbb{E}_x[\hat{f}(x)\nabla_\theta \log p(x;\theta) + \nabla_\theta f(x;\theta)] \quad (3.2) \\
&= \mathbb{E}_x[g_{\text{SL}}(x;\theta)].
\end{aligned}$$

The corresponding SCG is depicted at the top of Figure 1. Comparing (3.1) and (3.2), note that the first term, $\hat{f}(x)$ has lost its functional dependency on θ, as indicated by the hat notation and the lack of a θ argument. While these terms evaluate to the same estimate of the first order gradient, the lack of the dependency yields a discrepancy between the exact derivation of the second order gradient and a second

application of SL:

$$\begin{aligned}\mathrm{SL}(g_{\mathrm{SL}}(x;\theta)) &= \log p(x;\theta)\hat{g}_{\mathrm{SL}}(x) + g_{\mathrm{SL}}(x;\theta)\\ (\nabla^2_\theta \mathcal{L})_{\mathrm{SL}} &= \mathbb{E}_x[\nabla_\theta \mathrm{SL}(g_{\mathrm{SL}})]\\ &= \mathbb{E}_x[\hat{g}_{\mathrm{SL}}(x)\nabla_\theta \log p(x;\theta) + \nabla_\theta g_{\mathrm{SL}}(x;\theta)].\end{aligned} \tag{3.3}$$

By contrast, the exact derivation of $\nabla^2_\theta \mathcal{L}$ results in the following expression:

$$\begin{aligned}\nabla^2_\theta \mathcal{L} &= \nabla_\theta \mathbb{E}_x[g(x;\theta)]\\ &= \mathbb{E}_x[g(x;\theta)\nabla_\theta \log p(x;\theta) + \nabla_\theta g(x;\theta)].\end{aligned} \tag{3.4}$$

Since $g_{\mathrm{SL}}(x;\theta)$ differs from $g(x;\theta)$ only in its dependencies on θ, g_{SL} and g are identical when *evaluated*. However, due to the missing dependencies in g_{SL}, the *gradients* w.r.t. θ, which appear in the higher order gradient estimates in (3.3) and (3.4), differ:

$$\begin{aligned}\nabla_\theta g(x;\theta) &= \nabla_\theta f(x;\theta)\nabla_\theta \log(p(x;\theta))\\ &\quad + f(x;\theta)\nabla^2_\theta \log(p(x;\theta))\\ &\quad + \nabla^2_\theta f(x;\theta),\\ \nabla_\theta g_{\mathrm{SL}}(x;\theta) &= \hat{f}(x)\nabla^2_\theta \log(p(x;\theta))\\ &\quad + \nabla^2_\theta f(x;\theta).\end{aligned}$$

We lose the term $\nabla_\theta f(x;\theta)\nabla_\theta \log(p(x;\theta))$ in the second order SL gradient because $\nabla_\theta \hat{f}(x) = 0$ (see left part of Figure 1). This issue occurs immediately in the second order gradients when f depends directly on θ. However, as $g(x;\theta)$ always depends on θ, the SL approach always fails to produce correct third or higher order gradient estimates even if f depends only indirectly on θ.

3.3. Example

Here is a toy example to illustrate a possible failure case. Let $x \sim \mathrm{Ber}(\theta)$ and $f(x,\theta) = x(1-\theta) + (1-x)(1+\theta)$. For this simple example we can exactly evaluate all terms:

$$\begin{aligned}\mathcal{L} &= \theta(1-\theta) + (1-\theta)(1+\theta)\\ \nabla_\theta \mathcal{L} &= -4\theta + 1\\ \nabla^2_\theta \mathcal{L} &= -4\end{aligned}$$

Evaluating the expectations for the SL gradient estimators analytically results in the following terms, with an incorrect second-order estimate:

$$\begin{aligned}(\nabla_\theta \mathcal{L})_{\mathrm{SL}} &= -4\theta + 1\\ (\nabla^2_\theta \mathcal{L})_{\mathrm{SL}} &= -2\end{aligned}$$

If, for example, the Newton-Raphson method was used to optimise $\mathcal{L}$, the solution could be found in a single iteration with the correct Hessian. In contrast, the wrong estimates from the SL approach would require damping to approach the optimum at all, and many more iterations would be needed.

The failure mode seen in this toy example appears whenever the objective includes a regularisation term that depends on θ, and is also impacted by the stochastic samples. One example in a practical algorithm is soft Q-learning for RL (Schulman et al., 2017), which regularises the policy by adding an entropy penalty to the rewards. This penalty encourages the agent to maintain an exploratory policy, reducing the probability of getting stuck in local optima. Clearly the penalty depends on the policy parameters θ. However, the policy entropy also depends on the states visited, which in turn depend on the stochastically sampled actions. As a result, the entropy regularised RL objective in this algorithm has the exact property leading to the failure of the SL approach shown above. Unlike our toy analytic example, the consequent errors do not just appear as a rescaling of the proper higher order gradients, but depend in a complex way on the parameters θ. Any second order methods with such a regularised objective therefore requires an alternate strategy for generating gradient estimators, even setting aside the awkwardness of repeatedly generating new surrogate objectives.

4. Correct Gradient Estimators with DiCE

In this section, we propose the *Infinitely **Di**fferentiable Monte-**C**arlo **E**stimator* (DiCE), a practical algorithm for programatically generating correct gradients of any order in arbitrary SCGs. The naive option is to recursively apply the update rules in (3.1) that map from $f(x;\theta)$ to the estimator of its derivative $g(x;\theta)$. However, this approach has two deficiencies. First, by defining gradients directly, it fails to provide an objective that can be used in standard deep learning libraries. Second, these naive gradient estimators violate the auto-diff paradigm for generating further estimators by repeated differentiation since in general $\nabla_\theta f(x;\theta) \neq g(x;\theta)$. Our approach addresses these issues, as well as fixing the missing terms from the SL approach.

As before, $\mathcal{L} = \mathbb{E}[\sum_{c\in\mathcal{C}} c]$ is the objective in an SCG. The correct expression for a gradient estimator that preserves all required dependencies for further differentiation is:

$$\nabla_\theta \mathcal{L} = \mathbb{E}\left[\sum_{c\in\mathcal{C}}\left(c\sum_{w\in\mathcal{W}_c}\nabla_\theta \log p(w \mid \mathrm{DEPS}_w) + \nabla_\theta c(\mathrm{DEPS}_c)\right)\right], \tag{4.1}$$

where $\mathcal{W}_c = \{w \mid w \in \mathcal{S}, w \prec c, \theta \prec w\}$, *i.e.*, the set of stochastic nodes that depend on θ and influence the cost c.

For brevity, from here on we suppress the DEPS notation, assuming all probabilities and costs are conditioned on their relevant ancestors in the SCG.

Note that (4.1) is the generalisation of (3.1) to arbitrary SCGs. The proof is given by Schulman et al. (2015, Lines 1-10, Appendix A). Crucially, in Line 11 the authors then replace c by $\hat{c}$, severing the dependencies required for correct higher order gradient estimators. As described in Section 2.2, this was done so that the SL approach reproduces the score function estimator after a single differentiation and can thus be used as an objective for backpropagation in a deep learning library.

To support correct higher order gradient estimators, we propose DiCE, which relies heavily on a novel operator, MagicBox ($\boxdot$). MagicBox takes a set of stochastic nodes $\mathcal{W}$ as input and has the following two properties by design:

1. $\boxdot(\mathcal{W}) \rightarrowtail 1$,
2. $\nabla_\theta \boxdot(\mathcal{W}) = \boxdot(\mathcal{W}) \sum_{w \in \mathcal{W}} \nabla_\theta \log(p(w;\theta))$.

Here, $\rightarrowtail$ indicates "*evaluates to*" in contrast to full equality, $=$, which includes equality of all gradients. In the autodiff paradigm, $\rightarrowtail$ corresponds to a forward pass evaluation of a term. Meanwhile, the behaviour under differentiation in property (2) indicates the new graph nodes that will be constructed to hold the gradients of that object. Note that that $\boxdot(\mathcal{W})$ reproduces the dependency of the gradient on the sampling distribution under differentiation through the requirements above. Using $\boxdot$, we can next define the DiCE objective, $\mathcal{L}_{\boxdot}$:

$$\mathcal{L}_{\boxdot} = \sum_{c \in \mathcal{C}} \boxdot(\mathcal{W}_c) c. \tag{4.2}$$

Below we prove that the DiCE objective indeed produces correct arbitrary order gradient estimators under differentiation.

Theorem 1. $\mathbb{E}[\nabla_\theta^n \mathcal{L}_{\boxdot}] \rightarrowtail \nabla_\theta^n \mathcal{L}, \forall n \in \{0, 1, 2, \dots\}$.

Proof. For each cost node $c \in \mathcal{C}$, we define a sequence of nodes, $c^n, n \in \{0, 1, \dots\}$ as follows:

$$\begin{aligned} c^0 &= c, \\ \mathbb{E}[c^{n+1}] &= \nabla_\theta \mathbb{E}[c^n]. \end{aligned} \tag{4.3}$$

By induction it follows that $\mathbb{E}[c^n] = \nabla_\theta^n \mathbb{E}[c]\ \forall n$, *i.e.*, c^n is an estimator of the nth order derivative of the objective $\mathbb{E}[c]$.

We further define $c^n_{\boxdot} = c^n \boxdot(\mathcal{W}_{c^n})$. Since $\boxdot(x) \rightarrowtail 1$, clearly $c^n_{\boxdot} \rightarrowtail c^n$. Therefore $\mathbb{E}[c^n_{\boxdot}] \rightarrowtail \mathbb{E}[c^n] = \nabla_\theta^n \mathbb{E}[c]$, *i.e.*, $c^n_{\boxdot}$ is also a valid estimator of the nth order derivative of the objective. Next, we show that $c^n_{\boxdot}$ can be generated by differentiating $c^0_{\boxdot}$ n times. This follows by induction, if $\nabla_\theta c^n_{\boxdot} = c^{n+1}_{\boxdot}$, which we prove as follows:

$$\begin{aligned} \nabla_\theta c^n_{\boxdot} &= \nabla_\theta (c^n \boxdot(\mathcal{W}_{c^n})) \\ &= c^n \nabla_\theta \boxdot(\mathcal{W}_{c^n}) + \boxdot(\mathcal{W}_{c^n}) \nabla_\theta c^n \\ &= c^n \boxdot(\mathcal{W}_{c^n}) \left(\sum_{w \in \mathcal{W}_{c^n}} \nabla_\theta \log(p(w;\theta)) \right) \\ &\quad + \boxdot(\mathcal{W}_{c^n}) \nabla_\theta c^n \\ &= \boxdot(\mathcal{W}_{c^n}) \left(\nabla_\theta c^n + c^n \sum_{w \in \mathcal{W}_{c^n}} \nabla_\theta \log(p(w;\theta)) \right) \end{aligned} \tag{4.4}$$

$$= \boxdot(\mathcal{W}_{c^{n+1}}) c^{n+1} = c^{n+1}_{\boxdot}. \tag{4.5}$$

To proceed from (4.4) to (4.5), we need two additional steps. First, we require an expression for c^{n+1}. Substituting $\mathcal{L} = \mathbb{E}[c^n]$ into (4.1) and comparing to (4.3), we find the following map from c^n to c^{n+1}:

$$c^{n+1} = \nabla_\theta c^n + c^n \sum_{w \in \mathcal{W}_{c^n}} \nabla_\theta \log p(w;\theta). \tag{4.6}$$

The term inside the brackets in (4.4) is identical to c^{n+1}. Secondly, note that (4.6) shows that c^{n+1} depends only on c^n and $\mathcal{W}_{c^n}$. Therefore, the stochastic nodes which influence c^{n+1} are the same as those which influence c^n. So $\mathcal{W}_{c^n} = \mathcal{W}_{c^{n+1}}$, and we arrive at (4.5).

To conclude the proof, recall that c^n is the estimator for the nth derivative of c, and that $c^n_{\boxdot} \rightarrowtail c^n$. Summing over $c \in \mathcal{C}$ then gives the desired result. □

Implementation of DiCE. DiCE is easy to implement in standard deep learning libraries [2]:

$$\begin{aligned} \boxdot(\mathcal{W}) &= \exp\big(\tau - \perp(\tau)\big), \\ \tau &= \sum_{w \in \mathcal{W}} \log(p(w;\theta)), \end{aligned}$$

where $\perp$ is an operator that sets the gradient of the operand to zero, so $\nabla_x \perp(x) = 0$.[3]

Since $\perp(x) \rightarrowtail x$, clearly $\boxdot(\mathcal{W}) \rightarrowtail 1$. Furthermore:

$$\begin{aligned} \nabla_\theta \boxdot(\mathcal{W}) &= \nabla_\theta \exp\big(\tau - \perp(\tau)\big) \\ &= \exp\big(\tau - \perp(\tau)\big) \nabla_\theta (\tau - \perp(\tau)) \\ &= \boxdot(\mathcal{W})(\nabla_\theta \tau + 0) \\ &= \boxdot(\mathcal{W}) \sum_{w \in \mathcal{W}} \nabla_\theta \log(p(w;\theta)). \end{aligned}$$

[2] A previous version of tf.contrib.bayesflow authored by Josh Dillon also used this implementation trick.

[3] This operator exists in PyTorch as `detach` and in TensorFlow as `stop_gradient`.

With this implementation of the $\boxdot$-operator, it is now straightforward to construct $\mathcal{L}_{\boxdot}$ as defined in (4.7). This procedure is demonstrated in Figure 2, which shows a reinforcement learning use case. In this example, the cost nodes are rewards that depend on stochastic actions, and the total objective is $J = \mathbb{E}[\sum r_t]$. We construct a DiCE objective $J_{\boxdot} = \sum_t \boxdot(\{a_{t'}, t' \leq t\}) r_t$. Now $\mathbb{E}[J_{\boxdot}] \rightarrowtail J$ and $\mathbb{E}[\nabla^n_\theta J_{\boxdot}] \rightarrowtail \nabla^n_\theta J$, so $J_{\boxdot}$ can both be used to estimate the return and to produce estimators for any order gradients under auto-diff, which can be used for higher order methods.

Note that DiCE can be equivalently expressed with $\boxdot(\mathcal{W}) = \tilde{p}/\perp(\tilde{p}), \tilde{p} = \prod_{w \in \mathcal{W}} p(w; \theta)$. We use the exponentiated form to emphasise the generator-like functionality of the operator and to ensure numerical stability.

Causality. The SL approach handles causality by summing over stochastic nodes, w, and multiplying $\nabla \log(p(w))$ for each stochastic node with a sum of the *downstream* costs, $\hat{Q}_w$. In contrast, the DiCE objective sums over costs, c, and multiplies each cost with a sum over the gradients of log-probabilities from *upstream* stochastic nodes, $\sum_{w \in \mathcal{W}_c} \nabla \log(p(w))$.

In both cases, integrating causality into the gradient estimator leads to reduction of variance compared to the naive approach of multiplying the full sum over costs with the full sum over grad-log-probabilities.

However, the second formulation leads to greatly reduced conceptual complexity when calculating higher order terms, which we exploit in the definition of the DiCE objective. This is because each further gradient estimator maintains the same backward looking dependencies for each term in the original sum over costs, *i.e.*, $\mathcal{W}_{c^n} = \mathcal{W}_{c^{n+1}}$.

In contrast, the SL approach is centred around the stochastic nodes, which each become associated with a growing number of downstream costs after each differentiation. Consequently, we believe that our DiCE objective is more intuitive, as it is conceptually centred around the original objective and remains so under repeated differentiation.

Variance Reduction. We can include a baseline term in the definition of the DiCE objective:

$$\mathcal{L}_{\boxdot} = \sum_{c \in \mathcal{C}} \boxdot(\mathcal{W}_c) c + \sum_{w \in \mathcal{S}} (1 - \boxdot(\{w\})) b_w. \tag{4.7}$$

The baseline b_w is a design choice and can be any function of nodes not influenced by w. As long as this condition is met, the baseline does not change the expectation of the gradient estimates, but can considerably reduce the variance. A common choice is the average cost.

Since $(1 - \boxdot(\{w\})) \rightarrowtail 0$, this implementation of the baseline leaves the evaluation of the estimator $\mathcal{L}_{\boxdot}$ of the original objective unchanged.

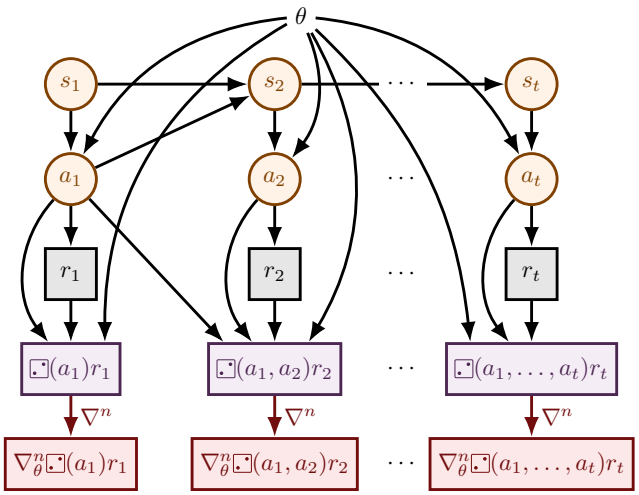

Figure 2. DiCE applied to a reinforcement learning problem. A stochastic policy conditioned on s_t and θ produces actions, a_t, which lead to rewards r_t and next states, s_{t+1}. Associated with each reward is a DiCE objective that takes as input the set of all causal dependencies that are functions of θ, *i.e.*, the actions. Arrows from θ, a_i and r_i to gradient estimators omitted for clarity.

Hessian-Vector Product. The Hessian-vector, $v^\top H$, is useful for a number of algorithms, such as estimation of eigenvectors and eigenvalues of H (Pearlmutter, 1994). Using DiCE, $v^\top H$ can be implemented efficiently without having to compute the full Hessian. Assuming v does not depend on θ and using $^\top$ to indicate the transpose:

$$\begin{aligned} v^\top H &= v^\top \nabla^2 \mathcal{L}_{\boxdot} \\ &= v^\top (\nabla^\top \nabla \mathcal{L}_{\boxdot}) \\ &= \nabla^\top (v^\top \nabla \mathcal{L}_{\boxdot}). \end{aligned}$$

In particular, $(v^\top \nabla \mathcal{L}_{\boxdot})$ is a scalar, making this implementation well suited for auto-diff.

5. Case Studies

While the main contribution of this paper is to provide a novel general approach for any order gradient estimation in SCGs, we also provide a proof-of-concept empirical evaluation for a set of case studies, carried out on the *iterated prisoner's dilemma* (IPD). In IPD, two agents iteratively play matrix games with two possible actions: (C)ooperate and (D)efect. The possible outcomes of each game are DD, DC, CD, CC with the corresponding first agent payoffs, -2, 0, -3, -1, respectively. This setting is useful because (1) it has a nontrivial but analytically calculable value function, allowing for verification of gradient estimates, and (2) differentiating through the learning steps of other agents in multi-agent RL is a highly relevant application of higher order policy gradient estimators in RL (Foerster et al., 2018).

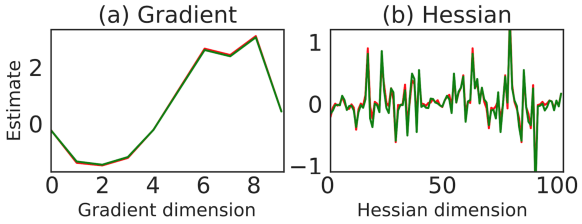

Figure 3. For the iterated prisoner's dilemma, shown is the flattened true (red) and estimated (green) gradient (left) and Hessian (right) using the first and second derivative of DiCE and the exact value function respectively. The correlation coefficients are 0.999 for the gradients and 0.97 for the Hessian; the sample size is 100k.

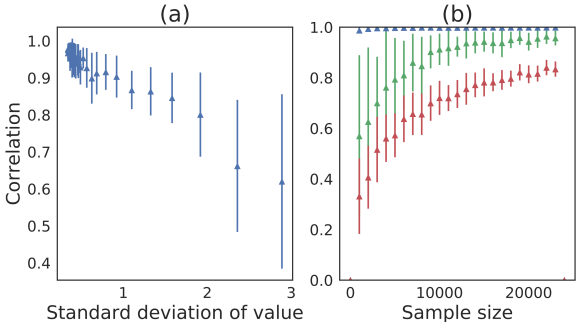

Figure 4. Shown in (a) is the correlation of the gradient estimator (averaged across agents) as a function of the estimation error of the baseline when using a sample size of 128 and in (b) as a function of sample size when using a converged baseline (in blue) and no baseline (in green) for gradients and in red for Hessian. Errors bars indicate standard deviation on both plots.

Empirical Verification. We first verify that DiCE recovers gradients and Hessians in stochastic computation graphs. To do so, we use DiCE to estimate gradients and Hessians of the expected return for fixed policies in IPD.

As shown in Figure 3, we find that indeed the DiCE estimator matches both the gradients (a) and the Hessians (b) for both agents accurately. Furthermore, Figure 4 shows how the estimate of the gradient improve as the value function becomes more accurate during training, in (a). Also shown is the quality of the gradient estimation as a function of sample size with and without a baseline, in (b). Both plots show that the baseline is a key component of DiCE for accurate estimation of gradients.

DiCE for multi-agent RL. In *learning with opponent-learning awareness* (LOLA), Foerster et al. (2018) show that agents that differentiate through the learning step of their opponent converge to Nash equilibria with higher social welfare in the IPD.

Since the standard policy gradient learning step for one agent has no dependency on the parameters of the other agent (which it treats as part of the environment), LOLA relies on a Taylor expansion of the expected return in combination with an analytical derivation of the second order gradients to be able to differentiate through the expected return after the opponent's learning step.

Here, we take a more direct approach, made possible by DiCE. Let π_{θ_1} be the policy of the LOLA agent and let π_{θ_2} be the policy of its opponent and vice versa. Assuming that the opponent learns using policy gradients, LOLA-DiCE agents learn by directly optimising the following stochastic objective w.r.t. θ_1:

$$\begin{aligned}\mathcal{L}^1(\theta_1,\theta_2)_{\text{LOLA}} &= \mathbb{E}_{\pi_{\theta_1},\pi_{\theta_2+\Delta\theta_2(\theta_1,\theta_2)}}\left[\mathcal{L}^1\right], \text{where} \\ \Delta\theta_2(\theta_1,\theta_2) &= \alpha_2\nabla_{\theta_2}\mathbb{E}_{\pi_{\theta_1},\pi_{\theta_2}}\left[\mathcal{L}^2\right],\end{aligned} \tag{5.1}$$

where α_2 is a scalar step size and $\mathcal{L}^i = \sum_{t=0}^{T}\gamma^t r_t^i$ is the sum of discounted returns for agent i.

To evaluate these terms directly, our variant of LOLA unrolls the learning process of the opponent, which is functionally similar to model-agnostic meta-learning (MAML, Finn et al., 2017). In the MAML formulation, the gradient update of the opponent, $\Delta\theta_2$, corresponds to the inner loop (typically the training objective) and the gradient update of the agent itself to the outer loop (typically the test objective). Algorithm 1 describes the procedure we use to compute updates for the agent's parameters.

Using the following DiCE-objective to estimate gradient steps for agent i, we are able to preserve all dependencies:

$$\mathcal{L}^i_{\boxdot(\theta_1,\theta_2)} = \sum_t \boxdot\left(\left\{a_{j\in\{1,2\}}^{t'\leq t}\right\}\right)\gamma^t r_t^i, \tag{5.2}$$

where $\left\{a_{j\in\{1,2\}}^{t'\leq t}\right\}$ is the set of all actions taken by both agents up to time t. To save computation, we cache the $\Delta\theta_i$ of the inner loop when unrolling the outer loop policies in order to avoid recalculating them at every time step.

Algorithm 1 LOLA-DiCE: policy gradient update for θ_1

input Policy parameters of the agent, θ_1, and of the opponent, θ_2
Initialize: $\theta_2' \leftarrow \theta_2$
for k in $1 \ldots K$ **do** // inner loop lookahead steps
  Rollout trajectories τ_k under $(\pi_{\theta_1}, \pi_{\theta_2'})$
  Update: $\theta_2' \leftarrow \theta_2' + \alpha_2\nabla_{\theta_2'}\mathcal{L}^2_{\boxdot(\theta_1,\theta_2')}$ // lookahead update
end for
Rollout trajectories τ under $(\pi_{\theta_1}, \pi_{\theta_2'})$.
Update: $\theta_1' \leftarrow \theta_1 + \alpha_1\nabla_{\theta_1}\mathcal{L}^1_{\boxdot(\theta_1,\theta_2')}$ // PG update
output θ_1'.

Using DiCE, differentiating through $\Delta\theta_2$ produces the correct higher order gradients, which is critical for LOLA. By contrast, simply differentiating through the SL-based first order gradient estimator multiple times, as was done for MAML (Finn et al., 2017), results in omitted gradient

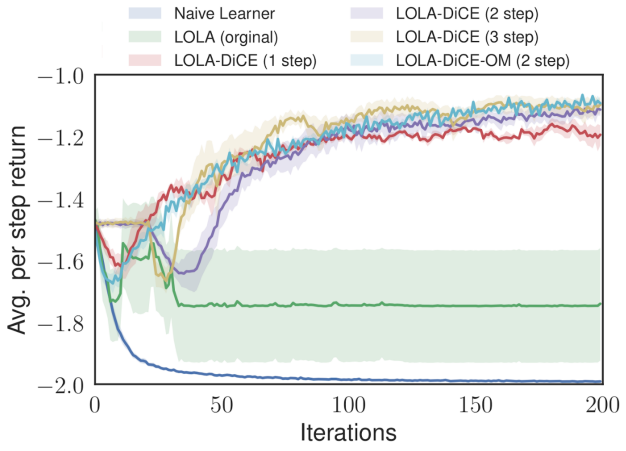

Figure 5. Joint average per step returns for different training methods. Comparing Naive Learning with the original LOLA algorithm and LOLA-DiCE with a varying number of look-ahead steps. Shaded areas represent the 95% confidence intervals based on 5 runs. All agents used batches of size 64, which is more than 60 times smaller than the size required in the original LOLA paper.

terms and a biased gradient estimator, as pointed out by Al-Shedivat et al. (2017) and Stadie et al. (2018).

Figure 5 shows a comparison between the LOLA-DiCE agents and the original formulation of LOLA. In our experiments, we use a time horizon of 150 steps and a reduced batch size of 64; the lookahead gradient step, α, is set to 1 and the learning rate is 0.3. Importantly, given the approximation used, the LOLA method was restricted to a single step of opponent learning. In contrast, using DiCE we can unroll and differentiate through an arbitrary number of the opponent learning steps.

The original LOLA implemented via second order gradient corrections shows no stable learning, as it requires much larger batch sizes ($\sim$4000). By contrast, LOLA-DiCE agents discover strategies of high social welfare, replicating the results of the original LOLA paper in a way that is both more direct, efficient and establishes a common formulation between MAML and LOLA.

6. Related Work

Gradient estimation is well studied, although many methods have been named and explored independently in different fields, and the primary focus has been on first order gradients. Fu (2006) provides an overview of methods from the point of view of simulation optimization.

The score function (SF) estimator, also referred to as the likelihood ratio estimator or REINFORCE, has received considerable attention in many fields. In reinforcement learning, policy gradient methods (Williams, 1992) have proven highly successful, especially when combined with variance reduction techniques (Weaver & Tao, 2001; Grondman et al., 2012). The SF estimator has also been used in the analysis of stochastic systems (Glynn, 1990), as well as for variational inference (Wingate & Weber, 2013; Ranganath et al., 2014). Kingma & Welling (2013) and Rezende et al. (2014) discuss Monte-Carlo gradient estimates in the case where the stochastic parts of a model can be reparameterised.

These approaches are formalised for arbitrary computation graphs by Schulman et al. (2015), but to our knowledge our paper is the first to present a practical and correct approach for generating higher order gradient estimators utilising auto-diff. To easily make use of these estimates for optimising neural network models, automatic differentiation for backpropagation has been widely used (Baydin et al., 2015).

One rapidly growing application area for such higher order gradient estimates is meta-learning for reinforcement learning. Finn et al. (2017) compute a loss after a number of policy gradient learning steps, differentiating through the learning step to find parameters that can be quickly fine-tuned for different tasks. Li et al. (2017) extend this work to also meta-learn the fine-tuning step direction and magnitude. Al-Shedivat et al. (2017) and Stadie et al. (2018) derive the proper higher order gradient estimators for their work by reapplying the score function trick. Foerster et al. (2018) use a multi-agent version of the same higher order gradient estimators in combination with a Taylor expansion of the expected return. None present a general strategy for constructing higher order gradient estimators for arbitrary stochastic computation graphs.

7. Conclusion

We presented DiCE, a general method for computing any order gradient estimators for stochastic computation graphs. DiCE resolves the deficiencies of current approaches for computing higher order gradient estimators: analytical calculation is error-prone and incompatible with auto-diff, while repeated application of the surrogate loss approach is cumbersome and, as we show, leads to incorrect estimators in many cases. We prove the correctness of DiCE estimators, introduce a simple practical implementation of DiCE for use in deep learning frameworks, and validate its correctness and utility in a multi-agent reinforcement learning problem. We believe DiCE will unlock further exploration and adoption of higher order learning methods in meta-learning, reinforcement learning, and other applications of stochastic computation graphs. As a next step we will extend and improve the variance reduction of DiCE in order to provide a simple end-to-end solution for higher order gradient estimation. In particular we hope to include solutions such as REBAR (Tucker et al., 2017) in the DiCE operator.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement number 637713) and National Institute of Health (NIH R01GM114311). It was also supported by the Oxford Google DeepMind Graduate Scholarship and the UK EPSRC CDT in Autonomous Intelligent Machines and Systems. We would like to thank Misha Denil, Brendan Shillingford and Wendelin Boehmer for providing feedback on the manuscript.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P. A., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and Zheng, X. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, Savannah, GA, USA, November 2-4, 2016.*, pp. 265–283, 2016.

Al-Shedivat, M., Bansal, T., Burda, Y., Sutskever, I., Mordatch, I., and Abbeel, P. Continuous adaptation via meta-learning in nonstationary and competitive environments. *CoRR*, abs/1710.03641, 2017.

Baydin, A. G., Pearlmutter, B. A., and Radul, A. A. Automatic differentiation in machine learning: a survey. *CoRR*, abs/1502.05767, 2015.

Dennis, Jr, J. E. and Moré, J. J. Quasi-newton methods, motivation and theory. *SIAM review*, 19(1):46–89, 1977.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 1126–1135, 2017.

Foerster, J. N., Chen, R. Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., and Mordatch, I. Learning with opponent-learning awareness. In *AAMAS*, 2018.

Fu, M. C. Gradient estimation. *Handbooks in operations research and management science*, 13:575–616, 2006.

Furmston, T., Lever, G., and Barber, D. Approximate newton methods for policy search in markov decision processes. *Journal of Machine Learning Research*, 17(227): 1–51, 2016.

Glynn, P. W. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM*, 33 (10):75–84, 1990.

Grondman, I., Busoniu, L., Lopes, G. A., and Babuska, R. A survey of actor-critic reinforcement learning: Standard and natural policy gradients. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(6):1291–1307, 2012.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

Li, Z., Zhou, F., Chen, F., and Li, H. Meta-sgd: Learning to learn quickly for few shot learning. *CoRR*, abs/1707.09835, 2017.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch. 2017.

Pearlmutter, B. A. Fast exact multiplication by the hessian. *Neural computation*, 6(1):147–160, 1994.

Ranganath, R., Gerrish, S., and Blei, D. M. Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*, pp. 814–822, 2014.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. pp. 1278–1286, 2014.

Schulman, J., Heess, N., Weber, T., and Abbeel, P. Gradient Estimation Using Stochastic Computation Graphs. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 3528–3536, 2015.

Schulman, J., Abbeel, P., and Chen, X. Equivalence between policy gradients and soft q-learning. *CoRR*, abs/1704.06440, 2017.

Stadie, B., Yang, G., Houthooft, R., Chen, X., Duan, Y., Wu, Y., Abbeel, P., and Sutskever, I. Some considerations on learning to explore via meta-reinforcement learning, 2018. URL https://openreview.net/forum?id=Skk3Jm96W.

Tucker, G., Mnih, A., Maddison, C. J., Lawson, J., and Sohl-Dickstein, J. Rebar: Low-variance, unbiased gradient estimates for discrete latent variable models. In *Advances in Neural Information Processing Systems*, pp. 2624–2633, 2017.

Weaver, L. and Tao, N. The optimal reward baseline for gradient-based reinforcement learning. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pp. 538–545. Morgan Kaufmann Publishers Inc., 2001.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Reinforcement Learning*, pp. 5–32. Springer, 1992.

Wingate, D. and Weber, T. Automated variational inference in probabilistic programming. *CoRR*, abs/1301.1299, 2013.