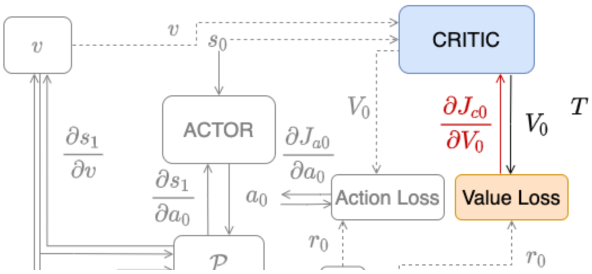
# Derivative Computation for ORCHID: Optimisation of Robotic Control and Hardware In Design using Reinforcement Learning.

Lucy Jackson[1], Celyn Walters [1], Steve Eckersley[2], Pete Senior [2] and Simon Hadfield[1]



$v$
$v$
$s_0$
CRITIC
ACTOR
$V_0$
$\frac{\partial J_{c0}}{\partial V_0}$
$V_0$
$T = 0$
$\frac{\partial s_1}{\partial v}$
$\frac{\partial J_{a0}}{\partial a_0}$
$\frac{\partial s_1}{\partial a_0}$
$a_0$
Action Loss
Value Loss
$r_0$
$r_0$
$\mathcal{P}$

**Fig. 1:** Path required for critic update in time step 0.

### T=0

Throughout training all parameters are constrained using only two loss functions. One constrains the critic network ($J_{ct}$) and the second constrains both the actor network and the hardware parameters ($J_{at}$). For simplicity, the derivatives are defined in terms of the network outputs not their trainable parameters since these are architecture dependant. In the case of the actor this is $a_t$ and for the critic, it is $V_t$. Figure 1 highlights the path used to calculate the critic loss for the first time step ($J_{c0}$). The derivative of this function defines the first update used to optimise the critic network:

$$\frac{\partial J_{c0}}{\partial V_0}. \tag{1}$$

From previous definition we know that

$$J_{C0} = \left(\gamma^0 r_0 + \mu_\phi(s_1, v) - \mu_\phi(s_0, v)\right)^2, \tag{2}$$

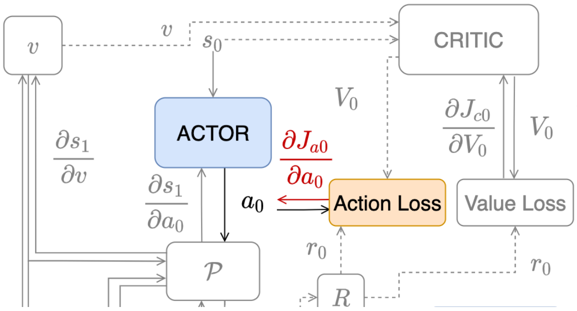
so the final equation used to constrain the critic at time $T = 0$ is

$$\frac{\partial J_{c0}}{\partial V_0} = -2\left(r_0 + \mu_\phi(s_1, v) - \mu_\phi(s_0, v)\right). \tag{3}$$

The action loss is used to constrain the parameters of the robot $v$ and the weights of the actor network, $\theta$. However, for the first time step there is no differentiable pathway from the loss to $v$ and only a single step derivative of $J_{a0}$, $\frac{\partial J_{a0}}{\partial a_0}$, as shown in Figure 2. It is know that

$$J_{a0} = -\hat{A}_0 log(\pi_\theta(a_0|s_0)) - H_0(\pi_\theta(a_0|s_0)). \tag{4}$$

We take the partial derivative of this w.r.t. the action $a_0$ by applying the product rule to the first term and noting that the differential of entropy is $log(\pi_\theta(a_0|s_0)) + 1$:



$v$
$v$
$s_0$
CRITIC
ACTOR
$V_0$
$\frac{\partial J_{c0}}{\partial V_0}$
$V_0$
$T = 0$
$\frac{\partial s_1}{\partial v}$
$\frac{\partial J_{a0}}{\partial a_0}$
$\frac{\partial s_1}{\partial a_0}$
$a_0$
Action Loss
Value Loss
$r_0$
$r_0$
$\mathcal{P}$
$R$

**Fig. 2:** Derivative path for first actor update.

$$\frac{\partial J_{a0}}{\partial a_0} = -\left[\frac{\partial \hat{A}_0}{\partial a_0} log(\pi_\theta(a_0|s_0)) + \hat{A}_0 \frac{\partial log(\pi_\theta(a_0|s_0))}{\partial a_0}\right] - \left[log(\pi_\theta(a_0|s_0)) + 1\right] \tag{5}$$

We additionally note that since the advantage is defined as

$$\hat{A}_t = \sum_t^T \gamma^t r_t - V(s_t, v), \tag{6}$$

and the reward $r_t$ is non-differentiable, $\frac{\partial \hat{A}_0}{\partial a_0} = 0$. This leaves us with the final equation

$$\frac{\partial J_{a0}}{\partial a_0} = -\frac{\hat{A}_0}{\pi_\theta(a_0|s_0)} - log(\pi_\theta(a_0|s_0)) - 1, \tag{7}$$

which is used to constrain the actor network at $T = 0$.

### T=1

At this time step there exists still only one derivative path for the critic update (as seen in Figure 3) giving the same output as equation 3 but for the current time step ($\frac{\partial J_{c1}}{\partial V_1}$),

$$J_{C1} = \left(\gamma^1 r_1 + \mu_\phi(s_2, v) - \mu_\phi(s_1, v)\right)^2, \tag{8}$$

$$\frac{\partial J_{c1}}{\partial V_1} = -2\left(\gamma r_1 + \mu_\phi(s_2, v) - \mu_\phi(s_1, v)\right). \tag{9}$$

There also exists a second path for $\frac{\partial J_{a1}}{\partial a_1}$, similar to that in equation 7. A visualisation of this can be seen in Figure 4. The derivative can then be defined for this time step,

$$\frac{\partial J_{a1}}{\partial a_1} = -\frac{\hat{A}_1}{\pi_\theta(a_1|s_1)} - log(\pi_\theta(a_1|s_1)) - 1 \tag{10}$$

There also now exists a non-trivial path for $\frac{\partial J_{a1}}{\partial v}$, shown in Figure 5 and defined as,

$$\frac{\partial J_{a1}}{\partial v} = \frac{\partial J_{a1}}{\partial a_1}\frac{\partial a_1}{\partial s_1}\frac{\partial s_1}{\partial v}. \tag{11}$$

*This work was supported by Surrey Satellite Technology Ltd.

[1] University of Surrey, GU2. l.jackson@surrey.ac.uk, s.hadfield@surrey.ac.uk, celyn.walters@surrey.ac.uk

[2]Surrey Satellite Technology Limited. Tycho House, GU2. s.eckersley@sstl.co.uk, p.senior@sstl.co.uk

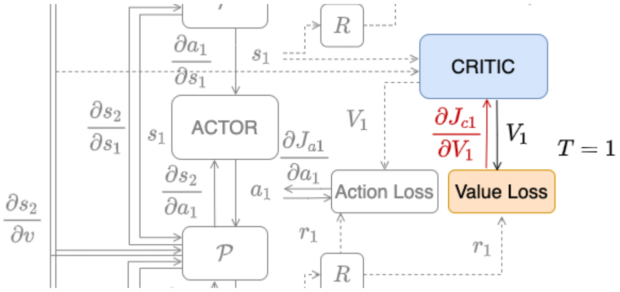


**Fig. 3:** Critic update for $T = 1$

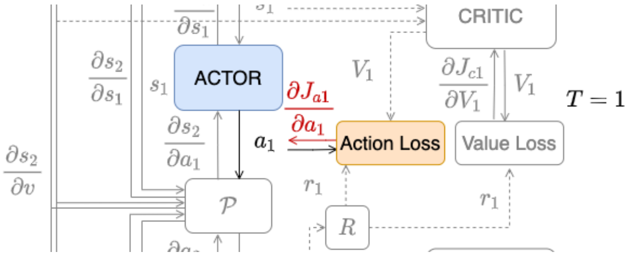


**Fig. 4:** Derivative path for second actor update.

It is here that the importance of the differentiable transition function becomes apparent, otherwise there would be no information flow from the action loss to the morphology parameter ($v$). Note that dotted lines represent non-differentiable paths.

The first term in equation 11 has already been defined in equation 10. The second term ($\frac{\partial a1}{\partial s1}$) is network specific, and can be easily obtained via auto-differentiation across the entire network w.r.t. the network inputs rather than the network parameters. The final term $\frac{\partial s1}{\partial v}$ must be obtained from the differentiable simulator and is specific to the environment being used.

## T=2

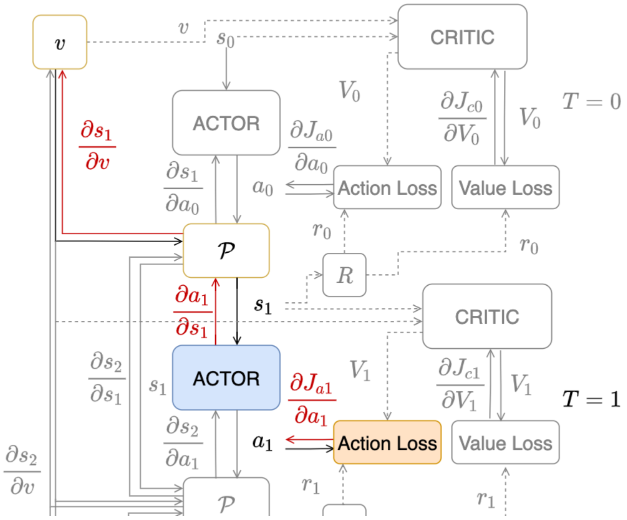


**Fig. 5:** Derivative path for morphology constraint at T=1.

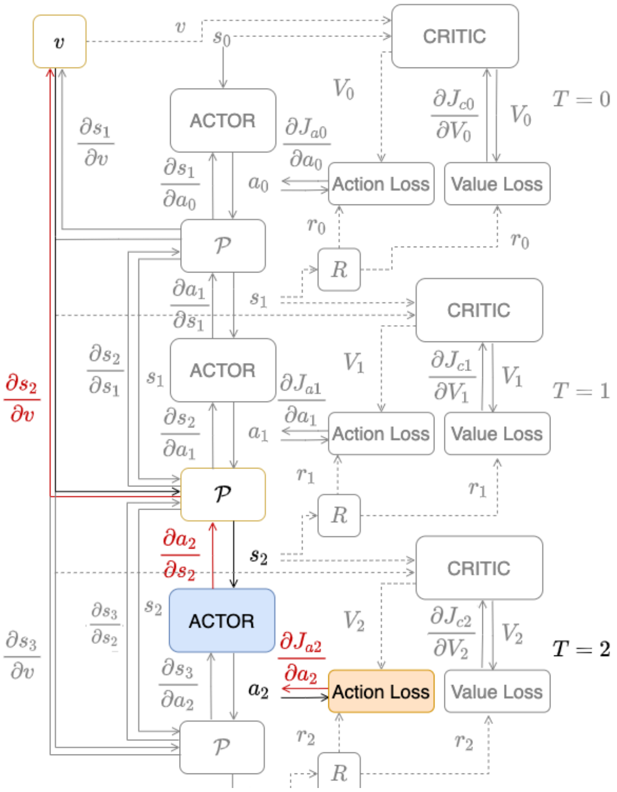


**Fig. 6:** First differential path to constrain $v$ at $T = 2$

The equations used to constrain the actor and critic networks, $\frac{\partial J_{a2}}{\partial a_2}$ and $\frac{\partial J_{c2}}{\partial V_2}$ respectively at this time step can be found by re-defining equation 10 and 3 for the current time step (omitted here for brevity). However, there now exists a complex combination of paths for information flow from the actor loss to $v$. The first path is highlighted in Figure 6 and mirrors that used in the previous time step,

$$\frac{\partial J_{a2}}{\partial v} = \frac{\partial J_{a2}}{\partial a_2}\frac{\partial a_2}{\partial s_2}\frac{\partial s_2}{\partial v}. \tag{12}$$

The second allows for gradients to back propagate through time, shown in Figure 7 and defined as,

$$\frac{\partial J_{a2}}{\partial v} = \frac{\partial J_{a2}}{\partial a_2}\frac{\partial a_2}{\partial s_2}\frac{\partial s_2}{\partial a_1}\frac{\partial a_1}{\partial s_1}\frac{\partial s_1}{\partial v}. \tag{13}$$

The last potential path of information flow facilities transfer through only the state transition function. This can be seen in Figure 8 and is defined as,

$$\frac{\partial J_{a2}}{\partial v} = \frac{\partial J_{a2}}{\partial a_2}\frac{\partial a_2}{\partial s_2}\frac{\partial s_2}{\partial s_1}\frac{\partial s_1}{\partial v}. \tag{14}$$

Combining the possible derivative paths gives the final update used at $T = 2$ for $v$:

$$\frac{\partial J_{a2}}{\partial v} = \frac{\partial J_{a_2}}{\partial a_2}\frac{\partial a_2}{\partial s_2}\frac{\partial s_2}{\partial v} + \frac{\partial J_{a_2}}{\partial a_2}\frac{\partial a_2}{\partial s_2}\left(\frac{\partial s_2}{\partial s_1} + \frac{\partial s_2}{\partial a_1}\frac{\partial a_1}{\partial s_1}\right)\frac{\partial s_1}{\partial v} \tag{15}$$

It is important to notice that this update applies only if $J_{a2}$ is used to constrain $v$ at $T = 2$, if $J_{a1}$ is also used then the derivative path becomes

$$\begin{aligned}\frac{\partial J_{a_1} + \partial J_{a_2}}{\partial v} = {} & \frac{\partial J_{a_1}}{\partial a_1}\frac{\partial a_1}{\partial s_1}\frac{\partial s_1}{\partial v} + \frac{\partial J_{a_2}}{\partial a_2}\frac{\partial a_2}{\partial s_2}\frac{\partial s_2}{\partial v} \\ & + \frac{\partial J_{a_2}}{\partial a_2}\frac{\partial a_2}{\partial s_2}\left(\frac{\partial s_2}{\partial s_1} + \frac{\partial s_2}{\partial a_1}\frac{\partial a_1}{\partial s_1}\right)\frac{\partial s_1}{\partial v}.\end{aligned} \tag{16}$$

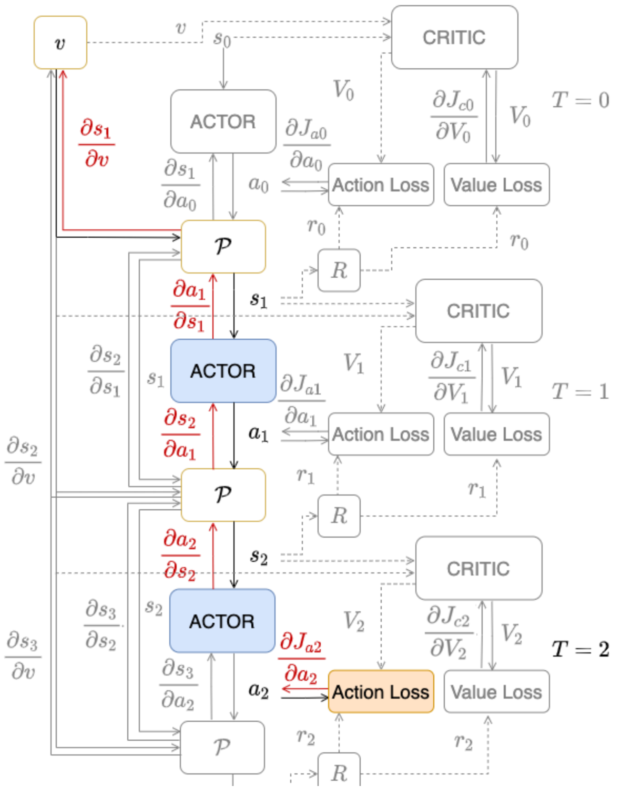

$v$
$v$
$s_0$
CRITIC
ACTOR
$V_0$
$\frac{\partial J_{c0}}{\partial V_0}$
$V_0$
$T=0$
$\frac{\partial s_1}{\partial v}$
$\frac{\partial J_{a0}}{\partial a_0}$
$\frac{\partial s_1}{\partial a_0}$
$a_0$
Action Loss
Value Loss
$\mathcal{P}$
$r_0$
$r_0$
$R$
$\frac{\partial a_1}{\partial s_1}$
$s_1$
CRITIC
$\frac{\partial s_2}{\partial s_1}$
$s_1$
ACTOR
$V_1$
$\frac{\partial J_{a1}}{\partial a_1}$
$\frac{\partial J_{c1}}{\partial V_1}$
$V_1$
$T=1$
$\frac{\partial s_2}{\partial v}$
$\frac{\partial s_2}{\partial a_1}$
$a_1$
Action Loss
Value Loss
$r_1$
$r_1$
$\mathcal{P}$
$R$
$\frac{\partial a_2}{\partial s_2}$
$s_2$
CRITIC
$\frac{\partial s_3}{\partial s_2}$
$s_2$
ACTOR
$V_2$
$\frac{\partial J_{c2}}{\partial V_2}$
$V_2$
$\frac{\partial s_3}{\partial v}$
$\frac{\partial J_{a2}}{\partial a_2}$
$T=2$
$\frac{\partial s_3}{\partial a_2}$
$a_2$
Action Loss
Value Loss
$r_2$
$r_2$
$\mathcal{P}$
$R$

**Fig. 7:** Second differential path to constrain $v$ at $T = 2$

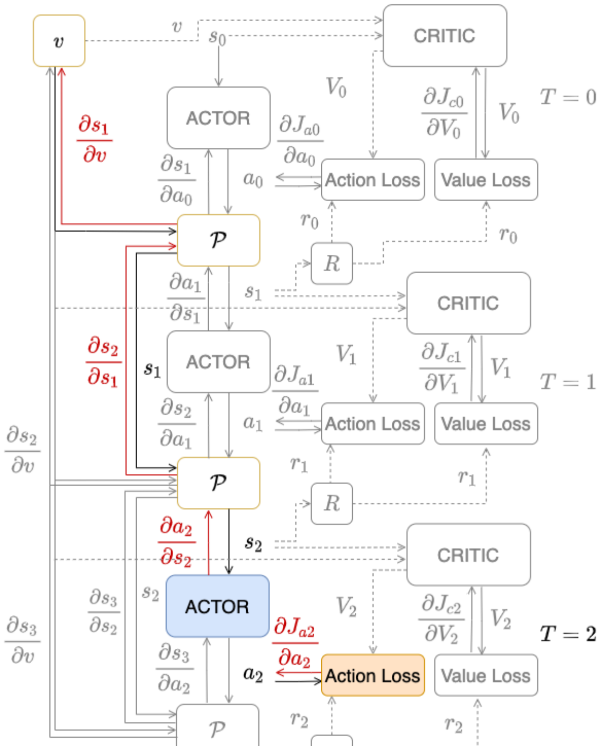

$v$
$v$
$s_0$
CRITIC
ACTOR
$V_0$
$\frac{\partial J_{c0}}{\partial V_0}$
$V_0$
$T=0$
$\frac{\partial s_1}{\partial v}$
$\frac{\partial J_{a0}}{\partial a_0}$
$\frac{\partial s_1}{\partial a_0}$
$a_0$
Action Loss
Value Loss
$r_0$
$r_0$
$\mathcal{P}$
$R$
$\frac{\partial a_1}{\partial s_1}$
$s_1$
CRITIC
$\frac{\partial s_2}{\partial s_1}$
$s_1$
ACTOR
$V_1$
$\frac{\partial J_{a1}}{\partial a_1}$
$\frac{\partial J_{c1}}{\partial V_1}$
$V_1$
$T=1$
$\frac{\partial s_2}{\partial v}$
$\frac{\partial s_2}{\partial a_1}$
$a_1$
Action Loss
Value Loss
$r_1$
$r_1$
$\mathcal{P}$
$R$
$\frac{\partial a_2}{\partial s_2}$
$s_2$
CRITIC
$\frac{\partial s_3}{\partial s_2}$
$s_2$
ACTOR
$V_2$
$\frac{\partial J_{c2}}{\partial V_2}$
$V_2$
$\frac{\partial s_3}{\partial v}$
$\frac{\partial J_{a2}}{\partial a_2}$
$T=2$
$\frac{\partial s_3}{\partial a_2}$
$a_2$
Action Loss
Value Loss
$r_2$
$r_2$
$\mathcal{P}$

**Fig. 8:** Last differential path to constrain $v$ at $T = 2$

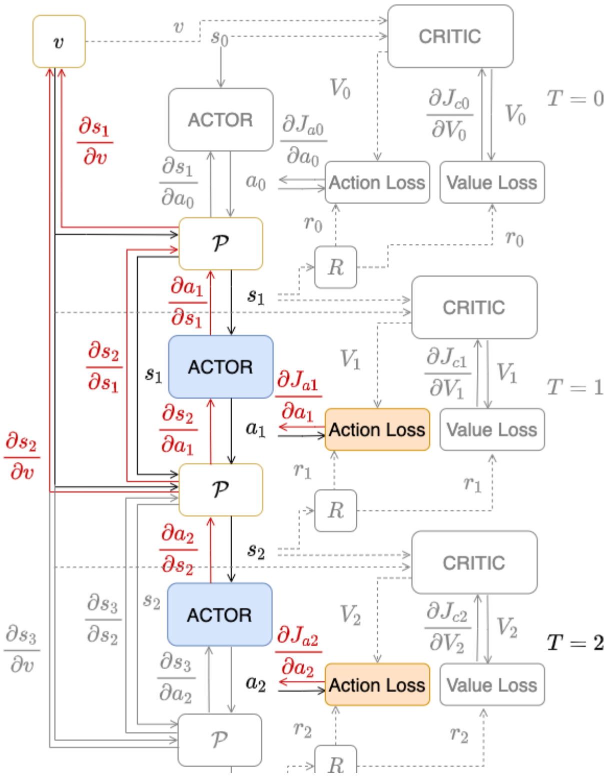
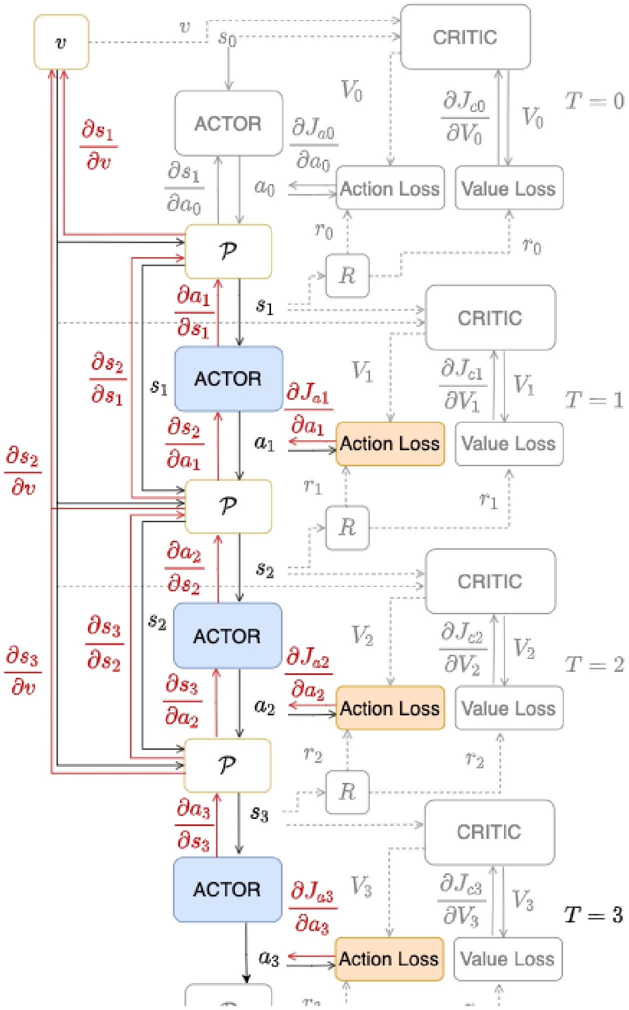

$v$
$v$
$s_0$
CRITIC
ACTOR
$V_0$
$\frac{\partial J_{c0}}{\partial V_0}$
$V_0$
$T=0$
$\frac{\partial s_1}{\partial v}$
$\frac{\partial J_{a0}}{\partial a_0}$
$\frac{\partial s_1}{\partial a_0}$
$a_0$
Action Loss
Value Loss
$r_0$
$r_0$
$\mathcal{P}$
$R$
$\frac{\partial a_1}{\partial s_1}$
$s_1$
CRITIC
$\frac{\partial s_2}{\partial s_1}$
$s_1$
ACTOR
$V_1$
$\frac{\partial J_{c1}}{\partial V_1}$
$V_1$
$T=1$
$\frac{\partial J_{a1}}{\partial a_1}$
$\frac{\partial s_2}{\partial v}$
$\frac{\partial s_2}{\partial a_1}$
$a_1$
Action Loss
Value Loss
$r_1$
$r_1$
$\mathcal{P}$
$R$
$\frac{\partial a_2}{\partial s_2}$
$s_2$
CRITIC
$\frac{\partial s_3}{\partial s_2}$
$s_2$
ACTOR
$V_2$
$\frac{\partial J_{c2}}{\partial V_2}$
$V_2$
$T=2$
$\frac{\partial s_3}{\partial v}$
$\frac{\partial J_{a2}}{\partial a_2}$
$\frac{\partial s_3}{\partial a_2}$
$a_2$
Action Loss
Value Loss
$r_2$
$r_2$
$\mathcal{P}$
$R$

**Fig. 9:** Full path of information flow for the morphology update at $T = 2$

This full path is shown in Figure 9.

### T=3

As with $T = 0$, $T = 1$ and $T = 2$ the derivative used to constrain the critic network parameters remains the same at $T = 3$, but defined for the this time step. The same theory applies to the actor network parameter update, as with the standard RL pipelines. However, there now exists a complex path for information to flow on the backpass to allow for informed updates to $v$. The full path for $T = 3$ can be seen in Figure 10. A brief comparison of this figure and Figure 9 shows each new time step introduces exponentially more differential paths. In the case of $T = 3$ there are 7 potential paths for $\frac{\partial J_{a3}}{\partial v}$, without the inclusion of $J_{a2}$ and $J_{a1}$. This increase leads to high computation demands. As a result there exists a compact version of ORCHID where gradients are detached in between time steps. This leads to a derivative path similar to that shown in Figure 5 and 6 existing at each time step. This limits the memory demands throughout training.



$v$
$v$
$s_0$
CRITIC
ACTOR
$V_0$
$\frac{\partial J_{c0}}{\partial V_0}$
$V_0$
$T = 0$
$\frac{\partial s_1}{\partial v}$
$\frac{\partial J_{a0}}{\partial a_0}$
$\frac{\partial s_1}{\partial a_0}$
$a_0$
Action Loss
Value Loss
$r_0$
$r_0$
$\mathcal{P}$
$R$
$\frac{\partial a_1}{\partial s_1}$
$s_1$
CRITIC
$\frac{\partial s_2}{\partial s_1}$
ACTOR
$s_1$
$V_1$
$\frac{\partial J_{c1}}{\partial V_1}$
$V_1$
$T = 1$
$\frac{\partial J_{a1}}{\partial a_1}$
$\frac{\partial s_2}{\partial v}$
$\frac{\partial s_2}{\partial a_1}$
$a_1$
Action Loss
Value Loss
$r_1$
$r_1$
$\mathcal{P}$
$R$
$\frac{\partial a_2}{\partial s_2}$
$s_2$
CRITIC
ACTOR
$s_2$
$\frac{\partial s_3}{\partial s_2}$
$V_2$
$\frac{\partial J_{c2}}{\partial V_2}$
$V_2$
$T = 2$
$\frac{\partial s_3}{\partial v}$
$\frac{\partial J_{a2}}{\partial a_2}$
$\frac{\partial s_3}{\partial a_2}$
$a_2$
Action Loss
Value Loss
$r_2$
$r_2$
$\mathcal{P}$
$R$
$\frac{\partial a_3}{\partial s_3}$
$s_3$
CRITIC
ACTOR
$V_3$
$\frac{\partial J_{a3}}{\partial a_3}$
$\frac{\partial J_{c3}}{\partial V_3}$
$V_3$
$T = 3$
$a_3$
Action Loss
Value Loss

**Fig. 10:** Full derivative path for $T = 3$