Generalized Advantage Estimation

Actor-Critic with Eligibility Traces

Reinforcement Learning

Sai Sampath Kedari
sampath@umich.edu

Contents

1 From Advantage Estimation to GAE

The advantage function is

$$A^{\pi_\theta}(S_t, A_t) = Q^{\pi_\theta}(S_t, A_t) - V^{\pi_\theta}(S_t) = \mathbb{E}^{\pi_\theta}[G_t \mid S_t, A_t] - V^{\pi_\theta}(S_t).$$

Since $V^{\pi_\theta}(S_t)$ is fixed under conditioning on (S_t, A_t),

$$A^{\pi_\theta}(S_t, A_t) = \mathbb{E}^{\pi_\theta}[G_t - V^{\pi_\theta}(S_t) \mid S_t, A_t]\,.$$

1.1 One-Step Advantage Estimator

$$Q^{\pi_\theta}(S_t, A_t) = \mathbb{E}^{\pi_\theta}[R_{t+1} + \gamma V^{\pi_\theta}(S_{t+1}) \mid S_t, A_t] \qquad \text{; Bellman equation for } Q^{\pi_\theta}$$

$$A^{\pi_\theta}(S_t, A_t) = \mathbb{E}^{\pi_\theta}[R_{t+1} + \gamma V^{\pi_\theta}(S_{t+1}) - V^{\pi_\theta}(S_t) \mid S_t, A_t] \qquad \text{; subtract } V^{\pi_\theta}(S_t)$$

Define the one-step TD advantage estimator

$$\delta_t := R_{t+1} + \gamma V^{\pi_\theta}(S_{t+1}) - V^{\pi_\theta}(S_t), \qquad A^{\pi_\theta}(S_t, A_t) = \mathbb{E}^{\pi_\theta}[\delta_t \mid S_t, A_t]\,.$$

1.2 n-Step Advantage Estimators

For $n \geq 1$, define the n-step bootstrapped return

$$G_t^{(n)} := \sum_{k=0}^{n-1} \gamma^k R_{t+k+1} + \gamma^n V^{\pi_\theta}(S_{t+n}).$$

Then

$$A^{\pi_\theta}(S_t, A_t) = \mathbb{E}^{\pi_\theta}\left[G_t^{(n)} - V^{\pi_\theta}(S_t) \mid S_t, A_t\right] \qquad \text{; } n\text{-step Bellman expansion}$$

$$= \mathbb{E}^{\pi_\theta}\left[\sum_{k=0}^{n-1} \gamma^k R_{t+k+1} + \gamma^n V^{\pi_\theta}(S_{t+n}) - V^{\pi_\theta}(S_t) \;\middle|\; S_t, A_t\right]. \qquad \text{; expand } G_t^{(n)}$$

Define the n-step advantage estimator

$$\delta_t^{(n)} := \sum_{k=0}^{n-1} \gamma^k R_{t+k+1} + \gamma^n V^{\pi_\theta}(S_{t+n}) - V^{\pi_\theta}(S_t), \qquad A^{\pi_\theta}(S_t, A_t) = \mathbb{E}^{\pi_\theta}\left[\delta_t^{(n)} \mid S_t, A_t\right].$$

With the true value function V^{π_θ}, every n-step estimator is conditionally unbiased for the advantage.

1.3 The Family of Estimators

$$
\begin{array}{lll}
n=1 \ \text{(TD)} & \delta_t^{(1)} = R_{t+1} + \gamma V^{\pi_\theta}(S_{t+1}) - V^{\pi_\theta}(S_t) & \textbf{TD(0)} \\
n=2 & \delta_t^{(2)} = R_{t+1} + \gamma R_{t+2} + \gamma^2 V^{\pi_\theta}(S_{t+2}) - V^{\pi_\theta}(S_t) & \text{low variance} \\
n=3 & \delta_t^{(3)} = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 V^{\pi_\theta}(S_{t+3}) - V^{\pi_\theta}(S_t) & \text{high bias} \\
\vdots & \vdots & \downarrow \\
n & \delta_t^{(n)} = \sum_{k=0}^{n-1} \gamma^k R_{t+k+1} + \gamma^n V^{\pi_\theta}(S_{t+n}) - V^{\pi_\theta}(S_t) & \\
\vdots & \vdots & \text{high variance} \\
 & & \text{low bias} \\
n=\infty \ \text{(MC)} & \delta_t^{(\infty)} = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} - V^{\pi_\theta}(S_t) = G_t - V^{\pi_\theta}(S_t) & \textbf{MC}
\end{array}
$$

Here "bias" refers to the bias induced once the oracle V^{π_θ} is replaced by a learned critic (Section 1.6): small n leans harder on the bootstrap, large n leans harder on the sampled return.

1.4 n-Step TD Error Telescoping

Lemma 1.1.

$$
\delta_t^{(n)} = \sum_{k=0}^{n-1} \gamma^k \delta_{t+k}, \qquad \delta_{t+k} = R_{t+k+1} + \gamma V^{\pi_\theta}(S_{t+k+1}) - V^{\pi_\theta}(S_{t+k}).
$$

Proof.

$$
\begin{aligned}
\sum_{k=0}^{n-1} \gamma^k \delta_{t+k} &= \sum_{k=0}^{n-1} \gamma^k \left[R_{t+k+1} + \gamma V^{\pi_\theta}(S_{t+k+1}) - V^{\pi_\theta}(S_{t+k}) \right] && \text{; substitute } \delta_{t+k} \\
&= \sum_{k=0}^{n-1} \gamma^k R_{t+k+1} + \sum_{k=0}^{n-1} \gamma^{k+1} V^{\pi_\theta}(S_{t+k+1}) - \sum_{k=0}^{n-1} \gamma^k V^{\pi_\theta}(S_{t+k}) && \text{; split sums} \\
&= \sum_{k=0}^{n-1} \gamma^k R_{t+k+1} + \sum_{j=1}^{n} \gamma^j V^{\pi_\theta}(S_{t+j}) - \sum_{j=0}^{n-1} \gamma^j V^{\pi_\theta}(S_{t+j}) && \text{; reindex } j = k+1 \\
&= \sum_{k=0}^{n-1} \gamma^k R_{t+k+1} + \gamma^n V^{\pi_\theta}(S_{t+n}) - V^{\pi_\theta}(S_t) && \text{; telescope} \\
&= \delta_t^{(n)}. && \text{; definition}
\end{aligned}
$$

□

An n-step advantage estimator is a discounted sum of one-step TD errors from t to $t+n-1$.

1.5 Geometric Mixture of the n-Step Estimators

Rather than commit to a single n, take a linear combination of *all* the estimators, weighted by the geometric distribution $(1-\lambda)\lambda^{n-1}$ of Figure 2.

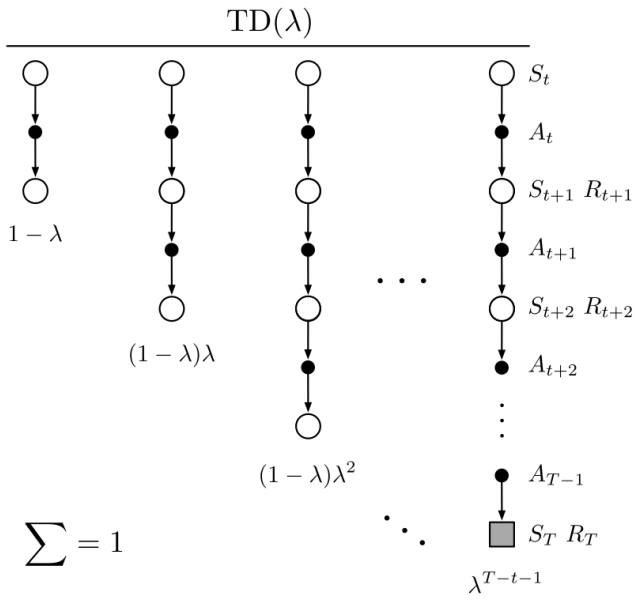

Figure 1

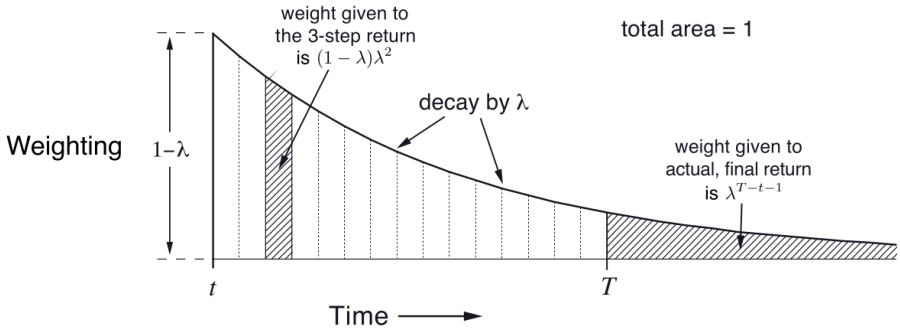

Figure 2

$$
\begin{aligned}
& (1-\lambda) && \delta_t^{(1)} = R_{t+1} + \gamma V^{\pi_\theta}(S_{t+1}) - V^{\pi_\theta}(S_t) \\
+\ & (1-\lambda)\,\lambda && \delta_t^{(2)} = R_{t+1} + \gamma R_{t+2} + \gamma^2 V^{\pi_\theta}(S_{t+2}) - V^{\pi_\theta}(S_t) \\
+\ & (1-\lambda)\,\lambda^2 && \delta_t^{(3)} = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 V^{\pi_\theta}(S_{t+3}) - V^{\pi_\theta}(S_t) \\
& + && \vdots \qquad\qquad \vdots \\
+\ & (1-\lambda)\,\lambda^{n-1} && \delta_t^{(n)} = \sum_{k=0}^{n-1} \gamma^k R_{t+k+1} + \gamma^n V^{\pi_\theta}(S_{t+n}) - V^{\pi_\theta}(S_t) \\
& + && \vdots \qquad\qquad \vdots \\
\hline
A_t^{\mathrm{GAE}(\gamma,\lambda)} & := && (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1}\, \delta_t^{(n)}
\end{aligned}
$$

Using the telescoping identity,

$$
\begin{aligned}
A_t^{\mathrm{GAE}(\gamma,\lambda)} &= (1-\lambda)\sum_{n=1}^{\infty}\lambda^{n-1}\sum_{k=0}^{n-1}\gamma^k\delta_{t+k} && \text{; replace } \delta_t^{(n)} \\
&= (1-\lambda)\sum_{k=0}^{\infty}\gamma^k\delta_{t+k}\sum_{n=k+1}^{\infty}\lambda^{n-1} && \text{; swap sums, collect fixed } k \\
&= (1-\lambda)\sum_{k=0}^{\infty}\gamma^k\delta_{t+k}\,\frac{\lambda^k}{1-\lambda} && \text{; } \textstyle\sum_{n=k+1}^{\infty}\lambda^{n-1}=\frac{\lambda^k}{1-\lambda} \\
&= \sum_{k=0}^{\infty}(\gamma\lambda)^k\delta_{t+k}. && \text{; cancel } (1-\lambda)
\end{aligned}
$$

Therefore

$$A_t^{\mathrm{GAE}(\gamma,\lambda)} = G_t^{\lambda} - V^{\pi_\theta}(S_t) = \sum_{k=0}^{\infty}(\gamma\lambda)^k\delta_{t+k}.$$

The identity $G_t^{\lambda} - V^{\pi_\theta}(S_t) = \sum_{k=0}^{\infty}(\gamma\lambda)^k\delta_{t+k}$ was proven in Report 3.

1.6 Replacing the Oracle Value by a Critic

Every estimator above is exactly unbiased for A^{π_θ} when it uses the true value V^{π_θ}. Bias enters only when V^{π_θ} is replaced by a learned critic, $V^{\pi_\theta}(s) \rightsquigarrow \hat{v}(s,w)$. Define the approximate one-step TD error

$$\hat{\delta}_t = R_{t+1} + \gamma\hat{v}(S_{t+1},w) - \hat{v}(S_t,w),$$

and the practical GAE estimator, over an infinite and a length-T rollout,

$$\hat{A}_t^{\mathrm{GAE}(\gamma,\lambda)} = \sum_{k=0}^{\infty}(\gamma\lambda)^k\hat{\delta}_{t+k}, \qquad \hat{A}_t^{\mathrm{GAE}(\gamma,\lambda)} = \sum_{k=0}^{T-t-1}(\gamma\lambda)^k\hat{\delta}_{t+k}.$$

Here λ controls how much of the critic's bias is propagated against how much Monte-Carlo variance is admitted.

1.7 Endpoint Cases

With the critic form, the two extremes of λ are

$$\lambda = 0: \quad \hat{A}_t^{\mathrm{GAE}(\gamma,0)} = \hat{\delta}_t \quad \text{(one-step TD advantage)},$$

$$\lambda = 1: \quad \hat{A}_t^{\mathrm{GAE}(\gamma,1)} = \sum_{k=0}^{\infty}\gamma^k\hat{\delta}_{t+k} = G_t - \hat{v}(S_t,w) \quad \text{(Monte-Carlo advantage)}.$$

2 From the Update to Eligibility Traces

2.1 The Update with Each Advantage Estimator

The one-step actor-critic of Report 16 updates actor and critic with the shared scalar $\hat{\delta}_t$:

$$\theta_{t+1} = \theta_t + \alpha_t^{\theta}\,\hat{\delta}_t\,\nabla_\theta \log \pi_{\theta_t}(A_t \mid S_t),$$

$$w_{t+1} = w_t + \alpha_t^{w}\,\hat{\delta}_t\,\nabla_w \hat{v}(S_t,w_t).$$

$$\Downarrow \qquad \text{replace the multiplier} \;\; \hat{\delta}_t \rightsquigarrow \hat{A}_t^{\mathrm{GAE}}$$

$$\theta_{t+1} = \theta_t + \alpha_t^{\theta}\, \hat{A}_t^{\mathrm{GAE}}\, \nabla_\theta \log \pi_{\theta_t}(A_t \mid S_t),$$
$$w_{t+1} = w_t + \alpha_t^{w}\, \hat{A}_t^{\mathrm{GAE}}\, \nabla_w \hat{v}(S_t, w_t).$$

$$\Downarrow \qquad \text{expand} \;\; \hat{A}_t^{\mathrm{GAE}} = \textstyle\sum_{k=0}^{\infty} (\gamma\lambda)^k \hat{\delta}_{t+k}$$

$$\theta_{t+1} = \theta_t + \alpha_t^{\theta} \Big[\sum_{k=0}^{\infty} (\gamma\lambda)^k \hat{\delta}_{t+k} \Big] \nabla_\theta \log \pi_{\theta_t}(A_t \mid S_t),$$

$$w_{t+1} = w_t + \alpha_t^{w} \Big[\sum_{k=0}^{\infty} (\gamma\lambda)^k \hat{\delta}_{t+k} \Big] \nabla_w \hat{v}(S_t, w_t).$$

Each update at time t needs the *future* errors $\hat{\delta}_t, \hat{\delta}_{t+1}, \ldots$, so it cannot be applied online as written. The tables below reorganize the same total update by the time each $\hat{\delta}$ arrives.

3 Forward View

Laid out as a table, row τ is one fixed gradient, the columns are the errors $\hat{\delta}_1, \hat{\delta}_2, \ldots$, and entry (τ, t) is $(\gamma\lambda)^{t-\tau}\hat{\delta}_t$ for $t \geq \tau$, else 0. The **row sum** is the forward update: a fixed gradient collects all its future errors.

3.1 Actor

	$\hat{\delta}_1$		$\hat{\delta}_2$		$\hat{\delta}_3$	$\cdots$	**row sum** $= \Delta\theta_\tau/\alpha^\theta$
$\nabla_\theta \log \pi_{\theta_1}$	$\hat{\delta}_1$	$+$	$\gamma\lambda\,\hat{\delta}_2$	$+$	$(\gamma\lambda)^2\hat{\delta}_3$	$\cdots$	$\hat{A}_1^{\mathrm{GAE}}\,\nabla_\theta \log \pi_{\theta_1}$
$\nabla_\theta \log \pi_{\theta_2}$	0		$\hat{\delta}_2$	$+$	$\gamma\lambda\,\hat{\delta}_3$	$\cdots$	$\hat{A}_2^{\mathrm{GAE}}\,\nabla_\theta \log \pi_{\theta_2}$
$\nabla_\theta \log \pi_{\theta_3}$	0		0		$\hat{\delta}_3$	$\cdots$	$\hat{A}_3^{\mathrm{GAE}}\,\nabla_\theta \log \pi_{\theta_3}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\ddots$	$\vdots$

$$\Delta\theta_\tau = \alpha^\theta \Big[\textstyle\sum_{t\geq\tau} (\gamma\lambda)^{t-\tau} \hat{\delta}_t \Big] \nabla_\theta \log \pi_{\theta_\tau}(A_\tau \mid S_\tau) = \alpha^\theta\, \hat{A}_\tau^{\mathrm{GAE}}\, \nabla_\theta \log \pi_{\theta_\tau}(A_\tau \mid S_\tau).$$

3.2 Critic

	$\hat{\delta}_1$		$\hat{\delta}_2$		$\hat{\delta}_3$	$\cdots$	**row sum** $= \Delta w_\tau / \alpha^w$
$\nabla_w \hat{v}(S_1)$	$\hat{\delta}_1$	$+$	$\gamma\lambda\,\hat{\delta}_2$	$+$	$(\gamma\lambda)^2\hat{\delta}_3$	$\cdots$	$\hat{A}_1^{\mathrm{GAE}}\,\nabla_w \hat{v}(S_1)$
$\nabla_w \hat{v}(S_2)$	0		$\hat{\delta}_2$	$+$	$\gamma\lambda\,\hat{\delta}_3$	$\cdots$	$\hat{A}_2^{\mathrm{GAE}}\,\nabla_w \hat{v}(S_2)$
$\nabla_w \hat{v}(S_3)$	0		0		$\hat{\delta}_3$	$\cdots$	$\hat{A}_3^{\mathrm{GAE}}\,\nabla_w \hat{v}(S_3)$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\ddots$	$\vdots$

$$\Delta w_\tau = \alpha^w \Big[\textstyle\sum_{t\geq\tau}(\gamma\lambda)^{t-\tau}\hat{\delta}_t\Big]\nabla_w \hat{v}(S_\tau, w_\tau) = \alpha^w\, \hat{A}_\tau^{\mathrm{GAE}}\, \nabla_w \hat{v}(S_\tau, w_\tau).$$

4 Backward View

The forward tables wait for future errors. But entry $(\tau, t) = (\gamma\lambda)^{t-\tau}$ depends only on how long ago S_τ was visited, not on the future: the whole column $\hat{\delta}_t$ is known the instant $\hat{\delta}_t$ is computed. So process the table column by column instead of row by row. When $\hat{\delta}_t$ arrives, apply its contribution to every past gradient at once. The cells now carry the gradient, and the **column sum** is the online update.

4.1 Actor

	$\hat{\delta}_1$	$\hat{\delta}_2$	$\hat{\delta}_3$ $\cdots$
$\nabla_\theta \log \pi_{\theta_1}$	$\nabla_\theta \log \pi_{\theta_1}$	$\gamma\lambda\,\nabla_\theta \log \pi_{\theta_1}$	$(\gamma\lambda)^2\nabla_\theta \log \pi_{\theta_1}$
	$+$	$+$	$+$
$\nabla_\theta \log \pi_{\theta_2}$	0	$\nabla_\theta \log \pi_{\theta_2}$	$\gamma\lambda\,\nabla_\theta \log \pi_{\theta_2}$
		$+$	$+$
$\nabla_\theta \log \pi_{\theta_3}$	0	0	$\nabla_\theta \log \pi_{\theta_3}$
column sum z_t^θ	z_1^θ	z_2^θ	z_3^θ

The column sum is the *actor trace* z_t^θ, and the update from $\hat{\delta}_t$ is

$$\Delta\theta \text{ from } \hat{\delta}_t = \alpha^\theta\, \hat{\delta}_t \underbrace{\sum_{\tau\leq t}(\gamma\lambda)^{t-\tau}\nabla_\theta \log \pi_{\theta_\tau}(A_\tau \mid S_\tau)}_{z_t^\theta} = \alpha^\theta\, \hat{\delta}_t\, z_t^\theta.$$

4.2 Critic

	$\hat{\delta}_1$	$\hat{\delta}_2$	$\hat{\delta}_3 \quad \cdots$
$\nabla_w \hat{v}(S_1)$	$\nabla_w \hat{v}(S_1)$	$\gamma\lambda\, \nabla_w \hat{v}(S_1)$	$(\gamma\lambda)^2 \nabla_w \hat{v}(S_1)$
	$+$	$+$	$+$
$\nabla_w \hat{v}(S_2)$	0	$\nabla_w \hat{v}(S_2)$	$\gamma\lambda\, \nabla_w \hat{v}(S_2)$
		$+$	$+$
$\nabla_w \hat{v}(S_3)$	0	0	$\nabla_w \hat{v}(S_3)$
column sum z_t^w	z_1^w	z_2^w	z_3^w

The column sum is the *critic trace* z_t^w, and the update from $\hat{\delta}_t$ is

$$\Delta w \text{ from } \hat{\delta}_t = \alpha^w\, \hat{\delta}_t \underbrace{\sum_{\tau \le t} (\gamma\lambda)^{t-\tau} \nabla_w \hat{v}(S_\tau, w_\tau)}_{z_t^w} = \alpha^w\, \hat{\delta}_t\, z_t^w.$$

5 Eligibility Traces

5.1 Definition

The trace is the current column: the discounted running sum of past gradients,

$$z_t^\theta = \sum_{\tau=0}^{t} (\gamma\lambda)^{t-\tau} \nabla_\theta \log \pi_{\theta_\tau}(A_\tau \mid S_\tau), \qquad z_t^w = \sum_{\tau=0}^{t} (\gamma\lambda)^{t-\tau} \nabla_w \hat{v}(S_\tau, w_\tau).$$

This is the FA analogue of the tabular trace: the state indicator $\mathbf{1}\{S_\tau = s\}$ of Report 3 is replaced by the gradient of the object being learned, the score for the actor and the value gradient for the critic.

5.2 The Recursion

Peel off the $\tau = t$ term (shown for the actor; the critic is identical):

$$
\begin{aligned}
z_t^\theta &= \sum_{\tau=0}^{t} (\gamma\lambda)^{t-\tau} \nabla_\theta \log \pi_{\theta_\tau}(A_\tau \mid S_\tau) \\
&= (\gamma\lambda)^0 \, \nabla_\theta \log \pi_{\theta_t}(A_t \mid S_t) + \sum_{\tau=0}^{t-1} (\gamma\lambda)^{t-\tau} \nabla_\theta \log \pi_{\theta_\tau}(A_\tau \mid S_\tau) && \text{; split off } \tau = t \\
&= \nabla_\theta \log \pi_{\theta_t}(A_t \mid S_t) + \gamma\lambda \sum_{\tau=0}^{t-1} (\gamma\lambda)^{(t-1)-\tau} \nabla_\theta \log \pi_{\theta_\tau}(A_\tau \mid S_\tau) && \text{; factor } \gamma\lambda \text{ from the tail} \\
&= \nabla_\theta \log \pi_{\theta_t}(A_t \mid S_t) + \gamma\lambda \, z_{t-1}^\theta. && \text{; tail is } z_{t-1}^\theta
\end{aligned}
$$

So each step decays the running sum by $\gamma\lambda$ and adds the current gradient:

$$
z_t^\theta = \gamma\lambda \, z_{t-1}^\theta + \nabla_\theta \log \pi_{\theta_t}(A_t \mid S_t), \qquad z_t^w = \gamma\lambda \, z_{t-1}^w + \nabla_w \hat{v}(S_t, w_t), \qquad z_{-1}^\theta = z_{-1}^w = 0.
$$

5.3 Endpoints

$\lambda = 0$ collapses the trace to the current gradient, $z_t^\theta = \nabla_\theta \log \pi_{\theta_t}$, recovering the one-step actor-critic of Report 16. $\lambda = 1$ keeps every past gradient, decayed only by γ, approaching the Monte-Carlo update.

5.4 The Online Update

At each step: observe the transition, compute $\hat{\delta}_t$, decay-and-add both traces, then apply the shared error to each,

$$
\begin{aligned}
\hat{\delta}_t &= R_{t+1} + \gamma \, \hat{v}(S_{t+1}, w_t) - \hat{v}(S_t, w_t), \\
z_t^\theta &= \gamma\lambda \, z_{t-1}^\theta + \nabla_\theta \log \pi_{\theta_t}(A_t \mid S_t), & \theta_{t+1} &= \theta_t + \alpha_t^\theta \, \hat{\delta}_t \, z_t^\theta, \\
z_t^w &= \gamma\lambda \, z_{t-1}^w + \nabla_w \hat{v}(S_t, w_t), & w_{t+1} &= w_t + \alpha_t^w \, \hat{\delta}_t \, z_t^w.
\end{aligned}
$$

5.5 Algorithm

Algorithm 1: Actor-Critic with GAE Eligibility Traces, Rollout Form

Input: a differentiable policy parameterization $\pi_\theta(a \mid s)$
Input: a differentiable state-value parameterization $\hat{v}(s, w)$
Input: step sizes $\alpha_t^\theta, \alpha_t^w > 0$; discount γ; trace decay λ
initialize policy parameter $\theta_0 \in \mathbb{R}^{d'}$ and state-value weights $w_0 \in \mathbb{R}^d$
repeat
  sample/reset $S_0 \sim \rho_0$
  $S_t \leftarrow S_0; \quad z^\theta \leftarrow 0; \quad z^w \leftarrow 0$
  for $t = 0, 1, \ldots, T-1$ *or until* S_t *terminal* **do**
    Sample $A_t \sim \pi_{\theta_t}(\cdot \mid S_t)$
    Observe $S_{t+1}, R_{t+1} \sim p(\cdot, \cdot \mid S_t, A_t)$ `// environment dynamics`
    $\hat{\delta}_t \leftarrow R_{t+1} + \gamma\, \hat{v}(S_{t+1}, w_t) - \hat{v}(S_t, w_t)$
    $z^\theta \leftarrow \gamma\lambda\, z^\theta + \nabla_\theta \log \pi_{\theta_t}(A_t \mid S_t)$
    $z^w \leftarrow \gamma\lambda\, z^w + \nabla_w \hat{v}(S_t, w_t)$
    $\theta_{t+1} \leftarrow \theta_t + \alpha_t^\theta\, \hat{\delta}_t\, z^\theta$
    $w_{t+1} \leftarrow w_t + \alpha_t^w\, \hat{\delta}_t\, z^w$
    $S_t \leftarrow S_{t+1}$
until *converged*

Setting $\lambda = 0$ zeroes both traces except their current term and recovers the one-step actor-critic of Report 16; $\lambda = 1$ accumulates the full discounted gradient history, giving the Monte-Carlo advantage.