# Actor-Critic with a Baseline

The Advantage

Reinforcement Learning

**Sai Sampath Kedari**
sampath@umich.edu

---

## Contents

# 1 Actor-Critic with a Baseline: The Advantage

## 1.1 Starting Point from QAC

Report 15 carried the exact policy gradient down to a per-step oracle estimator [report 15]:

**Key Result**

$$\widehat{g}_t^{\,Q,\mathrm{oracle}}(\theta) \;=\; \gamma^t\, Q^{\pi_\theta}(S_t, A_t)\, \nabla_\theta \log \pi_\theta(A_t \mid S_t).$$

This report changes only the scalar multiplier of the score:

$$Q^{\pi_\theta}(S_t, A_t) \quad \rightsquigarrow \quad Q^{\pi_\theta}(S_t, A_t) - b(S_t).$$

*A state-only baseline $b = b_\theta(S)$ is allowed; $b(S, A)$ is not generally allowed. A baseline never changes the mean; a good baseline reduces variance. The whole report is the search for the multiplier that keeps the gradient exact while shrinking its variance.*

The $\gamma^t$-weighted time-marginal machinery that produced this estimator is not repeated; it is identical to Report 15 with the multiplier carried through, and is recorded in Appendix A.

Everything downstream of the policy gradient theorem is a contest over one scalar: the multiplier on $\nabla_\theta \log \pi_\theta(A_t \mid S_t)$. Report 15 used $Q^{\pi_\theta}$. Its flaw is not bias but variance. $Q^{\pi_\theta}(S_t, A_t)$ carries the full return magnitude, most of which is shared across all actions at $S_t$ and says nothing about which action to prefer. This report subtracts that shared, action-independent part.

## 1.2 Baseline Policy-Gradient Identity

**Theorem 1.1** (Baseline Variance Reduction)**.** *For any state-only baseline $b(S)$, i.e. a function of the state but not the action,*

$$\begin{aligned}\nabla_\theta J(\theta) \;&=\; \frac{1}{1-\gamma}\, \mathbb{E}_{S\sim d^{\pi_\theta,\rho_0}}\Big[\, \mathbb{E}_{A\sim\pi_\theta(\cdot|S)}\big[\, Q^{\pi_\theta}(S, A)\, \nabla_\theta \log \pi_\theta(A \mid S)\,\big]\Big] \\ &=\; \frac{1}{1-\gamma}\, \mathbb{E}_{S\sim d^{\pi_\theta,\rho_0}}\Big[\, \mathbb{E}_{A\sim\pi_\theta(\cdot|S)}\big[\, \big(Q^{\pi_\theta}(S, A) - b(S)\big)\, \nabla_\theta \log \pi_\theta(A \mid S)\,\big]\Big].\end{aligned}$$

*Proof.* It suffices to show the baseline term contributes nothing:

$$\mathbb{E}_{A\sim\pi_\theta(\cdot|s)}\big[\, b(s)\, \nabla_\theta \log \pi_\theta(A \mid s)\,\big] \;=\; 0 \qquad \forall\, s.$$

Expanding the inner action expectation at a fixed state $s$,

$$\begin{aligned}
\mathbb{E}_{A\sim\pi_\theta(\cdot|s)}\big[\, b(s)\,\nabla_\theta \log \pi_\theta(A \mid s)\,\big] &= \sum_a \pi_\theta(a \mid s)\, b(s)\,\nabla_\theta \log \pi_\theta(a \mid s) && \text{; definition of } \mathbb{E}_A \\
&= b(s)\sum_a \pi_\theta(a \mid s)\,\frac{\nabla_\theta \pi_\theta(a \mid s)}{\pi_\theta(a \mid s)} && \text{; } \nabla_\theta \log \pi = \nabla_\theta \pi/\pi \\
&= b(s)\sum_a \nabla_\theta \pi_\theta(a \mid s) && \text{; cancel } \pi_\theta(a \mid s) \\
&= b(s)\,\nabla_\theta\Big(\sum_a \pi_\theta(a \mid s)\Big) && \text{; gradient out of finite sum} \\
&= b(s)\,\nabla_\theta(1) \;=\; 0. && \text{; } \textstyle\sum_a \pi_\theta(a \mid s) = 1
\end{aligned}$$

Subtracting a term of zero mean from the multiplier leaves the outer expectation unchanged, which is the second form. □

**Key Result**

$$\nabla_\theta J(\theta) \;=\; \frac{1}{1-\gamma}\,\mathbb{E}_{S\sim d^{\pi_\theta},\rho_0}\Big[\,\mathbb{E}_{A\sim\pi_\theta(\cdot|S)}\big[\,\big(Q^{\pi_\theta}(S,A) - b(S)\big)\,\nabla_\theta \log \pi_\theta(A \mid S)\,\big]\Big].$$

The step $\sum_a \nabla_\theta \pi_\theta(a \mid s) = \nabla_\theta(1) = 0$ is the same expected-score-zero identity used in the policy-gradient preliminaries: the score has mean zero under $\pi_\theta$, so any state-only weight $b(S)$ washes out.

### 1.3 Value Baseline and Advantage

Choose

$$b(S) \;=\; V^{\pi_\theta}(S).$$

Then

$$\begin{aligned}
Q^{\pi_\theta}(S,A) - b(S) &= Q^{\pi_\theta}(S,A) - V^{\pi_\theta}(S) && \text{; value baseline} \\
&= A^{\pi_\theta}(S,A). && \text{; definition of advantage}
\end{aligned}$$

Therefore,

**Key Result**

$$\nabla_\theta J(\theta) \;=\; \frac{1}{1-\gamma}\,\mathbb{E}_{S\sim d^{\pi_\theta},\rho_0}\Big[\,\mathbb{E}_{A\sim\pi_\theta(\cdot|S)}\big[\,A^{\pi_\theta}(S,A)\,\nabla_\theta \log \pi_\theta(A \mid S)\,\big]\Big].$$

*Two expectations of the same return, one conditioned on the action, one averaging it out:*

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

*The advantage measures how much taking $A_t$ shifts the expected return relative to the policy's average at $S_t$. The score direction is now signed by this comparison, not by the raw magnitude of $Q^{\pi_\theta}$.*

Equivalently, the advantage is $Q^{\pi_\theta}$ *centered* to zero mean under $\pi_\theta$:

$$\mathbb{E}_{A\sim\pi_\theta(\cdot|S)}\big[\, A^{\pi_\theta}(S, A) \,\big] \;=\; \mathbb{E}_{A\sim\pi_\theta(\cdot|S)}\big[\, Q^{\pi_\theta}(S, A) \,\big] - V^{\pi_\theta}(S) \;=\; 0.$$

This is why $V^{\pi_\theta}$ is the *natural* baseline, not merely *a* baseline: it is the state-only offset that removes exactly the action-independent part, leaving a mean-zero multiplier.

Carrying the new multiplier through the same per-step form as Report 15 gives the advantage oracle estimator:

**Key Result**

$$\begin{aligned}\widehat{g}_t^{\,A,\mathrm{oracle}}(\theta) \;&=\; \gamma^t\, A^{\pi_\theta}(S_t, A_t)\, \nabla_\theta \log \pi_\theta(A_t \mid S_t) \\ &=\; \gamma^t \,\big[\, Q^{\pi_\theta}(S_t, A_t) - V^{\pi_\theta}(S_t) \,\big]\, \nabla_\theta \log \pi_\theta(A_t \mid S_t).\end{aligned}$$

### 1.4 From Two Critics to One

The advantage needs two learned functions,

$$A^{\pi_\theta}(S_t, A_t) = Q^{\pi_\theta}(S_t, A_t) - V^{\pi_\theta}(S_t) \;\rightsquigarrow\; \hat{q}(S_t, A_t, w^q) - \hat{v}(S_t, w^v),$$

two critics and two parameter vectors. The next two subsections collapse this to a single $V$-critic: $Q^{\pi_\theta}$ is rewritten through its Bellman equation so that the whole advantage is estimated by one TD error.

### 1.5 Expected SARSA Bellman Form

**Key Result**

$$Q^{\pi_\theta}(s, a) \;=\; \mathbb{E}^{\pi_\theta}\big[\, R_{t+1} + \gamma\, V^{\pi_\theta}(S_{t+1}) \mid S_t = s,\; A_t = a \,\big].$$

This is the Expected SARSA Bellman form for $Q^{\pi_\theta}$.

The next action is already averaged inside $V^{\pi_\theta}$:

$$V^{\pi_\theta}(s') \;=\; \mathbb{E}_{A'\sim\pi_\theta(\cdot|s')}\big[\, Q^{\pi_\theta}(s', A') \,\big].$$

Two targets for the same $Q^{\pi_\theta}$:

$$R_{t+1} + \gamma\, Q^{\pi_\theta}(S_{t+1}, A_{t+1}) \qquad \text{SARSA target: samples } A_{t+1},$$

$$R_{t+1} + \gamma\, V^{\pi_\theta}(S_{t+1}) \qquad \text{Expected SARSA target: no } A_{t+1} \text{ sample.}$$

*In QAC the SARSA bootstrap samples the next action $A_{t+1}$ to form the target. The Expected SARSA target bootstraps with $V^{\pi_\theta}(S_{t+1})$, the exact policy average over $A_{t+1}$: the next action is* integrated out analytically *rather than sampled. This is Rao-Blackwellization, replacing a sample by its conditional expectation can only lower variance, never raise it. One fewer sampled variable in the target, and since the advantage never needs $Q^{\pi_\theta}$ alone but only its deviation from $V^{\pi_\theta}$, a single $V$-critic suffices. That is the estimation advantage of this report over QAC.*

### 1.6 TD Error as One-Sample Advantage

Define the true TD error:

$$\delta_t^\pi \;:=\; R_{t+1} + \gamma\, V^{\pi_\theta}(S_{t+1}) - V^{\pi_\theta}(S_t).$$

**Theorem 1.2** (TD Error is a One-Sample Advantage)**.** *Under $\pi_\theta$,*

$$\mathbb{E}^{\pi_\theta}\big[\,\delta_t^\pi \mid S_t = s,\; A_t = a\,\big] \;=\; A^{\pi_\theta}(s,a).$$

*Proof.*

$$\begin{aligned}
\mathbb{E}^{\pi_\theta}\big[\,\delta_t^\pi \mid s,a\,\big] &= \mathbb{E}^{\pi_\theta}\big[\,R_{t+1} + \gamma\, V^{\pi_\theta}(S_{t+1}) - V^{\pi_\theta}(S_t) \mid S_t = s, A_t = a\,\big] && \text{; definition}\\
&= \mathbb{E}^{\pi_\theta}\big[\,R_{t+1} + \gamma\, V^{\pi_\theta}(S_{t+1}) \mid S_t = s, A_t = a\,\big] - V^{\pi_\theta}(s) && \text{; } S_t = s\\
&= Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s) && \text{; Expected SARSA Bellman form}\\
&= A^{\pi_\theta}(s,a). && \text{; definition of advantage}
\end{aligned}$$

□

The advantage never had to be built from two critics: one sample of the TD error is already an unbiased sample of it. This gives the oracle estimator

$$\widehat{g}_t^{\,\delta,\mathrm{oracle}}(\theta) \;=\; \gamma^t\, \delta_t^\pi\, \nabla_\theta \log \pi_\theta(A_t \mid S_t).$$

### 1.7 Exact Identity vs Approximation

The exact identity

$$\mathbb{E}^{\pi_\theta}\big[\,\delta_t^\pi \mid S_t, A_t\,\big] \;=\; A^{\pi_\theta}(S_t, A_t)$$

holds for

$$\delta_t^\pi \;=\; R_{t+1} + \gamma\, V^{\pi_\theta}(S_{t+1}) - V^{\pi_\theta}(S_t).$$

In practice,

$$\delta_t \;=\; R_{t+1} + \gamma\, \hat{v}(S_{t+1}, w_t) - \hat{v}(S_t, w_t).$$

Then

$$\mathbb{E}^{\pi_\theta}\big[\,\delta_t \mid S_t, A_t\,\big] \;\approx\; A^{\pi_\theta}(S_t, A_t), \qquad \text{with equality only if } \hat{v}(\cdot, w_t) = V^{\pi_\theta}(\cdot).$$

## 1.8 One-Critic Actor-Critic

$$V^{\pi_\theta}(s) \;\rightsquigarrow\; \hat{v}(s, w).$$

**Key Result**

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

## 1.9 Bias and Variance, Layer by Layer

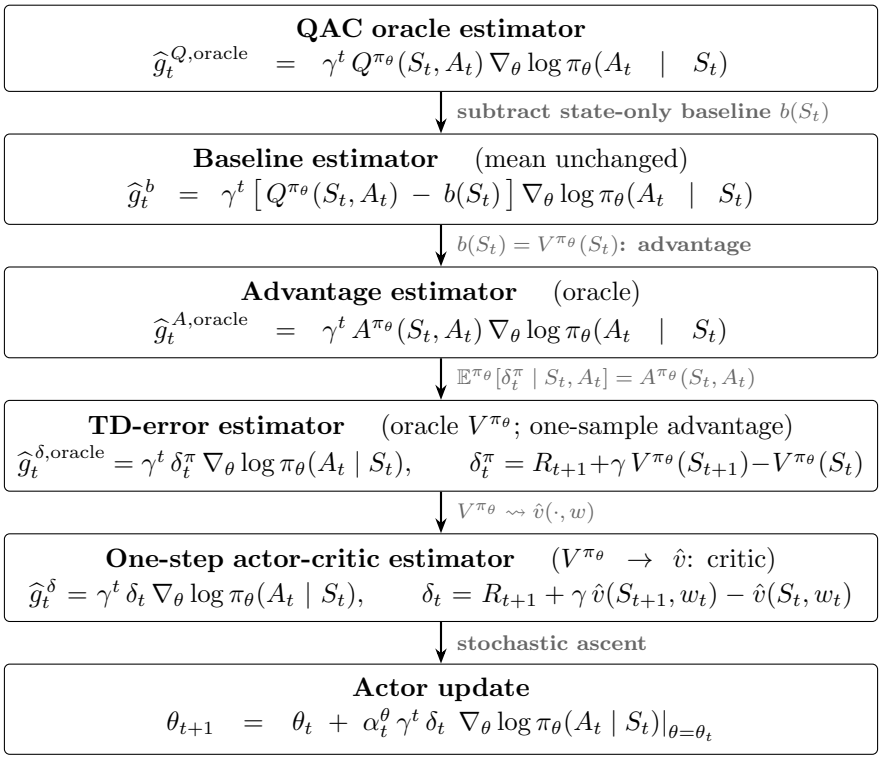
| | Object | Bias | Variance |
|---|---|---|---|
| A | $\widehat{g}_t^{Q,\text{oracle}}$ | none, oracle $Q^\pi$ | high |
| B | $Q^\pi(S_t, A_t) - b(S_t)$ | none | changed |
| C | $Q^\pi(S_t, A_t) - V^\pi(S_t) = A^\pi(S_t, A_t)$ | none | usually lower |
| D | $\delta_t^\pi$ | none as an advantage sample | TD target sampling |
| E | $\delta_t = R_{t+1} + \gamma\hat{v}(S_{t+1}, w_t) - \hat{v}(S_t, w_t)$ | critic approximation/tracking | TD target sampling |

## 1.10 Algorithm: One-Step Actor-Critic

**Algorithm 1:** One-Step Actor-Critic, 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 $\gamma$
initialize policy parameter $\theta_0 \in \mathbb{R}^{d'}$ and state-value weights $w_0 \in \mathbb{R}^d$
**repeat**
  sample/reset $S_0 \sim \rho_0$
  $S_t \leftarrow S_0$
  **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`
    $\delta_t \leftarrow R_{t+1} + \gamma\, \hat{v}(S_{t+1}, w_t) - \hat{v}(S_t, w_t)$
    $\theta_{t+1} \leftarrow \theta_t + \alpha_t^\theta\, \gamma^t\, \delta_t\, \nabla_\theta \log \pi_{\theta_t}(A_t \mid S_t)$
    $w_{t+1} \leftarrow w_t + \alpha_t^w\, \delta_t\, \nabla_w \hat{v}(S_t, w_t)$
    $S_t \leftarrow S_{t+1}$
**until** *converged*

### 1.11 The Ladder




### 1.12 What We Bought

QAC carried one critic ($Q^{\pi_\theta}$) and paid for it in variance. This report carries one critic ($V^{\pi_\theta}$) and less variance, from two moves that cost nothing in structural bias: subtracting the action-independent part of the signal (the baseline), and integrating out the next action analytically (Expected SARSA). Every estimator on the ladder is exactly unbiased; bias enters only in the last row of the table, and only through critic approximation $\hat{v} \neq V^{\pi_\theta}$, never through the estimator structure itself. That the bias column reads "none" until that final row is the whole point.

### 1.13 Scope

This report is one-step actor-critic. $n$-step actor-critic, eligibility traces, and GAE come next.

## A Time-Marginal Form

The per-step estimators above inherit the $\gamma^t$-weighted time-marginal identity of Report 15, with the multiplier ($Q^{\pi_\theta}$, or $A^{\pi_\theta}$) carried through unchanged. For the advantage multiplier:

$$d^{\pi_\theta,\rho_0}(s) \;=\; (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \Pr_\theta(S_t = s).$$

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \frac{1}{1-\gamma} \sum_s d^{\pi_\theta,\rho_0}(s) \sum_a \pi_\theta(a \mid s)\, A^{\pi_\theta}(s,a)\, \nabla_\theta \log \pi_\theta(a \mid s) && \text{; advantage PG} \\
&= \frac{1}{1-\gamma} \sum_s \Big[ (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \Pr_\theta(S_t = s) \Big] \\
&\qquad \cdot \sum_a \pi_\theta(a \mid s)\, A^{\pi_\theta}(s,a)\, \nabla_\theta \log \pi_\theta(a \mid s) && \text{; substitute } d^{\pi_\theta,\rho_0} \\
&= \sum_{t=0}^{\infty} \gamma^t \sum_s \Pr_\theta(S_t = s) \sum_a \pi_\theta(a \mid s)\, A^{\pi_\theta}(s,a)\, \nabla_\theta \log \pi_\theta(a \mid s) && \text{; cancel and rearrange} \\
&= \sum_{t=0}^{\infty} \gamma^t\, \mathbb{E}_{S_t \sim \Pr_\theta} \Big[ \mathbb{E}_{A_t \sim \pi_\theta(\cdot \mid S_t)} \big[ A^{\pi_\theta}(S_t, A_t)\, \nabla_\theta \log \pi_\theta(A_t \mid S_t) \big] \Big], && \text{; time-}t\text{ marginals}
\end{aligned}
$$

so

$$
\widehat{g}_T^{A,\text{oracle}}(\theta) \;=\; \sum_{t=0}^{T-1} \gamma^t\, A^{\pi_\theta}(S_t, A_t)\, \nabla_\theta \log \pi_\theta(A_t \mid S_t).
$$

# B Estimator Naming Ladder

The four multipliers of the score, in order of refinement:

$$
\begin{aligned}
\widehat{g}_t^{Q,\text{oracle}} &= \gamma^t\, Q^{\pi_\theta}(S_t, A_t)\, \nabla_\theta \log \pi_\theta(A_t \mid S_t) \\
\widehat{g}_t^{A,\text{oracle}} &= \gamma^t\, A^{\pi_\theta}(S_t, A_t)\, \nabla_\theta \log \pi_\theta(A_t \mid S_t) \\
\widehat{g}_t^{\hat{A}} &= \gamma^t \big[\, \hat{q}(S_t, A_t, w_t^q) - \hat{v}(S_t, w_t^v) \,\big]\, \nabla_\theta \log \pi_\theta(A_t \mid S_t) \\
\widehat{g}_t^{\delta} &= \gamma^t\, \delta_t\, \nabla_\theta \log \pi_\theta(A_t \mid S_t)
\end{aligned}
$$