Actor-Critic

Reinforcement Learning

Sai Sampath Kedari
sampath@umich.edu

Contents

1 Generalized Policy Iteration and Actor-Critic

1.1 GPI in Dynamic Programming

Policy evaluation: estimate V^{π_k},

$$V^{i+1} = T_{\pi_k} V^i : \qquad V^0 \xrightarrow{T_{\pi_k}} V^1 \xrightarrow{T_{\pi_k}} V^2 \xrightarrow{T_{\pi_k}} \cdots \longrightarrow V^{\pi_k}, \qquad \pi_k \xrightarrow{E} V^{\pi_k}.$$

Policy improvement: greedy,

$$\pi_{k+1}(s) \in \arg\max_a \Big[r(s,a) + \gamma \sum_{s'} p(s' \mid s,a)\, V^{\pi_k}(s') \Big], \qquad V^{\pi_k} \xrightarrow{I} \pi_{k+1}.$$

Overall:

$$\pi_0 \xrightarrow{E} V^{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} V^{\pi_1} \xrightarrow{I} \pi_2 \xrightarrow{E} \cdots \xrightarrow{I} \pi_* \xrightarrow{E} V^*.$$

1.2 GPI in Model-Free Control

Greedy improvement over V^{π_k} needs the model p, r. Move to Q-space.

Policy evaluation: estimate Q^{π_k},

$$Q^{i+1} = T_{\pi_k} Q^i : \qquad Q^0 \xrightarrow{T_{\pi_k}} Q^1 \xrightarrow{T_{\pi_k}} Q^2 \xrightarrow{T_{\pi_k}} \cdots \longrightarrow Q^{\pi_k}, \qquad \pi_k \xrightarrow{E} Q^{\pi_k}.$$

Policy improvement: ε-greedy,

$$\pi_{k+1} = \varepsilon\text{-greedy}(Q^{\pi_k}), \qquad Q^{\pi_k} \xrightarrow{I} \pi_{k+1}.$$

Overall:

$$\pi_0 \xrightarrow{E} Q^{\pi_0} \xrightarrow{I} \pi_1 \xrightarrow{E} Q^{\pi_1} \xrightarrow{I} \pi_2 \xrightarrow{E} \cdots \xrightarrow{I} \pi_* \xrightarrow{E} q_*.$$

1.3 GPI in Actor-Critic

The improvement step is still an $\arg\max_a Q^{\pi_k}(s,a)$: an optimization over actions at every state, intractable for large or continuous $\mathcal{A}$. Replace the tabular policy by a parameterized family π_θ and improve by a gradient step on θ.

Policy evaluation (Critic): estimate $Q_w \approx Q^{\pi_{\theta_k}}$,

$$\pi_{\theta_k} \xrightarrow{E} Q_w.$$

Policy improvement (Actor): gradient step,

$$\theta_{k+1} = \theta_k + \alpha\, \nabla_\theta J(\theta_k), \qquad Q_w \xrightarrow{I} \pi_{\theta_{k+1}}.$$

Overall:

$$\pi_{\theta_0} \xrightarrow{E} Q_w \xrightarrow{I} \pi_{\theta_1} \xrightarrow{E} Q_w \xrightarrow{I} \pi_{\theta_2} \xrightarrow{E} \cdots \longrightarrow \pi_{\theta_*}.$$

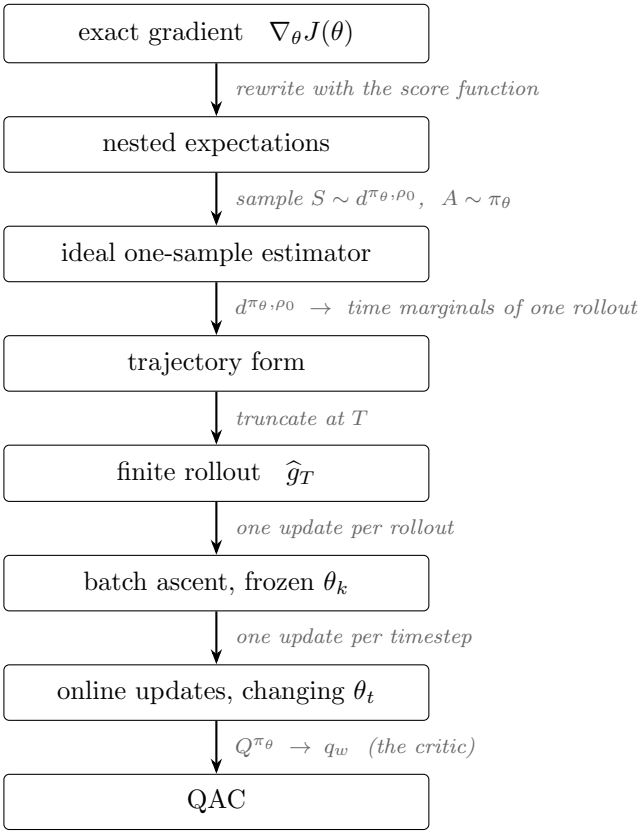
Critic = policy evaluation of the fixed policy (T_π, no max). Actor = policy improvement by gradient, replacing the arg max.

The next section derives the actor exactly; the critic of this cycle is Step 8 there.

2 QAC: Derivation from the Exact Policy Gradient

2.1 Roadmap

One exact quantity is turned into one running algorithm through the following sequence of transformations: algebraic rewritings, unbiased sampling, and implementation approximations. Each arrow is a single move, and each move is one subsection.



The first three rows are exact or unbiased. Bias enters only at the truncation, the online switch, and the critic.

2.2 Starting Point

The exact discounted policy-gradient theorem:

$$\nabla_\theta J(\theta) \;=\; \frac{1}{1-\gamma} \sum_s d^{\pi_\theta,\rho_0}(s) \sum_a \nabla_\theta \pi_\theta(a \mid s)\, Q^{\pi_\theta}(s,a) \qquad (\in \mathbb{R}^{d'}),$$

with objective and discounted visitation

$$J(\theta) \;=\; \mathbb{E}_{S_0 \sim \rho_0}\big[V^{\pi_\theta}(S_0)\big],$$

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

d^{π_θ,ρ_0} is not the state distribution at any single time; it is the normalized discounted mixture of all time marginals (policy gradient theorem report).

If the gradient were computable, gradient ascent would finish the job:

$$\theta_{k+1} \;=\; \theta_k + \alpha_k \, \nabla_\theta J(\theta_k).$$

It is not computable. Two objects in it are unknown:

1. $d^{\pi_\theta,\rho_0}(s)$ depends on the dynamics p and the entire visitation pattern;
2. $Q^{\pi_\theta}(s,a)$ the true action-value.

The rest of this section replaces these exact quantities by sampled ones, one move at a time.

2.3 Step 1: Rewrite as Nested Expectations

Apply the score-function identity $\nabla_\theta \pi_\theta(a \mid s) = \pi_\theta(a \mid s)\, \nabla_\theta \log \pi_\theta(a \mid s)$:

$$\begin{aligned}
\nabla_\theta J(\theta) &= \frac{1}{1-\gamma} \sum_s d^{\pi_\theta,\rho_0}(s) \sum_a \nabla_\theta \pi_\theta(a \mid s)\, Q^{\pi_\theta}(s,a) \\
&= \frac{1}{1-\gamma} \sum_s d^{\pi_\theta,\rho_0}(s) \sum_a \pi_\theta(a \mid s)\, Q^{\pi_\theta}(s,a)\, \nabla_\theta \log \pi_\theta(a \mid s) && \text{; score-function identity} \\
&= \frac{1}{1-\gamma}\, \mathbb{E}_{S\sim d^{\pi_\theta,\rho_0}} \Big[\mathbb{E}_{A\sim\pi_\theta(\cdot\mid S)} \big[\, Q^{\pi_\theta}(S,A)\, \nabla_\theta \log \pi_\theta(A \mid S) \,\big] \Big] && \text{; sums} \to \text{expectations}
\end{aligned}$$

Key Result

$$\nabla_\theta J(\theta) \;=\; \frac{1}{1-\gamma}\, \underbrace{\mathbb{E}_{S\sim d^{\pi_\theta,\rho_0}}}_{\text{(1) state}} \Big[\, \underbrace{\mathbb{E}_{A\sim\pi_\theta(\cdot\mid S)}}_{\text{(2) action}} \big[\, \underbrace{Q^{\pi_\theta}(S,A)}_{\text{(3) return}}\, \nabla_\theta \log \pi_\theta(A \mid S) \,\big] \Big]$$

Three expectations, replaced one at a time in what follows:

1. the state expectation $\mathbb{E}_{S\sim d^{\pi_\theta,\rho_0}}$ (Steps 2 and 3);
2. the action expectation $\mathbb{E}_{A\sim\pi_\theta}$ (free: we own the policy);
3. the return expectation inside $Q^{\pi_\theta} = \mathbb{E}^{\pi_\theta}[\, G_t \mid S_t, A_t \,]$ (the critic, Step 8).

Until Step 8, Q^{π_θ} is oracle-given.

2.4 The Estimator Principle

Nearly all of statistics is one move: replacing an expectation by an average of samples. We want a number defined as an expectation,

$$m \;=\; \mathbb{E}[X],$$

which we cannot integrate, but whose distribution we can draw from. Draw n independent copies $X_1, \ldots, X_n \overset{d}{=} X$ and average:

$$\widehat{m}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\mathbb{E}[\widehat{m}_n] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] \;=\; m \qquad \text{; unbiased for every } n$$

$$\mathrm{Var}(\widehat{m}_n) = \frac{\mathrm{Var}(X)}{n} \qquad \text{; variance shrinks with } n$$

Unbiasedness holds already at $n = 1$; the sample size only controls variance. In RL, data is one trajectory or one transition per update, so the working case throughout this report is the one-sample estimator,

$$\widehat{m} \;=\; X_1, \qquad \mathbb{E}[\widehat{m}] \;=\; m.$$

Whenever an expectation is replaced by samples below, the resulting random quantity is read through the decomposition of the preliminaries report,

$$\widehat{g}_k \;=\; g_k + b_k + \varepsilon_k, \qquad \text{unbiased} \iff b_k = 0.$$

The purely algebraic rewritings have neither bias nor variance.

2.5 Step 2: Ideal One-Sample Estimator

Pretend for one subsection that $S \sim d^{\pi_\theta,\rho_0}$ can be drawn. Replace the two outer expectations by samples, one at a time.

(i) Sample the state, $S \sim d^{\pi_\theta,\rho_0}$:

$$\widehat{g}_{\text{state}}(\theta) \;=\; \frac{1}{1-\gamma}\,\mathbb{E}_{A\sim\pi_\theta(\cdot|S)}\big[\,Q^{\pi_\theta}(S,A)\,\nabla_\theta \log \pi_\theta(A \mid S)\,\big].$$

(ii) Sample the action, $A \sim \pi_\theta(\cdot \mid S)$:

$$\widehat{g}^{\,\text{ideal}}(\theta) \;=\; \frac{1}{1-\gamma}\,Q^{\pi_\theta}(S,A)\,\nabla_\theta \log \pi_\theta(A \mid S).$$

Unbiasedness, by taking the expectation back in the same order:

$$\mathbb{E}\big[\widehat{g}^{\,\text{ideal}}\big] = \mathbb{E}_{S\sim d^{\pi_\theta,\rho_0}}\Big[\,\mathbb{E}_{A\sim\pi_\theta(\cdot|S)}\big[\,\tfrac{1}{1-\gamma}\,Q^{\pi_\theta}(S,A)\,\nabla_\theta \log \pi_\theta(A \mid S)\,\big]\Big] \qquad \text{; tower rule over } A, \text{ then } S$$

$$= \nabla_\theta J(\theta) \qquad \text{; nested-expectation form above}$$

Key Result

$$\widehat{g}^{\,\text{ideal}}(\theta) = \frac{1}{1-\gamma}\,Q^{\pi_\theta}(S,A)\,\nabla_\theta \log \pi_\theta(A \mid S), \qquad S \sim d^{\pi_\theta,\rho_0},\; A \sim \pi_\theta(\cdot \mid S),$$

$$\mathbb{E}\big[\widehat{g}^{\,\text{ideal}}\big] = \nabla_\theta J(\theta).$$

Bias: 0.

Variance: yes; S and A are random.

Two impossible pieces remain: drawing $S \sim d^{\pi_\theta,\rho_0}$, and knowing Q^{π_θ}. The next step removes the first.

2.6 Step 3: From d^{π_θ,ρ_0} to One Trajectory

Substitute the definition of d^{π_θ,ρ_0} into the state expectation and let the sum over time do the sampling. Two measures appear below and are written explicitly: $\mathbb{E}_{S_t\sim \Pr_\theta}$ is expectation over the time-t state marginal $\Pr_\theta(S_t = \cdot)$, and $\mathbb{E}_{\tau\sim\pi_\theta}$ is expectation over the full trajectory law of π_θ started from ρ_0. Throughout, abbreviate the inner action expectation

$$\phi_\theta(s) \;:=\; \mathbb{E}_{A\sim\pi_\theta(\cdot|s)}\big[\, Q^{\pi_\theta}(s,A)\,\nabla_\theta \log \pi_\theta(A \mid s)\,\big] \;\in\; \mathbb{R}^{d'}.$$

$$\begin{aligned}
\nabla_\theta J(\theta) &= \frac{1}{1-\gamma}\sum_s d^{\pi_\theta,\rho_0}(s)\;\phi_\theta(s) \\
&= \frac{1}{1-\gamma}\sum_s \Big[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t\,\Pr_\theta(S_t=s)\Big]\,\phi_\theta(s) && \text{; definition of } d^{\pi_\theta,\rho_0} \\
&= \sum_s\sum_{t=0}^{\infty}\gamma^t\,\Pr_\theta(S_t=s)\;\phi_\theta(s) && \text{; cancel } (1-\gamma) \\
&= \sum_{t=0}^{\infty}\gamma^t\sum_s \Pr_\theta(S_t=s)\;\phi_\theta(s) && \text{; interchange; abs. convergent} \\
&= \sum_{t=0}^{\infty}\gamma^t\,\mathbb{E}_{S_t\sim\Pr_\theta}\big[\,\phi_\theta(S_t)\,\big] && \text{; inner sum is } \mathbb{E} \text{ over the time-}t \text{ marginal}
\end{aligned}$$

Key Result

$$\nabla_\theta J(\theta) \;=\; \sum_{t=0}^{\infty}\gamma^t\,\mathbb{E}_{S_t\sim\Pr_\theta}\Big[\mathbb{E}_{A_t\sim\pi_\theta(\cdot|S_t)}\big[\,Q^{\pi_\theta}(S_t,A_t)\,\nabla_\theta\log\pi_\theta(A_t\mid S_t)\,\big]\Big]$$

The interchange is licensed by absolute convergence: finite $\mathcal{S}$, $\mathcal{A}$, bounded rewards, $\gamma < 1$.

One expectation over $S \sim d^{\pi_\theta,\rho_0}$ has become a γ^t-weighted sum of expectations over ordinary trajectory states $S_t \sim \Pr_\theta(S_t = \cdot)$. No single S_t is distributed as d^{π_θ,ρ_0}; the discounted sum over all t reconstructs it exactly.

Bias: none introduced; this is exact algebra.

2.7 Step 4: Fixed-θ Trajectory Estimator

Freeze θ and roll out one trajectory under π_θ:

$$\begin{aligned} S_0 &\sim \rho_0, \\ A_t &\sim \pi_\theta(\cdot \mid S_t), \\ S_{t+1} &\sim p(\cdot \mid S_t, A_t) \\ \Longrightarrow \qquad S_t &\sim \Pr_\theta(S_t = \cdot). \end{aligned}$$

Replace each conditional expectation by its sampled random variable, $\mathbb{E}_{S_t}[\,\cdot\,] \rightsquigarrow S_t$ and $\mathbb{E}_{A_t|S_t}[\,\cdot\,] \rightsquigarrow A_t$:

Key Result

$$\widehat{g}(\theta) \;=\; \sum_{t=0}^{\infty} \gamma^t \, Q^{\pi_\theta}(S_t, A_t) \, \nabla_\theta \log \pi_\theta(A_t \mid S_t)$$

Unbiasedness, by reversing the two replacements:

$$\begin{aligned} \mathbb{E}_{\tau\sim\pi_\theta}\big[\widehat{g}(\theta)\big] &= \sum_{t=0}^{\infty} \gamma^t \, \mathbb{E}_{\tau\sim\pi_\theta}\big[\, Q^{\pi_\theta}(S_t, A_t) \, \nabla_\theta \log \pi_\theta(A_t \mid S_t) \,\big] && \text{; } \mathbb{E} \text{ through sum; dominated conv.} \\ &= \sum_{t=0}^{\infty} \gamma^t \, \mathbb{E}_{S_t\sim\Pr_\theta}\Big[\mathbb{E}_{A_t\sim\pi_\theta}\big[\, Q^{\pi_\theta}(S_t, A_t) \, \nabla_\theta \log \pi_\theta(A_t \mid S_t) \,\big]\Big] && \text{; condition on } S_t \\ &= \nabla_\theta J(\theta) && \text{; Step 3 identity} \end{aligned}$$

Moving $\mathbb{E}_{\tau\sim\pi_\theta}$ through the infinite sum is again licensed by dominated convergence under the same bounds.

Bias: 0 (fixed θ, exact Q^{π_θ}).

Variance: yes; one trajectory, one action per step.

2.8 Step 5: Finite Rollout

An infinite trajectory cannot be run. Keep the first T terms:

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

The bias is exactly the discarded tail:

$$\begin{aligned} b_T(\theta) &= \mathbb{E}_{\tau\sim\pi_\theta}\big[\widehat{g}_T(\theta)\big] - \nabla_\theta J(\theta) && \text{; definition of bias} \\ &= -\sum_{t=T}^{\infty} \gamma^t \, \mathbb{E}_{\tau\sim\pi_\theta}\big[\, Q^{\pi_\theta}(S_t, A_t) \, \nabla_\theta \log \pi_\theta(A_t \mid S_t) \,\big] && \text{; first } T \text{ terms cancel} \end{aligned}$$

Two readings of T:

1. **true terminal time** (absorbing state, zero reward and value after): every tail term is identically zero, so $b_T = 0$ and the finite sum *is* the infinite sum;
2. **artificial cutoff** in a continuing task: $b_T \neq 0$, tail truncation bias of size $O(\gamma^T)$.

Bias: tail truncation, case 2 only.

Variance: unchanged.

2.9 Step 6: Batch Stochastic Ascent, Frozen θ_k

At iteration k: freeze θ_k, roll out one trajectory under π_{θ_k}, update once.

Key Result

$$\widehat{g}_T(\theta_k) = \sum_{t=0}^{T-1} \gamma^t \, Q^{\pi_{\theta_k}}(S_t, A_t) \;\; \nabla_\theta \log \pi_\theta(A_t \mid S_t)|_{\theta=\theta_k} \,,$$

$$\theta_{k+1} \;=\; \theta_k + \alpha_k \, \widehat{g}_T(\theta_k).$$

Every symbol inside is evaluated at the same frozen θ_k: the sampling policy, the values, the scores.

Bias: truncation only.

Variance: sampling.

2.10 Step 7: Online Updates, Changing θ_t

The practical algorithm updates with the *current* parameter at every step:

Key Result

$$\theta_{t+1} \;=\; \theta_t + \alpha_t \, \gamma^t \, Q^{\pi_{\theta_t}}(S_t, A_t) \;\; \nabla_\theta \log \pi_\theta(A_t \mid S_t)|_{\theta=\theta_t}$$

The trajectory is now generated by a changing sequence of policies,

$$\begin{aligned} A_0 &\sim \pi_{\theta_0}(\cdot \mid S_0), \\ A_1 &\sim \pi_{\theta_1}(\cdot \mid S_1), \\ A_2 &\sim \pi_{\theta_2}(\cdot \mid S_2), \\ &\vdots \end{aligned}$$

so that:

1. S_t was produced by $\pi_{\theta_0}, \ldots, \pi_{\theta_{t-1}}$, not by the single frozen π_{θ_t};
2. hence the per-step direction is *not* an exactly unbiased sample of $\nabla_\theta J(\theta_t)$.

This is the online policy-drift approximation. Its justification is stochastic approximation, not a new policy-gradient identity: with small step sizes the policy drifts slowly, and the sample stream

is treated as approximately generated by the current policy. Note also that the oracle itself now moves: the update at time t consumes $Q^{\pi_{\theta_t}}$, the action-value of the *current* policy, so any critic replacing the oracle must track π_{θ_t} as it drifts. That tracking requirement is the two-timescale problem, taken up with the critic.

Drift error: yes, relative to the fixed-θ_t estimator.

Variance: yes.

2.11 Remark: The γ^t Factor

The factor γ^t in the online update is not decoration. It is what implements the discounted visitation weighting of the exact gradient: summing the per-step updates reconstructs the γ^t-weighted time-marginal identity of Step 3, hence the expectation over d^{π_θ,ρ_0}. This report derives the exact estimator, so γ^t stays.

Many implementations drop it and update with $\alpha_t\, Q^{\pi_{\theta_t}}(S_t, A_t)\, \nabla_\theta \log \pi_\theta(A_t \mid S_t)|_{\theta=\theta_t}$ instead. The dropped version weights states by undiscounted occupancy rather than by d^{π_θ,ρ_0}, and it is not the gradient of *any* objective (Thomas 2014; Nota and Thomas 2019); its gap to $\nabla_\theta J$ is bounded by $O(1-\gamma)$ (Nota 2022). That practical variant and its bias are outside the scope of this exact derivation.

2.12 Step 8: The Critic

Steps 1 through 7 treated Q^{π_θ} as known. It is the last expectation left to estimate:

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

Different critic targets, already derived in this repo:

1. SARSA(0) [report 04]:

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

2. n-step SARSA [report 05]:

$$Q^{\pi_\theta}(s,a) = \mathbb{E}^{\pi_\theta}\Big[\, \textstyle\sum_{i=0}^{n-1} \gamma^i R_{t+i+1} + \gamma^n\, Q^{\pi_\theta}(S_{t+n}, A_{t+n}) \mid S_t = s,\; A_t = a \,\Big];$$

3. SARSA(λ) [report 06]: geometric mixture of the n-step targets over n.

We take SARSA(0). The one-sample target and tabular update:

$$U_t \;=\; R_{t+1} + \gamma\, Q(S_{t+1}, A_{t+1}), \qquad A_{t+1} \sim \pi_\theta(\cdot \mid S_{t+1}),$$

$$Q(S_t, A_t) \;\leftarrow\; Q(S_t, A_t) + \alpha_t^w \,\big[\, U_t - Q(S_t, A_t) \,\big].$$

Parameterize $\hat{q}(s,a,w) \approx Q^{\pi_\theta}(s,a)$, $w \in \mathbb{R}^d$, and take the semi-gradient step with the bootstrap target $U_t = R_{t+1} + \gamma\, \hat{q}(S_{t+1}, A_{t+1}, w_t)$ held fixed [report 08]:

Key Result

$$\delta_t \;=\; R_{t+1} + \gamma\, \hat{q}(S_{t+1}, A_{t+1}, w_t) - \hat{q}(S_t, A_t, w_t),$$
$$w_{t+1} \;=\; w_t + \alpha_t^w\, \delta_t\, \nabla_w \hat{q}(S_t, A_t, w_t).$$

This fills the critic line:

$$Q^{\pi_\theta}(S_t, A_t) \quad \rightsquigarrow \quad \hat{q}(S_t, A_t, w_t).$$

Bias: bootstrap bias while $\hat{q} \neq Q^{\pi_\theta}$; function-approximation error if Q^{π_θ} is not representable; tracking error because the actor moves.

Variance: yes; the Bellman expectation is replaced by one sampled transition and one sampled next action.

With $\hat{q} = Q^{\pi_\theta}$ and the actor frozen during sampling, Step 4's unbiasedness is recovered exactly.

2.13 Bias and Variance, Layer by Layer

	Layer	Bias	Variance
A	exact gradient $\nabla_\theta J(\theta)$	none	none
B	nested expectations	none	none
C	ideal one-sample, $S \sim d^{\pi_\theta, \rho_0}$	none	state, action
D	time-marginal identity	none	none
E	fixed-θ trajectory, infinite	none	trajectory, action
F	finite rollout $\widehat{g}_T$	tail, if artificial cutoff	trajectory, action
G	batch ascent, frozen θ_k	as F	as F
H	per-step accumulation, frozen θ_k	as F	as F
I	online, changing θ_t	policy drift	sampling
J	critic $\hat{q}(\cdot, \cdot, w)$ replaces oracle	critic approximation / tracking error	target sampling

A, B, D are identities. C and E introduce variance but no bias. Bias first appears at F, the fixed-θ analysis stops literally applying at I, and J adds the critic's own bias and variance on top of every layer above. In the tabular fixed-policy case, SARSA(0) is consistent under its usual coverage and step-size assumptions. In QAC the actor moves, so the critic must track $Q^{\pi_{\theta_t}}$; this is the two-timescale issue.

2.14 QAC Actor Update and Algorithm

Replace the oracle $Q^{\pi_{\theta_t}}$ by the critic $\hat{q}(\cdot, \cdot, w_t)$:

Key Result

$$A_t \sim \pi_{\theta_t}(\cdot \mid S_t), \qquad S_{t+1}, R_{t+1} \sim p(\cdot, \cdot \mid S_t, A_t), \qquad A_{t+1} \sim \pi_{\theta_t}(\cdot \mid S_{t+1}),$$

$$\theta_{t+1} \;=\; \theta_t + \alpha_t^{\theta}\, \gamma^t\, \hat{q}(S_t, A_t, w_t)\; \nabla_\theta \log \pi_\theta(A_t \mid S_t)|_{\theta=\theta_t}$$

Both updates read the same pre-update pair (θ_t, w_t) and the same transition. The algorithm below is written for the start-state discounted objective: the index t and the factor γ^t are measured from the beginning of a rollout sampled from ρ_0. The MDP may be episodic or continuing, but the estimator is organized into rollout segments.

Algorithm 1: Q Actor-Critic with SARSA(0) Critic, Rollout Form

Input: a differentiable policy parameterization $\pi_\theta(a \mid s)$
Input: a differentiable action-value parameterization $\hat{q}(s, a, w)$
Input: step sizes $\alpha_t^{\theta}, \alpha_t^{w} > 0$; discount γ
initialize policy parameter $\theta_0 \in \mathbb{R}^{d'}$ and action-value weights $w_0 \in \mathbb{R}^{d}$
repeat
  sample/reset $S_0 \sim \rho_0$; sample $A_0 \sim \pi_{\theta_0}(\cdot \mid S_0)$
  $S_t \leftarrow S_0$; $A_t \leftarrow A_0$
  for $t = 0, 1, \ldots, T-1$ *or until* S_t *terminal* **do**
    Observe $S_{t+1}, R_{t+1} \sim p(\cdot, \cdot \mid S_t, A_t)$ `// environment dynamics`
    Sample $A_{t+1} \sim \pi_{\theta_t}(\cdot \mid S_{t+1})$
    $\delta_t \leftarrow R_{t+1} + \gamma\, \hat{q}(S_{t+1}, A_{t+1}, w_t) - \hat{q}(S_t, A_t, w_t)$
    $\theta_{t+1} \leftarrow \theta_t + \alpha_t^{\theta}\, \gamma^t\, \hat{q}(S_t, A_t, w_t)\, \nabla_\theta \log \pi_{\theta_t}(A_t \mid S_t)$
    $w_{t+1} \leftarrow w_t + \alpha_t^{w}\, \delta_t\, \nabla_w \hat{q}(S_t, A_t, w_t)$
    $S_t \leftarrow S_{t+1}$
    $A_t \leftarrow A_{t+1}$
until *converged*

The action A_{t+1} is sampled from the pre-update policy θ_t: it forms the SARSA target for $Q^{\pi_{\theta_t}}$ and is also the action executed at step $t+1$. Executing an action from the slightly stale policy is part of the online drift approximation of Step 7. The update is online within a rollout, but the estimator is organized by rollout starts from ρ_0.

2.15 The Ladder