Deterministic Policy Gradient

Deep Reinforcement Learning

Sai Sampath Kedari
sampath@umich.edu

Contents

1 Deterministic Policy Gradient Theorem

1.1 Problem Formulation

Consider the discounted Markov decision process

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r, \gamma), \qquad \mathcal{S} \subseteq \mathbb{R}^n, \qquad \mathcal{A} \subseteq \mathbb{R}^m, \qquad \gamma \in [0, 1),$$

where $p(s' \mid s, a)$ is the environment dynamics and $r(s, a)$ is the expected one-step reward.

Let

$$\mu_\theta : \mathcal{S} \to \mathcal{A}, \qquad \theta \in \mathbb{R}^{d'},$$

be a differentiable deterministic policy. The state-action process induced by μ_θ is

$$S_0 \sim \rho, \qquad A_t = \mu_\theta(S_t), \qquad S_{t+1} \sim p(\cdot \mid S_t, A_t), \tag{1}$$

where ρ is the initial-state density.

Assume the required value-function derivatives exist and are bounded/integrable so that differentiation, integration, and infinite summation may be interchanged.

Since $\mu_\theta(s)$ is vector-valued, its derivative with respect to θ is technically a Jacobian. Following the deterministic policy-gradient notation, we write

$$\nabla_\theta \mu_\theta(s) \in \mathbb{R}^{d' \times m}, \qquad d\mu_\theta(s) = \nabla_\theta \mu_\theta(s)^T d\theta.$$

The performance objective is

$$J(\theta) := \mathbb{E}_{S_0 \sim \rho}\left[V^{\mu_\theta}(S_0)\right] = \int_{s \in \mathcal{S}} \rho(s) V^{\mu_\theta}(s)\, ds. \tag{2}$$

For a deterministic policy,

$$V^{\mu_\theta}(s) = Q^{\mu_\theta}\left(s, \mu_\theta(s)\right).$$

1.2 Gradient of the State-Value Function

Fix $s \in \mathcal{S}$, and define

$$q(\theta, a) := Q^{\mu_\theta}(s, a).$$

Then

$$\begin{aligned} dq(\theta, a) &= D_\theta q(\theta, a)[d\theta] + D_a q(\theta, a)[da] \\ &= \nabla_\theta q(\theta, a)^T d\theta + \nabla_a q(\theta, a)^T da. \end{aligned} \tag{3}$$

For

$$a = \mu_\theta(s), \qquad da = \nabla_\theta \mu_\theta(s)^T d\theta,$$

Equation (3) gives

$$\begin{aligned} dV^{\mu_\theta}(s) &= \nabla_\theta Q^{\mu_\theta}(s, a)^T\big|_{a=\mu_\theta(s)}\, d\theta \\ &\quad + \nabla_a Q^{\mu_\theta}(s, a)^T\big|_{a=\mu_\theta(s)}\, \nabla_\theta \mu_\theta(s)^T d\theta \\ &= \Big[\nabla_\theta Q^{\mu_\theta}(s, a)\big|_{a=\mu_\theta(s)} + \nabla_\theta \mu_\theta(s)\, \nabla_a Q^{\mu_\theta}(s, a)\big|_{a=\mu_\theta(s)}\Big]^T d\theta. \end{aligned} \tag{4}$$

Since

$$dV^{\mu_\theta}(s) = \nabla_\theta V^{\mu_\theta}(s)^T d\theta,$$

$$\boxed{\nabla_\theta V^{\mu_\theta}(s) = \nabla_\theta \mu_\theta(s)\ \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} + \nabla_\theta Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)}\,.} \tag{5}$$

Here,

$$\nabla_\theta Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)}$$

denotes differentiation with respect to θ while holding a fixed, followed by evaluation at $a = \mu_\theta(s)$.

Define

$$\boxed{\phi_\theta(s) := \nabla_\theta \mu_\theta(s)\ \nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)}\,.} \tag{6}$$

The dimensions are

$$\underbrace{\nabla_\theta \mu_\theta(s)}_{d' \times m}\ \underbrace{\nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)}}_{m \times 1} = \underbrace{\phi_\theta(s)}_{d' \times 1}.$$

Thus,

$$\nabla_\theta V^{\mu_\theta}(s) = \phi_\theta(s) + \nabla_\theta Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)}\,. \tag{7}$$

1.3 Recursive Form of the Value Gradient

The Bellman equation is

$$Q^{\mu_\theta}(s,a) = r(s,a) + \gamma \int_{s' \in \mathcal{S}} p(s' \mid s,a) V^{\mu_\theta}(s')\, ds'.$$

For fixed s and a, $r(s,a)$ and $p(s' \mid s,a)$ are properties of the MDP and do not depend on θ. Hence,

$$\begin{aligned}\nabla_\theta Q^{\mu_\theta}(s,a) &= \nabla_\theta \left[r(s,a) + \gamma \int_{s' \in \mathcal{S}} p(s' \mid s,a) V^{\mu_\theta}(s')\, ds' \right] \\ &= \gamma \int_{s' \in \mathcal{S}} p(s' \mid s,a) \nabla_\theta V^{\mu_\theta}(s')\, ds',\end{aligned} \tag{8}$$

where the derivative and integral are exchanged by the Leibniz integral rule.

Evaluating at $a = \mu_\theta(s)$ and substituting into Equation (7),

$$\begin{aligned}\nabla_\theta V^{\mu_\theta}(s) &= \phi_\theta(s) + \nabla_\theta Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)} \\ &= \phi_\theta(s) + \gamma \int_{s' \in \mathcal{S}} p(s' \mid s, \mu_\theta(s)) \nabla_\theta V^{\mu_\theta}(s')\, ds'.\end{aligned} \tag{9}$$

Thus,

$$\boxed{\nabla_\theta V^{\mu_\theta}(s) = \phi_\theta(s) + \gamma \int_{s' \in \mathcal{S}} p(s' \mid s, \mu_\theta(s)) \nabla_\theta V^{\mu_\theta}(s')\, ds'.} \tag{10}$$

1.4 Policy-Induced Transition Densities

For $t \geq 1$, let

$$p(s \to s', t, \mu_\theta)$$

denote the density of reaching s' after t transitions, starting from s, while following μ_θ.

The one-step, two-step, and general t-step transition densities are

$$p(s \to s', 1, \mu_\theta) = p(s' \mid s, \mu_\theta(s)), \tag{11}$$

$$p(s \to s', 2, \mu_\theta) = \int_{\bar{s} \in \mathcal{S}} p(s \to \bar{s}, 1, \mu_\theta) p(\bar{s} \to s', 1, \mu_\theta)\, d\bar{s}, \tag{12}$$

$$p(s \to s', t+1, \mu_\theta) = \int_{\bar{s} \in \mathcal{S}} p(s \to \bar{s}, t, \mu_\theta) p(\bar{s} \to s', 1, \mu_\theta)\, d\bar{s}. \tag{13}$$

Equivalently,

$$\boxed{p(s \to s', t, \mu_\theta) = p_{\mu_\theta}\left(S_t = s' \mid S_0 = s\right).} \tag{14}$$

For $t = 0$,

$$p(s \to s', 0, \mu_\theta) = \delta(s' - s).$$

Equation (9) may therefore be written as

$$\nabla_\theta V^{\mu_\theta}(s) = \phi_\theta(s) + \gamma \int_{s' \in \mathcal{S}} p(s \to s', 1, \mu_\theta) \nabla_\theta V^{\mu_\theta}(s')\, ds'. \tag{15}$$

1.5 Unrolling the Recursion

Iterating Equation (15),

$$\begin{aligned}
\nabla_\theta V^{\mu_\theta}(s) &= \phi_\theta(s) + \gamma \int_{s' \in \mathcal{S}} p(s \to s', 1, \mu_\theta) \nabla_\theta V^{\mu_\theta}(s')\, ds' \\
&= \phi_\theta(s) + \gamma \int_{s' \in \mathcal{S}} p(s \to s', 1, \mu_\theta) \left[\phi_\theta(s') + \gamma \int_{s'' \in \mathcal{S}} p(s' \to s'', 1, \mu_\theta) \nabla_\theta V^{\mu_\theta}(s'')\, ds'' \right] ds' \\
&= \phi_\theta(s) + \gamma \int_{s' \in \mathcal{S}} p(s \to s', 1, \mu_\theta) \phi_\theta(s')\, ds' \\
&\quad + \gamma^2 \int_{s' \in \mathcal{S}} \int_{s'' \in \mathcal{S}} p(s \to s', 1, \mu_\theta) p(s' \to s'', 1, \mu_\theta) \nabla_\theta V^{\mu_\theta}(s'')\, ds''\, ds' \\
&= \phi_\theta(s) + \gamma \int_{s' \in \mathcal{S}} p(s \to s', 1, \mu_\theta) \phi_\theta(s')\, ds' \\
&\quad + \gamma^2 \int_{s'' \in \mathcal{S}} p(s \to s'', 2, \mu_\theta) \nabla_\theta V^{\mu_\theta}(s'')\, ds'' \\
&= \phi_\theta(s) + \gamma \int_{s' \in \mathcal{S}} p(s \to s', 1, \mu_\theta) \phi_\theta(s')\, ds' \\
&\quad + \gamma^2 \int_{s'' \in \mathcal{S}} p(s \to s'', 2, \mu_\theta) \phi_\theta(s'')\, ds'' \\
&\quad + \gamma^3 \int_{s''' \in \mathcal{S}} p(s \to s''', 3, \mu_\theta) \nabla_\theta V^{\mu_\theta}(s''')\, ds''' \\
&\quad \vdots
\end{aligned} \tag{16}$$

After $N+1$ steps,

$$\nabla_\theta V^{\mu_\theta}(s) = \int_{s' \in \mathcal{S}} \sum_{t=0}^{N} \gamma^t p(s \to s', t, \mu_\theta) \phi_\theta(s')\, ds' + \gamma^{N+1} \int_{s' \in \mathcal{S}} p(s \to s', N+1, \mu_\theta) \nabla_\theta V^{\mu_\theta}(s')\, ds'. \tag{17}$$

If $\nabla_\theta V^{\mu_\theta}$ is bounded, the remainder satisfies

$$\left\| \gamma^{N+1} \int_{s' \in \mathcal{S}} p(s \to s', N+1, \mu_\theta) \nabla_\theta V^{\mu_\theta}(s')\, ds' \right\| \leq \gamma^{N+1} \sup_{s' \in \mathcal{S}} \left\| \nabla_\theta V^{\mu_\theta}(s') \right\| \longrightarrow 0,$$

since $\gamma < 1$. Taking $N \to \infty$,

$$\nabla_\theta V^{\mu_\theta}(s) = \int_{s' \in \mathcal{S}} \sum_{t=0}^{\infty} \gamma^t p(s \to s', t, \mu_\theta) \phi_\theta(s')\, ds'. \tag{18}$$

The reduction from the double integral to $p(s \to s'', 2, \mu_\theta)$ uses Fubini's theorem to exchange the order of integration.

Using

$$p(s \to s', t, \mu_\theta) = p_{\mu_\theta}\left(S_t = s' \mid S_0 = s\right),$$

we obtain

$$\boxed{\nabla_\theta V^{\mu_\theta}(s) = \sum_{t=0}^{\infty} \gamma^t \int_{s' \in \mathcal{S}} p_{\mu_\theta}\left(S_t = s' \mid S_0 = s\right) \phi_\theta(s')\, ds'.} \tag{19}$$

Thus, $\nabla_\theta V^{\mu_\theta}(s)$ is the discounted sum of the local contributions ϕ_θ over states reachable from s under μ_θ.

1.6 Gradient of the Performance Objective

For $S_0 \sim \rho$, define the state density at time t by

$$\boxed{p_{\mu_\theta, \rho}(S_t = s) := \int_{s_0 \in \mathcal{S}} \rho(s_0) p_{\mu_\theta}\left(S_t = s \mid S_0 = s_0\right)\, ds_0.} \tag{20}$$

Then, using Equation (19),

$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \int_{s_0 \in \mathcal{S}} \rho(s_0) V^{\mu_\theta}(s_0)\, ds_0 \\
&= \int_{s_0 \in \mathcal{S}} \rho(s_0) \nabla_\theta V^{\mu_\theta}(s_0)\, ds_0 \\
&= \int_{s_0 \in \mathcal{S}} \rho(s_0) \sum_{t=0}^{\infty} \gamma^t \int_{s \in \mathcal{S}} p_{\mu_\theta}\left(S_t = s \mid S_0 = s_0\right) \phi_\theta(s)\, ds\, ds_0 \\
&= \int_{s \in \mathcal{S}} \sum_{t=0}^{\infty} \gamma^t \left[\int_{s_0 \in \mathcal{S}} \rho(s_0) p_{\mu_\theta}\left(S_t = s \mid S_0 = s_0\right)\, ds_0 \right] \phi_\theta(s)\, ds \\
&= \int_{s \in \mathcal{S}} \sum_{t=0}^{\infty} \gamma^t p_{\mu_\theta, \rho}(S_t = s) \phi_\theta(s)\, ds.
\end{aligned} \tag{21}$$

The first derivative-integral interchange uses the Leibniz integral rule; the reordering of the state integrations uses Fubini's theorem.

Define the normalized discounted state-visitation density

$$\boxed{d^{\mu_\theta,\rho}(s) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t p_{\mu_\theta,\rho}(S_t = s).} \tag{22}$$

Therefore,

$$\begin{aligned}\nabla_\theta J(\theta) &= \frac{1}{1-\gamma}\int_{s\in\mathcal{S}} d^{\mu_\theta,\rho}(s)\phi_\theta(s)\,ds \\ &= \frac{1}{1-\gamma}\int_{s\in\mathcal{S}} d^{\mu_\theta,\rho}(s)\nabla_\theta\mu_\theta(s)\;\nabla_a Q^{\mu_\theta}(s,a)|_{a=\mu_\theta(s)}\;ds \\ &= \frac{1}{1-\gamma}\mathbb{E}_{S\sim d^{\mu_\theta,\rho}}\left[\nabla_\theta\mu_\theta(S)\;\nabla_a Q^{\mu_\theta}(S,a)|_{a=\mu_\theta(S)}\right].\end{aligned} \tag{23}$$

Hence,

$$\boxed{\nabla_\theta J(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{S\sim d^{\mu_\theta,\rho}}\left[\nabla_\theta\mu_\theta(S)\;\nabla_a Q^{\mu_\theta}(S,a)|_{a=\mu_\theta(S)}\right].} \tag{24}$$

2 Actor Objective from the Performance-Difference Identity

Let

$$\mu_{\text{old}} := \mu_{\theta_{\text{old}}}$$

denote the current deterministic policy, and let

$$\mu_\theta$$

denote a candidate deterministic policy.

2.1 Deterministic Performance-Difference Identity

Recall the stochastic performance-difference identity

$$J(\theta) - J(\theta_{\text{old}}) = \frac{1}{1-\gamma}\mathbb{E}_{S\sim d^{\pi_\theta,\rho}}\left[\mathbb{E}_{A\sim\pi_\theta(\cdot|S)}\left[A^{\pi_{\text{old}}}(S,A)\right]\right]. \tag{25}$$

For deterministic policies,

$$\pi_\theta(a\mid s) = \delta(a-\mu_\theta(s))\,, \qquad \pi_{\text{old}}(a\mid s) = \delta(a-\mu_{\text{old}}(s))\,.$$

Therefore,

$$\begin{aligned}J(\theta) - J(\theta_{\text{old}}) &= \frac{1}{1-\gamma}\mathbb{E}_{S\sim d^{\mu_\theta,\rho}}\left[\int_{a\in\mathcal{A}}\delta(a-\mu_\theta(S))\,A^{\mu_{\text{old}}}(S,a)\,da\right] \\ &= \frac{1}{1-\gamma}\mathbb{E}_{S\sim d^{\mu_\theta,\rho}}\left[A^{\mu_{\text{old}}}\left(S,\mu_\theta(S)\right)\right].\end{aligned} \tag{26}$$

Thus,

$$\boxed{J(\theta) - J(\theta_{\text{old}}) = \frac{1}{1-\gamma}\mathbb{E}_{S\sim d^{\mu_\theta,\rho}}\left[A^{\mu_{\text{old}}}\left(S,\mu_\theta(S)\right)\right].} \tag{27}$$

2.2 Local Actor Surrogate Objective

Construct the local surrogate by freezing the discounted state-visitation distribution at the current policy:

$$\begin{aligned}\widetilde{J}(\theta;\theta_{\text{old}}) &:= \frac{1}{1-\gamma}\mathbb{E}_{S\sim d^{\mu_{\text{old}},\rho}}\left[A^{\mu_{\text{old}}}\left(S,\mu_\theta(S)\right)\right] \\ &= \frac{1}{1-\gamma}\mathbb{E}_{S\sim d^{\mu_{\text{old}},\rho}}\left[Q^{\mu_{\text{old}}}\left(S,\mu_\theta(S)\right) - V^{\mu_{\text{old}}}(S)\right] \\ &= \frac{1}{1-\gamma}\mathbb{E}_{S\sim d^{\mu_{\text{old}},\rho}}\left[Q^{\mu_{\text{old}}}\left(S,\mu_\theta(S)\right)\right] \\ &\quad - \frac{1}{1-\gamma}\mathbb{E}_{S\sim d^{\mu_{\text{old}},\rho}}\left[V^{\mu_{\text{old}}}(S)\right]. \end{aligned} \tag{28}$$

Since the final term is independent of θ,

$$\arg\max_\theta \widetilde{J}(\theta;\theta_{\text{old}}) = \arg\max_\theta \frac{1}{1-\gamma}\mathbb{E}_{S\sim d^{\mu_{\text{old}},\rho}}\left[Q^{\mu_{\text{old}}}\left(S,\mu_\theta(S)\right)\right]. \tag{29}$$

Define the local actor surrogate objective

$$\boxed{J_{\text{actor}}(\theta;\theta_{\text{old}}) := \frac{1}{1-\gamma}\mathbb{E}_{S\sim d^{\mu_{\text{old}},\rho}}\left[Q^{\mu_{\text{old}}}\left(S,\mu_\theta(S)\right)\right].} \tag{30}$$

2.3 Gradient of the Local Actor Surrogate

With

$$d^{\mu_{\text{old}},\rho} \qquad \text{and} \qquad Q^{\mu_{\text{old}}}$$

fixed with respect to θ,

$$\begin{aligned}\nabla_\theta J_{\text{actor}}(\theta;\theta_{\text{old}}) &= \frac{1}{1-\gamma}\mathbb{E}_{S\sim d^{\mu_{\text{old}},\rho}}\left[\nabla_\theta Q^{\mu_{\text{old}}}\left(S,\mu_\theta(S)\right)\right] \\ &= \frac{1}{1-\gamma}\mathbb{E}_{S\sim d^{\mu_{\text{old}},\rho}}\left[\nabla_\theta \mu_\theta(S)\ \nabla_a Q^{\mu_{\text{old}}}(S,a)|_{a=\mu_\theta(S)}\right]. \end{aligned} \tag{31}$$

At

$$\theta = \theta_{\text{old}},$$

$$\begin{aligned}\nabla_\theta J_{\text{actor}}(\theta;\theta_{\text{old}})|_{\theta=\theta_{\text{old}}} &= \frac{1}{1-\gamma}\mathbb{E}_{S\sim d^{\mu_{\text{old}},\rho}}\left[\nabla_\theta \mu_\theta(S)|_{\theta=\theta_{\text{old}}}\ \nabla_a Q^{\mu_{\text{old}}}(S,a)|_{a=\mu_{\text{old}}(S)}\right] \\ &= \nabla_\theta J(\theta)|_{\theta=\theta_{\text{old}}}, \end{aligned} \tag{32}$$

where the final equality follows from the deterministic policy gradient theorem.

Hence,

$$\boxed{\nabla_\theta J_{\text{actor}}(\theta;\theta_{\text{old}})|_{\theta=\theta_{\text{old}}} = \nabla_\theta J(\theta)|_{\theta=\theta_{\text{old}}}.} \tag{33}$$

Therefore, $J_{\text{actor}}(\theta;\theta_{\text{old}})$ is a local actor surrogate rather than the true performance objective, but it is first-order exact at the current policy.

3 From the DPG Theorem to a Practical Actor Update

The deterministic policy gradient theorem requires the true state-action value function

$$Q^{\mu_\theta}(s,a),$$

which is generally unknown in a model-free problem. Introduce a differentiable function approximator

$$\boxed{\hat{q}(s,a,w) \approx Q^{\mu_\theta}(s,a),}$$

where w denotes the state-action value-function parameters.

Given states

$$S_i \sim d^{\mu_\theta,\rho}, \qquad i = 1,\dots,B,$$

define the finite-sample actor surrogate

$$\boxed{\widehat{J}_{\text{actor}}(\theta;w) := \frac{1}{(1-\gamma)B}\sum_{i=1}^{B}\hat{q}\left(S_i,\mu_\theta(S_i),w\right).} \tag{34}$$

During the actor update, the sampled states S_i and the state-action value-function parameters w are held fixed. Therefore,

$$\nabla_\theta \widehat{J}_{\text{actor}}(\theta;w) = \frac{1}{(1-\gamma)B}\sum_{i=1}^{B}\nabla_\theta\mu_\theta(S_i)\,\nabla_a\hat{q}(S_i,a,w)|_{a=\mu_\theta(S_i)}. \tag{35}$$

The factor $1/(1-\gamma)$ is constant with respect to θ and therefore only rescales the gradient. In implementation, it is usually omitted. Define the implementation actor objective

$$\boxed{\widehat{J}^{\text{impl}}_{\text{actor}}(\theta;w) := \frac{1}{B}\sum_{i=1}^{B}\hat{q}\left(S_i,\mu_\theta(S_i),w\right).} \tag{36}$$

Thus,

$$\widehat{J}^{\text{impl}}_{\text{actor}}(\theta;w) = (1-\gamma)\widehat{J}_{\text{actor}}(\theta;w), \tag{37}$$

and consequently

$$\boxed{\nabla_\theta \widehat{J}^{\text{impl}}_{\text{actor}}(\theta;w) = \frac{1}{B}\sum_{i=1}^{B}\nabla_\theta\mu_\theta(S_i)\,\nabla_a\hat{q}(S_i,a,w)|_{a=\mu_\theta(S_i)}.} \tag{38}$$

3.1 Automatic Differentiation

For each sampled state, the computational graph is

$$\theta \longrightarrow \mu_\theta(S_i) \longrightarrow \hat{q}\left(S_i,\mu_\theta(S_i),w\right).$$

With w fixed during the actor update, backpropagation through this graph automatically computes

$$\nabla_\theta\mu_\theta(S_i)\,\nabla_a\hat{q}(S_i,a,w)|_{a=\mu_\theta(S_i)}.$$

Hence, the policy Jacobian and action-value gradient do not need to be formed and multiplied explicitly; automatic differentiation produces the deterministic policy-gradient update directly.

4 From DPG to Deep Deterministic Policy Gradient

The deterministic policy gradient theorem specifies how to update the policy once a differentiable approximation

$$\hat{q}(s, a, w) \approx Q^{\mu_\theta}(s, a)$$

is available. It does not specify how this approximation should be learned.

A complete model-free algorithm must therefore learn the state-action value function from environment transitions

$$(S_t, A_t, R_{t+1}, S_{t+1}).$$

A second issue is exploration. Since

$$A_t = \mu_\theta(S_t)$$

is deterministic, the policy itself does not generate exploratory actions. An exploratory behaviour policy is therefore used during data collection, for example

$$A_t = \mu_\theta(S_t) + \varepsilon_t. \tag{39}$$

Deep Deterministic Policy Gradient combines these ingredients with

deterministic actor:	$\mu_\theta(s)$,
state-action value approximator:	$\hat{q}(s, a, w)$,
exploratory behaviour policy,	
experience replay,	
target networks.	

(40)

The resulting actor–critic algorithm, including the learning rule for $\hat{q}(s, a, w)$, off-policy replay-buffer training, and target-network updates, is developed in the following DDPG report.