1. • Policy gradient theorem
   Since our goal is to find optimal π_θ which maximizes expected return $J(\theta)$, we express it w.r.t.
   - reward $\mathcal{R}(s,a)$: expected reward we get at state s and take action a
   - (stationary) policy $\pi(s,a)$: probability of choosing action a at state s
   - discounted-aggregate state-visitation measure $\rho^\pi(s)$: the discounted state visitation probability in infinite horizon following policy π

$$
\begin{aligned}
J(\theta) &= \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t r_t\right] && \text{(Expected Return)}\\
&= \sum_{t=0}^{\infty}\gamma^t \mathbb{E}_\pi\left[r_t\right]\\
&= \sum_{t=0}^{\infty}\gamma^t \int_{\mathcal{S}}\left[\Pr(S_t=s)\int_{\mathcal{A}}\left[\Pr(A_t=a \mid S_t=s)\mathcal{R}_s^a\right]da\right]ds\\
&= \sum_{t=0}^{\infty}\gamma^t \int_{\mathcal{S}}\left[\Pr(S_t=s)\int_{\mathcal{A}}\left[\pi_\theta(s,a)\mathcal{R}_s^a\right]da\right]ds && (\pi_\theta \doteq \text{policy func approx})\\
&= \int_{\mathcal{S}}\left[\left(\sum_{t=0}^{\infty}\gamma^t \Pr(S_t=s)\right)\int_{\mathcal{A}}\left[\pi_\theta(s,a)\mathcal{R}_s^a\right]da\right]ds\\
&= \int_{\mathcal{S}}\left[\left(\sum_{t=0}^{\infty}\gamma^t \int_{\mathcal{S}} p_0(s_0)\Pr(s_0\to s,t,\pi)ds_0\right)\int_{\mathcal{A}}\left[\pi_\theta(s,a)\mathcal{R}_s^a\right]da\right]ds\\
&= \int_{\mathcal{S}}\left[\int_{\mathcal{S}}\left[\sum_{t=0}^{\infty}\gamma^t p_0(s_0)\Pr(s_0\to s,t,\pi)\right]ds_0\int_{\mathcal{A}}\left[\pi_\theta(s,a)\mathcal{R}_s^a\right]da\right]ds\\
&= \int_{\mathcal{S}}\rho^\pi(s)\int_{\mathcal{A}}\left[\pi_\theta(s,a)\mathcal{R}_s^a\right]da\cdot ds
\end{aligned}
$$

Then we naturally take the derivative of the above expression, which yields the policy gradient theorem:

Theorem 1. *Policy Gradient Theorem*

$$\nabla_\theta J(\theta) = \int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \left[\nabla_\theta \pi_\theta(s,a) Q^\pi(s,a)\right] da \cdot ds$$

The right-hand side can also be written as $\mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s,a) Q^\pi(s,a)\right]$

Proof. We notice the following facts:

(a)

$$J(\theta) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r_t\right] = \int_{\mathcal{S}} p_0(s) V^\pi(s) ds = \int_{\mathcal{S}} p_0(s) \int_{\mathcal{A}} \pi_\theta(s,a) Q^\pi(s,a) da \cdot ds$$

(b)

$$Q^\pi(s,a) = \mathcal{R}_s^a + \gamma \int_{\mathcal{S}} \mathcal{P}_{ss'}^a V^\pi(s') ds'$$

(c)

$$V^\pi(s) = \int_{\mathcal{A}} \pi(s,a) Q^\pi(s,a) da$$

(d)

$$\nabla_\theta \mathcal{R}_s^a = 0 \quad \nabla_\theta \mathcal{P}_{ss'}^a = 0$$

(e)

$$\Pr(s \to s', 1, \pi) = \int_{\mathcal{A}} \pi(s,a) \mathcal{P}_{ss'}^a da$$

(f)

$$\Pr(s_0 \to s_1, 1, \pi) \times \Pr(s_1 \to s_2, 1, \pi) = \Pr(s_0 \to s_2, 2, \pi)$$

(g)

$$\int_{\mathcal{S}} p_0(s) \Pr(s \to s, 0, \pi) ds = 1$$

(h)

$$\rho^\pi(s) \doteq \int_{\mathcal{S}} \left[\sum_{t=0}^{\infty} \gamma^t p_0(s_0) \Pr(s_0 \to s, t, \pi)\right] ds_0$$

Using the above facts we prove PGT as follows:

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \int_{\mathcal{S}} p_0(s_0) \int_{\mathcal{A}} \pi_\theta(s_0, a_0) \cdot Q^\pi(s_0, a_0) \cdot da_0 \cdot ds_0 && \text{(Fact a)} \\
&= \int_{\mathcal{S}} p_0(s_0) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(s_0, a_0) \cdot Q^\pi(s_0, a_0) \cdot da_0 \cdot ds_0 \\
&+ \int_{\mathcal{S}} p_0(s_0) \int_{\mathcal{A}} \pi_\theta(s_0, a_0) \cdot \nabla_\theta Q^\pi(s_0, a_0) \cdot da_0 \cdot ds_0 && (\star) \\
&= \int_{\mathcal{S}} p_0(s_0) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(s_0, a_0) \cdot Q^\pi(s_0, a_0) \cdot da_0 \cdot ds_0 \\
&+ \int_{\mathcal{S}} p_0(s_0) \int_{\mathcal{A}} \pi_\theta(s_0, a_0) \cdot \nabla_\theta \left(\int_{\mathcal{S}} \gamma \cdot \mathcal{P}_{s_0, s_1}^{a_0} \cdot V^\pi(s_1) \cdot ds_1 \right) \cdot da_0 \cdot ds_0 \\
& && \text{(Fact b, d)} \\
&= \int_{\mathcal{S}} p_0(s_0) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(s_0, a_0) \cdot Q^\pi(s_0, a) \cdot da_0 \cdot ds_0 \\
&+ \int_{\mathcal{S}} p_0(s_0) \int_{\mathcal{A}} \pi_\theta(s_0, a_0) \int_{\mathcal{S}} \gamma \cdot \mathcal{P}_{s_0, s_1}^{a_0} \cdot \nabla_\theta V^\pi(s_1) \cdot ds_1 \cdot da_0 \cdot ds_0 \\
& && \text{(Fact d)} \\
&= \int_{\mathcal{S}} p_0(s_0) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(s_0, a_0) \cdot Q^\pi(s_0, a_0) \cdot da_0 \cdot ds_0 \\
&+ \int_{\mathcal{S}} \left(\int_{\mathcal{S}} \gamma \cdot p_0(s_0) \int_{\mathcal{A}} \pi_\theta(s_0, a_0) \cdot \mathcal{P}_{s_0, s_1}^{a_0} \cdot da_0 \cdot ds_0 \right) \cdot \nabla_\theta V^\pi(s_1) \cdot ds_1 \\
& && \text{(change order of integration)} \\
&= \int_{\mathcal{S}} p_0(s_0) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(s_0, a_0) \cdot Q^\pi(s_0, a_0) \cdot da_0 \cdot ds_0 \\
&+ \int_{\mathcal{S}} \left(\int_{\mathcal{S}} \gamma \cdot p_0(s_0) \cdot \Pr(s_0 \to s_1, 1, \pi) \cdot ds_0 \right) \cdot \nabla_\theta V^\pi(s_1) \cdot ds_1 && \text{(Fact e)} \\
&= \int_{\mathcal{S}} p_0(s_0) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(s_0, a_0) \cdot Q^\pi(s_0, a_0) \cdot da_0 \cdot ds_0 \\
&+ \int_{\mathcal{S}} \left(\int_{\mathcal{S}} \gamma p_0(s_0) \Pr(s_0 \to s_1, 1, \pi) \cdot ds_0 \right) \cdot \nabla_\theta \left(\int_{\mathcal{A}} \pi_\theta(s_1, a_1) Q^\pi(s_1, a_1) \, da_1 \right) \cdot ds_1 \\
& && \text{(Fact c)}
\end{aligned}
$$

From here we observe the same structure with step $(\star)$, and thus we can expand this term in the same way and iterate, note that we need Fact f and g to merge some of the terms in each iteration, finally we arrive at:

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \sum_{t=0}^{\infty} \int_{\mathcal{S}} \int_{\mathcal{S}} \gamma^t \cdot p_0\left(s_0\right) \cdot \operatorname{Pr}\left(s_0 \rightarrow s_t, t, \pi\right) \cdot d s_0 \int_{\mathcal{A}} \nabla_\theta \pi_\theta\left(s_t, a_t\right) \cdot Q^\pi\left(s_t, a_t\right) \cdot d a_t \cdot d s_t \\
&= \int_{\mathcal{S}} \int_{\mathcal{S}} \sum_{t=0}^{\infty} \gamma^t \cdot p_0\left(s_0\right) \cdot \operatorname{Pr}\left(s_0 \rightarrow s, t, \pi\right) \cdot d s_0 \int_{\mathcal{A}} \nabla_\theta \pi_\theta(s, a) \cdot Q^\pi(s, a) \cdot d a \cdot d s \\
&= \int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(s, a) \cdot Q^\pi(s, a) \cdot d a \cdot d s \qquad \text{(Fact h)}
\end{aligned}
$$

The other expression mentioned in the theorem can be quickly derived as follows:

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(s, a) \cdot Q^\pi(s, a) \cdot d a \cdot d s \\
&= \int_{\mathcal{S}} \rho^\pi(s) \int_{\mathcal{A}} \pi_\theta(s, a) \nabla_\theta \log \pi_\theta(s, a) \cdot Q^\pi(s, a) \cdot d a \cdot d s \\
&= \int_{\mathcal{S}} \int_{\mathcal{A}} \left(\rho^\pi(s) \pi_\theta(s, a)\right)\left(\nabla_\theta \log \pi_\theta(s, a) \cdot Q^\pi(s, a)\right) \cdot d a \cdot d s \\
&= \mathbb{E}_{\pi_\theta}\left[\nabla_\theta \log \pi_\theta(s, a) Q^\pi(s, a)\right]
\end{aligned}
$$

■

- Score function $\nabla_\theta \log \pi_\theta(s,a)$
  - Softmax policy
    The way we parameterize π is to weight actions using linear combinations of features: $\theta^\intercal \phi(s,a)$

$$\pi_\theta(s,a) = \frac{e^{\theta^\intercal \phi(s,a)}}{\sum_{\mathcal{A}} e^{\theta^\intercal \phi(s,a)}}$$

    The score function is derived as follows:

$$\begin{aligned}
\nabla_\theta \log \pi_\theta(s,a) &= \phi(s,a) - \frac{\sum_{\mathcal{A}} \phi(s,a) e^{\theta^\intercal \phi(s,a)}}{\sum_{\mathcal{A}} e^{\theta^\intercal \phi(s,a)}} \\
&= \phi(s,a) - \sum_{\mathcal{A}} \phi(s,a) \frac{e^{\theta^\intercal \phi(s,a)}}{\sum_{\mathcal{A}} e^{\theta^\intercal \phi(s,a)}} \\
&= \phi(s,a) - \sum_{\mathcal{A}} \phi(s,a) \pi_\theta(s,a) \\
&= \phi(s,a) - \mathbb{E}_\pi[\phi(s,\cdot)]
\end{aligned}$$

  - Gaussian policy
    Here our action space is continuous and the action is sampled from a Gaussian distribution $\mathcal{N}(\theta^\intercal \phi(s), \sigma^2)$, note that we use state features in this case and assume fixed variance.
    The score function is derived as follows:

$$\begin{aligned}
\nabla_\theta \log \pi_\theta(s,a) &= \nabla_\theta \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(a-\theta^\intercal \phi(s))^2}{2\sigma^2}} \\
&= \nabla_\theta - \frac{(a-\theta^\intercal \phi(s))^2}{2\sigma^2} \\
&= \frac{(a-\theta^\intercal \phi(s))\phi(s)}{\sigma^2}
\end{aligned}$$