From Discounted MDP to Average MDP

1. Discounted MDP

$$
\begin{aligned}
&\pi \xrightarrow{P(s'|s,a)} P_\pi(s'|s) = \sum_a \pi(a|s)P(s'|s,a) \\
&\xrightarrow{p_0} MC_\pi = \{\tau = (s_0, a_0, s_1, a_1, s_2, a_2, \dots) : s_0 \sim p_0, a_t = \pi(\cdot|s_t), s_{t+1} \sim P(\cdot|s_t, a_t)\} \\
&\xrightarrow{\gamma, r} \rho_\gamma(\pi) = \mathbb{E}_{\tau \sim MC_\pi}\left[(1-\gamma)\sum_{t=0}^{\infty} \gamma^t r(s_t)\right] = \sum_s p_\gamma(s;\pi) r(s), \\
&where\ p_\gamma(s;\pi) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t p_t(s;\pi),\ p_t(s;\pi) = Pr(s_t = s;\pi).
\end{aligned}
$$

Matrix Form

We denote that $\mathcal{S} = \{s1, s2, \dots, sN\}$, then

$$
P_\pi = \begin{bmatrix}
P_\pi(s1|s1) & P_\pi(s2|s1) & \cdots & P_\pi(sN|s1) \\
P_\pi(s1|s2) & P_\pi(s2|s2) & \cdots & P_\pi(sN|s2) \\
\vdots & \vdots & \ddots & \vdots \\
P_\pi(s1|sN) & P_\pi(s2|sN) & \cdots & P_\pi(sN|sN)
\end{bmatrix}
$$

and

$$
p_1^\pi = P_\pi^T p_0, \quad p_2^\pi = (P_\pi^T)^2 p_0, \quad \dots, \quad p_t^\pi = (P_\pi^T)^t p_0.
$$

Now

$$
p_\gamma^\pi = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t (P_\pi^T)^t p_0.
$$

Policy gradient theorem

$$
\theta_\pi \to \pi \to \cdots \to \rho_\gamma(\pi) \to \rho_\gamma(\theta_\pi).
$$

$$
\frac{\mathrm{d}\,\rho_\gamma(\theta_\pi)}{\mathrm{d}\,\theta_\pi} = \sum_s p_\gamma^\pi(s) \sum_a \pi(a|s) \nabla_\theta \log \pi(a|s) Q_\gamma^\pi(s,a)
$$

$$
Q_\gamma^\pi(s,a) = \mathbb{E}_{\tau \sim MC_\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t \middle| s_0 = s, a_0 = a\right].
$$

Off-policy Settings

We only follows behavior policy to sample from environment

$$
MC_\mu = \{\tau = (s_0, a_0, r_0, s_1, a_1, r_1, s_2, a_2, r_2, \dots) : s_0 \sim p_0, a_t = \mu(\cdot|s_t), s_{t+1} \sim P(\cdot|s_t, a_t)\}.
$$

Here are three problems:

$$
\frac{\mathrm{d}\,\rho(\theta_\pi)}{\mathrm{d}\,\theta_\pi} = \mathbb{E}_{s \sim p_\gamma^\pi, a \sim \pi(s)}\{\nabla_\theta \log \pi(a|s) Q_\gamma^\pi(s,a)\}
$$

- From $p_\gamma^\mu = (1-\gamma)\sum_{t=0}^\infty \gamma^t (P_\mu^T)^t p_0$ to $p_\gamma^\pi = (1-\gamma)\sum_{t=0}^\infty \gamma^t (P_\pi^T)^t p_0$;
- From $\mu(a|s)$ to $\pi(a|s)$;
- From $Q_\gamma^\pi(s,a)$ to $Q_\gamma^\mu(s,a)$.

$$\frac{\mathrm{d}\,\rho(\theta_\pi)}{\mathrm{d}\,\theta_\pi} = \mathbb{E}_{s\sim p_\gamma^\mu, a\sim\mu(s)}\left\{\frac{p_\gamma^\pi(s)}{p_\gamma^\mu(s)}\frac{\pi(a|s)}{\mu(a|s)}\nabla_\theta \log\pi(a|s) Q_\gamma^\pi(s,a)\right\}.$$

2. From Discounted MDP to Average MDP

$$\begin{aligned}
&\pi \xrightarrow{P(s'|s,a)} P_\pi(s'|s) = \sum_a \pi(a|s)P(s'|s,a) \\
&\xrightarrow{p_0} MC_\pi = \{\tau = (s_0, a_0, s_1, a_1, s_2, a_2, \ldots) : s_0 \sim p_0, a_t = \pi(\cdot|s_t), s_{t+1} \sim P(\cdot|s_t, a_t)\} \\
&\xrightarrow{\gamma, r} \rho_\gamma(\pi) = \mathbb{E}_{\tau\sim MC_\pi}\left[(1-\gamma)\sum_{t=0}^\infty \gamma^t r(s_t)\right] = \sum_s p_\gamma(s;\pi) r(s), \\
&where\; p_\gamma(s;\pi) = (1-\gamma)\sum_{t=0}^\infty \gamma^t p_t(s;\pi),\; p_t(s;\pi) = Pr(s_t = s;\pi).
\end{aligned}$$

$$\begin{aligned}
&\pi \xrightarrow{P(s'|s,a), p_0(s), \gamma} P_{\pi,\gamma}(s'|s) = \gamma\sum_a \pi(a|s)P(s'|s,a) + (1-\gamma)p_0(s') \\
&\xrightarrow{p_0} MC_{\pi,\gamma} = \{\tau = (s_0, a_0, s_1, a_1, s_2, a_2, \ldots) : s_0 \sim p_0, a_t = \pi(\cdot|s_t), s_{t+1} \sim \gamma P(\cdot|s_t, a_t) + (1-\gamma)p_0\} \\
&\xrightarrow{r} \rho_{stationary}(\pi) = \mathbb{E}_{s\sim d_{\pi,\gamma}}[r(s)] = \sum_s d_{\pi,\gamma}(s) r(s),
\end{aligned}$$

where $d_{\pi,\gamma}$ is stationary distribution that satisfies $d_{\pi,\gamma} = P_{\pi,\gamma}^T d_{\pi,\gamma}$.

Matrix Form

The state transition matrix is

$$P_{\pi,\gamma} = \gamma P_\pi + (1-\gamma) e p_0^T.$$

> **Lemma 1**.
>
> $$d_{\pi,\gamma} = p_\gamma^\pi = (1-\gamma)\sum_{t=0}^\infty \gamma^t (P_\pi^T)^t p_0.$$
>
> This theorem also means $\rho_{stationary}(\pi) = \rho_\gamma(\pi)$.
>
> **proof**:
>
> $$\begin{aligned}
d_{\pi,\gamma} &= P_{\pi,\gamma}^T d_{\pi,\gamma} \\
&= [\gamma P_\pi^T + (1-\gamma)p_0 e^T] d_{\pi,\gamma} \\
(I - \gamma P_\pi^T) d_{\pi,\gamma} &= (1-\gamma) p_0 \\
d_{\pi,\gamma} &= (1-\gamma)(I - \gamma P_\pi^T)^{-1} p_0 \\
&= (1-\gamma)\sum^\infty \gamma^t (P_\pi^T)^t p_0
\end{aligned}$$

Average MDP

$$\rho_{avg}(\pi) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{\tau\sim MC_{\pi,\gamma}}[r_t].$$

$$\rho_{stationary}(\pi) = \mathbb{E}_{s\sim d_{\pi,\gamma}}[r(s)] = \sum_s d_{\pi,\gamma}(s) r(s)$$

> **Lemma 2**.

> $$\rho_{avg}(\pi) = \rho_{stationary}(\pi).$$

Policy Gradient Theorem

$$\frac{\mathrm{d}\,\rho_{avg}(\theta_\pi)}{\mathrm{d}\,\theta_\pi} = \sum_s d_{\pi,\gamma}(s) \sum_a \pi(a|s) \nabla_\theta \log \pi(a|s) Q^\pi_{avg}(s,a)$$

$$Q^\pi_{avg}(s,a) = \mathbb{E}_{\tau \sim MC_{\pi,\gamma}} \left[\sum_{t=0}^{\infty} (r_t - \rho(\pi)) \big| s_0 = s, a_0 = a \right]$$

Off-policy Settings

We only follows behavior policy to sample from environment

- $MC_{\mu,\gamma} = \{\tau = (s_0, a_0, r_0, s_1, a_1, r_1, s_2, a_2, r_2, \ldots) :$ $s_0 \sim p_0, a_t = \mu(\cdot|s_t), s_{t+1} \sim \gamma P(\cdot|s_t, a_t) + (1-\gamma) p_0 \}\,;$
- $MC2_{\mu,\gamma} = \{m = (s, a, r, s') : s \sim d_{\mu,\gamma}, a = \mu(\cdot|s), s' \sim \gamma P(\cdot|s, a) + (1-\gamma) p_0 \}\,.$

3. Off-policy Algorithms

3.1 COP-TD

The algorithm's key target is to get $c(s) = \frac{d_{\pi,\gamma}(s)}{d_\mu(s)}$ that satisfies

$$\begin{cases} d_{\pi,\gamma} = P^T_{\pi,\gamma} d_{\pi,\gamma} \\ d_{\pi,\gamma} = D_\mu c \\ D_\mu = diag(d_\mu) \end{cases} \Rightarrow D_\mu c = P^T_{\pi,\gamma} D_\mu c.$$

The loss of COP-TD algorithm is

$$\begin{cases} L(c) = \frac{1}{2} \| c - D_\mu^{-1} P^T_{\pi,\gamma} D_\mu c_{target} \|^2, \\ L(c_{target}) = \frac{1}{2} \| c_{target} - c \|^2. \end{cases}$$

$$\begin{aligned} d_{\pi,\gamma} =& P^T_{\pi,\gamma} d_{\pi,\gamma} \\ d_{\pi,\gamma}(s') =& \int \int [\gamma P(s'|s,a)\pi(a|s) + (1-\gamma) p_0(s')] d_{\pi,\gamma}(s) ds da \\ =& \gamma \int \int P(s'|s,a) \pi(a|s) d_{\pi,\gamma}(s) ds da + (1-\gamma) p_0(s') \end{aligned}$$

> **Algorithm 1.** (Discounted COP-TD algorithm)
>
> $$c(s') = c(s') + \alpha \left[\gamma \frac{\pi(a|s)}{\mu(a|s)} c(s) + (1-\gamma) - c(s') \right].$$
>
> - Target space: $MC_{\pi,\gamma} = \{\tau = (s_0, a_0, s_1, a_1, s_2, a_2, \ldots) : s_0 \sim d_\mu, a_t = \pi(\cdot|s_t), s_{t+1} \sim \gamma P(\cdot|s_t, a_t) + (1-\gamma) d_\mu \}$;
> - Sample space: $MC2_\mu = \{m = (s, a, r, s') : s \sim d_\mu, a \sim \mu(s), s' \sim P(s'|s, a)\}.$

3.2 GenDICE

$$\frac{\mathrm{d}\,\rho(\theta_\pi)}{\mathrm{d}\,\theta_\pi} = \mathbb{E}_{s \sim p^\mu_\gamma, a \sim \mu(s)} \left\{ \frac{p^\pi_\gamma(s)}{p^\mu_\gamma(s)} \frac{\pi(a|s)}{\mu(a|s)} \nabla_\theta \log \pi(a|s) Q^\pi_\gamma(s,a) \right\}.$$

A new target is finding the ratio function

$$r(s,a) = \frac{p_\gamma^\pi(s,a)}{p_\gamma^\pi(s,a)} = \frac{p_\gamma^\pi(s)\pi(a|s)}{p_\gamma^\mu(s)\mu(a|s)} = \frac{d_{\pi,\gamma}(s)\pi(a|s)}{d_{\mu,\gamma}(s)\mu(a|s)}$$

We need to find a new target equation:

$$\begin{aligned}
d_{\pi,\gamma} =& P_{\pi,\gamma}^T d_{\pi,\gamma} \\
d_{\pi,\gamma}(s') =& \int\int [\gamma P(s'|s,a)\pi(a|s) + (1-\gamma)p_0(s')] d_{\pi,\gamma}(s) dsda \\
=& \gamma \int\int P(s'|s,a)\pi(a|s) d_{\pi,\gamma}(s) dsda + (1-\gamma)p_0(s') \\
\pi(a'|s') d_{\pi,\gamma}(s') =& \gamma \int\int \pi(a'|s')P(s'|s,a)\pi(a|s) d_{\pi,\gamma}(s) dsda + (1-\gamma)\pi(a'|s')p_0(s') \\
d_{\pi,\gamma}(s',a') =& \gamma \int\int \pi(a'|s')P(s'|s,a) d_{\pi,\gamma}(s',a') dsda + (1-\gamma)\pi(a'|s')p_0(s') \\
=& \int\int \pi(a'|s')[\gamma P(s'|s,a) + (1-\gamma)p_0(s')] d_{\pi,\gamma}(s',a') dsda \\
=& \int\int P_{\pi,\gamma}(s',a'|s,a) d_{\pi,\gamma}(s',a') dsda \\
d_{\mu,\gamma}(s',a') r(s',a') =& \int\int P_{\pi,\gamma}(s',a'|s,a) d_{\mu,\gamma}(s,a) r(s,a) dsda \\
D_{\mu,\gamma} r =& P_{\pi,\gamma} D_{\mu,\gamma} r.
\end{aligned}$$

The loss of GenDICE is

$$\min_{r \succeq 0} D_\phi(P_{\pi,\gamma} D_{\mu,\gamma} r \| D_{\mu,\gamma} r), \quad s.t. \mathbb{E}_{d_{\mu,\gamma}}[r] = 1.$$

> Definition (f-divergence) For $\phi : \mathbb{R}_+ \to \mathbb{R}$ is convex function, lower-semicontinuous function with $\phi(1) = 0$
>
> $$D_\phi(p\|q) = \int q(x)\phi\left(\frac{p(x)}{q(x)}\right) dx$$

$$\begin{aligned}
&\min_r D_\phi(P_{\pi,\gamma} D_{\mu,\gamma} r \| D_{\mu,\gamma} r) \\
=& \min_r \int\int D_{\mu,\gamma} r(s,a) \phi\left(\frac{P_{\pi,\gamma} D_{\mu,\gamma} r(s,a)}{D_{\mu,\gamma} r(s,a)}\right) dsda \\
=& \min_r \int\int D_{\mu,\gamma} r(s,a) \max_{f(s,a)} \left(\frac{P_{\pi,\gamma} D_{\mu,\gamma} r(s,a)}{D_{\mu,\gamma} r(s,a)} f(s,a) - \phi^*(f(s,a))\right) dsda \\
=& \min_r \max_f \int\int P_{\pi,\gamma} D_{\mu,\gamma} r(s,a) f(s,a) - D_{\mu,\gamma} r(s,a) \phi^*(f(s,a)) dsda
\end{aligned}$$

3.3 GradientDICE

$$\min_{r \succeq 0} \frac{1}{2} \|P_{\pi,\gamma} D_{\mu,\gamma} r - D_{\mu,\gamma} r\|^2_{D^{-1}_{\mu,\gamma}}, \quad s.t. \mathbb{E}_{d_{\mu,\gamma}}[r] = 1.$$

$$\min_{r \succeq 0} \frac{1}{2} \|D^{-1}_{\mu,\gamma} P_{\pi,\gamma} D_{\mu,\gamma} r - r\|^2_{D_{\mu,\gamma}}, \quad s.t. \mathbb{E}_{d_{\mu,\gamma}}[r] = 1.$$