# Policy Gradient Methods - Track 1

## 1. Policy Gradient for a Mixture of Policies

Define

$$v(s,\theta,\omega) = \sum_o \mu(o|s,\theta) \sum_a \pi(a|s,o,\omega)\Big(r(s,a) + \gamma \sum_s{}' P(s'|s,a)v(s',\theta,\omega)\Big)$$

$$\begin{aligned}\frac{\partial v(s,\theta,\omega)}{\partial \omega_i} &= \sum_o \mu(o|s,\theta)\frac{\partial}{\partial \omega_i} \sum_a \pi(a|s,o,\omega)\Big(r(s,a) + \gamma \sum_s{}' P(s'|s,a)v(s',\theta,\omega)\Big) \\ &= \sum_o \mu(o|s,\theta)\frac{1}{1-\gamma}\mathbf{E}\left[\sum_a q(s,a)\frac{\partial \pi(a|s,o,\omega)}{\partial \omega_i}\right]\end{aligned}$$

Last line is obtained from policy gradient theorem.

$$\begin{gathered}\frac{\partial v(s,\theta,\omega)}{\partial \theta_i} = \\ \sum_o \frac{\partial \mu(o,s,\theta)}{\partial \theta_i} \sum_a \pi(a|s,o,\omega)\left(r(s,a) + \gamma \sum_{s'} P(s'|s,a)v(s',\theta,\omega)\right) + \\ \sum_o \mu(o|s,\theta) \sum_a \pi(a|s,o,\omega)\gamma \sum_{s'} P(s'|s,a)\frac{\partial v(s',\theta,\omega)}{\partial \theta_i}\end{gathered} \tag{1}$$

Let $g_{\theta_i} = \frac{\partial v(s,\theta,\omega)}{\partial \theta_i}$ Let $h_{\theta_i} = \sum_o \frac{\partial \mu(o,s,\theta)}{\partial \theta_i} \sum_a \pi(a|s,o,\omega)\left(r(s,a) + \gamma \sum_{s'} P(s'|s,a)v(s',\theta,\omega)\right)$

Both $g_{\theta_i}$ and $h_{\theta_i}$ are $R^{|S|*1}$

Define $P(s, s') = \sum_o \mu(o|s, \theta) \sum_a \pi(a|s, o, \omega) P(s'|s, a)$, where $P \in R^{|S|*|S|}$

Now we can write Equation (1) as

$$
\begin{aligned}
g_{\theta_i} &= h_{\theta_i} + \gamma P g_{\theta_i} \\
\equiv g_{\theta_i} &= (I - \gamma P)^{-1} h_{\theta_i}
\end{aligned}
$$

Since $\gamma$ is the discount factor and is less than 1. $\sigma(\gamma P) < 1$ and $(I - \gamma P)^{-1}$ exists. Moreover. $(I - \gamma P)^{-1} = \sum_{t=0}^{\infty} \gamma^t P$

$$
\begin{aligned}
\frac{\partial v(s, \theta, \omega)}{\partial \theta_i} &= g_{\theta_i} = (I - \gamma P)^{-1} h_{\theta_i} = \sum_{t=0}^{\infty} \gamma^t P h_{\theta_i} = \frac{1}{1-\gamma} \mathbf{E}[h_{\theta_i}] \\
&= \frac{1}{1-\gamma} \mathbf{E}\left[ \sum_o \frac{\partial \mu(o, s, \theta)}{\partial \theta_i} \sum_a \pi(a|s, o, \omega) \left( r(s, a) + \gamma \sum_{s'} P(s'|s, a) v(s', \theta, \omega) \right) \right]
\end{aligned}
$$

## 2. Policy Gradient Hessian

Derive a policy Hessian theorem for the discounted case. You can follow the same derivation as for the first order policy gradient theorem shown in class. First order policy gradient theorm:

$$
\frac{\partial V_\theta(S_0)}{\partial \theta_i} = \frac{1}{1-\gamma} \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta_i} Q_\theta(s, a) = \frac{1}{1-\gamma} \mathbb{E}\Big[ \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta_i} Q_\theta(s, a) \Big]
$$

Based on that, we can derive the Hessian matrix for policy gradient:

$$
\begin{aligned}
\frac{\partial V_\theta(S_0)}{\partial \theta_i \partial \theta_j} &= \frac{1}{1-\gamma} \frac{\partial}{\partial \theta_j} \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta_i} Q_\theta(s, a) \\
&= \frac{1}{1-\gamma} \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta_i \partial \theta_j} Q_\theta(s, a) + \frac{1}{1-\gamma} \sum_a \nabla_{\theta_i} \pi(a|s) \frac{\partial Q_\theta(s, a)}{\partial \theta_j}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial Q_\theta(s,a)}{\partial \theta_j} &= \frac{\partial}{\partial \theta_j}\Big[r(s,a) + \gamma \sum_{s'} P(s'|s,a)V_\theta(s')\Big] \\
&= \gamma \sum_{s'} P(s'|s,a)\frac{\partial}{\partial \theta_j}V_\theta(s')
\end{aligned}
$$

$$
\Rightarrow \frac{\partial V_\theta(S_0)}{\partial \theta_i \partial \theta_j} = \frac{1}{1-\gamma}\sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta_i \partial \theta_j}Q_\theta(s,a) + \frac{1}{1-\gamma}\sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta_i}\gamma \sum_{s'} P(s'|s,a)\frac{\partial}{\partial \theta_j}V_\theta(s')
$$

From policy gradient theorm we know that

$$
\frac{\partial V_\theta(S_0)}{\partial \theta_i \theta_j} \doteq g_\theta(s) = (I - \gamma P_\theta)^{-1} h_\theta
$$

where,

$$
\begin{aligned}
P_\theta &\doteq \sum_a \pi_\theta(a|s)P(s'|s,a) \\
h_\theta &\doteq \sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta_i \partial \theta_j}Q_\theta(s,a)
\end{aligned}
$$

Denote

$$
g_\theta(s) = \sum_{t=0}^{\infty} \gamma^t P h_{\theta_i} = \frac{1}{1-\gamma}\mathbb{E}\Big[\sum_a \frac{\partial \pi_\theta(a|s)}{\partial \theta_j}Q_\theta(s,a)\Big]
$$

$$\Rightarrow \frac{\partial V_\theta(S_0)}{\partial\theta_i\partial\theta_j} = \frac{1}{1-\gamma}\sum_a \frac{\partial\pi_\theta(a|s)}{\partial\theta_i\partial\theta_j}Q_\theta(s,a) + \frac{1}{1-\gamma}\sum_a \frac{\partial\pi_\theta(a|s)}{\partial\theta_i}\gamma\sum_{s'}P(s'|s,a)\frac{1}{1-\gamma}\mathbb{E}\Big[\sum_a \frac{\partial\pi_\theta(a|s)}{\partial\theta_j}Q_\theta(s,a)\Big]$$
$$= \frac{1}{1-\gamma}\sum_a \frac{\partial\pi_\theta(a|s)}{\partial\theta_i\partial\theta_j}Q_\theta(s,a) + \frac{1}{1-\gamma}\gamma\mathbb{E}\Big[\sum_a \frac{\partial\pi_\theta(a|s)}{\partial\theta_j}\Big]\frac{1}{1-\gamma}\mathbb{E}\Big[\sum_a \frac{\partial\pi_\theta(a|s)}{\partial\theta_j}Q_\theta(s,a)\Big]$$
$$= \frac{1}{1-\gamma}\mathbb{E}\Big[\sum_a \frac{\partial\pi_\theta(a|s)}{\partial\theta_i\partial\theta_j}Q_\theta(s,a) + \frac{\gamma}{1-\gamma}\sum_a \frac{\partial\pi_\theta(a|s)}{\partial\theta_i}\sum_a \frac{\partial\pi_\theta(a|s)}{\partial\theta_j}Q_\theta(s,a)\Big]$$

### 3. Constrained Optimization / Intrinsic Rewards

$$J_\alpha(\theta) = \mathbf{E}_{s_0\sim\alpha,\theta}\left[\sum_{t=0}^{\infty}\gamma^t r(S_t,A_t)\right] - \eta\mathbf{E}_{s_0\sim\alpha,\theta}\left[\sum_{t=0}^{\infty}\gamma^t c(S_t,A_t)\right]$$

$$J_\alpha(\theta) = \mathbf{E}_{s_0\sim\alpha,\theta}\left[\sum_{t=0}^{\infty}\gamma^t\left(r(S_t,A_t) - \eta c(S_t,A_t)\right)\right]$$
$$= \mathbf{E}_{s_0\sim\alpha,\theta}\left[r(S_0,A_0) - \eta c(S_0,A_0) + \sum_{t=1}^{\infty}\gamma^t\left(r(S_t,A_t) - \eta c(S_t,A_t)\right)\right]$$
$$= \mathbf{E}_{s_0\sim\alpha,\theta}\left[r(S_0,A_0) - \eta c(S_0,A_0) + \gamma\sum_{t=0}^{\infty}\gamma^t\left[r(S_{t+1},A_{t+1}) - \eta c(S_{t+1},A_{t+1})\right]\right]$$

Since our objective is to maximize the reward gained by starting from $S_0$, we can write $J_\alpha(\theta)$ as
$$J_\alpha(\theta) = V(S_0) = \mathbf{E}_{s_0\sim\alpha,\theta}\left[r(S_0,A_0) - \eta c(S_0,A_0) + \gamma V(S_1)\right]$$

Take the gradient of $V(S_0)$ with respect to $\theta_i$
$$\frac{\partial V(S_0)}{\partial\theta_i} = \mathbf{E}_{s_0\sim\alpha,\theta}\left[\sum_{A_0}\frac{\partial\pi(A_0|S_0,\theta))}{\partial\theta_i}\left[r(S_0,A_0) - \eta c(S_0,A_0) + \gamma\sum_{S'}P(S'|S_0,A_0)V_\pi(S')\right] + \gamma\frac{\partial V(S_1)}{\partial\theta_i}\right]$$

Last term in the above equation is $\gamma \frac{\partial V(S_1)}{\partial \theta_i}$. By recursing on it, we obtain

$$\frac{\partial V(S_0)}{\partial \theta_i} = \sum_{t=0}^{\infty} \gamma^t \mathbf{E}_{s_t,\theta} \left[ \sum_{A_t} \frac{\partial \pi(A_t|S_t,\theta))}{\partial \theta_i} [r(S_t,A_t) - \eta c(S_t,A_t) + \gamma \sum_{S'} P(S'|S_t,A_t) V_\pi(S')] \right]$$