Let us start with the defined objective function $J(\theta)$. We can expand the expectation as:

$$
\begin{aligned}
J(\theta) &= \mathbb{E}[\sum_{t=0}^{T-1} r_{t+1}|\pi_\theta] \\
&= \sum_{t=0}^{T-1} P(s_t, a_t|\tau)R(s_t, a_t)
\end{aligned}
$$

Differentiate both sides with respect to policy parameter θ:

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \sum_{t=0}^{T-1} \nabla_\theta P(s_t, a_t|\tau) r_{t+1} \\
&= \sum_{t=0}^{T-1} P(s_t, a_t|\tau)\nabla_\theta log P(s_t, a_t|\tau) r_{t+1} \\
&= \mathbb{E}[\sum_{t=0}^{T-1} \nabla_\theta log P(s_t, a_t|\tau) r_{t+1}]
\end{aligned}
$$

However, during, learning, we take random samples of episodes instead of computing the expectation, so

$$
\nabla_\theta J(\theta) \sim \sum_{t=0}^{T-1} \nabla_\theta log P(s_t, a_t|\tau) r_{t+1}
$$

From here, let us take a more careful look into $\nabla_\theta log P(s_t, a_t|\tau)$.

First, by definition,

$$
\begin{aligned}
P(s_t, a_t|\tau) &= P(s_0, a_0, s_1, a_2, ..., s_{t-1}, a_{t-1}, s_t|\pi_\theta) \\
&= P(s_0)\pi_\theta(a_1|s_0)P(s_1|s_0, a_0)\pi_\theta(a_2|s_1) \\
&\quad P(s_2|s_1, a_1)\pi_\theta(a_3|s_2)...P(s_{t-1}|s_{t-2}, a_{t-2})\pi_\theta(a_{t-1}|s_{t-2})P(s_t)
\end{aligned}
$$

If we *log* both sides,

$$
\begin{aligned}
log P(s_t, a_t|\tau) &= log(P(s_0)\pi_\theta(a_1|s_0)P(s_1|s_0, a_0)\pi_\theta(a_2|s_1)P(s_2|s_1, a_1)\pi_\theta(a_3|s_2)... \\
&\quad P(s_{t-1}|s_{t-2}, a_{t-2})\pi_\theta(a_{t-1}|s_{t-2})P(s_t)) \\
&= log P(s_0) + log\pi_\theta(a_1|s_0) + log P(s_1|s_0, a_0) + log\pi_\theta(a_2|s_1) \\
&\quad + log P(s_2|s_1, a_1) + log\pi_\theta(a_3|s_2) + ... \\
&\quad + log P(s_{t-1}|s_{t-2}, a_{t-2}) + log\pi_\theta(a_{t-1}|s_{t-2}) + log P(s_t)
\end{aligned}
$$

Then, differentiating $logP(s_t, a_t|\tau)$ with respect to θ yields:

$$\begin{aligned}
\nabla_\theta logP(s_t, a_t|\tau) = &\nabla_\theta logP(s_0) + \nabla_\theta log\pi_\theta(a_1|s_0) + \nabla_\theta logP(s_1|s_0, a_0) \\
&+ \nabla_\theta log\pi_\theta(a_2|s_1) + \nabla_\theta logP(s_2|s_1, a_1) + \nabla_\theta log\pi_\theta(a_3|s_2)+ \\
&... + \nabla_\theta logP(s_{t-1}|s_{t-2}, a_{t-2}) + \nabla_\theta log\pi_\theta(a_{t-1}|s_{t-2}) + \nabla_\theta logP(s_t))
\end{aligned}$$

However, note that the $P(s_t|s_{t-1}, a_{t-1})$ is not dependent on the policy parameter θ, and is solely dependant on the environment on which the reinforcement learning is acting on; it is assumed that the state transition is unknown to the agent in model free reinforcement learning. Thus, the gradient of it with respect to θ will be 0. How convenient!

So,

$$\begin{aligned}
\nabla_\theta logP(s_t, a_t|\tau) &= 0 + \nabla_\theta log\pi_\theta(a_1|s_0) + 0 + \nabla_\theta log\pi_\theta(a_2|s_1) + 0 + \nabla_\theta log\pi_\theta(a_3|s_2)+ \\
&\quad ... + 0 + \nabla_\theta log\pi_\theta(a_{t-1}|s_{t-2}) + 0 \\
&= \nabla_\theta log\pi_\theta(a_1|s_0) + \nabla_\theta log\pi_\theta(a_2|s_1) + \nabla_\theta log\pi_\theta(a_3|s_2)+ \\
&\quad ... + \nabla_\theta log\pi_\theta(a_{t-1}|s_{t-2}) \\
&= \sum_{t'=0}^{t} \nabla_\theta log\pi_\theta(a_{t'}|s_{t'})
\end{aligned}$$

Plugging this into our $\nabla_\theta J(\theta)$ yields:

$$\begin{aligned}
\nabla_\theta J(\theta) &= \sum_{t=0}^{T-1} r_{t+1} \nabla_\theta P(s_t, a_t|\tau) \\
&= \sum_{t=0}^{T-1} r_{t+1} (\sum_{t'=0}^{t} \nabla_\theta log\pi_\theta(a_{t'}|s_{t'}))
\end{aligned}$$

Lets play around that with a bit. Say, $T = 4$. Then,

$$\begin{aligned}
\nabla_\theta J(\theta) = &\sum_{t=0}^{3} r_{t+1}(\sum_{t'=0}^{t} \nabla_\theta log\pi_\theta(a_{t'}|s_{t'})) \\
&= r_1(\sum_{t'=0}^{0} \nabla_\theta log\pi_\theta(a_{t'}|s_{t'})) + r_2(\sum_{t'=0}^{1} \nabla_\theta log\pi_\theta(a_{t'}|s_{t'})) \\
&\quad + r_3(\sum_{t'=0}^{2} \nabla_\theta log\pi_\theta(a_{t'}|s_{t'})) + r_4(\sum_{t'=0}^{3} \nabla_\theta log\pi_\theta(a_{t'}|s_{t'})) \\
&= r_1 \nabla_\theta log\pi_\theta(a_0|s_0) + r_2(\nabla_\theta log\pi_\theta(a_0|s_0) + \nabla_\theta log\pi_\theta(a_1|s_1)) \\
&\quad + r_3(\nabla_\theta log\pi_\theta(a_0|s_0) + \nabla_\theta log\pi_\theta(a_1|s_1) + \nabla_\theta log\pi_\theta(a_2|s_2)) \\
&\quad + r_4(\nabla_\theta log\pi_\theta(a_0|s_0) + \nabla_\theta log\pi_\theta(a_1|s_1) + \nabla_\theta log\pi_\theta(a_2|s_2) \\
&\quad + \nabla_\theta log\pi_\theta(a_3|s_3)) \\
&= \nabla_\theta log\pi_\theta(a_0|s_0)(r_1 + r_2 + r_3 + r_4) \\
&\quad + \nabla_\theta log\pi_\theta(a_1|s_1)(r_2 + r_3 + r_4) \\
&\quad + \nabla_\theta log\pi_\theta(a_2|s_2)(r_3 + r_4) + \nabla_\theta log\pi_\theta(a_3|s_3) r_4 \\
&= \sum_{t=0}^{3} \nabla_\theta log\pi_\theta(a_t|s_t)(\sum_{t'=t}^{3} r_{t'+1})
\end{aligned}$$

So, in general,

$$\nabla_\theta J(\theta) = \sum_{t=0}^{T-1} \nabla_\theta log\pi_\theta(a_t|s_t)(\sum_{t'=t}^{T-1} r_{t'+1})$$

Incorporating the discount factor $\gamma \in [0, 1]$ for future rewards,

$$\nabla_\theta J(\theta) = \sum_{t=0}^{T-1} \nabla_\theta log\pi_\theta(a_t|s_t)(\sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t'+1})$$

For simplicity, we will denote $\sum_{t'=t}^{T-1} \gamma^{t'} r_{t'-1}$ as G_t, the discounted cumulative future reward. Replacing $\sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t+1}$ with G_t, we derive the policy gradient,

$$\nabla_\theta J(\theta) = \sum_{t=0}^{T-1} \nabla_\theta log\pi_\theta(a_t|s_t) G_t$$

Then, we update the policy paramter θ as:

$$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$$

where α is the learning rate in $[0, 1]$