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

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

where i is an arbitrary starting point in a trajectory, $P(s_t, a_t|\tau)$ is the probability of the occurrence of s_t, a_t given the trajectory τ.

Differentiate both sides with respect to policy parameter θ:

$$\text{Using} \quad \frac{d}{dx} log f(x) = \frac{f'(x)}{f(x)},$$

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

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

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

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

First, by definition, representing the starting point of the sequence in a trajectory as s_0,

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

If we log both sides,

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

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

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

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 expand that!

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

Simplifying the term $\sum_{t'=t+1}^{T} r_{t'}$ to G_t, we can derive the policy gradient

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

In corporating the discount factor $\gamma \in [0, 1]$ into our objective (in order to weight immediate rewards more than future rewards):

$J(\theta) = \mathbb{E}[\gamma^0 r_1 + \gamma^1 r_2 + \gamma^2 r_3 + ... + \gamma^{T-1} r_T | \pi_\theta]$

We can perform a similar derivation to obtain

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

and simplifying $\sum_{t'=t+1}^{T} \gamma^{t'-t-1} r_{t'}$ to G_t,

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