Starting from the objective of PG

> θ is the parameter of a policy

The policy objective functions are:

- In episodic environments: $J_1(\theta) = V^\theta(s_1) = \mathbb{E}_\theta[v_1]$
- In continuing environments:
  - If we use state value: $J_V(\theta) = \sum_s \tau^\theta(s) V^\theta(s)$
  - If we use state-action (reward) value: $J_R(\theta) = \sum_s \tau^\theta(s) \sum_a \pi_\theta(a|s) \mathcal{R}_s^a$

    > $\tau^\theta(s)$ is the stationary distribution of state s for policy $\theta : S \to \mathbb{R}$
    >
    > $\pi_\theta(s, a)$ is the probability of taking action a at state s according to parameterized policy θ.
    >
    > θ directly influence π and indirectly influence τ.

If we denote the trajectory as $\Upsilon : (s_1, a_1, s_2, a_2, \ldots, s_{T-1}, a_{T-1}, s_T)$, then the objective function can be re-written as:

$$
\begin{aligned}
J(\theta) &= \mathbb{E}_{\pi_\theta}[\sum_{t=0}^{T-1} R(s_t, a_t); \pi_\theta] \\
&= \mathbb{E}_{\Upsilon \sim p_\theta(\Upsilon)}[\sum_{t=0}^{T-1} R(s_t, a_t)] \\
&= \sum_{\Upsilon} P(\Upsilon|\theta) R(\Upsilon) = \int p_\theta(\Upsilon) r(\Upsilon) d\Upsilon
\end{aligned}
$$

which means, the expected rewards equals the sum of the probability of a trajectory $\times$ corresponding rewards.

$p_\theta(\Upsilon) = p_\theta(s_1, a_1, \ldots, s_T) = p(s_1) \prod_{t=1}^{T-1} \pi_\theta(a_t|s_t) p(s_T|s_{T-1}, a_{T-1})$ is the probability of a trajectory under policy π_θ.

Policy-based RL is an optimization problem which finds θ that directly optimizes the goal $J(\theta)$.

Policy gradient search for a local maximum in $J(\theta)$ by ascending the gradient of the policy w.r.t. θ:

$$\Delta\theta = \alpha \Delta_\theta J(\theta)$$

With the policy gradient theorem, we have:

> For any differentiable policy $\pi_\theta(s, a)$, for any of the policy objective functions $J = J_1, J_V, J_R$, the policy gradient is

$$\Delta_\theta J(\theta) = \mathbb{E}_{\pi_\theta, \tau^\theta}[\Delta_\theta \ log \ \pi_\theta(s,a) Q^{\pi_\theta}(s,a)]$$
$$[model \ based] = \int \tau^\theta(s) \int \ \pi_\theta(a|s) * \Delta_\theta \ log \ \pi_\theta(a|s) * Q^{\pi_\theta}(s,a) \ da \ ds$$
$$[model \ free] = \int \ \pi_\theta(a|s) * \Delta_\theta \ log \ \pi_\theta(a|s) * Q^{\pi_\theta}(s,a) \ da$$

For better understanding, we have the corresponding gradient w.r.t Υ :

$$\Delta_\theta J(\theta) = \Delta_\theta \int p_\theta(\Upsilon) r(\Upsilon) d\Upsilon = \int p_\theta(\Upsilon) * \Delta_\theta log \ p_\theta(\Upsilon) * r(\Upsilon) d\Upsilon$$
$$= \mathbb{E}_{\Upsilon \sim p_\theta(\Upsilon)}[\Delta_\theta log \ p_\theta(\Upsilon) * r(\Upsilon)]$$

Take the log:

$$log \ p_\theta(\Upsilon) = \log \ p(s_1) \prod_{t=1}^{T-1} \pi_\theta(a_t|s_t) p(s_T|s_{T-1}, a_{T-1})$$
$$= \log \ p(s_1) + \sum_{t=1}^{T-1} log \ \pi_\theta(a_t|s_t) + log \ p(s_T|s_{T-1}, a_{T-1})$$

Take the derivative over θ :

$$\Delta_\theta log \ p_\theta(\Upsilon) = \Delta_\theta \sum_{t=1}^{T-1} log \ \pi_\theta(a_t|s_t)$$

Then we have:

$$\Delta_\theta J(\theta) = \mathbb{E}_{\Upsilon \sim p_\theta(\Upsilon)}[\Delta_\theta log \ p_\theta(\Upsilon) * r(\Upsilon)]$$
$$= \mathbb{E}_{\Upsilon \sim p_\theta(\Upsilon)}[(\Delta_\theta \sum_{t=1}^{T-1} log \ \pi_\theta(a_t|s_t)) * r(\Upsilon)$$

This is good, since the gradient now is an expectation so that we can use sampling (of trajectories) to approximate it.

Monte-Carlo Policy Gradient (REINFORCE)

Monte-Carlo approach believes that if we draw samples from a distribution, we can estimate the expectation of that distribution well by everaging the samples.

I don't know why it is the following way, but all the materials I see use just one sample (trajectory) as the expectation of the distribution (maybe that's why the variance of the estimation is large) in a stochastic way (means a trajectory has a stochastic of training instances).

$$\Delta_\theta J(\theta) = \mathbb{E}_{\pi_\theta, \tau^\theta}[\Delta_\theta \ log \ \pi_\theta(a|s) Q^{\pi_\theta}(s,a)]$$
$$\approx \Delta_\theta \ log \ \pi_\theta(a_t|s_t) * V(s_t)$$

function REINFORCE
Initialise θ arbitrarily
for each episode $\{s_1, a_1, r_2, ..., s_{T-1}, a_{T-1}, r_T\} \sim \pi_\theta$ **do**
for $t = 1$ to $T - 1$ **do**
$\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(s_t, a_t) v_t$
end for
end for
return θ
end function

In a batch way (means a trajectory has a batch of training instances):

$$\begin{aligned}\Delta J(\theta) &= \mathbb{E}_{\pi_\theta, \tau^\theta}[\Delta_\theta \ log \ \pi_\theta(a|s) Q^{\pi_\theta}(s,a)] \\ &\approx \sum_{t \in T} \Delta_\theta \ log \ \pi_\theta(a|s) * V(s_t)\end{aligned}$$

Better understanding with trajectory objective, if we define $r(\Upsilon) = \sum_{t=1}^{T-1} r(s_t, a_t)$:

$$\Delta_\theta J(\theta) = \mathbb{E}_{\Upsilon \sim p_\theta(\Upsilon)}[(\sum_{t=1}^{T-1} \Delta_\theta log \ \pi_\theta(a_t|s_t)) * \sum_{t=1}^{T-1} r(s_t, a_t)]$$

$$[roll \ out \ once] \approx (\sum_{t=1}^{T-1} \Delta_\theta log \ \pi_\theta(a_t|s_t)) * \sum_{t=1}^{T-1} r(s_t, a_t)$$

REINFORCE with baselines

Take back the policy gradient theorem:

> $\Delta_\theta J(\theta) = \mathbb{E}_{\pi_\theta, \tau^\theta}[\Delta_\theta \ log \ \pi_\theta(a|s) Q^{\pi_\theta}(s,a)] \approx \sum_{t=1}^{T-1} \nabla_\theta \ log \ \pi_\theta(s,a) Q^{\pi_\theta}(s,a)$

Policy gradient methods maximize the expected total reward by repeatedly estimating the gradient $g := \nabla_\theta \mathbb{E}\left[\sum_{t=0}^{\infty} r_t\right]$. There are several different related expressions for the policy gradient, which have the form

$$g = \mathbb{E}\left[\sum_{t=0}^{\infty} \Psi_t \nabla_\theta \log \pi_\theta(a_t \mid s_t)\right], \tag{1}$$

where Ψ_t may be one of the following:

1. $\sum_{t=0}^{\infty} r_t$: total reward of the trajectory.
2. $\sum_{t'=t}^{\infty} r_{t'}$: reward following action a_t.
3. $\sum_{t'=t}^{\infty} r_{t'} - b(s_t)$: baselined version of previous formula.
4. $Q^\pi(s_t, a_t)$: state-action value function.
5. $A^\pi(s_t, a_t)$: advantage function.
6. $r_t + V^\pi(s_{t+1}) - V^\pi(s_t)$: TD residual.

The latter formulas use the definitions

$$V^\pi(s_t) := \mathbb{E}_{\substack{s_{t+1:\infty},\\ a_{t:\infty}}}\left[\sum_{l=0}^{\infty} r_{t+l}\right] \qquad Q^\pi(s_t, a_t) := \mathbb{E}_{\substack{s_{t+1:\infty},\\ a_{t+1:\infty}}}\left[\sum_{l=0}^{\infty} r_{t+l}\right] \tag{2}$$

$$A^\pi(s_t, a_t) := Q^\pi(s_t, a_t) - V^\pi(s_t), \quad \text{(Advantage function)}. \tag{3}$$

Now we have a **critic** (state-action value function Q^w) to tell us the goodness of its possible trajectory and approximate the $r(\Upsilon)$, then we finally introduce our baseline $b(s)$, which is independent of π_θ:

$$\Delta_\theta J(\theta) \approx \sum_{t=1}^{T-1} \nabla_\theta \; log \; \pi_\theta(s, a)(Q^w(s, a) - b(s))$$

so that,

- introducing b(s) will not introducing bias
- introducing b(s) will decrease the variance of the gradients

b(s) will not introducing bias

$$\mathbb{E}_{\pi_\theta,\tau^\theta}\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta,\tau^\theta}[\sum_{t=1}^{T-1}\nabla_\theta\ log\ \pi_\theta(s,a)(Q^w(s,a)-b(s))]$$

$$= \sum_{t=1}^{T-1}\mathbb{E}_{\pi_\theta,\tau^\theta}[\nabla_\theta\ log\ \pi_\theta(s,a)(Q^w(s,a)-b(s))]$$

$$= \sum_{t=1}^{T-1}\mathbb{E}_{\pi_\theta,\tau^\theta}[\nabla_\theta\ log\ \pi_\theta(s,a)Q^w(s,a)] - \sum_{t=1}^{T-1}\mathbb{E}_{\pi_\theta,\tau^\theta}[\nabla_\theta\ log\ \pi_\theta(s,a)b(s)]$$

$$= \sum_{t=1}^{T-1}\mathbb{E}_{\pi_\theta,\tau^\theta}[\nabla_\theta\ log\ \pi_\theta(s,a)Q^w(s,a)] - \sum_{t=1}^{T-1}\mathbb{E}_{\pi_\theta,\tau^\theta}[\nabla_\theta\ log\ \pi_\theta(s,a)] * b(s)$$

$$= \sum_{t=1}^{T-1}\mathbb{E}_{\pi_\theta,\tau^\theta}[\nabla_\theta\ log\ \pi_\theta(s,a)Q^w(s,a)] - \sum_{t=1}^{T-1}0 * b(s)$$

$$= \sum_{t=1}^{T-1}\mathbb{E}_{\pi_\theta,\tau^\theta}[\nabla_\theta\ log\ \pi_\theta(s,a)Q^w(s,a)]$$

b(s) will decrease the variance of the gradients (Actor-critic)

$$Var_{\pi_\theta,\tau^\theta}(\nabla_\theta J(\theta)) = \mathbb{E}_{\pi_\theta,\tau^\theta}[\nabla_\theta J(\theta)]^2 - \mathbb{E}^2_{\pi_\theta,\tau^\theta}[\nabla_\theta J(\theta)]$$

- For an actor-critic with a baseline:

$$Var[\Delta_\theta J(\theta)] \approx Var[\sum_{t=1}^{T-1}\nabla_\theta\ log\ \pi_\theta(s,a)(Q^w(s,a)-b(s))]$$

$$= \sum_{t=1}^{T-1}Var[\nabla_\theta\ log\ \pi_\theta(s,a)(Q^w(s,a)-b(s))]$$

$$= \sum_{t=1}^{T-1}\{E[\nabla_\theta\ log\ \pi_\theta(s,a)(Q^w(s,a)-b(s))]^2 - E^2[\nabla_\theta\ log\ \pi_\theta(s,a)(Q^w(s,a)-b(s))]\}$$

$$= \sum_{t=1}^{T-1}\{E[\nabla_\theta\ log\ \pi_\theta(s,a)(Q^w(s,a)-b(s))]^2 - E^2[\nabla_\theta\ log\ \pi_\theta(s,a)Q^w(s,a)]\}$$

Above we have several assumptions:

1. Several T are indipendent, which is pretty strong and sometimes wrong.
2. We assume independence among the values (since we are using anther Q-network Q^w that is independent on θ) involved in the expectation

$$E[\nabla_\theta\ log\ \pi_\theta(s,a)(Q^w(s,a)-b(s))]^2 \approx E[\nabla_\theta\ log\ \pi_\theta(s,a)]^2 * E[Q^w(s,a)-b(s)]^2$$

then (we negect t in the following):

$$Var[\Delta_\theta J(\theta)] \approx E[\nabla_\theta \, log \, \pi_\theta(s,a)]^2 * E[Q^w(s,a) - b(s)]^2 - E^2[\nabla_\theta \, log \, \pi_\theta(s,a) Q^w(s,a)]\}$$
$$= E[\nabla_\theta \, log \, \pi_\theta(s,a)]^2 * E[Q^w(s,a)^2 - 2 * Q^w(s,a) * b(s) + b(s)^2] - E^2[\nabla_\theta \, log \, \pi_\theta(s,a) Q^w(s,a)]\}$$

- For a PG without a baseline:

$$\hat{Var}[\Delta_\theta J(\theta)] \approx E[\nabla_\theta \, log \, \pi_\theta(s,a)]^2 * E[Q^{\pi_\theta}(s,a)^2] - E^2[\nabla_\theta \, log \, \pi_\theta(s,a) Q^{\pi_\theta}(s,a)]\}$$

Now the question becomes:

$$\sum_{t=1}^{T-1} E[-2 * Q^{\pi_\theta}(s,a) * b(s) + b(s)^2] < 0?$$

This value has the minimum value when $b(s) = Q^{\pi_\theta}(s,a)$