Chapter 3 Policy-based Methods

3.1 Policy Gradient

3.1.1 Direct Policy Differentiation

During previous discussions, we learned that trying to find $\theta^\star$ would help us get more reward.

$$\theta^\star = \arg\max_\theta E_{\tau\sim p_\theta(\tau)}\left[\sum_t r\left(\mathbf{s}_t, \mathbf{a}_t\right)\right] \tag{3.1.1}$$

$$p_\theta\left(\mathbf{s}_1, \mathbf{a}_1, \ldots, \mathbf{s}_T, \mathbf{a}_T\right) = p\left(\mathbf{s}_1\right)\prod_{t=1}\pi_\theta\left(\mathbf{a}_t \mid \mathbf{s}_t\right) p\left(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t\right) \tag{3.1.2}$$

We can rewrite the objective function as:

$$J(\theta) = E_{\tau\sim p_\theta(\tau)}\left[\sum_t r\left(\mathbf{s}_t, \mathbf{a}_t\right)\right] \approx \frac{1}{N}\sum_i\sum_t r\left(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}\right) \tag{3.1.3}$$

where $\sum_i$ means the sum over samples from π_θ. If we abbreviate $\sum_t r\left(\mathbf{s}_t, \mathbf{a}_t\right)$ as $r\left(\tau\right)$, then we have:

$$J(\theta) = E_{\tau\sim p_\theta(\tau)}[r(\tau)] = \int p_\theta(\tau) r(\tau) d\tau \tag{3.1.4}$$

To find the optimal θ, we calculate the policy differentiation:

$$\nabla_\theta J(\theta) = \int \nabla_\theta p_\theta(\tau) r(\tau) d\tau = \int p_\theta(\tau)\nabla_\theta \log p_\theta(\tau) r(\tau) d\tau = E_{\tau\sim p_\theta(\tau)}\left[\nabla_\theta \log p_\theta(\tau) r(\tau)\right] \tag{3.1.5}$$

The derivation from the second term to the third term of the equation might be a bit confusing. Here we use a convenient identity:

$$\underline{p_\theta(\tau)\nabla_\theta \log p_\theta(\tau)} = p_\theta(\tau)\frac{\nabla_\theta p_\theta(\tau)}{p_\theta(\tau)} = \underline{\nabla_\theta p_\theta(\tau)} \tag{3.1.6}$$

Also, note that $p_\theta(\tau)$ is given in (4.2) , and take *log* on both sides, then we have:

$$\log p_\theta(\tau) = \log p\left(\mathbf{s}_1\right) + \sum_{t=1}^{T}\log \pi_\theta\left(\mathbf{a}_t \mid \mathbf{s}_t\right) + \log p\left(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t\right) \tag{3.1.7}$$

Substituting into equation (4.5), we obtain the *direct policy differentiation.*

Theorem 3.1 (Direct policy differentiation).

$$\begin{aligned}\nabla_\theta J(\theta) &= E_{\tau\sim p_\theta(\tau)}\left[\left(\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta\left(\mathbf{a}_t \mid \mathbf{s}_t\right)\right)\left(\sum_{t=1}^{T} r\left(\mathbf{s}_t, \mathbf{a}_t\right)\right)\right] \\ &\approx \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta\left(\mathbf{a}_{i,t} \mid \mathbf{s}_{i,t}\right)\right)\left(\sum_{t=1}^{T} r\left(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}\right)\right)\end{aligned}$$

Algorithm 15 REINFORCE algorithm

Require: arbitrarily initialized π_θ
1: **repeat**
2: sample $\left\{\tau^i\right\}$ from $\pi_\theta\left(\mathbf{a}_t \mid \mathbf{s}_t\right)$ (run the policy)
3: $\nabla_\theta J(\theta) \approx \sum_i\left(\sum_t \nabla_\theta \log \pi_\theta\left(\mathbf{a}_t^i \mid \mathbf{s}_t^i\right)\right)\left(\sum_t r\left(\mathbf{s}_t^i, \mathbf{a}_t^i\right)\right)$
4: $\theta \leftarrow \theta+\alpha \nabla_\theta J(\theta)$
5: **until** convergence

Note. Markov property is not actually used in policy gradient. Therefore, you can use it in partially observed MDPs without modification.

3.1.2 Comparison to Maximum Likelihood

In policy gradient:

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^N\left(\sum_{t=1}^T \nabla_\theta \log \pi_\theta\left(\mathbf{a}_{i, t} \mid \mathbf{s}_{i, t}\right)\right)\left(\sum_{t=1}^T r\left(\mathbf{s}_{i, t}, \mathbf{a}_{i, t}\right)\right) \tag{3.1.8}$$

In maximum likelihood:

$$\nabla_\theta J_{\mathrm{ML}}(\theta) \approx \frac{1}{N} \sum_{i=1}^N\left(\sum_{t=1}^T \nabla_\theta \log \pi_\theta\left(\mathbf{a}_{i, t} \mid \mathbf{s}_{i, t}\right)\right) \tag{3.1.9}$$

Insight. Good stuff is made more likely; bad stuff is made less likely. What we just did in policy gradient simply formalize the notion of "through trial and error"!

3.1.3 The High Variance of Policy Gradient

One significant drawback of policy gradient is that the gradient is often noisy (i.e. the variance is high). Here are some methods of reducing variance.

Causality

Firstly, causality tells us that policy at time t' should not affect the reward at time t when $t \leq t'$. The modified gradient function is:

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T \nabla_\theta \log \pi_\theta\left(\mathbf{a}_{i, t} \mid \mathbf{s}_{i, t}\right)\left(\sum_{t'=t}^T r\left(\mathbf{s}_{i, t'}, \mathbf{a}_{i, t'}\right)\right) \tag{3.1.10}$$

The expression in brackets means "reward to go". And sometimes we also write it as $\hat{Q}_{i, t}^\pi$: estimate of expected reward if we take action $\mathbf{a}_{i,t}$ in state $\mathbf{s}_{i,t}$ ($\hat{Q}_{i, t}^\pi=\sum_{t'=1}^T r\left(\mathbf{s}_{i, t'}, \mathbf{a}_{i, t'}\right)$)

Baseline

The second approach is to introduce **baseline** to the reward part:

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^N \nabla_\theta \log p_\theta(\tau)[r(\tau)-b], \quad b=\frac{1}{N} \sum_{i=1}^N r(\tau) \tag{3.1.11}$$

We are allowed to do that because what we really care about is *expectation* ($E\left[\nabla_\theta \log p_\theta(\tau) b\right]$), and subtracting a baseline is *unbiased in expectation*:

$$E\left[\nabla_\theta \log p_\theta(\tau) b\right]=\int p_\theta(\tau) \nabla_\theta \log p_\theta(\tau) b d \tau=\int \nabla_\theta p_\theta(\tau) b d \tau=b \nabla_\theta \int p_\theta(\tau) d \tau=b \nabla_\theta 1=0 \tag{3.1.12}$$

Here we use the average reward as the baseline. It is not the best baseline (as to reducing variance), but it works pretty good. If you are not satisfied with it, we can try to find the best baseline:

$$\operatorname{Var}[x]=E\left[x^2\right]-E[x]^2 \tag{3.1.13}$$

$$\nabla_\theta J(\theta)=E_{\tau \sim p_\theta(\tau)}\left[\nabla_\theta \log p_\theta(\tau)(r(\tau)-b)\right] \tag{3.1.14}$$

$$\text{Var} = E_{\tau\sim p_\theta(\tau)}\left[\left(\nabla_\theta \log p_\theta(\tau)(r(\tau)-b)\right)^2\right] - E_{\tau\sim p_\theta(\tau)}\left[\nabla_\theta \log p_\theta(\tau)(r(\tau)-b)\right]^2 \tag{3.1.15}$$

$$\frac{d\,\text{Var}}{db} = \frac{d}{db}E\left[g(\tau)^2(r(\tau)-b)^2\right] = -2E\left[g(\tau)^2 r(\tau)\right] + 2bE\left[g(\tau)^2\right] = 0 \tag{3.1.16}$$

Solve this equation and you'll find the optimal baseline:

$$b = \frac{E\left[g(\tau)^2 r(\tau)\right]}{E\left[g(\tau)^2\right]} \tag{3.1.17}$$

Note that this is just expected reward, but weighted by gradient magnitudes!

3.1.4 More about reducing the variance

Before moving on, let's take a moment to review what it actually means to reduce the variance.

In a nutshell, in REINFORCE algorithm we are trying to optimize some function (i.e. the objective function $J(\theta)$) on a random variable (the return of some stochastic policy). What you need to do is sample a bunch of trajectories from the current policy, estimate the current expected value (remember, you are **sampling** trajectories so you don't know the expected value you can only **estimate** it), and update the parameters of the current policy in the direction that optimizes your estimate.

But you only have a finite amount of time to sample (at some point you have to update the parameters to make progress), so your estimate will be off of the true value and you will making updates based on that estimate. So let's say you are only going to sample 10 times, if the distribution of the values of that random variable has high variance, then sampling only 10 times will be a pretty poor estimate. If it has low variance, then 10 samples might be enough to approximate the value you want. Therefore we introduce the baseline to reduce the variance of the samples.

Another way of understanding *reducing the variance* is *reducing the aggressiveness of the updates*. Here is a concrete example: assuming we have a reward function that looks like

$$f(x) = \begin{cases} -x, & x < 0 \\ 0, & x \geq 0 \end{cases} \tag{3.1.18}$$

This can have a practical interpretation. For example, if you are in a fire, the only way to receive a reward of 0 is by escaping the fire in the positive x direction; otherwise, if you move in the negative x direction, you will get hurt and the reward will be negative.

In this scenario, if the baseline is not introduced, the gradient at 0 will be infinite, so one would not know if this policy would lead to some strange places. However, by introducing the baseline, a reasonable gradient will be obtained. This is why adding a baseline can also be interpreted as reducing the policy's aggressiveness.

3.2 Off-Policy Policy Gradients

If we revisit the derivation of policy differentiation, it's obvious that policy gradient is on-policy. The underlined part would be a trouble, because the neural networks change only a little bit with each gradient step, but you need to do the sampling all over again. Therefore, on-policy learning can be extremely inefficient!

$$\nabla_\theta J(\theta) = \underline{E_{\tau\sim p_\theta(\tau)}}\left[\nabla_\theta \log p_\theta(\tau) r(\tau)\right]$$

The question here is: what if we skip the sampling step? What if we don't have samples from $p_\theta(\tau)$ (we have samples from some $\bar{p}(\tau)$ instead)?

Before delving into this issue, let's talk about some maths first: *importance sampling*

Theorem 3.2 (Importance Sampling).

$$\begin{aligned} E_{x\sim p(x)}[f(x)] &= \int p(x)f(x)dx \\ &= \int \frac{q(x)}{q(x)}p(x)f(x)dx \\ &= \int q(x)\frac{p(x)}{q(x)}f(x)dx \\ &= E_{x\sim q(x)}\left[\frac{p(x)}{q(x)}f(x)\right] \end{aligned}$$

Its mathematical essence lies in using a *known distribution* (typically an easy-to-sample distribution) to estimate the expectation of another distribution (usually a difficult-to-sample distribution). Back to off-policy learning,

here we can turn the original objective function $J(\theta) = E_{\tau\sim p_\theta(\tau)}[r(\tau)]$, into a new objective function using importance sampling:

$$J(\theta) = E_{\tau\sim\bar{p}(\tau)}\left[\frac{p_\theta(\tau)}{\bar{p}(\tau)}r(\tau)\right] \tag{3.2.1}$$

Here we have:

$$p_\theta(\tau) = p(\mathbf{s}_1)\prod_{t=1}^{T}\pi_\theta(\mathbf{a}_t \mid \mathbf{s}_t)\,p(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t) \tag{3.2.2}$$

$$\frac{p_\theta(\tau)}{\bar{p}(\tau)} = \frac{p(\mathbf{s}_1)\prod_{t=1}^{T}\pi_\theta(\mathbf{a}_t \mid \mathbf{s}_t)\,p(\mathbf{s}_t, \mathbf{a}_t)}{p(\mathbf{s}_1)\prod_{t=1}^{T}\bar{\pi}(\mathbf{a}_t \mid \mathbf{s}_t)\,p(\mathbf{s}_t, \mathbf{a}_t)} = \frac{\prod_{t=1}^{T}\pi_\theta(\mathbf{a}_t \mid \mathbf{s}_t)}{\prod_{t=1}^{T}\bar{\pi}(\mathbf{a}_t \mid \mathbf{s}_t)} \tag{3.2.3}$$

for some new parameters θ', similarly we have:

$$\frac{p_{\theta'}(\tau)}{p_\theta(\tau)} = \frac{\prod_{t=1}^{T}\pi_{\theta'}(\mathbf{a}_t \mid \mathbf{s}_t)}{\prod_{t=1}^{T}\pi_\theta(\mathbf{a}_t \mid \mathbf{s}_t)} \tag{3.2.4}$$

therefore,

$$\nabla_{\theta'}J(\theta') = E_{\tau\sim p_\theta(\tau)}\left[\frac{\nabla_{\theta'}p_{\theta'}(\tau)}{p_\theta(\tau)}r(\tau)\right] = E_{\tau\sim p_\theta(\tau)}\left[\frac{p_{\theta'}(\tau)}{p_\theta(\tau)}\nabla_{\theta'}\log p_{\theta'}(\tau)r(\tau)\right] \tag{3.2.5}$$

Now estimate locally, at $\theta = \theta'$, we have:

$$\nabla_\theta J(\theta) = E_{\tau\sim p_\theta(\tau)}\left[\nabla_\theta \log p_\theta(\tau)r(\tau)\right] \tag{3.2.6}$$

When $\theta \neq \theta'$, then:

$$\begin{aligned}\nabla_{\theta'}J(\theta') &= E_{\tau\sim p_\theta(\tau)}\left[\frac{p_{\theta'}(\tau)}{p_\theta(\tau)}\nabla_{\theta'}\log\pi_{\theta'}(\tau)r(\tau)\right] \\ &= E_{\tau\sim p_\theta(\tau)}\left[\left(\prod_{t=1}^{T}\frac{\pi_{\theta'}(\mathbf{a}_t \mid \mathbf{s}_t)}{\pi_\theta(\mathbf{a}_t \mid \mathbf{s}_t)}\right)\left(\sum_{t=1}^{T}\nabla_{\theta'}\log\pi_{\theta'}(\mathbf{a}_t \mid \mathbf{s}_t)\right)\left(\sum_{t=1}^{T}r(\mathbf{s}_t, \mathbf{a}_t)\right)\right]\end{aligned} \tag{3.2.7}$$

Also, take causality into consideration, future actions don't affect current weight:

$$\nabla_{\theta'}J(\theta') = E_{\tau\sim p_\theta(\tau)}\left[\sum_{t=1}^{T}\nabla_{\theta'}\log\pi_{\theta'}(\mathbf{a}_t \mid \mathbf{s}_t)\left(\prod_{t'=1}^{t}\frac{\pi_{\theta'}(\mathbf{a}_{t'} \mid \mathbf{s}_{t'})}{\pi_\theta(\mathbf{a}_{t'} \mid \mathbf{s}_{t'})}\right)\left(\sum_{t'=t}^{T}r(\mathbf{s}_{t'}, \mathbf{a}_{t'})\underline{\left(\prod_{t''\neq t}^{t}\frac{\pi_{\theta'}(\mathbf{a}_{t''} \mid \mathbf{s}_{t''})}{\pi_\theta(\mathbf{a}_{t''} \mid \mathbf{s}_{t''})}\right)}\right)\right] \tag{3.2.8}$$

The underlined part in 3.2.8 is distribution mismatch between different policies. If we ignore this part (later we'll see why this is reasonable), we'll get the following:

$$\nabla_{\theta'}J(\theta') = E_{\tau\sim p_\theta(\tau)}\left[\sum_{t=1}^{T}\nabla_{\theta'}\log\pi_{\theta'}(\mathbf{a}_t \mid \mathbf{s}_t)\left(\prod_{t'=1}^{t}\frac{\pi_{\theta'}(\mathbf{a}_{t'} \mid \mathbf{s}_{t'})}{\pi_\theta(\mathbf{a}_{t'} \mid \mathbf{s}_{t'})}\right)\left(\sum_{t'=t}^{T}r(\mathbf{s}_{t'}, \mathbf{a}_{t'})\right)\right] \tag{3.2.9}$$

In the next section, we'll see that 3.2.9 is equivalent to policy iteration.

3.3 (Optional) Policy Gradient as Policy Iteration

Now we give further analysis of policy gradient as policy iteration. The trick here is to express the expected value under policy π_θ in terms of the expected value with repect to the trajectory τ under policy $\pi_{\theta'}$:

$$
\begin{aligned}
J\left(\theta'\right) - J(\theta) &= J\left(\theta'\right) - E_{\mathbf{s}_0 \sim p(\mathbf{s}_0)}\left[V^{\pi_\theta}\left(\mathbf{s}_0\right)\right] \\
&= J\left(\theta'\right) - E_{\tau \sim p_{\theta'}(\tau)}\left[V^{\pi_\theta}\left(\mathbf{s}_0\right)\right] \\
&= J\left(\theta'\right) - E_{\tau \sim p_{\theta'}(\tau)}\left[\sum_{t=0}^{\infty} \gamma^t V^{\pi_\theta}\left(\mathbf{s}_t\right) - \sum_{t=1}^{\infty} \gamma^t V^{\pi_\theta}\left(\mathbf{s}_t\right)\right] \\
&= J\left(\theta'\right) + E_{\tau \sim p_{\theta'}(\tau)}\left[\sum_{t=0}^{\infty} \gamma^t \left(\gamma V^{\pi_\theta}\left(\mathbf{s}_{t+1}\right) - V^{\pi_\theta}\left(\mathbf{s}_t\right)\right)\right] \\
&= E_{\tau \sim p_{\theta'}(\tau)}\left[\sum_{t=0}^{\infty} \gamma^t r\left(\mathbf{s}_t, \mathbf{a}_t\right)\right] + E_{\tau \sim p_{\theta'}(\tau)}\left[\sum_{t=0}^{\infty} \gamma^t \left(\gamma V^{\pi_\theta}\left(\mathbf{s}_{t+1}\right) - V^{\pi_\theta}\left(\mathbf{s}_t\right)\right)\right] \\
&= E_{\tau \sim p_{\theta'}(\tau)}\left[\sum_{t=0}^{\infty} \gamma^t \left(r\left(\mathbf{s}_t, \mathbf{a}_t\right) + \gamma V^{\pi_\theta}\left(\mathbf{s}_{t+1}\right) - V^{\pi_\theta}\left(\mathbf{s}_t\right)\right)\right] \\
&= E_{\tau \sim p_{\theta'}(\tau)}\left[\sum_{t=0}^{\infty} \gamma^t A^{\pi_\theta}\left(\mathbf{s}_t, \mathbf{a}_t\right)\right]
\end{aligned}
\tag{3.3.1}
$$

Writing out the expectation with respect to the trajectory τ explicitly:

$$
\begin{aligned}
E_{\tau \sim p_{\theta'}(\tau)}\left[\sum_t \gamma^t A^{\pi_\theta}\left(\mathbf{s}_t, \mathbf{a}_t\right)\right] &= \sum_t E_{\mathbf{s}_t \sim p_{\theta'}(\mathbf{s}_t)}\left[E_{\mathbf{a}_t \sim \pi_{\theta'}(\mathbf{a}_t|\mathbf{s}_t)}\left[\gamma^t A^{\pi_\theta}\left(\mathbf{s}_t, \mathbf{a}_t\right)\right]\right] \\
&= \sum_t E_{\mathbf{s}_t \sim p_{\theta'}(\mathbf{s}_t)}\left[E_{\mathbf{a}_t \sim \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)}\left[\frac{\pi_{\theta'}\left(\mathbf{a}_t \mid \mathbf{s}_t\right)}{\pi_\theta\left(\mathbf{a}_t \mid \mathbf{s}_t\right)} \gamma^t A^{\pi_\theta}\left(\mathbf{s}_t, \mathbf{a}_t\right)\right]\right]
\end{aligned}
\tag{3.3.2}
$$

where the second equation is obtained by using the importance sampling formula 3.2. Now the remaining question is: can we use p'_θ instead of p_θ in the expectation? The answer is yes, because we can actually bound the distribution gap between p_θ and p'_θ if the two policies π_θ and $\pi_{\theta'}$ are close enough.

3.4 (Optional) Bounding the Distribution Gap

3.5 (Optional) Advanced Policy Gradient Methods

In this section, we'll introduce the ideas of two advanced policy gradient methods. Feel free to skip this section if you are not interested in mathy details.

3.5.1 Trust Region Policy Optimization (TRPO)

Add TRPO algorithm

3.5.2 Proximal Policy Optimization (PPO)

Add PPO algorithm

3.6 Implementation Tips

There are some tips in implementing policy gradient:

1. Remember that the gradient has high variance. This isn't the same as supervised learning, actually the gradients would be really noisy.
2. Consider using much larger batches.
3. Tweaking learning rates is very hard.(We'll learn about policy gradient-specific learning rate adjustment methods later)

Chapter 4 Actor-Critic Methods

4.1 Introducing the Actor-Critic Methods

4.1.1 Recap

We mention Q function and V function here again as a recap.

$$Q^{\pi}\left(\mathbf{s}_t, \mathbf{a}_t\right) = \sum_{t'=t}^{T} E_{\pi_\theta}\left[r\left(\mathbf{s}_{t'}, \mathbf{a}_{t'}\right) \mid \mathbf{s}_t, \mathbf{a}_t\right] : \text{ total reward from taking } \mathbf{a}_t \text{ in } \mathbf{s}_t$$

$$V^{\pi}\left(\mathbf{s}_t\right) = E_{\mathbf{a}_t \sim \pi_\theta\left(\mathbf{a}_t \mid \mathbf{s}_t\right)}\left[Q^{\pi}\left(\mathbf{s}_t, \mathbf{a}_t\right)\right] : \text{ total reward from } \mathbf{s}_t$$

$$A^{\pi}\left(\mathbf{s}_t, \mathbf{a}_t\right) = Q^{\pi}\left(\mathbf{s}_t, \mathbf{a}_t\right) - V^{\pi}\left(\mathbf{s}_t\right) : \text{ how much better } \mathbf{a}_t \text{ is than average}$$

Policy gradient:

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta\left(\mathbf{a}_{i,t} \mid \mathbf{s}_{i,t}\right) Q\left(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}\right) \tag{4.1.1}$$

Adding a baseline:

$$\begin{aligned} \nabla_\theta J(\theta) &\approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta\left(\mathbf{a}_{i,t} \mid \mathbf{s}_{i,t}\right)\left(Q\left(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}\right) - V\left(\mathbf{s}_{i,t}\right)\right) \\ &\approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta\left(\mathbf{a}_{i,t} \mid \mathbf{s}_{i,t}\right) A^{\pi}\left(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}\right) \end{aligned} \tag{4.1.2}$$

4.1.2 Policy Evaluation

Let's go back to three basics steps of RL: **generate samples** (i.e. run the policy), **fit a model to estimate return**, and **improve the policy**. In policy gradient, we figure out how to calculate $\nabla_\theta J(\theta)$, which tells us how to improve the policy. But we still need to fit the model, which quantatively tell us how good the policy is. We have three possible quantities, Q^{π}, V^{π} or A^{π} , so the question is: *what* should we fit to *what*? Since the current reward of taking $\mathbf{a}_t$ at $\mathbf{s}_t$ is fixed, so we can rewrite $Q^{\pi}\left(\mathbf{s}_t, \mathbf{a}_t\right)$ and $A^{\pi}\left(\mathbf{s}_t, \mathbf{a}_t\right)$:

$$\begin{aligned} Q^{\pi}\left(\mathbf{s}_t, \mathbf{a}_t\right) &= r\left(\mathbf{s}_t, \mathbf{a}_t\right) + \sum_{t'=t+1}^{T} E_{\pi_\theta}\left[r\left(\mathbf{s}_{t'}, \mathbf{a}_{t'}\right) \mid \mathbf{s}_t, \mathbf{a}_t\right] \\ &\approx r\left(\mathbf{s}_t, \mathbf{a}_t\right) + V^{\pi}\left(\mathbf{s}_{t+1}\right) \end{aligned} \tag{4.1.3}$$

$$A^{\pi}\left(\mathbf{s}_t, \mathbf{a}_t\right) \approx r\left(\mathbf{s}_t, \mathbf{a}_t\right) + V^{\pi}\left(\mathbf{s}_{t+1}\right) - V^{\pi}\left(\mathbf{s}_t\right) \tag{4.1.4}$$

Therefore V^{π} would be a nice choice, since Q^{π} and A^{π} depend on both actions and states, while V^{π} depends on solely states. Of course this is not the only option in actor-critic algorithm. We'll talk about that later.

Actually calculating the value function is essentially calculating the expectation of reward under a given policy. That's why the process of fitting value fuction is also called policy evaluation.

Theorem 4.1 (Monte Carlo policy evaluation). Monte Carlo policy evaluation estimates the value of state-action pairs based on random sampling of experiences. It involves averaging the returns observed from multiple episodes to approximate the true value function for a given policy.

$$V^{\pi}\left(\mathbf{s}_t\right) \approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t'=t}^{T} r\left(\mathbf{s}_{t'}, \mathbf{a}_{t'}\right)$$

Note that in traditional Monte Carlo policy evaluation you need to generate samples from the same initial state, which means constantly resetting the simulator. Another way of doing that is training a neural network to estimate value function:

$$\text{training data: } \left\{\left(\mathbf{s}_{i,t}, \sum_{t'=t}^{T} r\left(\mathbf{s}_{i,t'}, \mathbf{a}_{i,t'}\right)\right)\right\}$$

$$\text{supervised regression: } \mathcal{L}(\phi) = \frac{1}{2}\sum_{i}\left\|\hat{V}_{\phi}^{\pi}\left(\mathbf{s}_i\right) - y_i\right\|^2$$

In fact, the value function is very intuitive. For example, when training an agent to play a chess-like game, if we set the probability of winning the game to 1 and the probability of losing to 0, the value function typically indicates the probability of eventually winning the game in the current state.

4.1.3 From Evaluation to Actor Critic

Now that we have learned policy evaluation and policy gradients, now we are able to put the two pieces together. And that makes an actor-critic algorithm:

Algorithm 16 Batch actor-critic algorithm

1: **repeat**
2: sample $\left\{\tau^i\right\}$ from $\pi_\theta\left(\mathbf{a}_t \mid \mathbf{s}_t\right)$ (run the policy)
3: fit $\hat{V}_{\phi}^{\pi}(\mathbf{s})$ to sampled reward sums
4: evaluate $\hat{A}^{\pi}\left(\mathbf{s}_i, \mathbf{a}_i\right) = r\left(\mathbf{s}_i, \mathbf{a}_i\right) + \hat{V}_{\phi}^{\pi}\left(\mathbf{s}_i'\right) - \hat{V}_{\phi}^{\pi}\left(\mathbf{s}_i\right)$
5: $\nabla_\theta J(\theta) \approx \sum_i \nabla_\theta \log \pi_\theta\left(\mathbf{a}_i \mid \mathbf{s}_i\right) \hat{A}^{\pi}\left(\mathbf{s}_i, \mathbf{a}_i\right)$
6: $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$
7: **until** convergence

Note that while fitting the value function, we are fitting the sum of reward in a given episode length T. What if T is ∞ ? In many cases $\hat{V}_{\phi}^{\pi}$ can get infinitely large. A simple trick here is to all discount factor γ, which would tell the policy that its better to get rewards sooner than later.

Without discount factors, the target function and loss function of the neural network is:

$$\begin{aligned} y_{i,t} &\approx r\left(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}\right) + \hat{V}_{\phi}^{\pi}\left(\mathbf{s}_{i,t+1}\right) \\ \mathcal{L}(\phi) &= \tfrac{1}{2}\textstyle\sum_i \left\|\hat{V}_{\phi}^{\pi}\left(\mathbf{s}_i\right) - y_i\right\|^2 \end{aligned} \tag{4.1.5}$$

Now adding the discount factors, we'll have:

$$y_{i,t} \approx r\left(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}\right) + \gamma \hat{V}_{\phi}^{\pi}\left(\mathbf{s}_{i,t+1}\right), \quad \gamma \in [0,1] (0.99 \text{ works well }) \tag{4.1.6}$$

Insight. One way of understanding discount factors is that we change the MDP by adding an extra state of death. Once we enter the death state we never leave, and the reward is zero. In each state the agent has the probability of $1-\gamma$ of falling into the death state.

4.1.4 Aside: the Discount Factor

Then how do we introduce discount factors to (Monte Carlo) policy gradients? Actually there seems to be two ways of doing this.

$$\begin{aligned} \text{option 1:} \quad & \nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta\left(\mathbf{a}_{i,t} \mid \mathbf{s}_{i,t}\right)\left(\sum_{t'=t}^{T} \gamma^{t'-t} r\left(\mathbf{s}_{i,t'}, \mathbf{a}_{i,t'}\right)\right) \\ \text{option 2:} \quad & \nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta\left(\mathbf{a}_{i,t} \mid \mathbf{s}_{i,t}\right)\right)\left(\sum_{t=1}^{T} \gamma^{t-1} r\left(\mathbf{s}_{i,t'}, \mathbf{a}_{i,t'}\right)\right) \end{aligned} \tag{4.1.7}$$

The only difference is whether we introduce the discount factors before or after we use the causality rules.

To showcase the differences more effectively, we can rewrite the expression of option 2 in the form of option 1.

$$\begin{aligned}\nabla_\theta J(\theta) &\approx \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta\left(\mathbf{a}_{i,t} \mid \mathbf{s}_{i,t}\right)\right)\left(\sum_{t=1}^{T}\gamma^{t-1} r\left(\mathbf{s}_{i,t'}, \mathbf{a}_{i,t'}\right)\right) \\ &= \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta\left(\mathbf{a}_{i,t} \mid \mathbf{s}_{i,t}\right)\left(\sum_{t'=t}^{T}\gamma^{t'}\left(\mathbf{s}_{i,t'}, \mathbf{a}_{i,t'}\right)\right) \\ &= \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\gamma^{t-1}\nabla_\theta \log \pi_\theta\left(\mathbf{a}_{i,t} \mid \mathbf{s}_{i,t}\right)\left(\sum_{t'=t}^{T}\gamma^{t'-t} r\left(\mathbf{s}_{i,t'}, \mathbf{a}_{i,t'}\right)\right)\end{aligned} \tag{4.1.8}$$

Note that γ^{t-1} comes before $\nabla_\theta \log \pi_\theta$. Then it becomes much clearer. Option 2 implies that we not only care less about the rewards in the future, we also care less about the decisions in the future. As a result you are discounting the gradients. In other word, making right decision in the first time step is more important than making right decision in future time steps. If you are solving a real discount problem, this is exactly what we are trying to do. Because remember the modified MDP settings, later steps don't matter when you are dead. **But**, in reality, this is often not quite what we want. The version that we actually use is **option 1**.

Take some time to review why we introduced the discount factor. For tasks with particularly long time episodes, we introduced the discount factor to artificially reduce the rewards for future actions in order to compute the value function. However, in practical tasks, we do not want to do this. For example, if we want a robot to run steadily forward, we actually want it to keep running. We want to learn a policy that can do the right thing for a long time, rather than just at nearby timesteps. Therefore, option 1 becomes more reasonable.

Algorithm 17 Batch actor-critic algorithm (with discount)

1: **repeat**
2: sample $\left\{\tau^i\right\}$ from $\pi_\theta\left(\mathbf{a}_t \mid \mathbf{s}_t\right)$ (run the policy)
3: fit $\hat{V}_\phi^\pi(\mathbf{s})$ to sampled reward sums
4: evaluate $\hat{A}^\pi\left(\mathbf{s}_i, \mathbf{a}_i\right) = r\left(\mathbf{s}_i, \mathbf{a}_i\right) + \gamma \hat{V}_\phi^\pi\left(\mathbf{s}_i'\right) - \hat{V}_\phi^\pi\left(\mathbf{s}_i\right)$
5: $\nabla_\theta J(\theta) \approx \sum_i \nabla_\theta \log \pi_\theta\left(\mathbf{a}_i \mid \mathbf{s}_i\right) \hat{A}^\pi\left(\mathbf{s}_i, \mathbf{a}_i\right)$
6: $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$
7: **until** convergence

The only difference that in step 3 we introduce the discount factor. In previous discussions we have been using policy gradients in episodic batch mode settings. If we modify a little bit, we'll get the online actor-critic algorithm:

Algorithm 18 Online actor-critic algorithm (with discount)

1: **repeat**
2: take action $\mathbf{a} \sim \pi_\theta(\mathbf{a} \mid \mathbf{s})$, get $(\mathbf{s}, \mathbf{a}, \mathbf{s}', r)$
3: update $\hat{V}_\phi^\pi(\mathbf{s})$ using target $r + \gamma \hat{V}_\phi^\pi\left(\mathbf{s}'\right)$
4: evaluate $\hat{A}^\pi\left(\mathbf{s}, \mathbf{a}\right) = r\left(\mathbf{s}, \mathbf{a}\right) + \gamma \hat{V}_\phi^\pi\left(\mathbf{s}'\right) - \hat{V}_\phi^\pi\left(\mathbf{s}\right)$
5: $\nabla_\theta J(\theta) \approx \nabla_\theta \log \pi_\theta(\mathbf{a} \mid \mathbf{s}) \hat{A}^\pi(\mathbf{s}, \mathbf{a})$
6: $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$
7: **until** convergence

Insight. In actor-critic algorithm, **actor** is the policy, and **critic** is the value function. It can be viewed as a improved version of policy gradient, with reduced variance.

4.2 Further Analysis

4.2.1 Critics as Baselines

Let's make a comparison here between the actor-critic algorithm and the policy gradient algorithm:

$$\begin{aligned}&\text{Actor-critic:}\quad \nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta\left(\mathbf{a}_{i,t} \mid \mathbf{s}_{i,t}\right)\left(r\left(\mathbf{s}_{i,t},\mathbf{a}_{i,t}\right)+\gamma\hat{V}_\phi^\pi\left(\mathbf{s}_{i,t+1}\right)-\hat{V}_\phi^\pi\left(\mathbf{s}_{i,t}\right)\right)\\&\text{Policy gradient:}\quad \nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta\left(\mathbf{a}_{i,t} \mid \mathbf{s}_{i,t}\right)\left(\left(\sum_{t'=t}^{T}\gamma^{t'-t}r\left(\mathbf{s}_{i,t'},\mathbf{a}_{i,t'}\right)\right)-b\right)\end{aligned} \tag{4.2.1}$$

In actor-critic, we have lower variance since we have the critic, but it is not unbiased if the critic is not perfect. While in policy gradient, it is unbiased but the variance would be high due to the single-sample estimate.

So intuitively a nature step we can take is to use $\hat{V}_\phi^\pi$ while still keep the estimater unbiased:

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta\left(\mathbf{a}_{i,t} \mid \mathbf{s}_{i,t}\right)\left(\left(\sum_{t'=t}^{T}\gamma^{t'-t}r\left(\mathbf{s}_{i,t'},\mathbf{a}_{i,t'}\right)\right)-\hat{V}_\phi^\pi\left(\mathbf{s}_{i,t}\right)\right) \tag{4.2.2}$$

Up to now we are all using the state-dependent functions as baselines. So can we use functions that involves both action and state as baselines? Here we come to control variates, which means baselines that are action-dependent. Here we are going to talk about that.

Baselines that use both actions and states are also called *control variates*. The true advantage function is:

$$\begin{aligned}A^\pi\left(\mathbf{s}_t,\mathbf{a}_t\right) &= Q^\pi\left(\mathbf{s}_t,\mathbf{a}_t\right)-V^\pi\left(\mathbf{s}_t\right)\\&=\sum_{t'=t}^{T}E_{\pi_\theta}\left[r\left(\mathbf{s}_{t'},\mathbf{a}_{t'}\right) \mid \mathbf{s}_t,\mathbf{a}_t\right]-E_{\mathbf{a}_t\sim\pi_\theta\left(\mathbf{a}_t\mid\mathbf{s}_t\right)}\left[Q^\pi\left(\mathbf{s}_t,\mathbf{a}_t\right)\right]\end{aligned} \tag{4.2.3}$$

While the approximate advantage function we use in policy gradient is:

$$\hat{A}^\pi\left(\mathbf{s}_t,\mathbf{a}_t\right)=\sum_{t'=t}^{\infty}\gamma^{t'-t}r\left(\mathbf{s}_{t'},\mathbf{a}_{t'}\right)-V_\phi^\pi\left(\mathbf{s}_t\right) \tag{4.2.4}$$

This is unbiased and has a lower variance (but still a higer variance than the actor-critic because of the single sample estimate). But we can make the variance even lower if we substract the Q value.

$$\hat{A}^\pi\left(\mathbf{s}_t,\mathbf{a}_t\right)=\sum_{t'=t}^{\infty}\gamma^{t'-t}r\left(\mathbf{s}_{t'},\mathbf{a}_{t'}\right)-Q_\phi^\pi\left(\mathbf{s}_t,\mathbf{a}_t\right) \tag{4.2.5}$$

This version has a nice property that it goes to zero in expectation if critic is correct. Unfortunately, it does not work if you simply plug it into the policy gradient, since there is an error term you have to compensate for. Now taking that error term into accounts, we have:

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta\left(\mathbf{a}_{i,t} \mid \mathbf{s}_{i,t}\right)\left(\hat{Q}_{i,t}-Q_\phi^\pi\left(\mathbf{s}_{i,t},\mathbf{a}_{i,t}\right)\right)+\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\nabla_\theta E_{\mathbf{a}\sim\pi_\theta\left(\mathbf{a}_t\mid\mathbf{s}_{i,t}\right)}\left[Q_\phi^\pi\left(\mathbf{s}_{i,t},\mathbf{a}_t\right)\right] \tag{4.2.6}$$

The second term represents the gradient of expectation value under the policy of the baseline. Note that this equation is valid even when the baseline depends merely on the state functions, but in that case the second term equals to zero. This kind of trick can provide for a very low variance policy gradient.

We can also take a look at the two different versions of advantage function:

$$\begin{aligned}&\hat{A}_{\mathrm{C}}^\pi\left(\mathbf{s}_t,\mathbf{a}_t\right)=r\left(\mathbf{s}_t,\mathbf{a}_t\right)+\gamma\hat{V}_\phi^\pi\left(\mathbf{s}_{t+1}\right)-\hat{V}_\phi^\pi\left(\mathbf{s}_t\right) && \text{+ lower variance}\\&&& \text{- higher bias if value is wrong (it always is)}\\&\hat{A}_{\mathrm{MC}}^\pi\left(\mathbf{s}_t,\mathbf{a}_t\right)=\textstyle\sum_{t'=t}^{\infty}\gamma^{t'-t}r\left(\mathbf{s}_{t'},\mathbf{a}_{t'}\right)-\hat{V}_\phi^\pi\left(\mathbf{s}_t\right) && \text{+ no bias}\\&&& \text{- higher variance (because single-sample estimate)}\end{aligned} \tag{4.2.7}$$

So can we find something in the middle, combine these two, to control bias/variance tradeoff? The trick here is that instead of unsing an infinite time horizon, you cut it off before the variance goes too high, given that the variance is generally small in the near future and goes higher in the far future. Here is what an n-step return estimater does:

$$\hat{A}_n^\pi(\mathbf{s}_t, \mathbf{a}_t) = \sum_{t'=t}^{t+n} \gamma^{t'-t} r(\mathbf{s}_{t'}, \mathbf{a}_{t'}) - \hat{V}_\phi^\pi(\mathbf{s}_t) + \gamma^n \hat{V}_\phi^\pi(\mathbf{s}_{t+n}) \tag{4.2.8}$$

Generally the larger n you choose, the bias goes smaller and the variance goes larger. The first and the third term contributes to variance, and the second term contributes to bias. The n here we use in n-step return estimater is fixed. Actually we do not have to choose a single n. Instead, we can take all possible n at one time, which is the generalized advantage estimation.

$$\hat{A}_{\text{GAE}}^\pi(\mathbf{s}_t, \mathbf{a}_t) = \sum_{n=1}^{\infty} w_n \hat{A}_n^\pi(\mathbf{s}_t, \mathbf{a}_t) \tag{4.2.9}$$

The generalized advantage estimation is actually a weighted combination of all n-step returns. In terms of determining the weights, we mostly prefer cutting earlier because it brings less variance. Therefore, a descent choice is to use exponential falloff ($w_n \propto \lambda^{n-1}$).

$$\hat{A}_{\text{GAE}}^\pi(\mathbf{s}_t, \mathbf{a}_t) = \sum_{t'=t}^{\infty} (\gamma\lambda)^{t'-t} \delta_{t'} \quad \delta_{t'} = r(\mathbf{s}_{t'}, \mathbf{a}_{t'}) + \gamma \hat{V}_\phi^\pi(\mathbf{s}_{t'+1}) - \hat{V}_\phi^\pi(\mathbf{s}_{t'}) \tag{4.2.10}$$

4.3 (Optional) Advanced Actor-Critic Methods

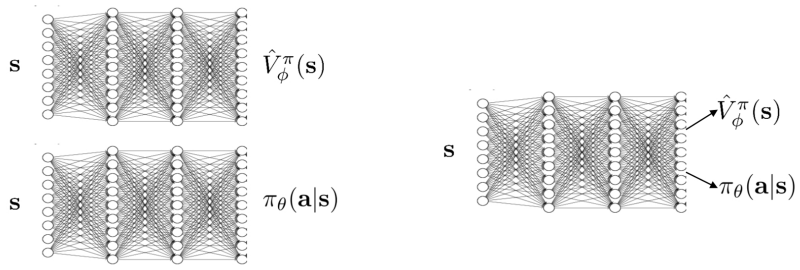
4.3.1 Soft Actor-Critic (SAC)

Add SAC

4.4 Implementations Tips

4.4.1 Architecture Design

There are two options of the neural network architecture design: two network design and shared network design.



(a) two network design (b) shared network design

The two-network design is simple and stable, but lacks efficient feature sharing between the actor and critic. In contrast, the shared network design allows for potential feature sharing efficiency, but it has more hyperparameters and can be more complex and less stable during training.

4.4.2 Parallel Actor-Critic

In a typical online actor-critic algorithm, step2 and step4 work best with a batch (e.g., parallel workers). There are also two types of parallel actor-critic design: **sychronized parallel actor-critic** and **asynchronous parallel actor-critic**.

Sychronized parallel actor-critic: multiple agents learn simultaneously, but they must synchronously update parameters at each step to ensure consistent behavior across all agents. However different agents would use different random seeds so their actions would be a little bit different.

Asynchronous parallel actor-critic: different agents independently update parameters during learning without the need for synchronous updates, leading to improved learning efficiency. However, each agent independently update parameters at their own pace, therefore the parameters update may be inconsistent.

4.4.3 Off-Policy Actor-Critic Algorithm

The problem of sample effeciency gets us thinking about another problem: can we remove the on-policy assumption entirely? A primitive way is to use a replay buffer where we store the transitions we saw in prior time steps. Based on this we have an immature off policy actor-critic algorithm:

Algorithm 19 (Fake) Off-policy actor-critic algorithm

1: **repeat**
2: take action $\mathbf{a} \sim \pi_\theta(\mathbf{a} \mid \mathbf{s})$, get $(\mathbf{s}, \mathbf{a}, \mathbf{s}', r)$, store in $\mathcal{R}$
3: sample a batch $\{\mathbf{s}_i, \mathbf{a}_i, r_i, \mathbf{s}'_i\}$ from buffer $\mathcal{R}$
4: update $\hat{V}_\phi^\pi(\mathbf{s})$ using target $\underline{y_i = r_i + \gamma \hat{V}_\phi^\pi(\mathbf{s}'_i)}$ for each $\mathbf{s_i}$, $\mathcal{L}(\phi) = \frac{1}{N}\sum_i \left\| \hat{V}_\phi^\pi(\mathbf{s}_i) - y_i \right\|^2$
5: evaluate $\hat{A}^\pi(\mathbf{s}_i, \mathbf{a}_i) = r(\mathbf{s}_i, \mathbf{a}_i) + \gamma \hat{V}_\phi^\pi(\mathbf{s}'_i) - \hat{V}_\phi^\pi(\mathbf{s}_i)$
6: $\nabla_\theta J(\theta) \approx \underline{\sum_i \nabla_\theta \log \pi_\theta(\mathbf{a}_i \mid \mathbf{s}_i) \hat{A}^\pi(\mathbf{s}_i, \mathbf{a}_i)}$
7: $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$
8: **until** convergence

There are two fallacies lying in this immature algorithm. The first problem comes from the traget value. When you are loading transitions from the replay buffer, you are actually loading actions from the older version of policy, not the latest one. Therefore $\mathbf{s}'$ comes from the older actions, which is not we actually want.

The second issue comes from the same reason. Because the action $\mathbf{a_i}$ does not come from the latest policy, you cannot calculate policy gradient.

To fix the first problem, the method we take here is to substitute V-function with Q-function. Note the difference between these two functions:

$$\begin{aligned} V^\pi(\mathrm{s}_t) &= \textstyle\sum_{t'=t}^{T} E_{\pi_\theta}\left[r(\mathrm{s}_{t'}, \mathbf{a}_{t'}) \mid \mathbf{s}_t\right] \\ Q^\pi(\mathbf{s}_t, \mathbf{a}_t) &= \textstyle\sum_{t'=t}^{T} E_{\pi_\theta}\left[r(\mathbf{s}_{t'}, \mathbf{a}_{t'}) \mid \mathbf{s}_t, \mathbf{a}_t\right] \end{aligned} \tag{4.4.1}$$

Q-function cares about the total reward from taking $\mathbf{a}_t$ in $\mathbf{s}_t$, and then following the policy π_θ. However we do not restrict that action $\mathbf{a}_t$ must come from π_θ. It is a valid function for any action.

So here in the thrid step in Algorithm 6, instead of using $\hat{V}_\phi^\pi(\mathbf{s})$, we update $\hat{Q}_\phi^\pi(\mathbf{s})$ using target y_i. This time we have:

$$\begin{aligned} y_i &= r_i + \gamma \hat{V}_\phi^\pi(\mathbf{s}'_i) \\ &= r_i + \gamma \hat{Q}_\phi^\pi(\mathbf{s}'_i, \mathbf{a}'_i) \end{aligned} \tag{4.4.2}$$

The trick here is to use:

$$V^\pi(\mathbf{s}_t) = \sum_{t'=t}^{T} E_{\pi_\theta}\left[r(\mathbf{s}_{t'}, \mathbf{a}_{t'}) \mid \mathbf{s}_t\right] = E_{\mathbf{a}\sim\pi(\mathbf{a}_t|\mathbf{s}_t)}\left[Q(\mathbf{s}_t, \mathbf{a}_t)\right] \tag{4.4.3}$$

Note that action $\mathbf{a}'_i$ in (5.10) does not come from the replay buffer $\mathcal{R}$. Instead, it is the action an agent would have taken following policy π_θ under state $\mathbf{s}'_i$. This procedure does not require state $\mathbf{s}'_i$ to actually happens in a simulator, because all you need is to plug state $\mathbf{s}'_i$ into the neural network and get the output action.

We can use a similar approach to resolve the policy gradient issue. Instead of using $\mathbf{a_i}$ which comes from the replay buffer, we use $\mathbf{a}_i^\pi \sim \pi_\theta(\mathbf{a} \mid \mathbf{s}_i)$ that comes from the latest policy.

So now we have step 5 in Algorithm 6 as:

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_i \nabla_\theta \log \pi_\theta(\mathbf{a}_i^\pi \mid \mathbf{s}_i) \hat{A}^\pi(\mathbf{s}_i, \mathbf{a}_i^\pi) \tag{4.4.4}$$

One more thing to say. In practice, we don't actually use the advantage function in policy gradient. We simply use the Q-function here:

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_i \nabla_\theta \log \pi_\theta \left(\mathbf{a}_i^\pi \mid \mathbf{s}_i \right) \hat{Q}^\pi \left(\mathbf{s}_i, \mathbf{a}_i^\pi \right) \tag{4.4.5}$$

This would increase the variance because it does not involve a baseline. However, a high variance here is OK since we don't need to interact with the simulators to sample actions. So it's easy to lower the variance by generating more samples of actions $\mathbf{a}_i^\pi$, without generating states $\mathbf{s}_i$

So we have the fixed version of off-policy actor-critic algorithm here:

Algorithm 20 Off-policy actor-critic algorithm

1: **repeat**
2: take action $\mathbf{a} \sim \pi_\theta(\mathbf{a} \mid \mathbf{s})$, get $(\mathbf{s}, \mathbf{a}, \mathbf{s}', r)$, store in $\mathcal{R}$
3: sample a batch $\{\mathbf{s}_i, \mathbf{a}_i, r_i, \mathbf{s}_i'\}$ from buffer $\mathcal{R}$
4: update $\hat{Q}_\phi^\pi(\mathbf{s})$ using target $y_i = r_i + \gamma \hat{Q}_\phi^\pi \left(\mathbf{s}_i', \mathbf{a}_i' \right)$ for each $\mathbf{s_i}$, $\mathcal{L}(\phi) = \frac{1}{N} \sum_i \left\| \hat{Q}_\phi^\pi \left(\mathbf{s}_i \right) - y_i \right\|^2$
5: evaluate $\hat{A}^\pi \left(\mathbf{s}_i, \mathbf{a}_i \right) = r \left(\mathbf{s}_i, \mathbf{a}_i \right) + \gamma \hat{V}_\phi^\pi \left(\mathbf{s}_i' \right) - \hat{V}_\phi^\pi \left(\mathbf{s}_i \right)$
6: $\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_i \nabla_\theta \log \pi_\theta \left(\mathbf{a}_i^\pi \mid \mathbf{s}_i \right) \hat{Q}^\pi \left(\mathbf{s}_i, \mathbf{a}_i^\pi \right)$
7: $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$
8: **until** convergence

Chapter 5 Model-Based Reinforcement Learning

5.1 Introduction to Model-Based reinforcement learning

What we have covered so far can be categorized as "model-free" reinforcement learning. This is because in previous discussions the transition probabilities are unknown and we did not even attempt to learn the transition probabilities. In reinforcement learning, we have the objective of maximizing the expectation of reward along the trajectory given as follows:

$$\pi_\theta(\tau) = p_\theta\left(\mathbf{s}_1, \mathbf{a}_1, \ldots, \mathbf{s}_T, \mathbf{a}_T\right) = p\left(\mathbf{s}_1\right) \prod_{t=1}^{T} \pi_\theta\left(\mathbf{a}_t \mid \mathbf{s}_t\right) p\left(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t\right) \tag{5.1.1}$$

$$\theta^\star = \arg\max_\theta E_{\tau \sim p_\theta(\tau)} \left[\sum_t r\left(\mathbf{s}_t, \mathbf{a}_t\right) \right] \tag{5.1.2}$$

The transition probabilities $p\left(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t\right)$ is not known in all the model-free RL algorithms that we have learned such as Q-learning and policy gradients. But what if we know the transition dynamics? Recall that at the very beginning of the notes we drew an analogy of RL and control theory. In many cases, we do know the system's internal transition. For example, in games (e.g., Atari games, chess, Go), easily modeled systems (e.g., navigating a car), and simulated environments (e.g., simulated robots, video games), the transitions are given to us.

Moreover, it is not uncommon to learn the transition models: in classic robotics, system identification fits unknown parameters of a known model to learn how the system evolves, and one could also imagine a deep learning approach where we could potentially fit a general-purpose model to observed transition data for later use.

It holds true that knowing the transition dynamics does make things easier, and this is what we are going to deal with. In model-based reinforcement learning, we are going to learn the transition dynamics, and then figure out how to choose actions. In this section we will mainly focus on how can we make decisions if we *know* the dynamics, specifically the following two topics:

1. How can we choose actions under perfect knowledge of the system dynamics?
2. Optimal control, trajectory optimization, planning.

So in this chapter we assume the transition model is **known** or **approximated** to further help planning.

5.2 Optimal Control and Planning

5.2.1 Optimal Control

Optimal control is a task that we come across when we are well aware of the transition probabilities and we try to learn how to control the system optimally. In optimal control, there are two different categories of controller design: the first one is open-loop control, where we do not have any state feedbacks, and we roll out a sequence of actions based on the current state that we observe. The second one is called closed-loop control, where we determine the action at each time step based on the current state, and how we determine the action to apply is based on state feedbacks. In terms of transition model, there are also two categories depending on whether the

model is deterministic or stochastic. While in reinforcement learning, we generally deal with stochastic transition models and close loop control.

In a open loop controller, if we have a deterministic transition in our system such that $\mathbf{s}_{t+1} = f(\mathbf{s}_t, \mathbf{a}_t)$, then our action sequence should be determined by choosing those that can return the maximum rewards:

$$\mathbf{a}_1, \dots, \mathbf{a}_T = \arg\max_{\mathbf{a}_1,\dots,\mathbf{a}_T} \sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t) \quad \text{s.t. } \mathbf{s}_{t+1} = f(\mathbf{s}_t, \mathbf{a}_t) \tag{5.2.1}$$

In stochastic scenarios, the transition function is a probabilistic distribution, where we have $p(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t)$, and the action sequence should be chosen based on expectation of the rewards:

$$\begin{gathered} p_\theta(\mathbf{s}_1, \dots, \mathbf{s}_T \mid \mathbf{a}_1, \dots, \mathbf{a}_T) = p(\mathbf{s}_1) \prod_{t=1}^{T} p(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t) \\ \mathbf{a}_1, \dots, \mathbf{a}_T = \arg\max_{\mathbf{a}_1,\dots,\mathbf{a}_T} E\left[\sum_t r(\mathbf{s}_t, \mathbf{a}_t) \mid \mathbf{a}_1, \dots, \mathbf{a}_T\right] \end{gathered} \tag{5.2.2}$$

Note that this policy might be suboptimal, since future states may reveal more information helpful to decision making. While in closed-loop stochastic cases we roll out all actions to apply only based on the initial state marginal and we do not consider any state-feedback.

In a closed-loop controller, however, we keep interacting with the world, so we need a policy function that can tell us the action to apply if we input the current state: $\pi(\mathbf{a}_t \mid \mathbf{s}_t)$, which we call a state-feedback. We choose our policy function as follows:

$$\begin{gathered} p(\mathbf{s}_1, \mathbf{a}_1, \dots, \mathbf{s}_T, \mathbf{a}_T) = p(\mathbf{s}_1) \prod_{t=1}^{T} \pi(\mathbf{a}_t \mid \mathbf{s}_t)\, p(\mathbf{s}_{t+1} \mid \mathbf{s}_t, \mathbf{a}_t) \\ \pi = \arg\max_\pi E_{\tau \sim p(\tau)}\left[\sum_t r(\mathbf{s}_t, \mathbf{a}_t)\right] \end{gathered} \tag{5.2.3}$$

Note that the form of policy π may vary, including nerual nets, or simply takes a time-varying linear form: $\mathbf{K}_t \mathbf{s}_t + \mathbf{k}_t$.

5.2.2 Open-Loop Planning

We'll start from open-loop planning first. It has the minimal assumption about the dynamics: it does not care about whether the transition model is continuous or discrete, deterministic or stochastic, or whether it is differentiable.

Recall the objective of stochastic open-loop planning, we roll out a sequence of actions by doing arg max on the sum of rewards:

$$\mathbf{a}_1, \dots, \mathbf{a}_T = \arg\max_{\mathbf{a}_1,\dots,\mathbf{a}_T} \sum_{t=1}^{T} r(\mathbf{s}_t, \mathbf{a}_t) \text{ s.t. } \mathbf{a}_{t+1} = f(\mathbf{s}_t, \mathbf{a}_t) \tag{5.2.4}$$

We can abstract away the control and planning part:

$$\mathbf{a}_1, \dots, \mathbf{a}_T = \arg\max_{\mathbf{a}_1,\dots,\mathbf{a}_T} J(\mathbf{a}_1, \dots, \mathbf{a}_T) \tag{5.2.5}$$

Or compactly, since we do not care about whether the action is sequential or discrete, we can say:

$$\mathbf{A} = \arg\max_{\mathbf{A}} J(\mathbf{A}) \tag{5.2.6}$$

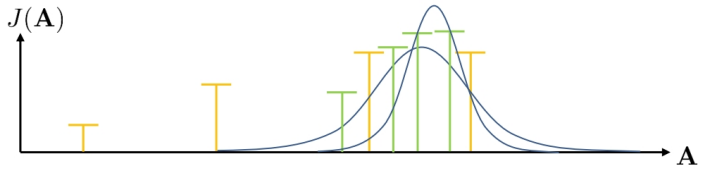
The simplest method is guess and check, or you can call it random shooting method:

1. pick $\mathbf{A}_1, \dots, \mathbf{A}_N$ from some distribution (e.g., uniform)
2. choose $\mathbf{A}_i$ based on $\arg\max_i J(\mathbf{A}_i)$

Random shooting method has the advantage that it is super simple to implement, though it could be highly inefficient in that we are not improving what we sample, so we might get stuck in some mediocre action sequence.

Cross-entropy method

We can dramatically improve this random shooting method by using cross-entropy method (CEM) Instead of randomly selecting samples, in CEM we first randomly initialize a set of samples and then evaluate the objective function for each sample. Next, we use these objective function values to update the sampling distribution, so that the samples are more likely to fall in regions with higher objective function values. This image illustrates the process of iteratively sampling from regions with higher objective function values.



Algorithm 21 Cross-entropy method

1: **repeat**
2: sample $\mathbf{A}_1, \ldots, \mathbf{A}_N$ from $p(\mathbf{A})$
3: evaluate $J(\mathbf{A}_1), \ldots, J(\mathbf{A}_N)$
4: pick the elites $\mathbf{A}_{i_1}, \ldots, \mathbf{A}_{i_M}$ with the highest value, where $M < N$
5: refit $p(\mathbf{A})$ to the elites $\mathbf{A}_{i_1}, \ldots, \mathbf{A}_{i_M}$
6: **until** Satisfactory result

When we "fit" a distribution, what we are actually doing is to iteratively search for the best parameters to find a distribution from which we can sample to give us the best cost-to-go value. Cross-entropy method is also very fast if parallelized, and also extremely simple to implement. However it has a harsh limit on dimensionality. And it only works for open-loop scenarios.

Monte Carlo tree search (MCTS)

In discrete scenarios, we use another method called Monte Carlo tree search, which is very popular in planning in stochastic games.

The gist of this method is that in discrete action space, we are essentially expanding out a tree, but we can not fullly explore the tree due to the exponential computational cost. One way to save the computational cost is to partially expand the tree (for example explore to a fixed depth in each branch) and then use a given policy (e.g., random distribution) to simulate a trajectory from the last expanded node.

Then the question would be: how do we decide on where to search first?

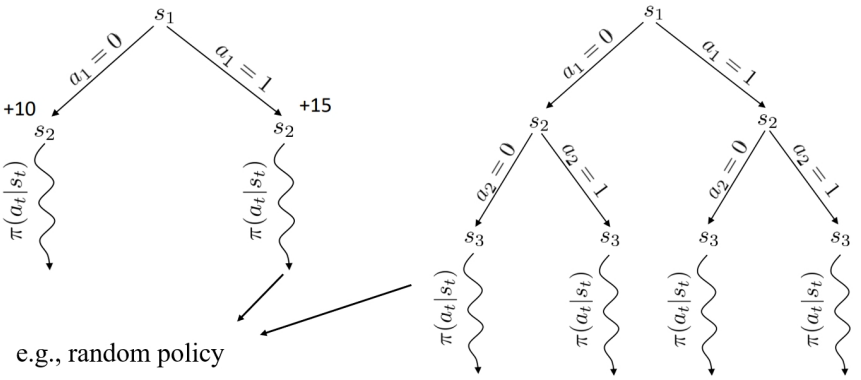


Figure 5.1: Monte Carlo tree search

Suppose we have two actions to take, $a = 0$ and $a = 1$. After taking either action, the agent follows a fixed policy to complete the episode, and the rewards obtained are 10 and 15, respectively.Note that the values of 10 and 15 are not the true values of the rewards, since this is a stochastic process and we have only executed it once.

Our intuition tells us that we should further exploring the action that gave us a higher reward, which is $a = 1$. Besides, if there is a third action that we have not yet explored, then we should also explore this unknown path. That is to say, we should choose nodes with best reward, but also prefer rarely visited nodes. This intuition actually gives the skectch of generic MCTS:

Algorithm 22 Generic MCTS algorithm

1: **repeat**
2: find a leaf s_l using TreePolicy (s_1)
3: evaluate the leaf using DefaultPolicy (s_l)
4: update all values in tree between s_1 and s_l
5: refit $p(\mathbf{A})$ to the elites $\mathbf{A}_{i_1}, \ldots, \mathbf{A}_{i_M}$
6: **until** take best action from s_1

A frequently used TreePolicy is called UCT TreePolicy: if s_t is not fully expanded, choose the remaining new action a_t; else, choose child branch with best score $S(s_{t+1})$, where the score is calculated by:

$$\text{Score}\left(s_t\right) = \frac{Q\left(s_t\right)}{N\left(s_t\right)} + 2C\sqrt{\frac{2\ln N\left(s_{t-1}\right)}{N\left(s_t\right)}} \tag{5.2.7}$$

Here Q for reward, and N for total steps of a chosen trajectory.

5.2.3 Trajectory Optimization with Derivatives

We are going to use the set of notation more commonly used in control therory for the following discussions. (using $\mathbf{u}_t$ for action and $\mathbf{x}_t$ for state, minimizing cost instead of maximizing reward)

$$\min_{\mathbf{u}_1,\ldots,\mathbf{u}_T} \sum_{t=1}^{T} c\left(\mathbf{x}_t, \mathbf{u}_t\right) \text{ s.t. } \mathbf{x}_t = f\left(\mathbf{x}_{t-1}, \mathbf{u}_{t-1}\right) \tag{5.2.8}$$

if we plug in the transition constraint, we have:

$$\min_{\mathbf{u}_1,\ldots,\mathbf{u}_T} c\left(\mathbf{x}_1, \mathbf{u}_1\right) + c\left(f\left(\mathbf{x}_1, \mathbf{u}_1\right), \mathbf{u}_2\right) + \cdots + c\left(f(f(\ldots)\ldots), \mathbf{u}_T\right) \tag{5.2.9}$$

5.3 Model-Based Reinforcement Learning

5.3.1 Basics

In this section, we are going to cover a rather simpler case of model-based RL. Specifically, we are going to talk about a technique to learn a model of the system first, and then use the optimal control technique we covered to improve the model. Furthermore, we will learn to address uncertainty in the model such as model mismatch and imperfection.

Why do we learn the model? We can learn the model so that we know $f(s_t, a_t) = s_{t+1}$ (in deterministic case) or $p(s_{t+1}|s_t, a_t)$ (in stochastic case), we could use the tools from optimal control to maximize our rewards.

Algorithm 23 Model-based Reinforcement Learning Version 0.5

Require: Some base policy for data collection π_0
1: Run base policy $\pi(a_t|s_t)$ (e.g. random policy) to collect $\mathcal{D} = \{(s, a, s')_i\}$ (supervised learning)
2: Learn dynamics model $f(s, a)$ to minimize $\sum_i ||f(s_i, a_i) - s_i'||^2$
3: Plan through $f(s, a)$ to choose actions

Our first attempt is naive: learn $f(s_t, a_t)$ from data, and then plan through it. We call this approach model-based RL version 0.5, or vanilla model-based RL, as shown in Alg. 23. This is essentially what people do in system identification, which is a technique used in classic robotics, and it is effective when we can hand-engineer a dynamics representation using our knowledge of physics, and fit just a few parameters. However, it does not work in general cases because of distribution mismatch, i.e., $p_{\pi_0}(s_t) \neq p_{\pi_f}(s_t)$. Furthermore, the distribution mismatch exacerbates as we use more expressive model classes (e.g., neural networks), since more complicated models would fit more tightly to the demonstration data (which might be inaccurate).

How to make $p_{\pi_0}(s_t) = p_{\pi_f}(s_t)$? We need to collect data from $p_{\pi_f}(s_t)$. We can do this by keep updating the dataset by running the current model, and then update the model accordingly. Take a look at the updated model-based RL algorithm in Alg. 24.

Algorithm 24 Model-based Reinforcement Learning Version 1.0

Require: Some base policy for data collection π_0
1: Run base policy $\pi(a_t|s_t)$ (e.g. random policy) to collect $\mathcal{D} = \{(s, a, s')_i\}$
2: **while** True **do**
3: Learn dynamics model $f(s, a)$ to minimize $\sum_i ||f(s_i, a_i) - s'_i||^2$
4: Plan through $f(s, a)$ to choose actions
5: Execute those actions and add the resulting data $\{(s, a, s')_j\}$ to $\mathcal{D}$

Version 1.0 addresses the model mismatch issue and drives the current model as close as possible to the true dynamics model. However, we are still blindly following a trajectory in step 5 of Alg. 24, and if we made a mistake, we would follow the wrong step which makes the mistake exacerbate. Therefore, we need to somehow adjust our plan as time goes on. One way to do this is to borrow some ideas from modern control theory: Model Predictive Control (MPC).

In MPC, we are given the system's dynamics model, and we are trying to design an adaptive controller by solving a finite time constrained optimal control problem at each time step, and take only the first action in the generated sequence of actions. Then we replan based on the new state. We essentially aims to take one action in the planned sequence and only observe one new state, and then append the observed transition to our dataset $\mathcal{D}$. The improvement is shown in Alg. 25.

Algorithm 25 Model-based Reinforcement Learning Version 1.5

Require: Some base policy for data collection π_0, hyperparameter N
1: Run base policy $\pi(a_t|s_t)$ (e.g. random policy) to collect $\mathcal{D} = \{(s, a, s')_i\}$
2: **while** True **do**
3: Learn dynamics model $f(s, a)$ to minimize $\sum_i ||f(s_i, a_i) - s'_i||^2$
4: **for** N times **do**
5: Plan through $f(s, a)$ to choose actions
6: Execute the first planned action, observe resulting state s' (MPC)
7: Append (s, a, s') to dataset $\mathcal{D}$

The inner loop in Alg. 25 refers to replanning in MPC, which is solving for an optimization problem at each time step after we take the first action planned. The outer loop means that we are periodically retraining the model in order to make it closer to the true underlying dynamics. Intuitively, the more frequently the agent replans, the less perfect each individual plan needs to be, because since we are frequently replanning, we are able to correct our mistakes made in previous plans more easily. Consequently, one is able to correct the plans as one increases the replanning frequency. Therefore, if we are frequently replanning, we could use shorter horizons.

5.3.2 Performance Gaps in Model-based RL

Sometimes model-based RL performs worse than model-free RL. The problem is from step 5 of Alg. 25. In this step, we plan through the model to choose actions, which means we are solving an optimization problem based on the data we collect. One could imagine that if we overfit the data, the agent might have some wrong belief about the model, thus generating wrong actions, as is illustrated in Fig. 5.2.

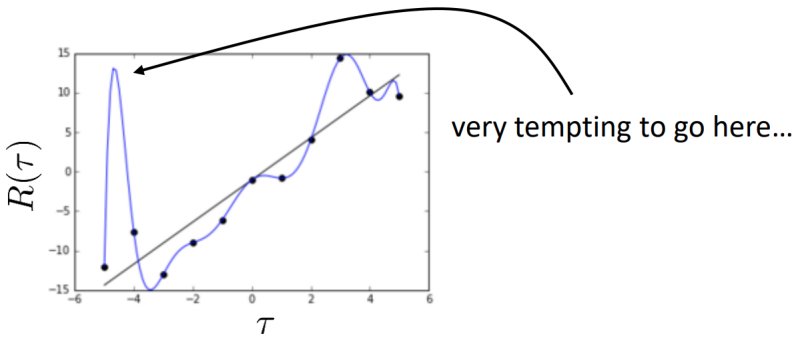


Figure 5.2: False belief about the model from overfitting

Will model-based learning run into the overfitting problem? The answer is yes, because the model does not see the true data distribution. In step 5 of Alg. 25, when we choose actions, we only take actions for which we think we will get high reward **in expectation** (with respect to uncertain dynamics). In the model-based setting, the model will quickly adapts to the distribution of these actions and converge, which avoids the model from exploring enough and results in bad actions planned.

Therefore, we need to explore to get more representative and holistic data of the model, thus preventing overfitting and false belief. Note that the expected value of the reward is not the same as optimistic or pessimistic, but it is often a good start.

5.3.3 Uncertainty-Aware Models

One way to deal with the problem of bad actions planned is to construct an uncertainty-aware model.

The uncertainty-aware model provides a way to quantitatively estimate the uncertainty in the model, so that we can assess the accuracy of the model and the planned actions. When we take uncertainty into account, the agent will be less likely to take actions that are highly uncertain, and will instead prefer actions that are both highly certain and have high reward. In the context of neural networks, this means that the agent will be less likely to take actions that are the result of overfitting, even if those actions have high predicted reward, because they are highly uncertain.

Remark. Why can uncertainty-aware models help alleviate the above problem?

Add more details

The first idea is to use entropy of output distribution, and as we know, higher entropy means higher uncertainty. We can estimate the entropy of $p(s_{t+1}|s_t, a_t)$. However, this is not enough because even when the model is wrong (e.g., an over-fitted model), we might still have low variance, thus low entropy, as long as the data points all satisfy the over-fitted model.

The reason why entropy of the output distribution alone is not expressive enough is that there are two types of uncertainty:

- Aleatoric (statistical) uncertainty, where *the data itself is noisy* .
- Epistemic (model) uncertainty, where *the model is certain about data, but we are not certain about model.*

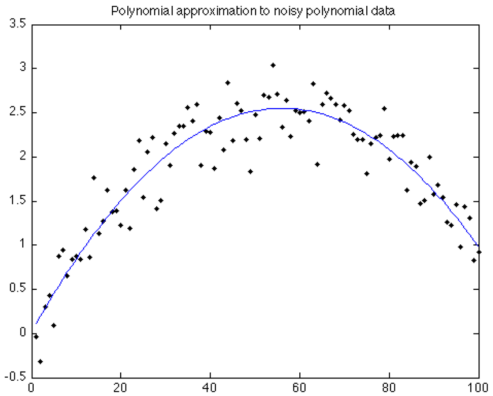


Figure 5.3: Aleatoric uncertainty

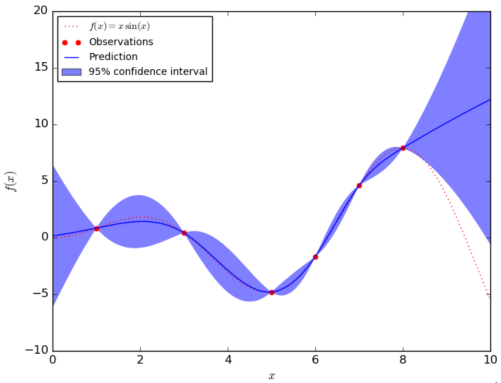


Figure 5.4: Epistemic uncertainty

The second idea is to estimate the epistemic (model) uncertainty, where we essentially estimate how uncertainty we are about the model.

Usually, we use maximum likelihood estimation, where

$$\arg\max_{\theta} \log p(\theta|\mathcal{D}) = \arg\max_{\theta} \log p(\mathcal{D}|\theta) \tag{5.3.1}$$

This is because when we are searching for the model parameters θ that are "most likely" there are actually two ways to interpret "most likely" in this context: (1) Given the model parameters θ, the probability of generating

the dataset $\mathcal{D}$ is maximized. (2) Given the dataset $\mathcal{D}$, the probability that this particular dataset was generated by the model with parameters θ is maximized. And these two interpretations essentially gives you the same parameters when performing arg max.

If we estimate the full distribution of data $p(\theta|\mathcal{D})$ instead of argmax, the entropy of the distribution gives us the model uncertainty from the data. In practice we can predict using

$$\int p(s_{t+1}|s_t, a_t, \theta)p(\theta|\mathcal{D})d\theta. \tag{5.3.2}$$

Bayesian Neural Networks

Bayesian Neural Networks

Bootstrap Ensembles

To learn the posterior distribution, we can also train bootstrap ensembles, where we use multiple networks to learn the same distribution. Formally, say we have N networks, each with a parameter θ_i to learn $p(s_{t+1}|s_t, a_t)$, we can then estimate the posterior by:

$$p(\theta|\mathcal{D}) \sim \frac{1}{N}\sum_i \delta(\theta_i) \tag{5.3.3}$$

where $\delta(\cdot)$ is the Dirac-delta function. To train it, we need to generate independent datasets to get independent models. One way to do this is to chop the original dataset into multiple subsets. Another way is to train θ_i on $\mathcal{D}_i$ sampled with replacement from $\mathcal{D}$. In reality, resampling with replacement is usually not necessary, because SGD and random initialization usually makes the model sufficiently independent.

With this ensemble of networks, we choose actions a little differently. Before, we choose actions by $J(a_1, \ldots, a_H) = \sum_{t=1}^{H} r(s_t, a_t)$, where $s_{t+1} = f(s_t, a_t)$, and now we average over the ensemble by $J(a_1, \ldots, a_H) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{H} r(s_{t,i}, a_{t,i})$, where $s_{t+1,i} = f(s_{t,i}, a_{t,i})$

In general, for candidate action sequence $a_1, \ldots, a_H$, we first sample $\theta \sim p(\theta|\mathcal{D})$, then at each time step t, we sample $s_{t+1} \sim p(s_{t+1}|s_t, a_t, \theta)$, then we calculate the reward $R = \sum_t r(s_t, a_t)$, and we repeat the aforementioned steps and accumulate the average reward.

5.3.4 Latent Space Model

In many cases, we are given very complex observations of states which we do not have full access to, such as pixel-based images. To learn the dynamics using observations, we need to learn from the latent space and infer the states from observations.

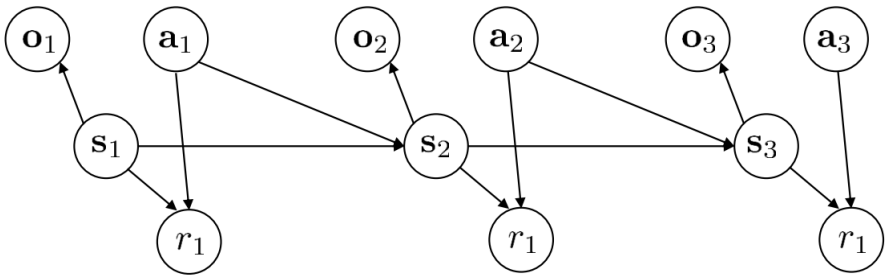


Figure 5.5: Latent space model

From Fig. 5.5, we can see that we need to learn the following models:

- $p(o_t|s_t)$, the observation model
- $p(s_{t+1}|s_t, a_t)$, the dynamics model
- $p(r_t|s_t, a_t)$, the reward model

Recall that in high level, model-based RL algorithms are basically doing a maximum likelihood estimation in training given fully observed states:

$$\max_{\phi} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \log p_{\phi}(s_{t+1,i}|s_{t,i}, a_{t,i}) \tag{5.3.4}$$

With latent models, then, we are not sure about the actual state, so we take the expected value:

$$\max_{\phi} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{E}\left[\log p_{\phi}(s_{t+1,i}|s_{t,i}, a_{t,i}) + \log p_{\phi}(o_{t,i}|s_{t,i})\right] \tag{5.3.5}$$

where the expectation is with respect to the distribution of $(s_t, s_{t+1}) \sim p(s_t, s_{t+1}|o_{1:T}, a_{1:T})$.

However, the posterior distribution $p(s_t, s_{t+1}|o_{1:T}, a_{1:T})$ is usually intractable if we have very complex dynamics. As a result, we could instead try to learn an approximate posterior $q_{\psi}(s_t|o_{1:t}, a_{1;t})$, which is called an **encoder**. We could also learn $q_{\psi}(s_t, s_{t+1}|o_{1:t}, a_{1;t})$ and $q_{\psi}(s_t|o_t)$. Learning the distribution $q_{\psi}(s_t|o_t)$ is crude, but it is simple to implement. Let us focus on $q_{\psi}(s_t|o_t)$ for now, and then the expectation becomes:

$$\max_{\phi} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{E}\left[\log p_{\phi}(s_{t+1,i}|s_{t,i}, a_{t,i}) + \log p_{\phi}(o_{t,i}|s_{t,i})\right]$$

where the expectation is with respect to $s_t \sim q_{\psi}(s_t|o_t)$, $s_{t+1} \sim q_{\psi}(s_{t+1}|o_{t+1})$.

For now, let us further assume that $q(s_t|o_t)$ is deterministic (because the stochastic case requires variational inference, which will be covered in-depth in a later chapter). In deterministic case, we are training a neural net $g_{\psi}(o_t) = s_t$ using a Dirac-delta function such that $q_{\psi}(s_t|o_t) = \delta(s_t = g_{\psi}(o_t))$. Then the expectation can be simplified as

$$\max_{\phi} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \log p_{\phi}(g_{\psi}(o_{t+1,i})|g_{\psi}(o_{t,i}), a_{t,i}) + \log p_{\phi}(o_{t,i}|g_{\psi}(o_{t,i})). \tag{5.3.6}$$

Now everything is differentiable, we can train using backpropagation.

Thus, we can slightly modify Alg. 25 so that we can deal with observations and latent space. We show the sketch of this slightly modified algorithm in Alg. 26. In line 3, we are respectively learning the dynamics, reward model, observation model, and encoder.

Algorithm 26 Model-based Reinforcement Learning with Latent States

Require: Some base policy for data collection π_0, hyperparameter N
1: Run base policy $\pi(a_t|s_t)$ (e.g. random policy) to collect $\mathcal{D} = \{(s, a, s')_i\}$
2: **while** True **do**
3: Learn dynamics model $p_{\phi}(s_{t+1}|s_t, a_t), p_{\phi}(r_t|s_t), p(o_t|s_t), g_{\psi}(o_t)$
4: **for** N times **do**
5: Plan through $f(s, a)$ to choose actions
6: Execute the first planned action, observe resulting state o' (MPC)
7: Append (o, a, o') to dataset $\mathcal{D}$

5.3.5 Further Reading

For more details on model-based RL, refer to Nagabandi et al. (2018), Chua et al. (2018), Feinberg et al. (2018) and Buckman et al. (2018).

For more details for latent space models, refer to Watter et al. (2015) and Zhang et al. (2019).

5.4 Model-Based Policy Learning

So far we have covered the basics of model-based RL that we first learn a model and use a model for control. We have seen that this approach does not work well in general because of the effect of distributional shift in model-based RL. We have also seen the method to quantify uncertainty in our model in order to alleviate this

issue. The methods we covered so far do not involve learning policies. In this chapter, we will cover model-based reinforcement learning of policies. Specifically, we will learn global policies and local policies, and combine local policies into global policies using guided policy search and policy distillation. We shall understand how and why we should use models to learn policies, global and local policy learning, and how local policies can be merged via supervised learning into a global policy.

We have seen the difference between a closed-loop and open-loop controller. We also discussed why an open-loop controller is suboptimal because we are rolling out a whole sequence of actions solely based on one state observation. Therefore, it would be more ideal if we could design a closed-loop controller where state feedbacks can help us correct the mistakes we make. Recall in a stochastic environment, we are optimizing over the policy as follows:

$$p(s_1, a_1, \ldots, s_T, a_T) = p(s_1) \prod_{t=1}^{T} \pi(a_t|s_t) p(s_{t+1}|s_t, a_t)$$

$$\pi = \arg\max_{\pi} \mathbb{E}_{\tau \sim p(\tau)} \left[\sum_t r(s_t, a_t) \right]$$

and π could take several forms: π can be a neural net, which we call a **global** policy, and it can also be a time-varying linear controller $K_t s_t + k_t$ as we saw in LQR, which we call a **local** policy.

5.5 Back-propagate into the Policy

Let us start with a simple solution for model-based policy learning. Ideally, we could build a computational graph in Tensorflow, and calculate the partial derivatives step by step so that we can backpropage into policy and optimize the policy, illustrated in Fig. 5.6.

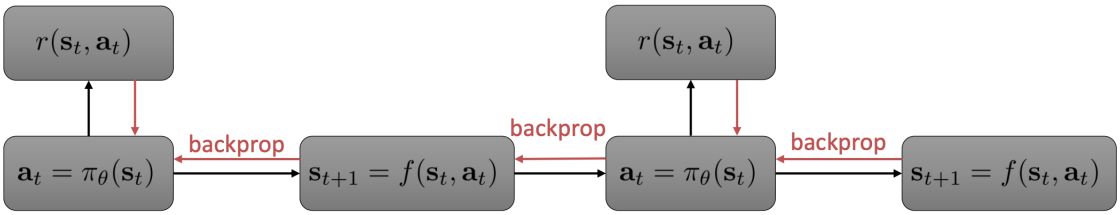


Figure 5.6: Back-propagate into policies

Then we can modify our model-based policy-free RL algorithm to accomodate this new policy learning process in Alg. 27.

Algorithm 27 Model-based Reinforcement Learning Version 1.5

Require: Some base policy for data collection π_0

1: Run base policy $\pi_0(a_t|s_t)$ (e.g. random policy) to collect $\mathcal{D} = \{(s, a, s')_i\}$
2: **while** True **do**
3: Learn dynamics model $f(s, a)$ to minimize $\sum_i ||f(s_i, a_i) - s'_i||^2$
4: Backpropagate through $f(s, a)$ into the policy to optimize $\pi_\theta(a_t|s_t)$
5: Run $\pi_\theta(a_t|s_t)$, appending the visited tuples (s, a, s') to $\mathcal{D}$.

5.5.1 Vanishing and Exploding Gradients

One problem with Alg. 27, or general gradient-based optimization is that as we progress into the time steps, we might encounter vanishing or exploding gradients. Because as we apply chain rule, the gradients get multiplied by each other, so we the product may get extremely big (exploding) or extremely small (vanishing), making optimization a lot harder. Furthermore, we have similar parameter sensitivity problems as shooting methods, but we no longer have convenient second order LQR-like method, because the policy function is extremely complicated and policy parameters couple all the time steps, so no dynamic programming.

So what can we do about it? First, we can use model-free RL algorithms with synthetic samples generated by the model. Essentially, we are using models to accelerate model-free RL. Second, we can use simpler policies than neural nets such as LQR, and train local policies to solve simple tasks, and then combine them into global policies via supervised learning.

5.6 Model-free Optimization with a Model

Model-free Optimization with a Model is yet to be written.

Recall the equation from policy gradients:

$$\nabla_\theta J(\theta) \simeq \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t})\hat{Q}^\pi_{i,t}$$

Note that we are not doing any backprop through time in policy gradient because we are calculating the gradient with respect to an expectation, so we can just take the derivative of the probability of the samples instead of the actual dynamics function.

Then we look at the regular backprop (pathwise) gradient, we see a more chain rule-like gradient:

$$\nabla_\theta J(\theta) = \sum_{t=1}^{T}\frac{dr_t}{ds_t}\prod_{t'=2}^{t}\frac{ds_{t'}}{da_{t'-1}}\frac{da_{t'-1}}{ds_{t'-1}}$$

The two gradients are different, because the policy gradient is for stochastic systems while the backprop policy is for deterministic systems. But using variational inference, we can prove that they are calculating the same gradient differently, thus having different tradeoffs. We will talk about variational inference more in-depth in the next chapter.

Actually, given more samples to reduce variance, policy gradient is more stable because it does not require multiplying many Jacobians. However, if our models are inaccurate, the samples we use from the wrong model will be incorrect, and the mistakes are likely to exacerbate as time goes on. So it would be nice to use such model-free optimizer and keep the rolled out samples' trajectory short. This is essentially what Dyna algorithm does.

5.6.1 Dyna

Dyna is an online Q-learning algorithm that performs model-free RL with a model.

Algorithm 28 Dyna

Require: Some exploration policy for data collection π_0

1: Given state s, pick action a using exploration policy
2: Observe s' and r, to get transition (s, a, s', r)
3: Update model $\hat{p}(s'|s, a)$ and $\hat{r}(s, a)$ using (s, a, s')
4: Q-update: $Q(s, a) \leftarrow Q(s, a) + \alpha\mathbb{E}_{s',r}\left[r + \max_{a'} Q(s', a') - Q(s, a)\right]$
5: **for** K times **do**
6: Sample $(s, a) \sim \mathcal{B}$ from buffer of past states and actions
7: Q-update: $Q(s, a) \leftarrow Q(s, a) + \alpha\mathbb{E}_{s',r}\left[r + \max_{a'} Q(s', a') - Q(s, a)\right]$

In step 3 of Alg. 28, we are updating the model and reward function using the observed transition. Then in step 6, we will sample some old state and action pairs and apply the model onto the sampled pair, so the s' in step 7 are synthetic next states. Intuitively, as the models get better, the expectation estimate in step 7 also gets more accurate. This algorithm seems arbitrary in many aspects, but the gist is to keep improving models and use models to improve Q-function estimation by taking expectations.

We can also generalize Dyna to see how this kind of general Dyna-style model-based RL algorithms work. The generalized algorithm is shown in Alg. 29.

As shown in Fig. 5.7, we choose some states (orange dots) from the buffer, simulate the next states using the learned model, and then train model-free RL with synthetic data (s, a, s', r) where s is from the experience

Algorithm 29 General Dyna

Require: Some exploration policy for data collection π_0
1: Collect some data, consisting of transitions (s, a, s', r)
2: Learn model $\hat{p}(s'|s, a)$ (and optionally, $\hat{r}(s, a)$)
3: **for** K times **do**
4: Sample $s \sim \mathcal{B}$ from buffer
5: Choose action a (from $\mathcal{B}$, from π, or random)
6: Simulate $s' \sim \hat{p}(s'|s, a)$ (and $r = \hat{r}(s, a)$)
7: Train on (s, a, s', r) with model-free RL
8: (optional) take N more model-based steps

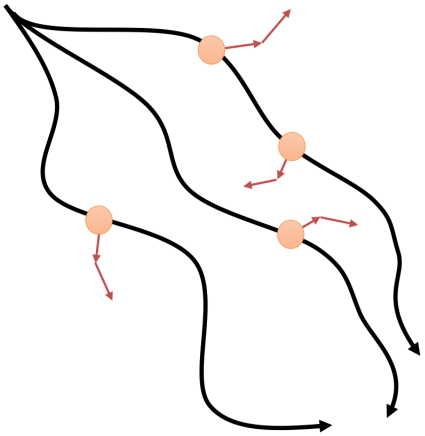


Figure 5.7: General Dyna training

buffer, s' is from the learned model. One could also take more than one step if one believes that the model is good enough for more steps.

This algorithm only requires very short (as few as one step) rollouts from model, so the mistakes will not exacerbate and accumulate much. Moreover, we explore well with a lot of samples because we still see diverse states.

5.7 Local and Global Models

Recall that in LQR, we can turn a constrianed optimization problem into an unconstrained problem:

$$\min_{u_1,\dots,u_T} c(x_1, u_1) + c(f(x_1, u_1), u_2) + \cdots + c(f(f(\dots)\dots), u_T)$$

Backpropagation is indeed a possible solution to solve this optimization problem, and we need $\frac{df}{dx_t}$, $\frac{df}{du_t}$, $\frac{dc}{dx_t}$, $\frac{dc}{du_t}$

5.7.1 Local Models

Since LQR gives us a state-feedback controller for a linear system, we can keep linearizing the system and iteratively apply LQR to generate local models. We fit $\frac{df}{dx_t}$, $\frac{df}{du_t}$ around the current trajectory or policy. Say the model is a Gaussian $p(x_{t+1}|x_t, u_t) = \mathcal{N}(f(x_t, u_t), \Sigma)$, then we can approximate the model as a linear function $f(x_t, u_t) \simeq A_t x_t + B_t u_t$, and we can use $\frac{df}{dx_t}$ as A_t, and $\frac{df}{du_t}$ as B_t.

Iterative LQR produces $\hat{x}_t, \hat{u}_t, K_t, k_t$, where $u_t = K_t(x_t - \hat{x}_t) + k_t + \hat{u}_t$. We can execute the controller using a Gaussian $p(u_t|x_t) = \mathcal{N}(K_t(x_t - \hat{x}_t) + k_t + \hat{u}_t, \Sigma_t)$ because we can add noise to the iLQR controller so that all samples do not look the same. Practically, we can set $\Sigma_t = Q_{u_t,u_t}^{-1}$. We can fit the model $p(s_{t+1}|s_t, a_t)$ using Bayesian linear regression, and use the global model as prior.

We also need to stay close to old controller if we go too far. If trajectory distribution is close, then dynamics will be close too. Close here means the KL-divergence is small $D_{KL}(p(\tau)||p(\bar{\tau})) \leq \epsilon$.

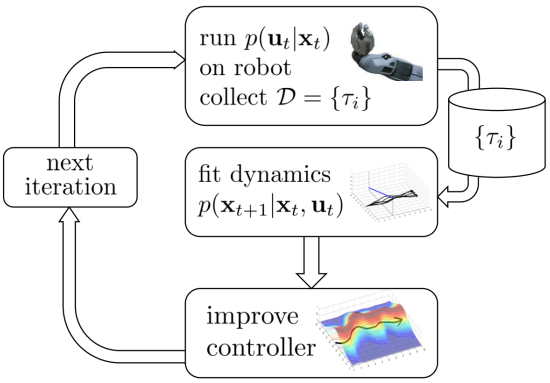


Figure 5.8: Local models fitting

Algorithm 30 Guided Policy Search

1: **while** True **do**
2: Optimize each local policy $\pi_{LQR,i}(u_t|x_t)$ on initial state $x_{0,i}$ with respect to $\tilde{c}_{k,i}(x_t, u_t)$
3: Use samples from the previous step to train $\pi_\theta(u_t|x_t)$ to mimic each $\pi_{LQR,i}(u_t|x_t)$
4: Update cost function $\tilde{c}_{k+1,i}(x_t, u_t) = c(x_t, u_t) + \lambda_{k+1} \log \pi_\theta(u_t|x_t)$

5.7.2 Guided Policy Search

The high level idea of guided policy search is to use some simpler local policy such as local LQR controller to help and guide the learning process of more complex global policy learner. Essentially, we would use the local models trajectories as the training data for a supervised learning neural net that can solve all the tasks.

However, one problem is that the local policies might not be able to be reproduced using a single neural net. Therefore, after training the global policy with supervised learning, we need to reoptimize the local policies using the global policy so that the policies are consistent with each other. The sketch of guided policy search is shown in Alg. 30. Note that the cost function $\tilde{c}_{k,i}$ is the modified cost function to keep π_{LQR} close to π_θ.

In Divide and Conquer RL, the idea is similar, except that we are replacing the local LQR controllers with local neural net.

5.7.3 Distillation

In RL, we borrow some ideas from supervised learning to achieve the task of learning a global policy from a bunch of local policies.

Recall in supervised learning, we use model ensemble to make our predictions more robust and accurate. However, keeping a lot of models is expensive during test time. Is there a way to train just one model that can behave as well as a meta-learner?

The idea, proposed by Hinton in **?**, is to train a model on the ensemble's predictions as "soft" targets using:

$$p_i = \frac{\exp(z_i/T)}{\sum_j (z_j/T)}$$

where T is called temperature. The new labels here can be intuitively explained using the example of MNIST dataset. For example, a handwritten digit "2" looks like a 2 and a backward 6. Therefore, the soft-labels that we use to train the distilled model is going to be "80% chance being 2 and 20% chance being 6".

In RL, to achieve multi-task global policy learning, we can use something similar called policy distillation. The idea is to train a global policy using a bunch of local tasks:

$$\mathcal{L} = \sum_a \pi_{E_i}(a|s) \log \pi_{AMN}(a|s)$$

where the meta-policy π_{AMN} can be trained in a supervised learning fashion.