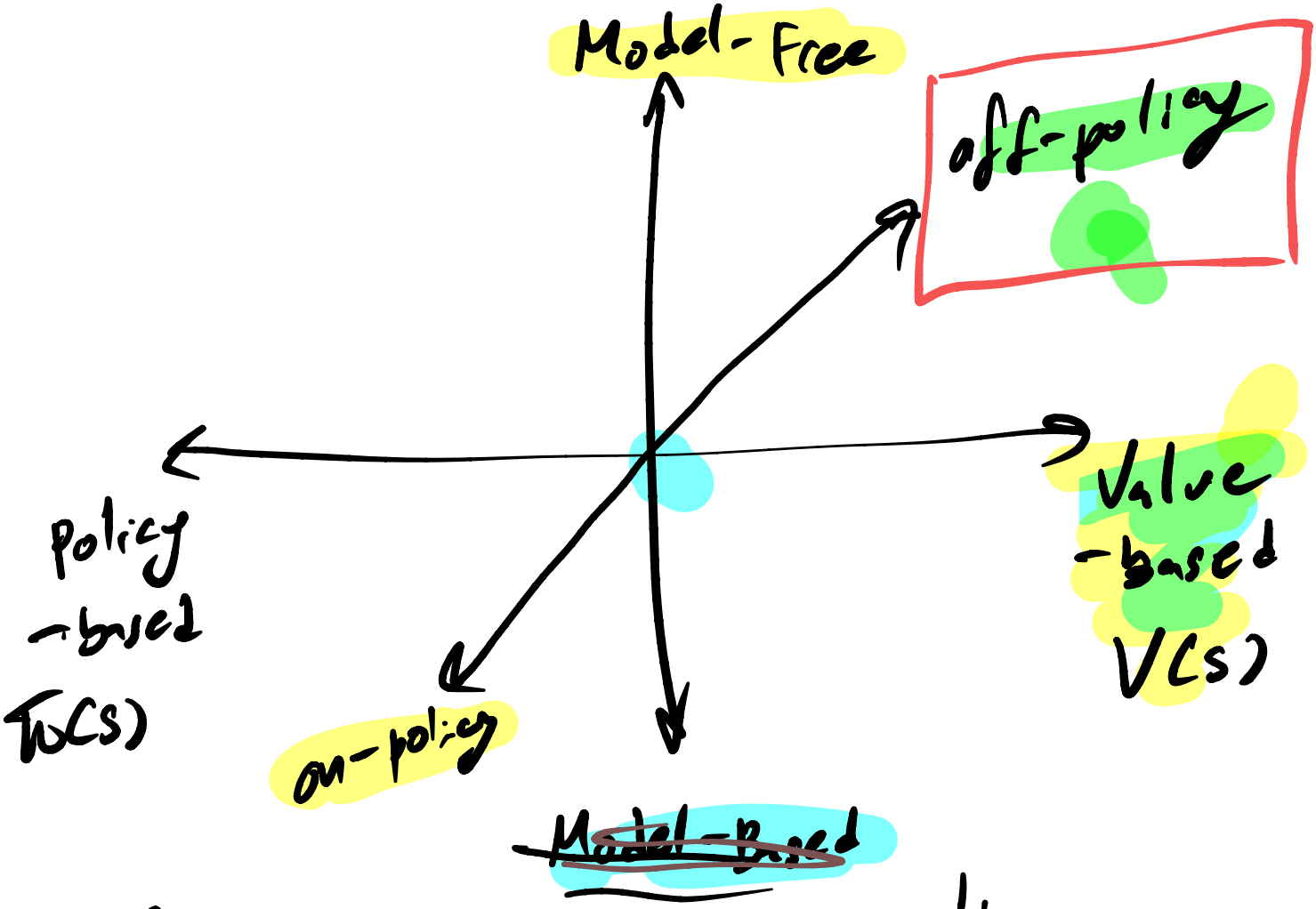


Off-policy vs On-policy

- Let's introduce two types of policy

1.) Behaviour policy

$b(A|S)$

2.) Target policy

$\pi(A|S)$

- We collect data with $b(A|S)$ and we use data to estimate Q^π, V^π where π is the target, (target)

- When $b(A|S)$ is $\pi(A|S)$ we say the algorithm is "on-policy"

	pros	cons
on-policy	- generally simpler - Many on-policy Algo are formulated for continuous action domains.	cannot do
off-policy	→ General + powerful → Better Exploration → Can learn from observation → Reuse Experiences → Decouple data-collection from learning	→ data has greater variance → Harder to be formulated

Q-learning vs SARSA

estimate q^*

Q-learning

- $Q^*(S,A) \leftarrow Q^*(S,A) + \alpha[R + \gamma \max_a Q(S',a) - Q(S,A)]$

→ S, A, R, S' ← $b(A|S)$ → π^* is the target policy

- SARSA

$$Q(S,A) \leftarrow Q(S,A) + \alpha[R + \gamma Q(S',A') - Q(S,A)]$$

A' is chosen from Q on-pol

∴ SARSA = on-pol

Off-policy prediction with Importance Sampling

- MC-control is on-policy
  - $Q^{\pi}(s,a)$
    because we use π to collect data.
- Can we estimate $Q^{\pi}(s,a)$ while collecting data from $b(A|S)$

$$Q^{\pi}(s,a) = \mathbb{E}^{\pi}[G \mid s,a]$$

MC-prediction $\Rightarrow \mathbb{E}^{\pi}[g_i] \approx \frac{1}{n}\sum_{n} g_i$

~~$\mathbb{E}[X] = \sum_{x} p(x) \cdot x$~~

$$Q^{\pi}(s,a) = \sum_{g_i} p^{\pi}(g_i)\, g_i \quad \leftarrow ①$$

$$= \sum_{g_i} p^{\pi}(g_i) \frac{p^{b}(g_i)}{p^{b}(g_i)} g_i \quad \leftarrow ②$$

$$= \sum_{g_i} p^{b}(g_i) \left[\frac{p^{\pi}(g_i)}{p^{b}(g_i)} \right] g_i$$

$$= \sum_{g_i} p^b(g_i) \left[\frac{p^\pi(g_i)}{p^b(g_i)} \cdot g_i \right]$$

$$= E^b \left[\frac{p^\pi(G)}{p^b(G)} G \mid s, a \right]$$

$$Q^\pi(s,a) = E^b [\rho G \mid s,a]$$

$$\rho = \frac{p^\pi}{p^b}$$ Importance weight

$$\rightarrow Q^\pi(s,a) = E^\pi [G \mid s,a]$$

$$\rho = \frac{p^{\pi}(g)}{p^{b}(g)}$$

trajectory

$$= \frac{\cancel{p(g|\tau)}\, p^{\pi}(\tau)}{\cancel{p(g|\tau)}\, p^{b}(\tau)}$$

$$\tau = \{ S_t, A_t, S_{t+1}, A_{t+1} \cdots \}$$

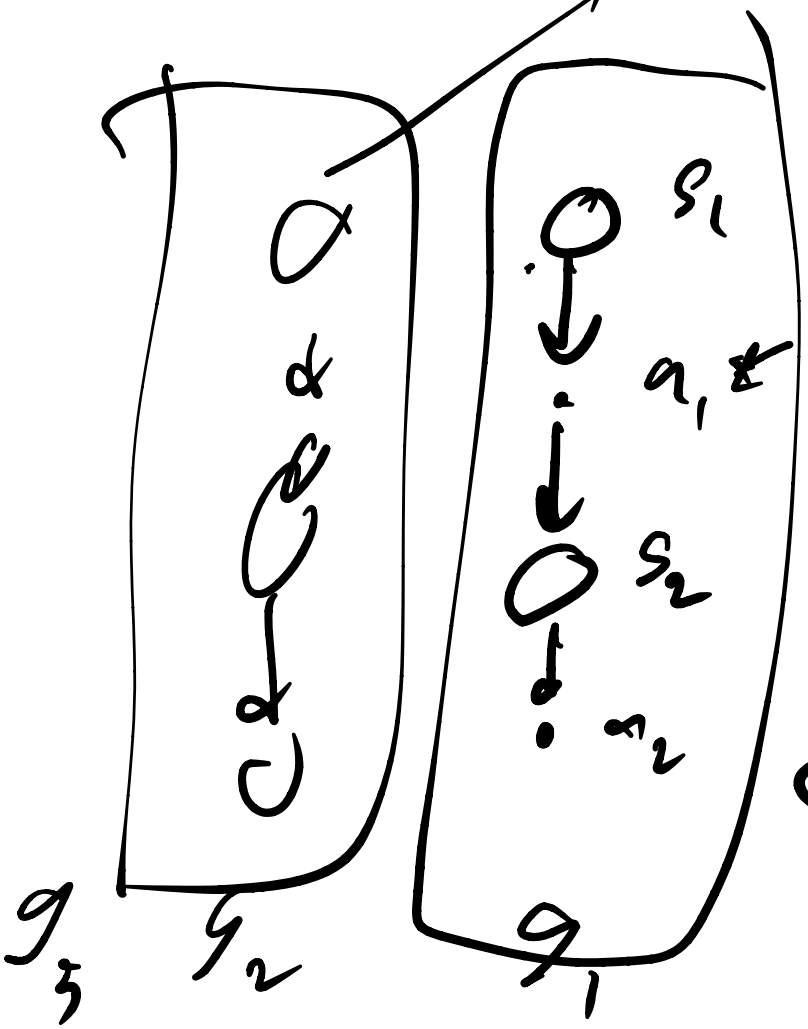
$$p(\tau) = p(S_t) \cdot \pi(A_t|S_t)\, p(S_{t+1}|S_t A_t) \cdot \pi(A_{t+1}|S_{t+1})\, p(S_{t+2}|S_{t+1} \cdots$$

$$p(\tau) = p(s_t) \prod_{k=t}^{T-1} \pi(A_k|S_k) \cdot p(S_{k+1}|S_k, A_k)$$

$$\rho = \frac{p^{\pi}(\tau)}{p^{b}(\tau)} = \frac{\cancel{p(s_t)} \prod_{k=t}^{T-1} \pi(A|S) \cancel{p(S_{k+1}|S_k, A_k)}}{\cancel{p(s_t)} \prod_{k=t}^{T-1} b(A|S) \cancel{p(S_{k+1}|S_k, A_k)}}$$

$$\boxed{\rho = \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{b(A_k|S_k)}}$$

$$Q^{\pi}(s,a) = E_{b}\left[G_t \cdot \prod_{n=t}^{T-1} \frac{\pi(A_n|S_n)}{b(A_n|S_n)} \right]$$



$\tau = \{s_1, a_1, r_2, s_2, a_2, \ldots\}$

$g = \sum \gamma^k r$

- We can use the same idea for n-step SARSA

$$Q^{\pi}(S,A) = E^{\pi}\left[R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + Q^{\pi}(S_t, A_t)\right]$$

$$= E^{b}\left[\rho\left(\quad\right)\right]$$

$$\rho = \prod_{k=t+1}^{t+n} \frac{\pi(A_k|S_k)}{b(A_k|S_k)}$$