策略 $\pi(a|s,\theta)$ 函数，参数 θ 的随机梯度能减小，

- 对于某一个 s，

$$\nabla J(\theta) = \sum_a \nabla_\theta \pi(a|s,\theta)\, q_\pi(s,a)$$
$$= \sum_a \pi(a|s,\theta)\left[\nabla_\theta \log \pi(a|s,\theta)\, q_\pi(s,a)\right]$$
$$= E_a\left[\nabla_\theta \log \pi(a|s,\theta)\, q_\pi(s,a)\right] \quad ①$$

依据证明可知，$q_\pi(s,a) - b(s)$ 后，$\nabla J(\theta)$ 不变，

$$\Longleftrightarrow E_a\left[\nabla_\theta \log \pi(a|s,\theta)\,(q_\pi(s,a) - b(s))\right] \quad ②$$

则：① 和 ② 两个随机梯度，具有相同的期望。

$\nabla_\theta \log \pi(a|s,\theta)$ 随方向增，$q_\pi(s,a) - b(s)$ 将降低方差。

- 减 baseline $b(s)$，使得 θ 的随机变化值更稳定。

$$\nabla J(\theta) = E_a[\ \theta \text{ 随机变化值}\ ]$$

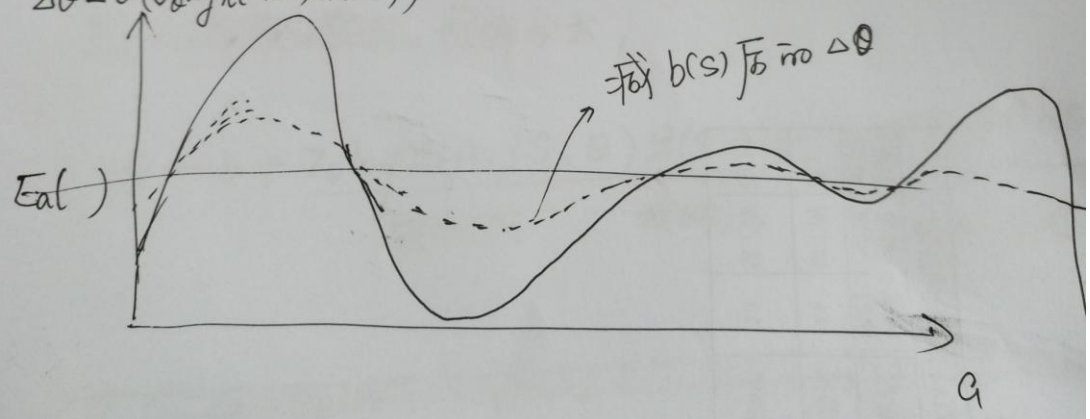
两方面进行讨论

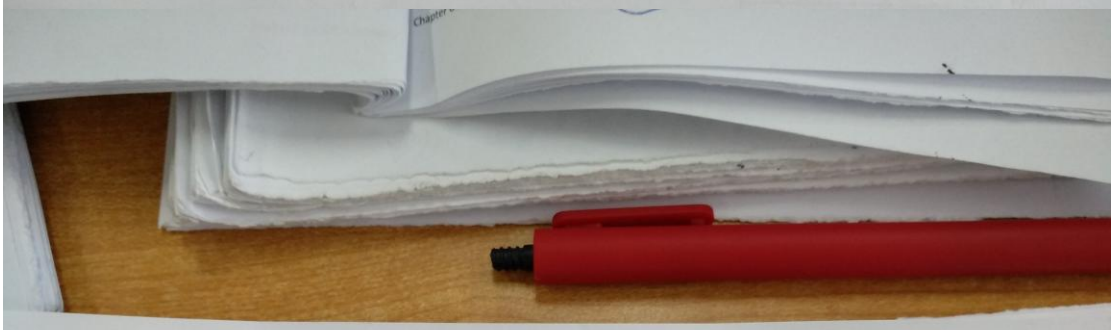
① 同一 s 情况

② 不同 s 情况

(1) 初同S情况

$\Delta\theta = \alpha(\nabla_\theta \log\pi(a|s,\theta) Q(s,a))$





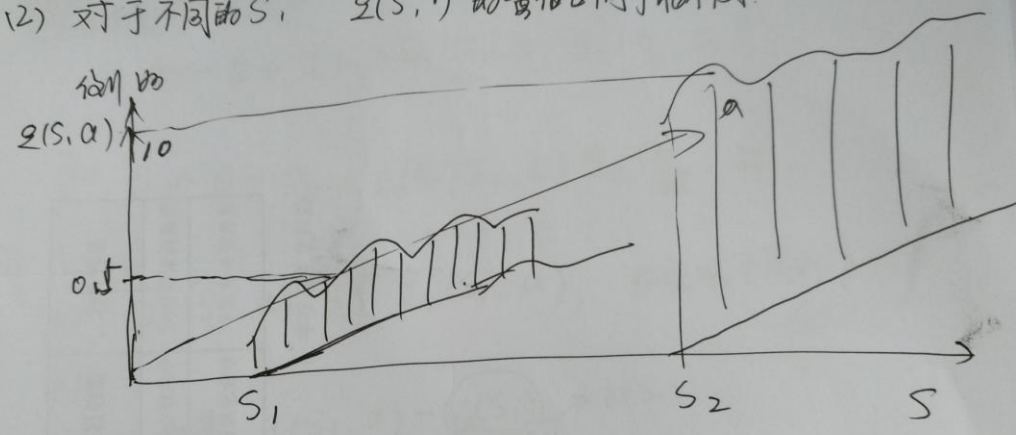
从效果上看，可理解为：如果 $Q(s,a_1) >> Q(s,a_2)$

那么，$\pi(a|s)$ 的分布趋向于 Q 较大，那么 (s,a_1) 被

采集的~~概率会变大~~，频率会大，

$$\theta \leftarrow \theta + \underbrace{\nabla_\theta \log\pi(a_1|s,\theta)}_{\text{朝该方向}} \underbrace{Q(s,a_1)}_{\text{增加很大}} - \underbrace{\nabla_\theta \log\pi(a_1|s,\theta) b(s)}_{\text{抑制作用}}$$

(2) 对于不同的 S，$q(S,\cdot)$ 的最值区间可能不同



所以建议 $b(S_1) \neq b(S_2)$

~~等效博客文本~~：不同的 S，对 θ 随机变化影响大。

$$\theta \leftarrow \theta + \nabla_\theta \log \pi(a_i | \underline{S_1}, \theta)\, q_\pi(\underline{S_1}, a_i)$$

和 $\theta \leftarrow \theta + \nabla_\theta \log \pi(a_i | S_2, \theta)\, q_\pi(\underline{S_2}, a_i)$

由于 $q_\pi(S_2, a) >> q_\pi(S_1, a)$，$\theta$ 会以不同的易级当 $q(S_1)$。

✓ 如果 $q_\pi(S_1, a) - V_\pi(S_1) \rightarrow b(S_1)$

$q_\pi(S_2, a) - V_\pi(S_2) \rightarrow b(S_2)$

则 $q(S_1,\cdot)$ 和 $q(S_2,\cdot)$ 都在 0 值附近。