Maintain $n(s)$. $V(s)$.

in a traj where init state is s & rand return is G.

$n(s) \leftarrow n(s)+1$

$$V(s) \leftarrow \frac{n(s)-1}{n(s)} V(s) + \frac{1}{n(s)} G.$$

$$\downarrow \quad \frac{1}{n(s)-1}$$

$$V(s) \Leftarrow (1-\alpha) V(s) + \alpha G.$$

$V(s)$: new. $V(s)$: old. G: fresh data. α: step-size / learning rate.

---

traj index $i_s^1, i_s^2, i_s^3, \dots$

s, $G^{i_s^j}$, $j = 1, 2, \dots, n_s$

$$V(s) = \frac{1}{n_s} \sum_{j=1}^{n_s} G^{i_s^j}$$

$$\hookrightarrow \underline{V(s)} = \operatorname*{argmin}_{x \in \mathbb{R}} \sum_j (x - G^{i_s^j})^2.$$

Δ

$$\Delta = \sum_j (x - G^{i_s^j})^2$$

$$\nabla_x \sum_i (x - G^{ij})$$

$$= \sum_j 2(x - G^{ij}).$$

$$V(s) \leftarrow V(s) - \alpha (V(s) - G^{ij})$$

$$= ((1-\alpha) V(s) + \alpha G^{ij}.$$

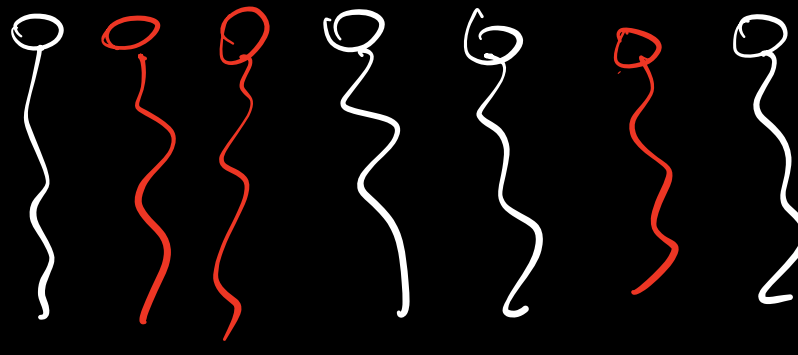
TD(0).

$$V(s_t) \leftarrow V(s_t) + \alpha (r_t + \gamma V(s_{t+1}) - V(s_t))$$

$V \to V^\pi(s) \quad \forall s.$

$$\mathbb{E}[r_t + \gamma V^\pi(s_{t+1}) - V^\pi(s_t) \mid s_t].$$

$$V^\pi(s) = \underbrace{R(s,\pi) - \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^\pi(s')]}_{\mathbb{E}[r + \gamma V^\pi(s') \mid s] \quad = 0}.$$

$$V(s) \leftarrow V(s) + \alpha \underbrace{(G - V(s))}_{\text{error}}.$$

Converges to: $V(s) = \mathbb{E}[G|s]$.