Lecture 23 RL

r_0 $r_1 - - - r_t$

current reward future

γ^t → smaller

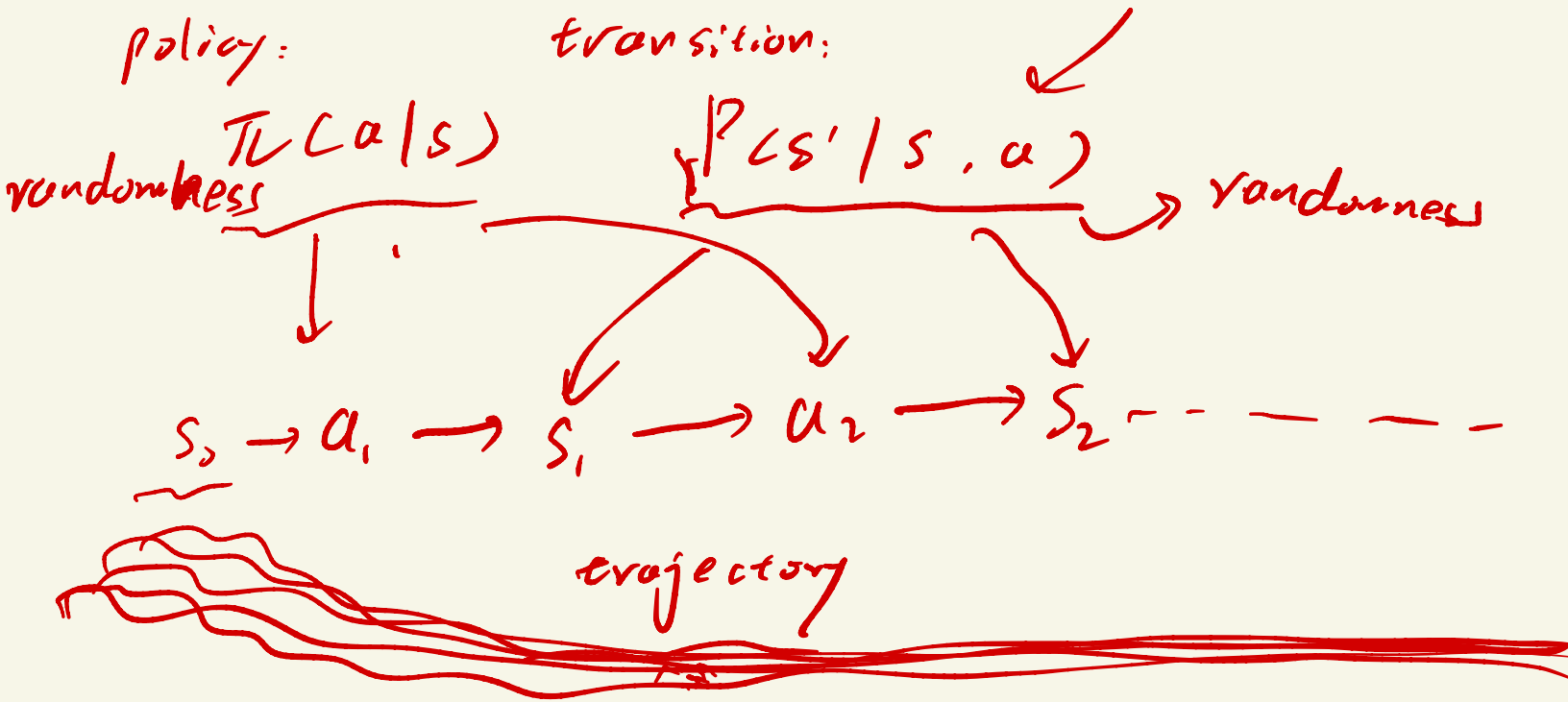
infinite time

if time is finite max $\sum_{t=0}^{T} r_t$ T steps

imitation

policy

efficiency vs effectiveness


policy:
transition:
$\pi(a|s)$
$P(s'|s,a)$
randomness
randomness
$s_0 \to a_1 \to s_1 \to a_2 \to s_2$
trajectory

$V^{\pi}(s)$

starting in s execute π

$$V^{\pi}(s) = \sum_{t=0}^{\infty} \gamma^t E[R(s_t, \pi(s_t))]$$

deterministic

starting from s

a_t

$$= R(s, \pi(s)) + \sum_{t=1}^{\infty} \gamma^t E[R(s_t, \pi(s_t))]$$

s_{t+1}

$$= R(s, \pi(s)) + \gamma \cdot \sum_{t=0}^{\infty} \gamma^t E[R(s_t, \pi(s_t))]$$

starting from the next state of s, s'

$$P(s' \mid s, \pi(s)) \cdot V(s') \quad \gamma \cdot E_{s'}[V(s')]$$

$$V^{\pi}(s) = \sum_{t=0}^{\infty} \gamma^t \, \mathbb{E}\left[R(s_t, \pi(s_t)) \mid s_0 = s \right]$$

$V^{\pi}(s)$: s

$\sum_{t=0}^{\infty} \gamma^t \, \mathbb{E}\left[R(s_t, \pi(s_t)) \mid s_0 = s \right]$: $V^{\pi}(s)$

$s_0 = s$: s

$S \in [1, 2, 3]$

Linear

1. $V^{\pi}(S=1) = R(S=1, \pi(S=1)) + \gamma \sum_{s_1 \in [1,2,3]} P(s_1 | S=1, \pi(S=1)) V^{\pi}(s_1)$

$$\begin{pmatrix} x = V^{\pi}(S=1) \\ y = V^{\pi}(S=2) \\ z = V^{\pi}(S=3) \end{pmatrix}$$

2. $V^{\pi}(S=2) = R(S=2) \cdots$

3. $V^{\pi}(S=3) \cdots$

$V^{\bar{\pi}}(s)$

$\pi(a|s)$

$\max_{a \in A}$

$\pi^*(a|s)$

V^*

$\bar{\pi}^*$

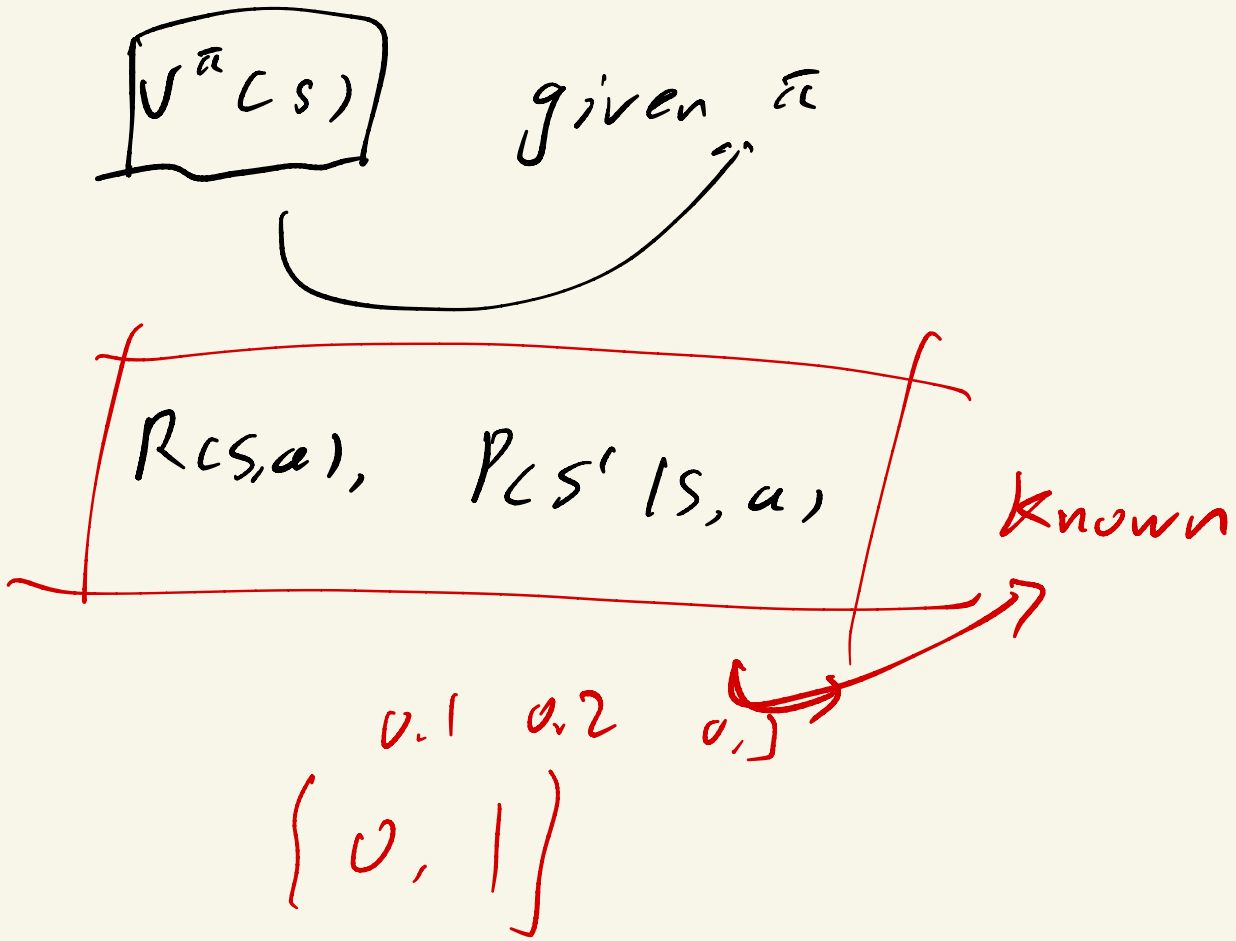
$V^{(t+1)}(s) \longrightarrow V^*$

$\pi(a|s)$ deterministic: $Q(s,a) = V(s)$

π^*

$Q(s,a) = R(s,a) + \gamma \cdots$

$V(s)$

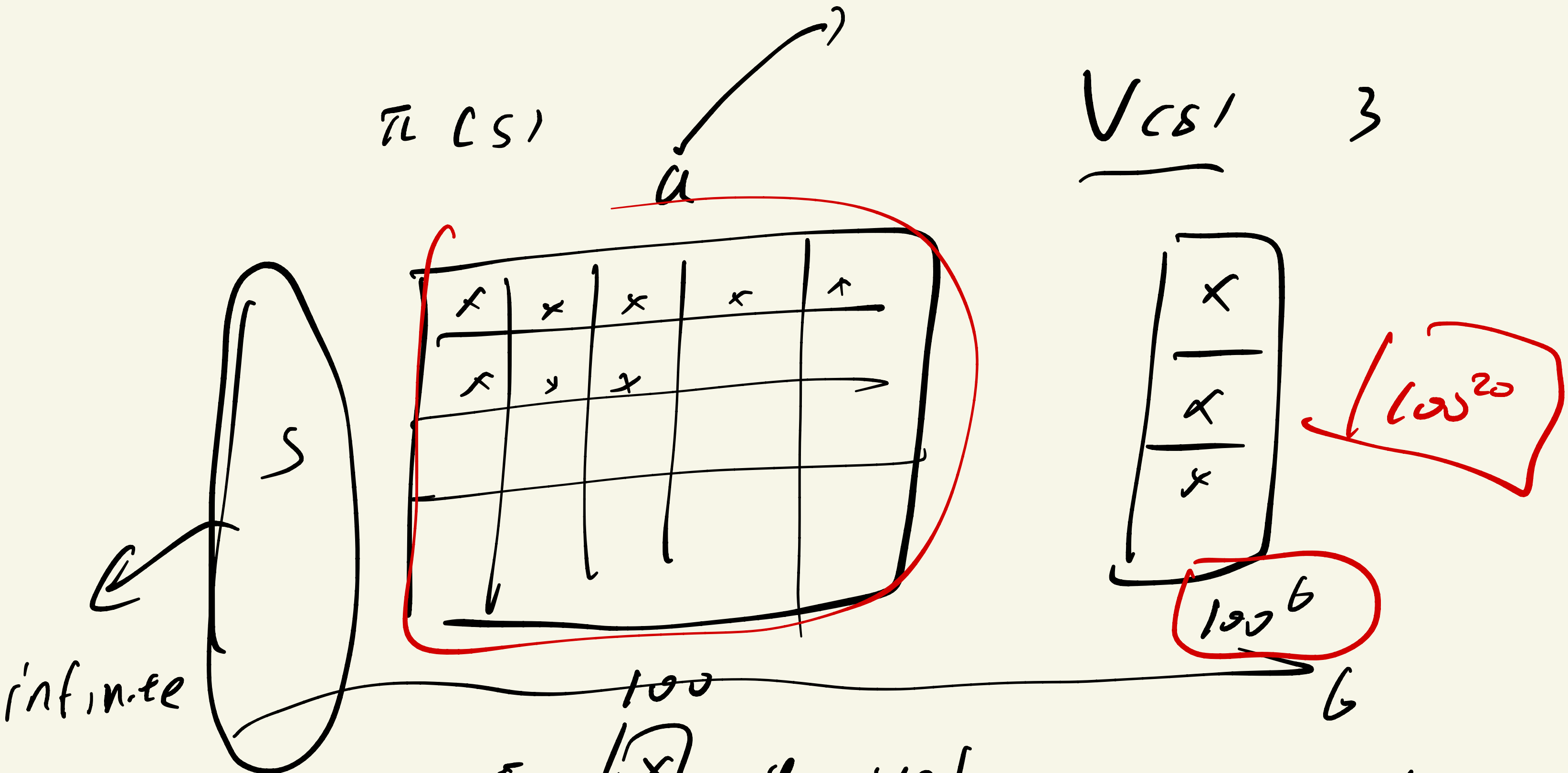

$V^{\pi}(s)$
given π
$R(s,a)$, $P(s'|s,a)$
known
a_1 a_2 a_3
$\{0, 1\}$

strength $\{1, 2, 3, 4, 5\}$

[100 times. 10 times chair fall down

$$R(\text{strength}=1) = \frac{10}{1000}$$

$$= 0.01$$



[x, y, velocity at x, velo... at y]

[s, V(s)]

$\pi(a|s)$

continuous

Policy model

$$\eta_\theta = \bar{E}_{\tau \sim p_\theta}\left[\sum_{t=0}^{T-1} \gamma^t R(s_t, a_t)\right]$$

unrelated to θ

- $\bar{E}_{z \sim q_\phi(z|x)}\, \bar{E}\, LBO(\theta, \phi)$

θ

$z \sim$ Gaussian

τ $s_0\, a_1\, s_1\, a_2\, s_2\, a_3$

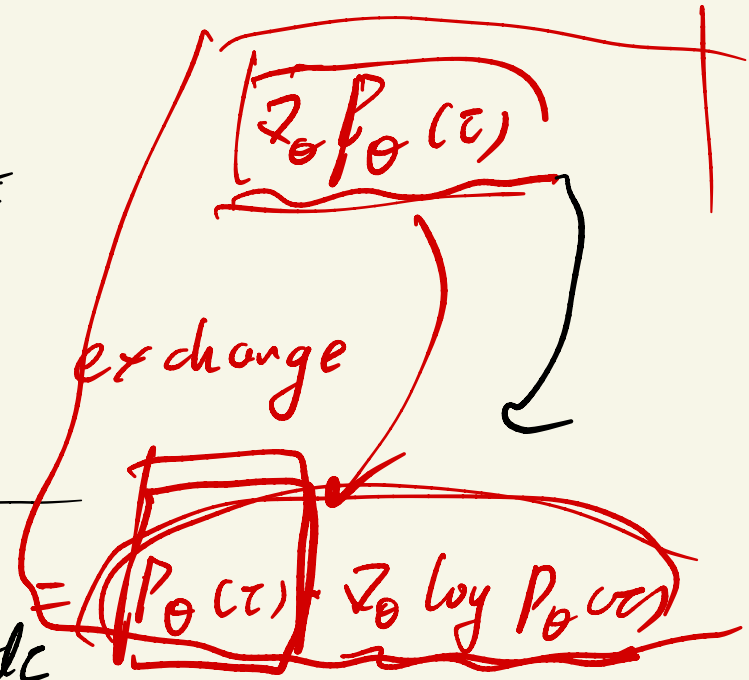
$$\nabla_\theta E_{\tau \sim p_\theta}[f(\tau)]$$

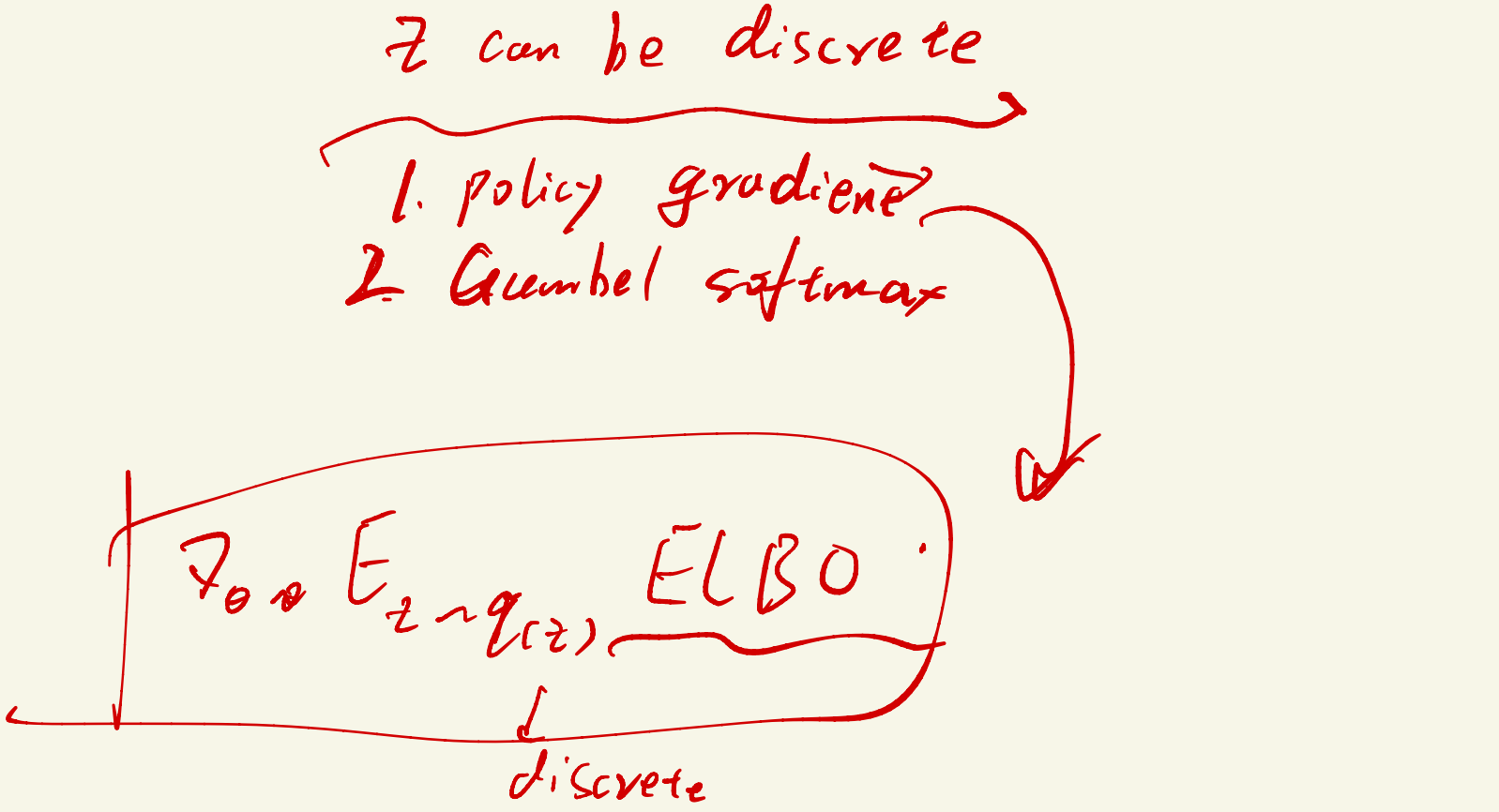
$$= \nabla_\theta \int p_\theta(\tau) f(\tau) \, d\tau$$

$$= \int_\tau \nabla_\theta p_\theta(\tau) f(\tau) \, d\tau$$

$$= \int_\tau p_\theta(\tau) \nabla_\theta \log p_\theta(\tau) f(\tau) \, d\tau$$

$$= E_{\tau \sim p_\theta}[\nabla_\theta \log p_\theta(\tau) f(\tau)]$$




z can be discrete
1. policy gradient
2. Gumbel softmax
$\nabla_\theta \mathbb{E}_{z \sim q(z)}$ ELBO
discrete