Policy Iteration (cont.)

$\pi_f(s) = \arg\max_{a \in A} f(s,a)$.

Initialize π_0.

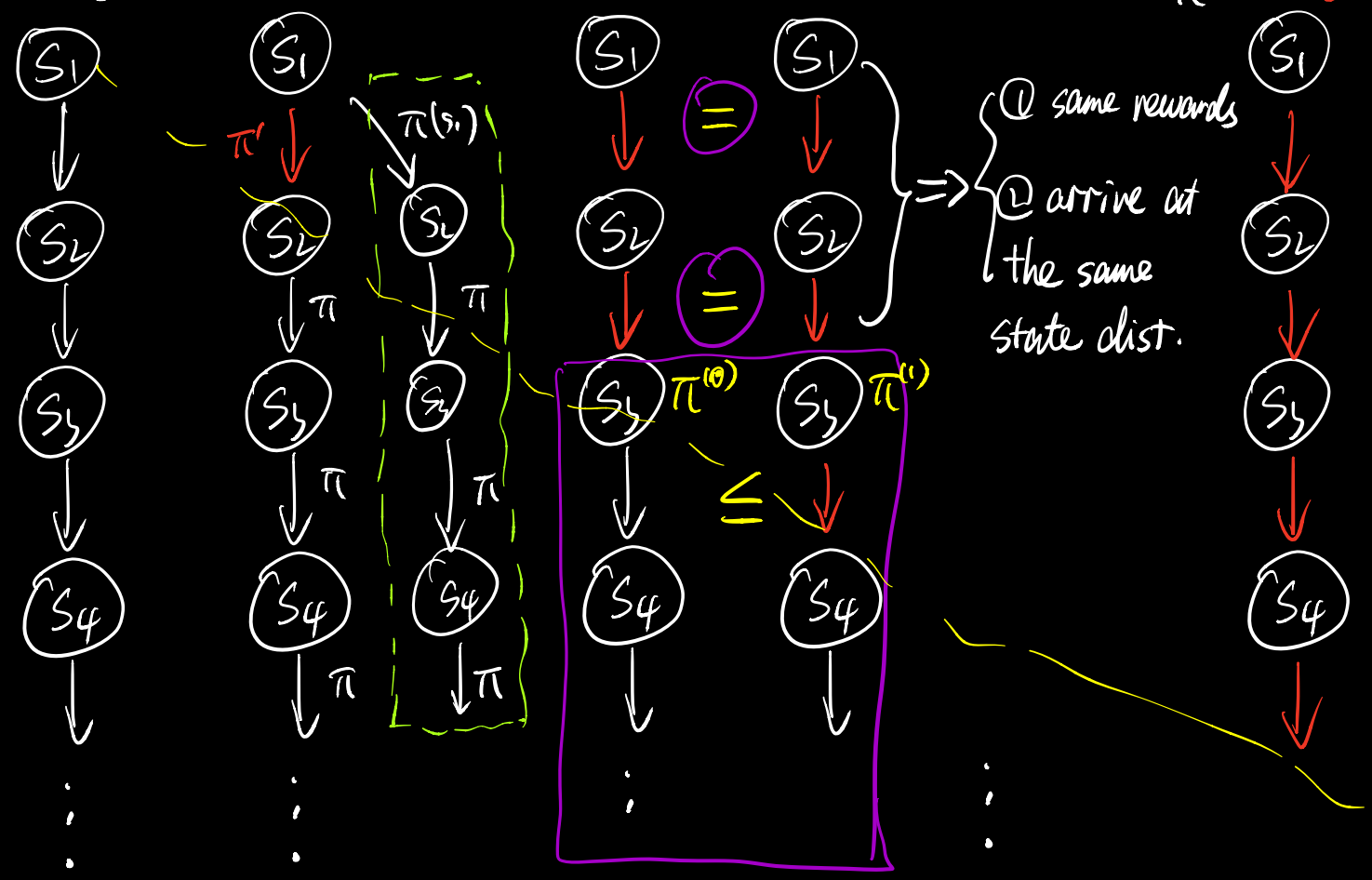
For $k = 1, 2, \cdots$,

- $\boxed{\text{Policy Evaluation}}$ Compute $Q^{\pi_{k-1}}$
- $\boxed{\text{Policy Improvement}}$ $\pi_k \leftarrow \pi_{\boxed{Q^{\pi_{k-1}}}}$

π' $(= \pi_k)$

π $(= \pi_{k-1})$

$\pi^{(0)} = \pi \le \pi^{(1)} \le \pi^{(2)} \le \pi^{(3)} \le \cdots \le \pi^{(\infty)} = \pi'$



$\Rightarrow$ ① same rewards ② arrive at the same state dist.

Q: if starting in $s_1 = s$, how good is $\pi^{(1)}$?

A: $Q^{\pi}(s, \pi') \ge Q^{\pi}(s, \pi) = V^{\pi}(s)$.

b/c: π' greedily maximizes $Q^{\pi}(s, \cdot)$

key: $\pi^{(1)}$ better than $\pi^{(0)}$ holds for all starting state

PD-lemma: $\forall \pi, \pi'$. $\forall s$. more general.

$$V^{\pi'}(s) - V^{\pi}(s) = \frac{1}{1-\gamma} \mathbb{E}_{\tilde{s} \sim d_s^{\pi'}} \left[Q^{\pi}(\tilde{s}, \pi') - V^{\pi}(\tilde{s}) \right].$$

Linear Programming (LP)

max. $x_1 + 2x_2$ ← linear objective.
subject. $3x_1 - x_2 \le 1$.
↖ linear (equality or inequality) constraints

Primal. LP for solving $V^* \in \mathbb{R}^{|S|}$

Choose any $d_0 \in \mathbb{R}^{|S|}$. s.t. $\forall s$, $d_0(s) > 0$.
$\sum_s d_0(s) = 1$

$$\boxed{\begin{aligned} \min_{V \in \mathbb{R}^{|S|}} \quad & d_0^T V. \\ \text{s.t.} \quad & \mathcal{T} V \le V. \end{aligned}}$$

nonlinear

① why this solves V^*?
② why this is LP?

recall: $\mathcal{T}^{\pi} V = R^{\pi} + \gamma P^{\pi} V$.

① $\mathcal{T} V \le V. \implies \mathcal{T}(\mathcal{T} V) \le \mathcal{T} V \le V. \implies \cdots$

monotone property of $\mathcal{T}$.

$\mathcal{T}^{\infty} V \le V \implies V^* \le V.$

original problem $\iff \boxed{\begin{aligned} \min \quad & d_0^T V \\ \text{s.t.} \quad & V \ge V^* \end{aligned}} \implies V = V^*.$

② why linear? obj is linear ✓.

constraint is: $\mathcal{T} V \le V. \iff \forall s, (\mathcal{T} V)(s) \le V(s).$

$\rightarrow \max_{a \in A} \left(R(s,a) + \gamma \langle P(\cdot|s,a), V \rangle \right) \le V(s).$ $\forall s$

$\Updownarrow$

$R(s,a) + \gamma \langle P(\cdot|s,a), V \rangle \le V(s).$ $\forall s, a$

Primal form: $\min_{V \in \mathbb{R}^{|S|}} d_0^T V$ s.t. $\underset{|S\times A|\times 1}{R} + \gamma \underset{|S\times A|\cdot|S|}{P} V \le V$

Dual form: $\max_{\mu \ge 0} \mu^T R$ → $|S\times A|\times 1$ reward vector

Bellman "flow" eq.

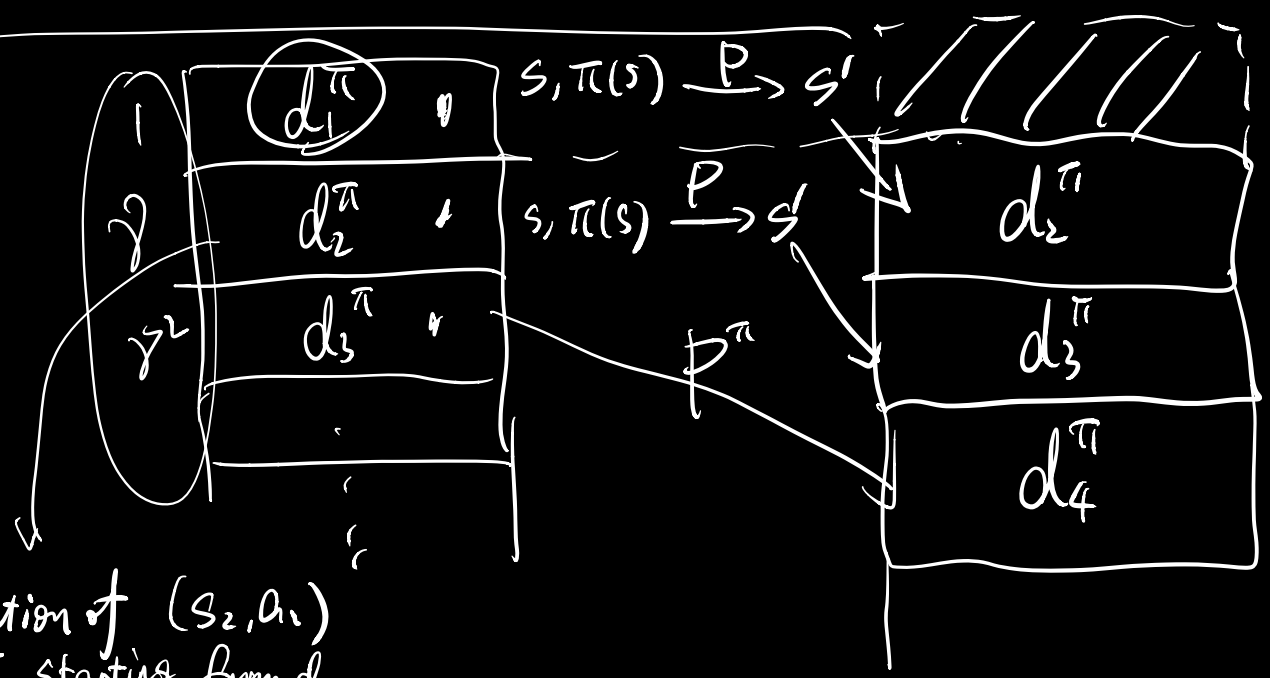
s.t. $\sum_{a'} \mu(s',a') = d_0(s') + \gamma \sum_{s,a} \mu(s,a) \cdot P(s'|s,a) \quad \forall s'$

Interpretation: any feasible $\mu = \tilde{d}^\pi$ for some stochastic policy π

↳ state-action occupancy (unnormalized)

$$\langle \tilde{d}^\pi, R \rangle = \mathbb{E}_{s\sim d_0}\left[V^\pi(s) \right]$$

$$\pi(a|s) = \frac{\mu(s,a)}{\sum_{a'} \mu(s,a')}$$



distribution of (s_2, a_2) under π, starting from d_0