1 マルコフ状態

$$P\left(S_{t+1}\,|S_t\right) = P\left(S_{t+1}\,|S_1, \ldots, S_t\right) \qquad (1)$$

- 未来と過去とは無関係

$$H_{1:t} \to S_t \to H_{t+1:\infty} \qquad (2)$$

  H: 履歴

- 一度状態 S が決まれば過去の履歴は不要
- 逆に言えば状態 S は未来に対する十分統計量
- 環境の状態 S_t^e はマルコフ性を持つ
- 歴史 H_t はマルコフ性を持つ

2 完全観測可能，部分観測可能 Full, partially observability

$$O_t = S_t^a = S_t^e \qquad (3)$$

- **部分観測可能なマルコフ決定過程 POMDP**: **P**artially **O**bservable **M**arkov **D**ecision **P**rocess
- **マルコフ決定過程** Markov decision processes: MDP
  - *ほぼ全ての強化学習はマルコフ決定過程として記述可能*

3 マルコフ過程 Markov Process

- **マルコフ過程** MP は
- **マルコフ過程** Markov Process (マルコフ連鎖 Markov Chain) : 状態 S と遷移行列 P
  - S: *状態の集合*
  - P: *状態間の遷移行列*
  - $P_{ss'} = P\left(S_{t+1} = s' \,|S_t = s\right)$

4 マルコフ過程決定過程 Markov Decision Process: MDP

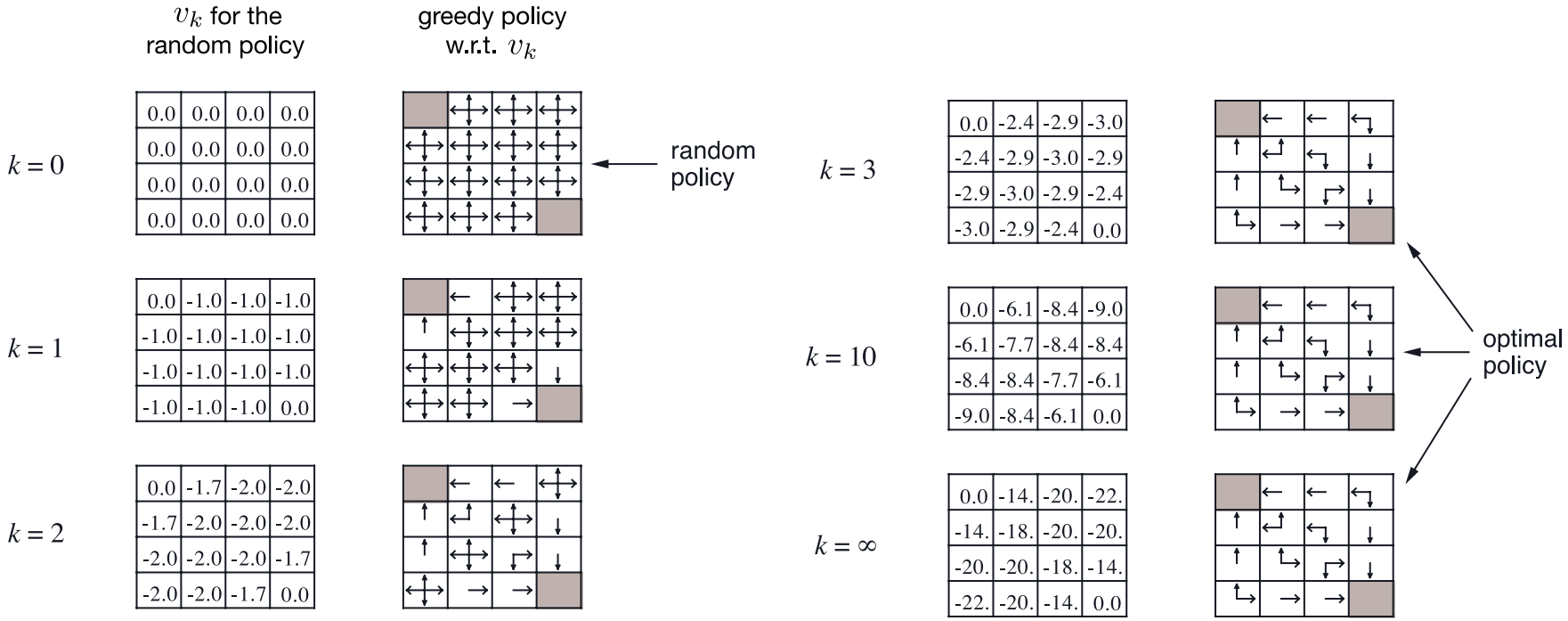
マルコフ決定過程 (MDP)

- MDP は，状態 S, 行動 A, 遷移確率 P, 報酬 R の組 $< S, A, P, R, \gamma >$

$$P_{ss'}^{a} = P\left(S_{t+1} = s' \,|S_t = s, A_t = a\right) \qquad (4)$$

- R は報酬関数 $R_s^a = \mathbf{E}\left[R_{t+1} \,|S_t = s, A_t = a\right]$
- γ は**割引率** discount factor $\gamma \in [0, 1]$

5 グリッドワールド



Sutton and Barto (1998) より

文献

Sutton, Richard S., and Andrew G. Barto. 1998. *Reinforcement Learning*. Cambridge, MA: MIT Press.