

马尔可夫决策过程

Markov decision process (MDP)

章阳 / 计算机学院
Email: yangzhang@whut.edu.cn

目 录

- 定义与背景介绍
 - 马尔可夫过程 (Markov Process)
 - 马尔可夫回报过程 (Markov Reward Process, MRP)
- 马尔可夫决策过程 (Markov Decision Process, MDP)
 - MDP的定义
 - MDP求最优解
- 应用实例与扩展
 - 简单例子
 - 移动无线充电系统的应用
 - MDP的扩展

目 录

- 定义与背景介绍
 - 马尔可夫过程 (Markov Process)
 - 马尔可夫回报过程 (Markov Reward Process, MRP)
- 马尔可夫决策过程 (Markov Decision Process, MDP)
 - MDP的定义
 - MDP求最优解
- 应用实例与扩展
 - 简单例子
 - 移动无线充电系统的应用
 - MDP的扩展

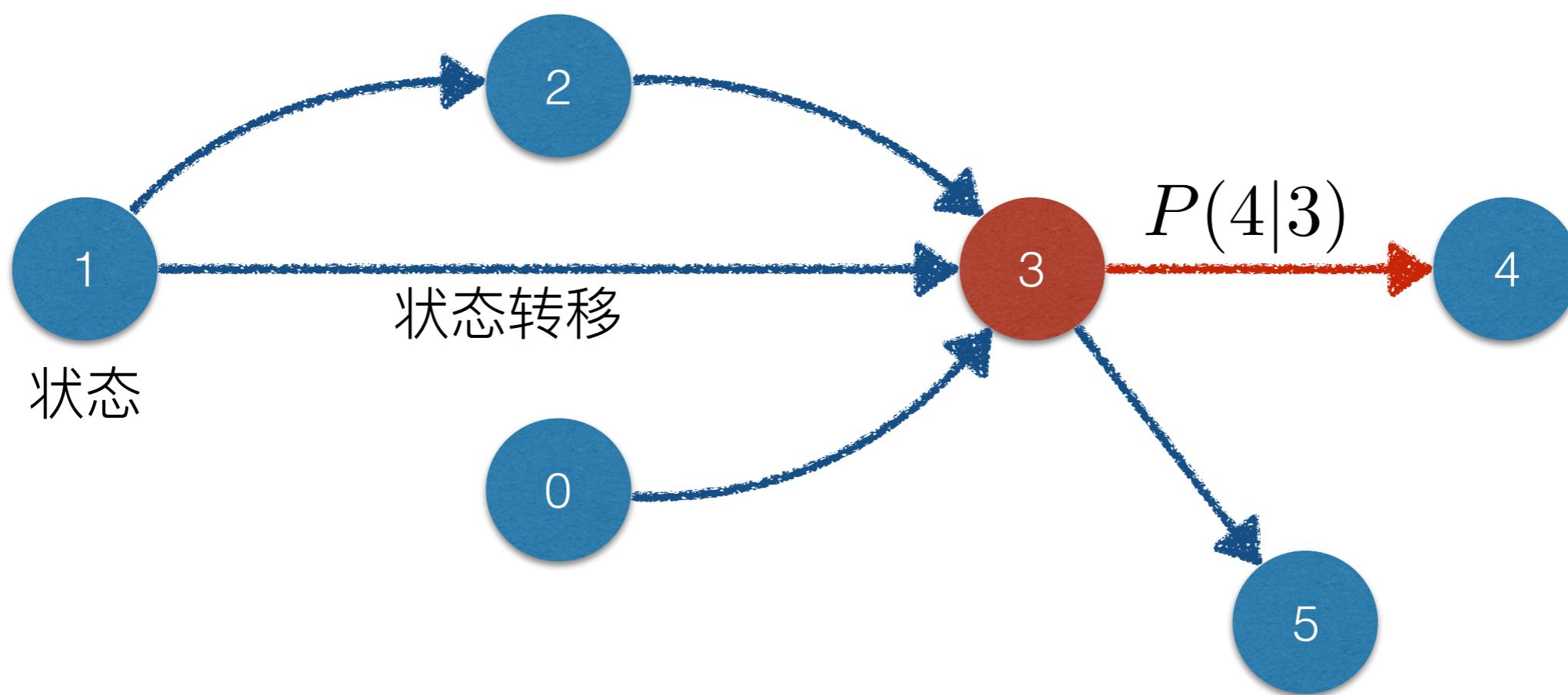
(离散事件的) 马尔可夫过程

- 定义：

在一个随机过程 $(X_t, t \in I)$ 中，已知当前的状态 X_t ，则将来状态 X_{t+1} 出现概率与过去状态 X_{t-1}, X_{t-2}, \dots 无关，即：

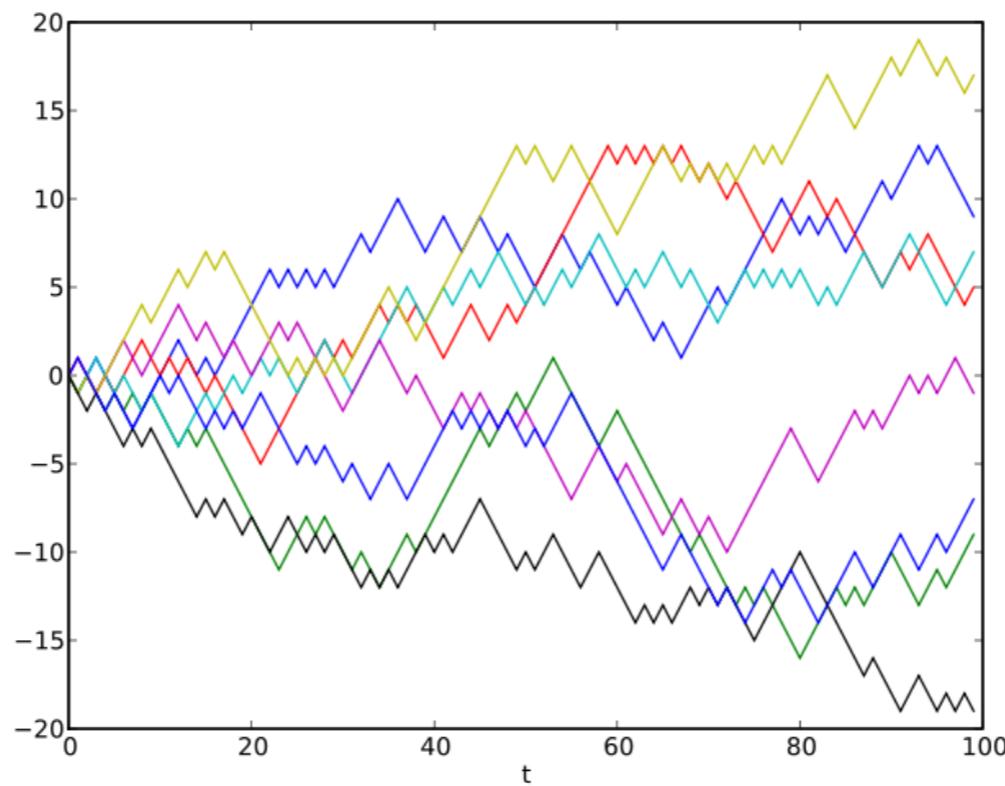
$$P(X_{t+1}|X_t, X_{t-1}, X_{t-2}, \dots) = P(X_{t+1}|X_t)$$

马尔可夫过程-马尔可夫链



马尔可夫过程：举例

- 公交站等车的人数
- 某仓库仓储量，某船只在各个港口载货量
- 布朗运动，随机游走 (Random walk) ，价格



马尔可夫过程：反例

- 天气变化：明天的天气状态可能与今天和昨天都有关

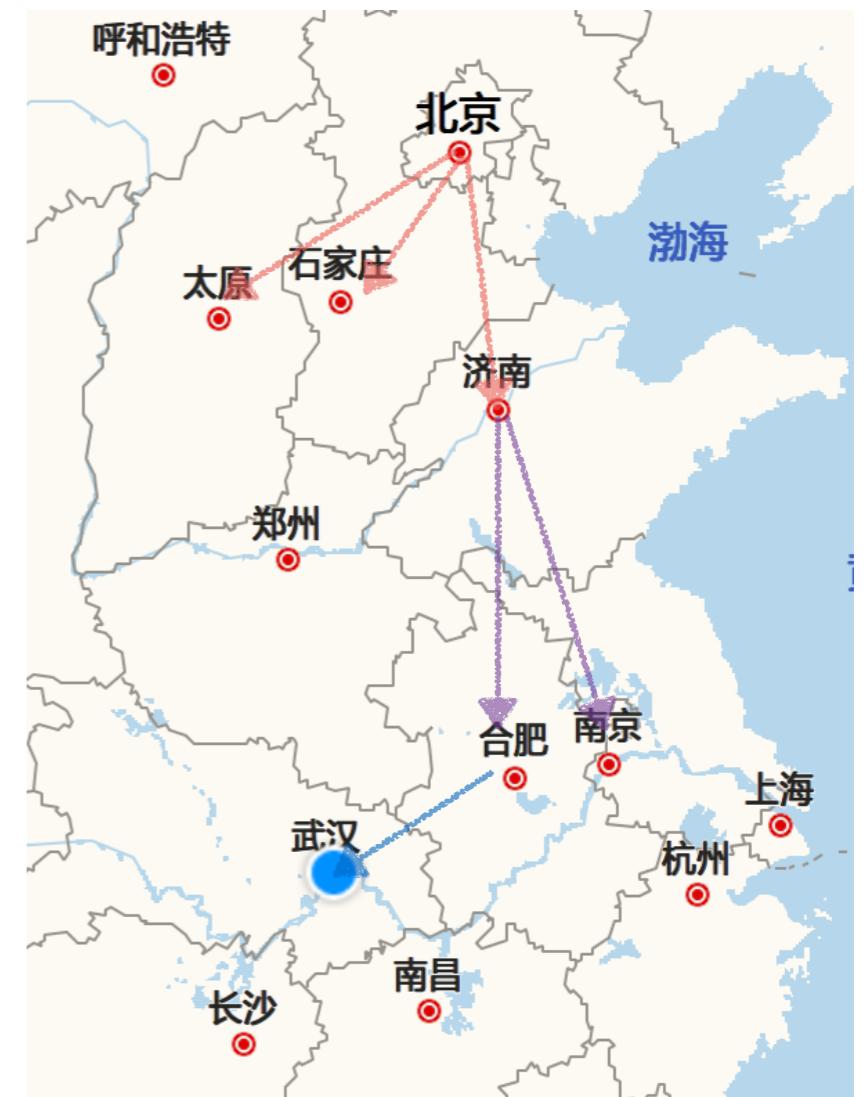
$$P(X_{t+1}|X_t, X_{t-1}, X_{t-2}, \dots) = P(X_{t+1}|X_t, \textcircled{X_{t-1}})$$

- 将该类非马尔可夫过程转化为马尔可夫过程
 - 重新定义一个状态 $S_t = (X_t, X_{t-1})$

$$S_t \in \{(s, s), (s, r), (r, s), (r, r)\}$$

马尔可夫过程的判定？

- 定义证明
- 直接假设或依据已有惯例
- 转化/映射为已有的马尔可夫过程
 - 例如映射成 random walk



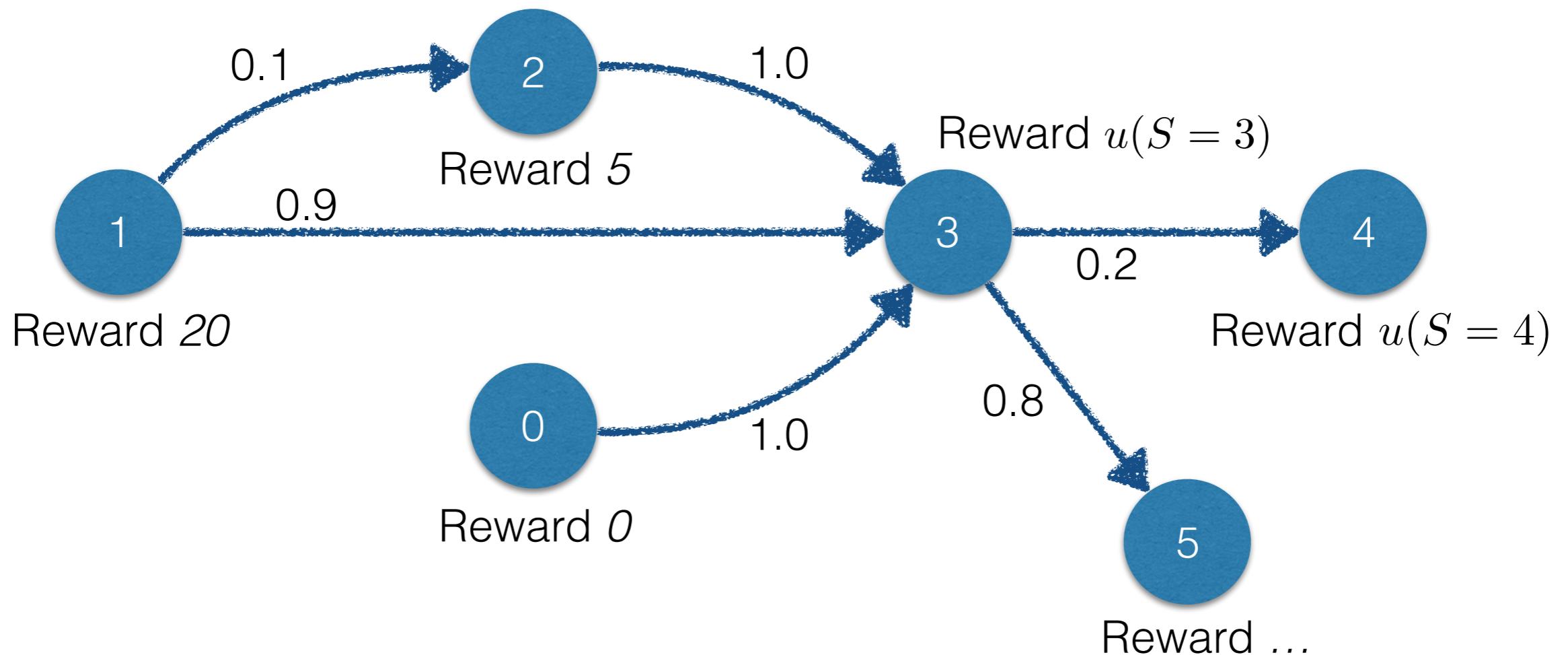
目 录

- 定义与背景介绍
 - 马尔可夫过程 (Markov Process)
 - 马尔可夫回报过程 (Markov Reward Process, MRP)
- 马尔可夫决策过程 (Markov Decision Process, MDP)
 - MDP的定义
 - MDP求最优解
- 应用实例与扩展
 - 简单例子
 - 移动无线充电系统的应用
 - MDP的扩展

马尔可夫回报过程

Markov reward process (MRP)

MDP = Markov process + reward/utility functions
马尔可夫过程 + 回报/效用函数



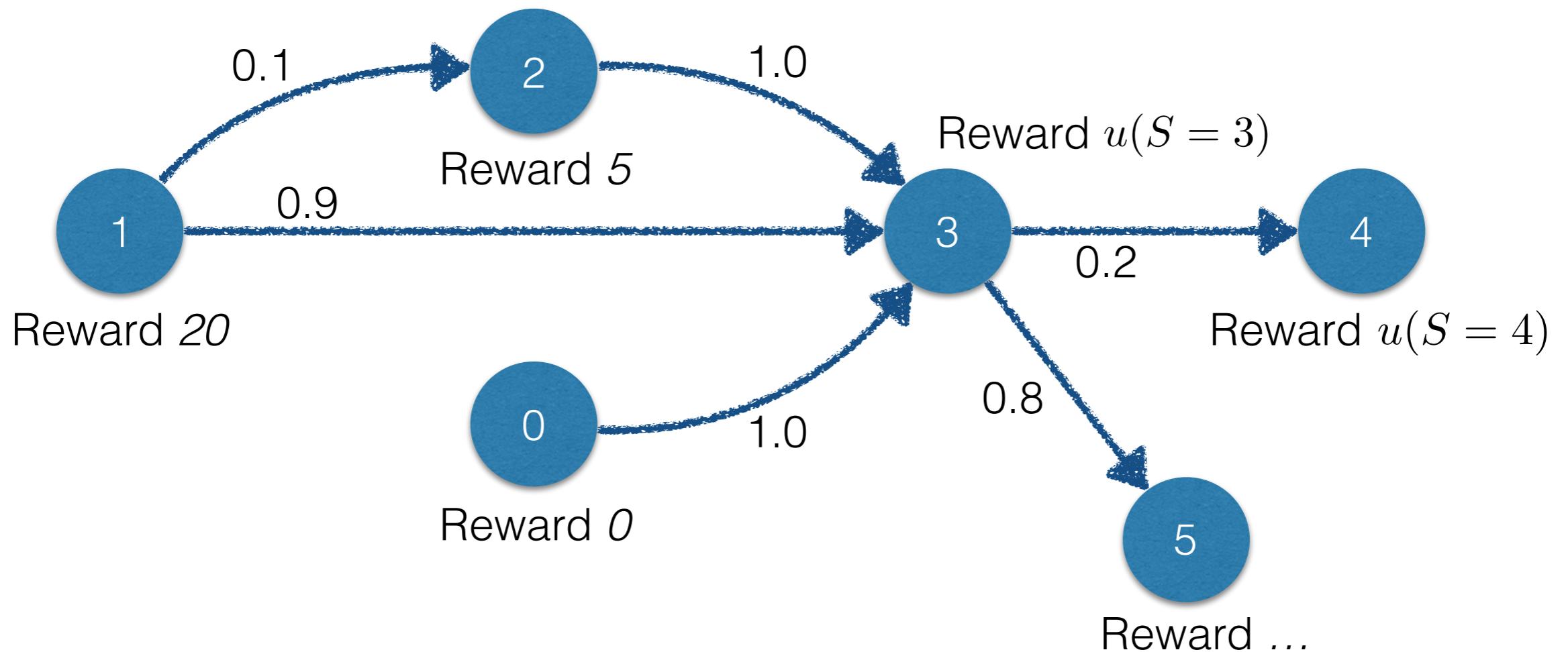
MRP形式化定义

- 一个MRP可以按如下描述：

$$\langle \mathbb{S}, \mathbf{P}, U, \gamma \rangle$$

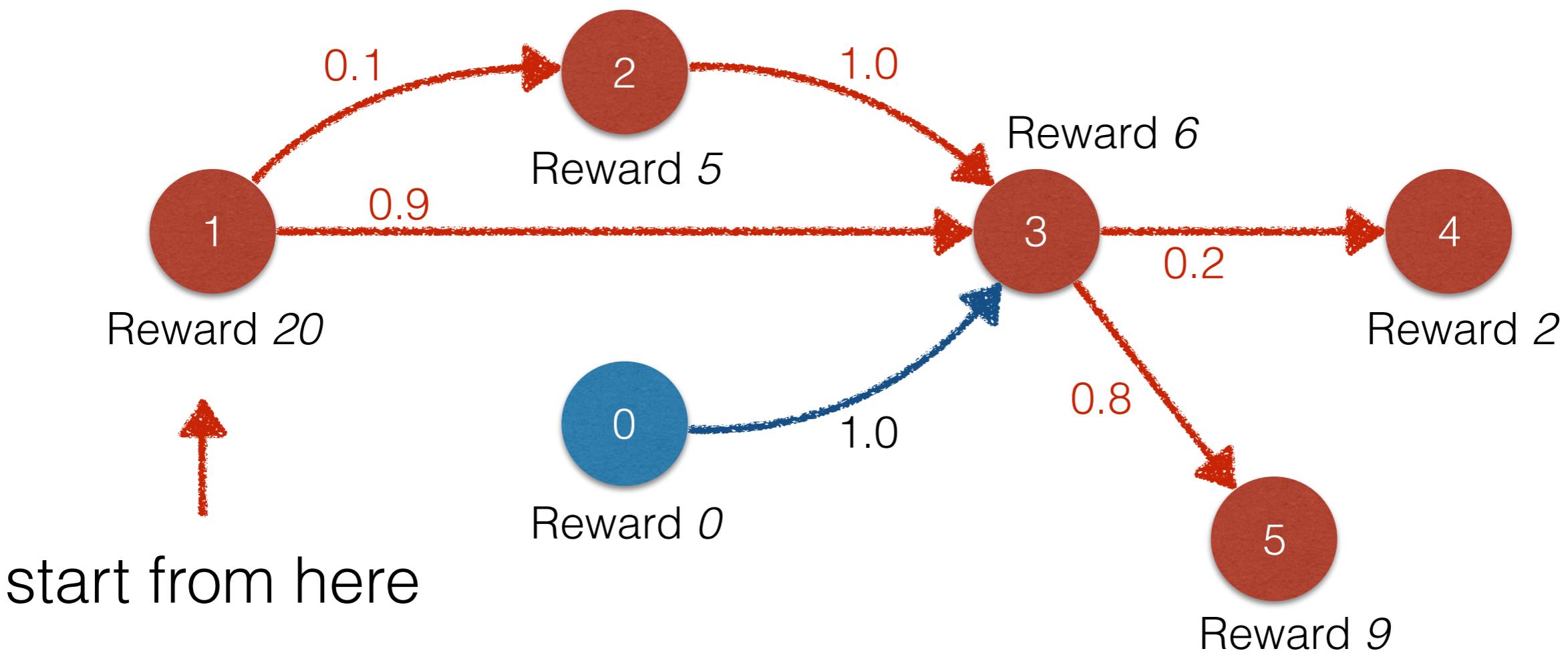
- \mathbb{S} 为所有的系统状态集合
- \mathbf{P} 状态转移概率 (state transition prob.) 矩阵
- U 为回报函数 (reward function)
- γ 为折扣/折旧因子 (discount factor)

MRP-求解的目标



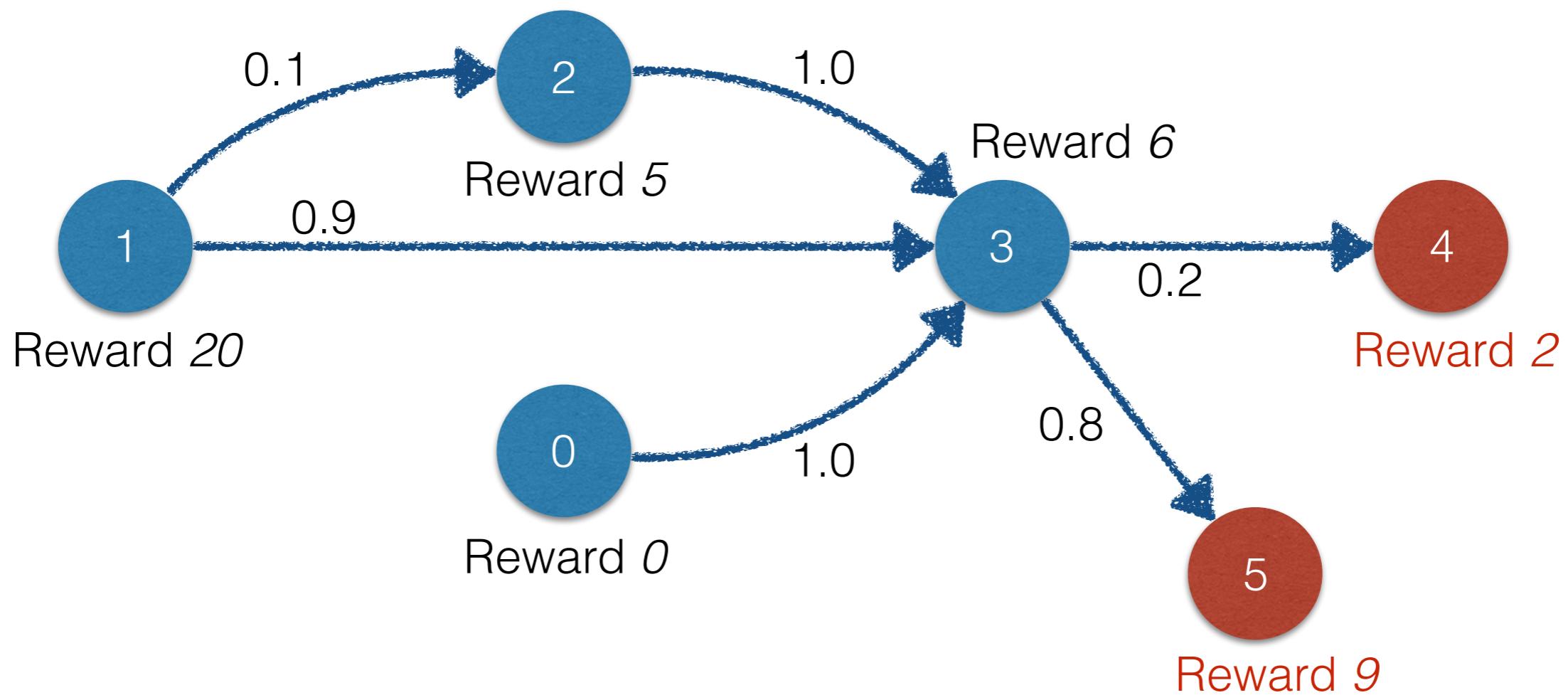
- 用户从处于任意一个系统状态起，持续运行，能获得的期望收益是多少

MRP



回报 (Reward) 为在当前 (immediate) 状态能获得的收益
我们计算“用户”从某个状态 S 开始总共能获得的期望回报 $H(S)$

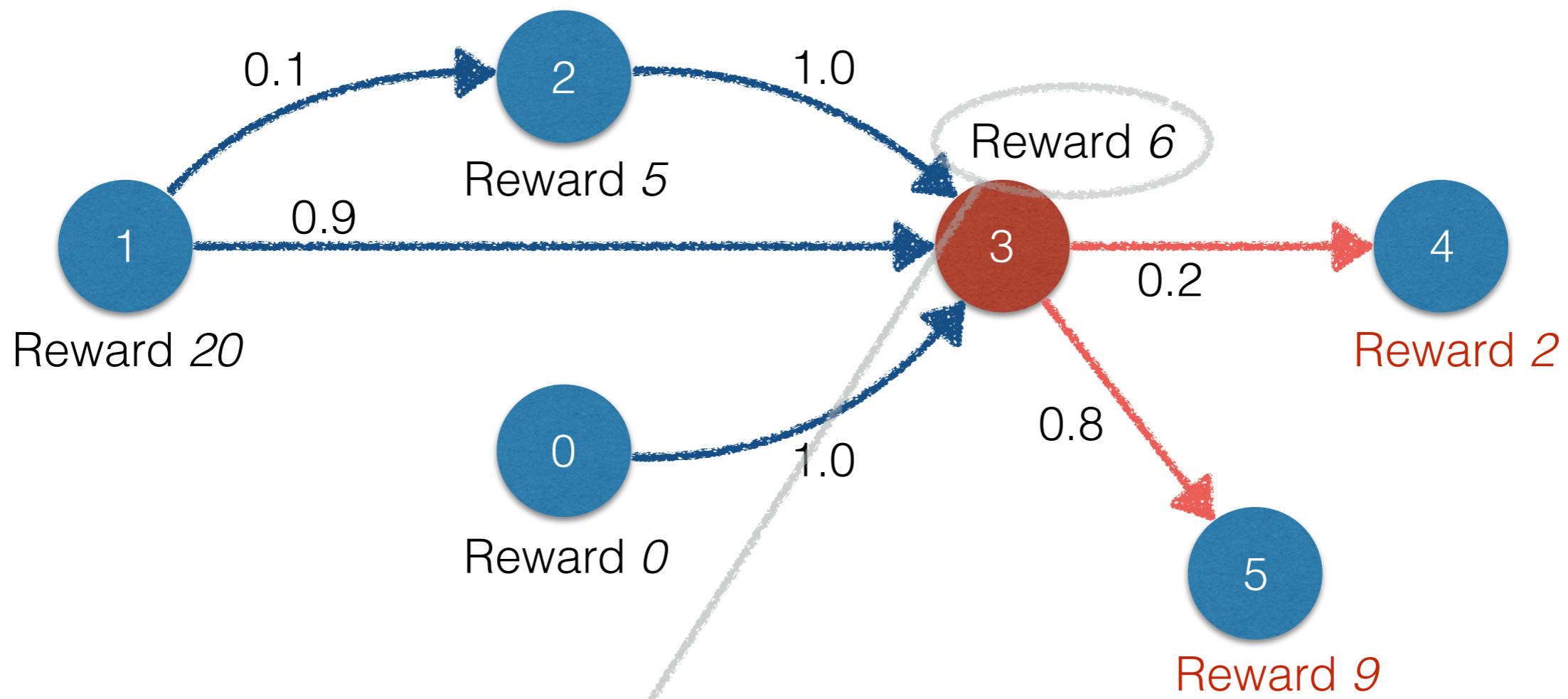
MRP求解 - 回溯法 Backward induction



$$H(S = 4) = u(S = 4) = 2$$

$$H(S = 5) = u(S = 5) = 9$$

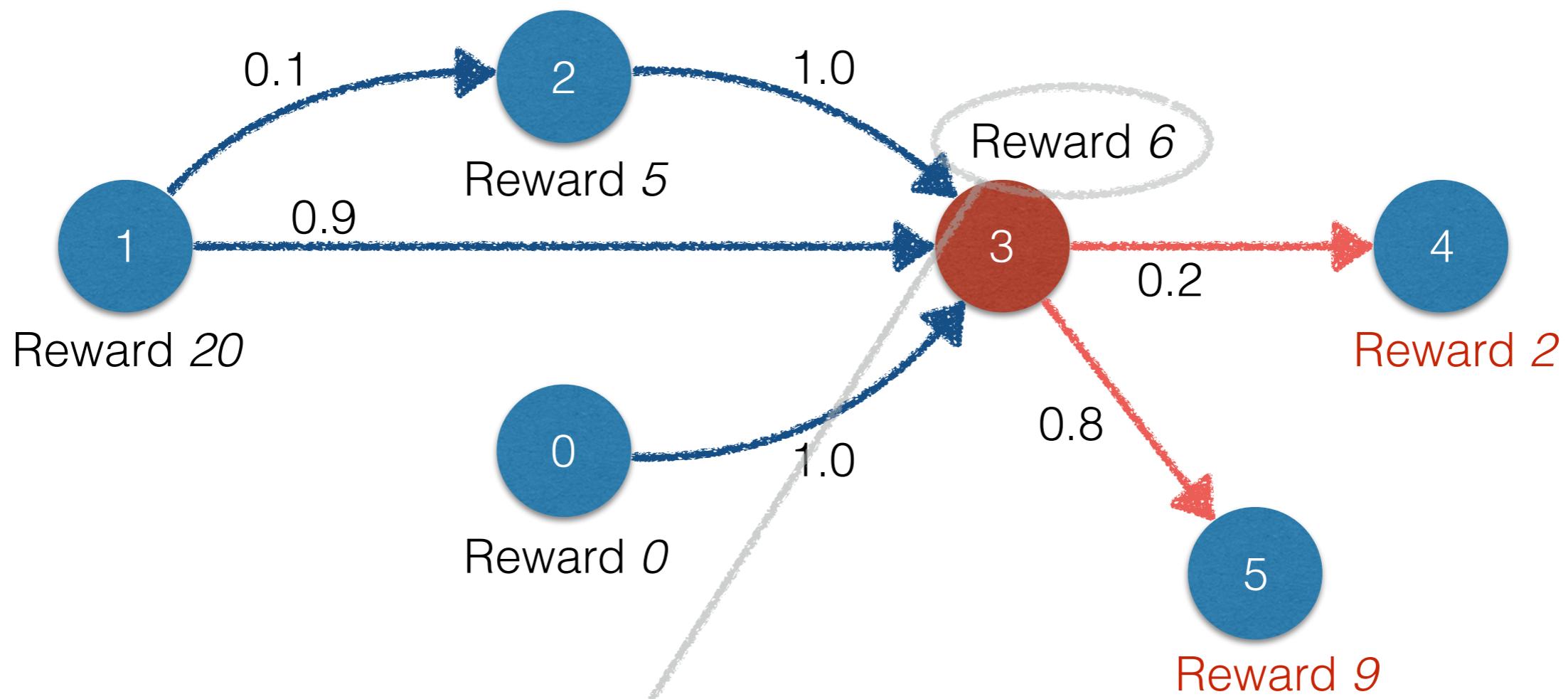
MRP求解 - 回溯法



$$\begin{aligned} H(S = 3) &= u(S = 3) + \gamma [0.2H(S = 4) + 0.8H(S = 5)] \\ &= 6 + \gamma[0.2 \cdot 2 + 0.8 \cdot 9] \end{aligned}$$

其中 $\gamma \in [0, 1)$ 表示“对未来收益的折价”，类似利率

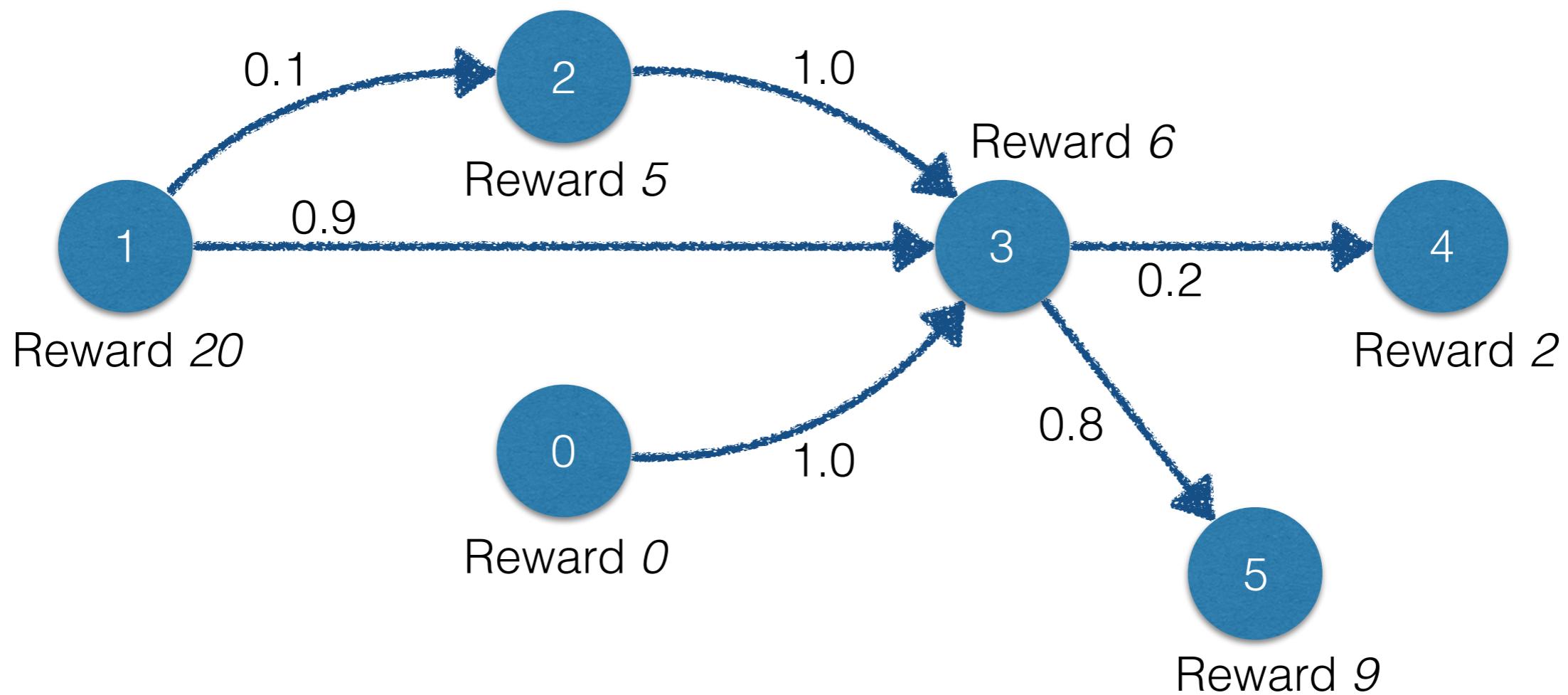
MRP求解 - 回溯法



$$\begin{aligned} H(S = 3) &= u(S = 3) + \gamma [0.2H(S = 4) + 0.8H(S = 5)] \\ &= 6 + \gamma [0.2 \cdot 2 + 0.8 \cdot 9] \end{aligned}$$

依此类推 $H(S=2), H(S=1), \dots$

MRP

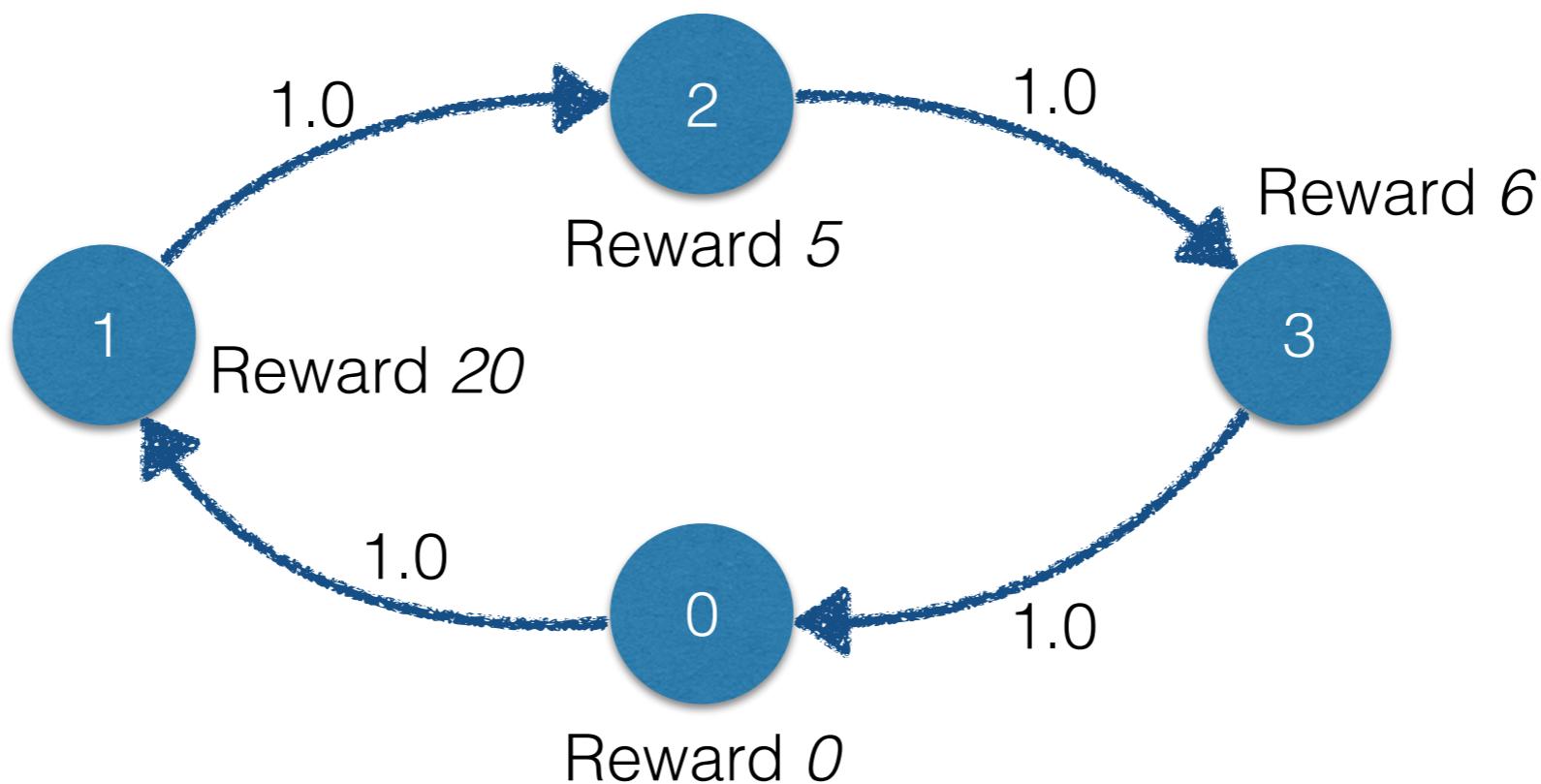


总回报的期望(迭代形式): $H(S_t) = \mathbb{E}\{u(S_t) + \gamma H(S_{t+1})\}$

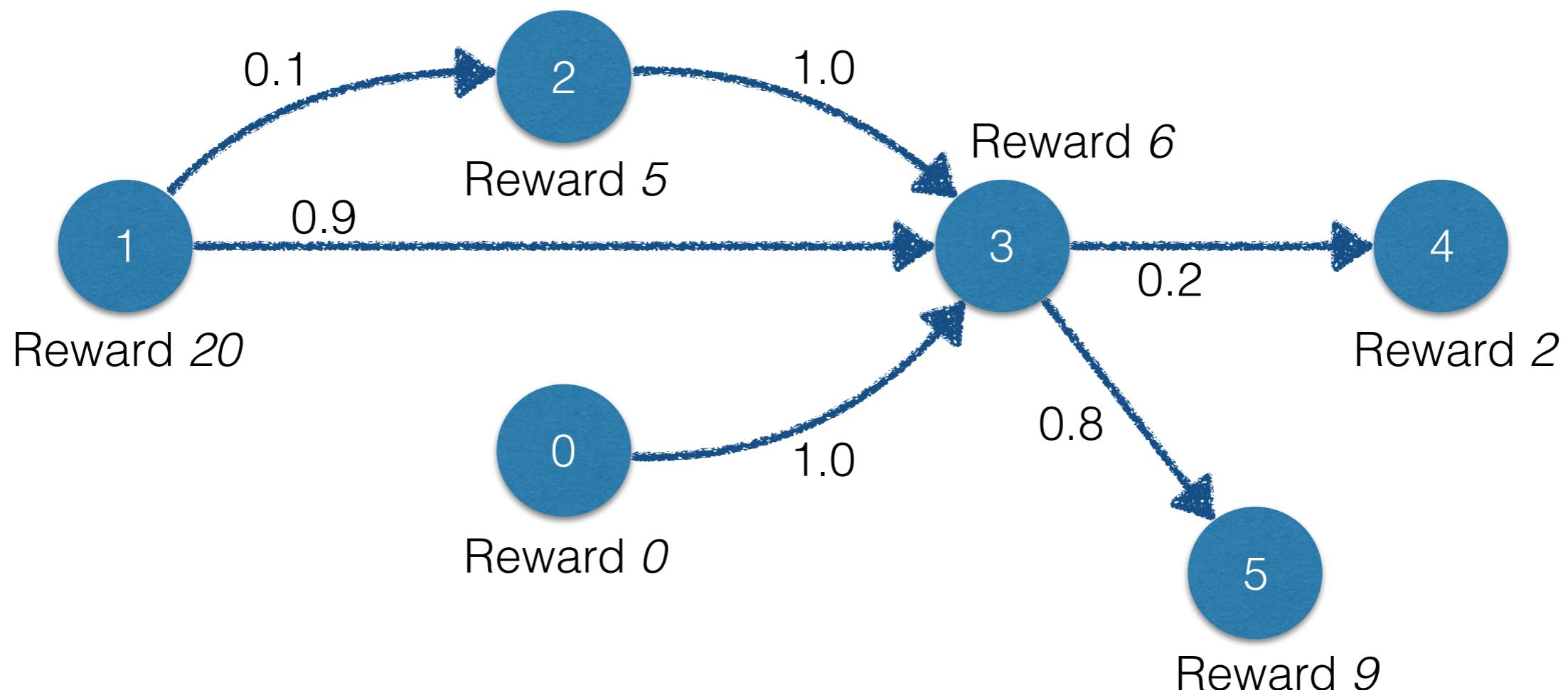
具体到本模型: $H(S) = u(S) + \gamma \sum_{S' \in \mathbb{S}} P(S, S') H(S')$

MRP - 迭代求解

- 适用于此类无线运行的、无终止状态或吸收状态 (absorbing state) 的马尔可夫过程
- 该类情况可能无法用回溯法



MRP - 迭代求解 Value iteration



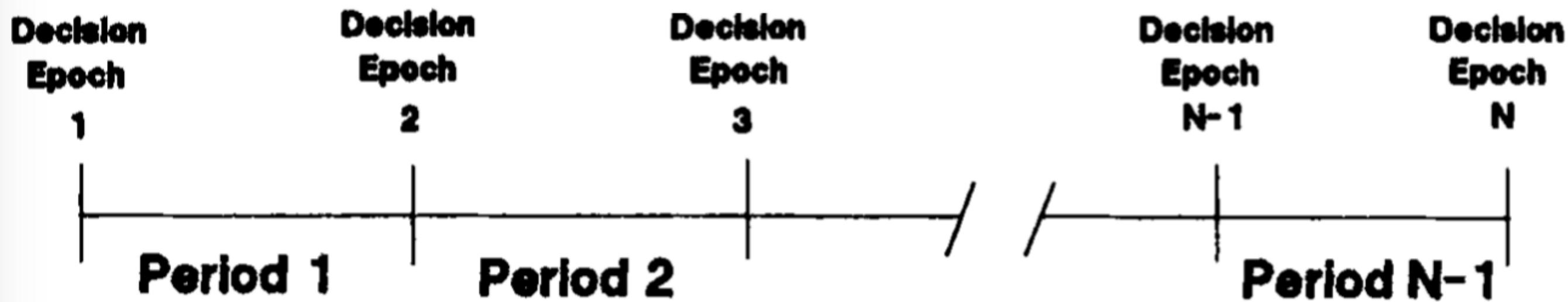
令 $H(S) \triangleq 0, \forall S \in \mathbb{S}$, 再用 $H(S) \triangleq u(S) + \gamma \sum_{S' \in \mathbb{S}} P(S, S')H(S')$

逐个更新每个状态的 $H(S)$ 直到全部收敛

目 录

- 定义与背景介绍
 - 马尔可夫过程 (Markov Process)
 - 马尔可夫回报过程 (Markov Reward Process, MRP)
- 马尔可夫决策过程 (Markov Decision Process, MDP)
 - MDP的定义
 - MDP求最优解
- 应用实例与扩展
 - 简单例子
 - 移动无线充电系统的应用
 - MDP的扩展

所针对的问题：决策

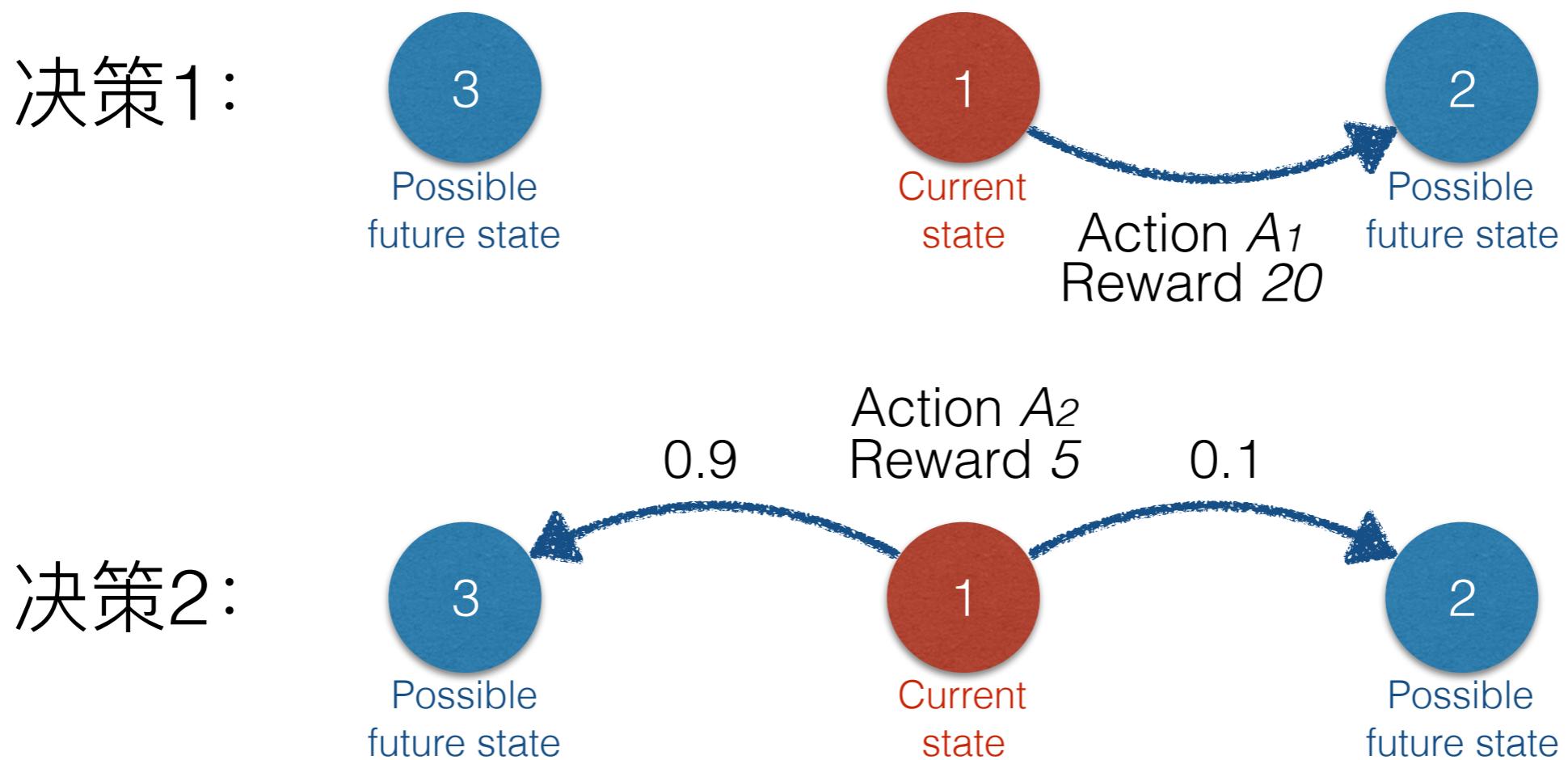


在每个决策时间点，用户观测并作出相应决策，系统继续运行

马尔可夫决策过程

Markov decision process (MDP)

MDP = Markov process + actions + reward functions
马尔可夫过程 + 决策 + 回报函数



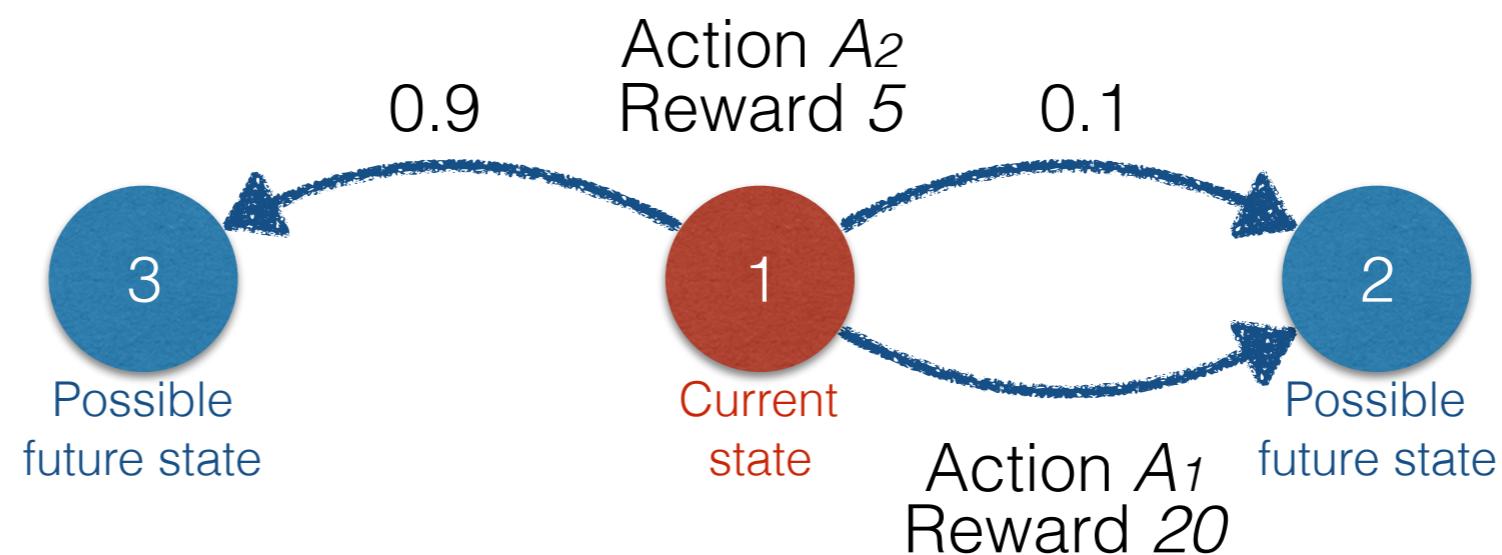
马尔可夫决策过程

Markov decision process (MDP)

MDP = Markov process + actions + reward functions
马尔可夫过程 + 决策 + 回报函数

决策2:

决策1:



MDP形式化定义

- 一个MDP可以按如下**描述**:

一种系统的描述与建模
不是一种解 (CMDP、POMDP)

$$\langle \mathbb{S}, \mathbb{A}, \mathbf{P}, U, \gamma \rangle$$

- \mathbb{S} 为所有的系统状态集合
- \mathbb{A} 为可能出现的决策 (action) 的集合
- \mathbf{P} 为状态转移概率 (state transition prob.) 矩阵
- U 为回报函数 (reward function)
- γ 为折扣/折旧因子 (discount factor)

MDP的意义

- MDP：指导决策者在随机系统中做出最优化决策的一套模型框架
- MDP在解决决策问题中的优势：同时考虑系统的“确定性”与“不确定性”
 - [确定] 系统现在的各种状态：直接观测
 - [不确定] 系统未来可能的状态：受到状态转移的影响，同时受到决策（action/decision）影响

MDP的意义

- MDP的形式化定义：

$$\langle \mathbb{S}, \mathbb{A}, \mathbf{P}, U, \gamma \rangle$$

- \mathbb{S} 当前能观测到的状态-确定性（是否存在观测的不确定性？）
- \mathbf{P} 为状态转移矩阵，包含系统的**所有**从当前状态转移到下一个未来状态的概率 - 从确定性到不确定性的桥梁（如何事先获得？）
- \mathbb{A} 可以影响系统状态的转移和当下收益reward-决策影响未来

目 录

- 定义与背景介绍
 - 马尔可夫过程 (Markov Process)
 - 马尔可夫回报过程 (Markov Reward Process, MRP)
- 马尔可夫决策过程 (Markov Decision Process, MDP)
 - MDP的定义
 - MDP求最优解
- 应用实例与扩展
 - 简单例子
 - 移动无线充电系统的应用
 - MDP的扩展

最优策略问题：MDP的最优解

- MDP的解是一个“策略” (Policy)
 - 给定任意当前的系统状态 S 时
 - 决策者采取决策 A
 - 即函数： $\phi : S \mapsto A$
 - 最优策略：能得到最大预期收益的策略 $\phi^* : S \mapsto A$

观察当前状态
做出最优决策

Bellman公式 Bellman equation

- 回忆：MRP中，一个状态下的期望回报：

$$H(S) = u(S) + \gamma \sum_{S' \in \mathbb{S}} P(S, S') H(S')$$

- MDP中，用户的决策影响状态变化与当下的回报

$$H(S, A) = u(S, A) + \gamma \sum_{S' \in \mathbb{S}} P(S, A, S') U(S')$$

其中 $U(S) = \max_{A \in \mathbb{A}} H(S, A)$

最优决策为 $\phi^*(S) = \arg \max_{A \in \mathbb{A}} H(A, S)$

Bellman
equation

求解Bellman equation

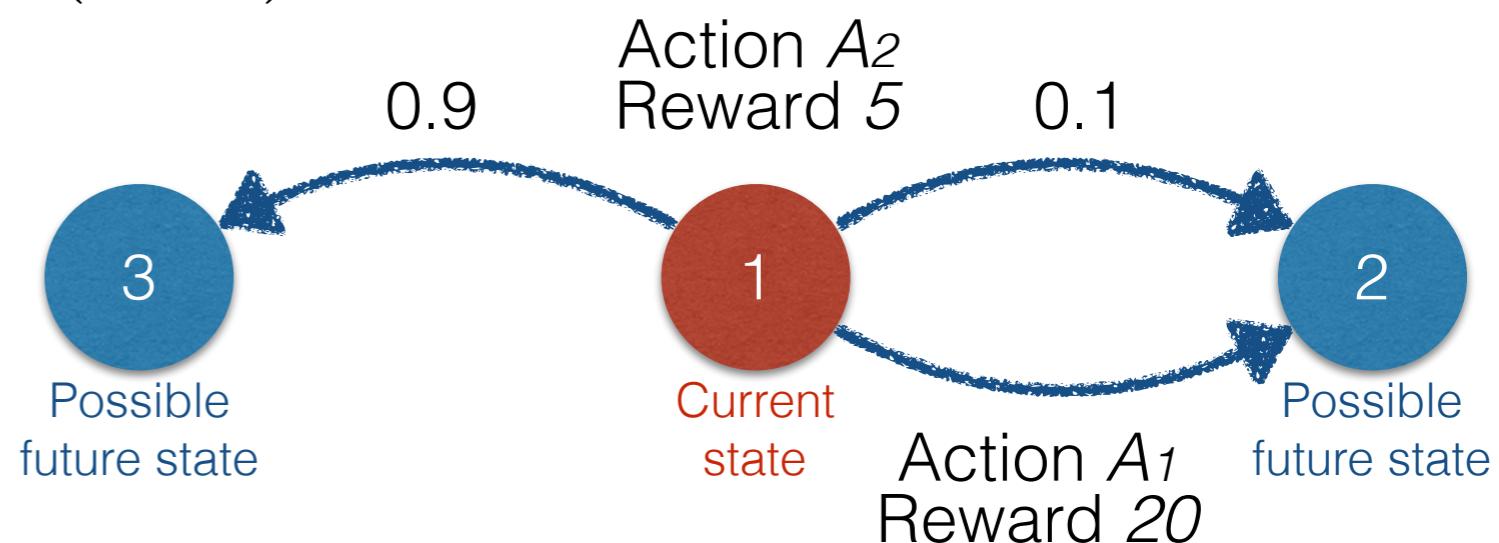
- 与MRP一样，可采用回溯法（backward induction）
 - 找到吸收态（absorbing state）或者系统终止状态
 - 从该状态向“前”回溯求解
 - 系统一直运行（如没有终止状态）则该解法不适用

求解Bellman equation

- 采用对**期望收益**的迭代算法 (**Value** iteration algorithm)
 - 先令每个状态的期望收益为0, 即 $U_0(S) \triangleq 0, \forall S \in \mathbb{S}$
 - 然后对每个状态用Bellman eqn迭代至 $U(S)$ 收敛, 即

$$H_{n+1}(S, A) = u(S, A) + \gamma \sum_{S' \in \mathbb{S}} P(S, A, S') U_n(S')$$

$$U_{n+1}(S) = \max_{A \in \mathbb{A}} H_{n+1}(S, A)$$



Value iteration algorithm

For each state S :

$$H_0(S) \triangleq 0$$

Repeat *until converge*:

For each state S :

For each action A :

$$\text{Compute } H_{n+1}(S, A) = u(S, A) + \gamma \sum_{S' \in \mathbb{S}} P(S, A, S') U_n(S')$$

$$\text{Compute and store } \phi_{n+1}^*(S) = \arg \max_A H_{n+1}(S, A)$$

$$\text{Compute and store } U_{n+1}(S) = \max_{A \in \mathbb{A}} H_{n+1}(S, A)$$

Return $\phi^*(S), U(S), \forall S \in \mathbb{S}$

求解Bellman equation

- 采用对**策略**的迭代算法 (**Policy** iteration algorithm)
 - 与Value iteration基本类似
 - 先（任意）设定一个策略 $\phi_0(S), \forall S \in \mathbb{S}$ ，带入 Bellman eqn
 - 遍历所有状态，更新策略 $\phi_{n+1}(S) : S \mapsto A, \forall S \in \mathbb{S}$
 - 直到策略收敛

求解Bellman equation

- 复杂度 $\mathcal{O}(|\mathbb{A}| \cdot |\mathbb{S}|^2)$
 - 其中 $|\mathbb{A}|$ 为可能的决策数量, $|\mathbb{S}|$ 为总的状态数量
- 收敛性-The principle of optimality
 - 即证明不动点 $f(x) = x$

This theorem provides a formal statement of “The Principle of Optimality,” a fundamental result of dynamic programming. An early verbal statement appeared in Bellman (1957, p. 83).

“An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.”

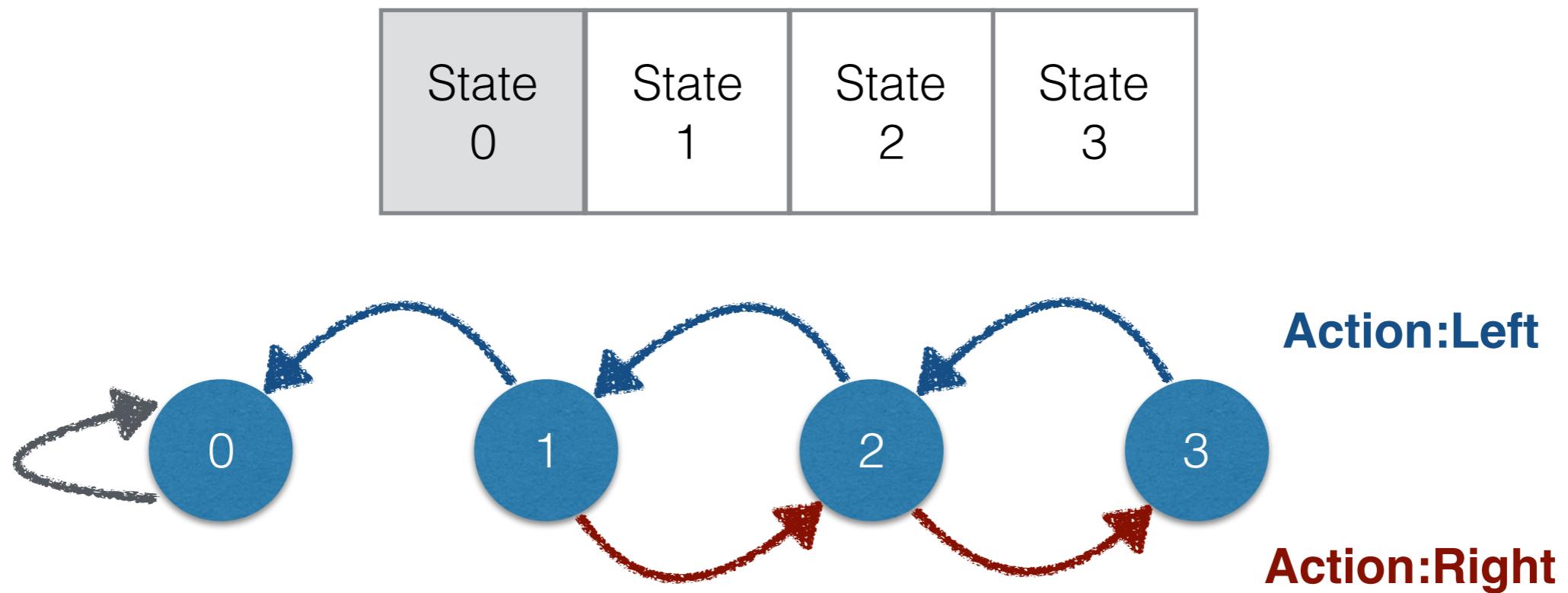
Denardo (1982, p. 15) provides a related statement.

“There exists a policy that is optimal for every state (at every stage).”

目 录

- 定义与背景介绍
 - 马尔可夫过程 (Markov Process)
 - 马尔可夫回报过程 (Markov Reward Process, MRP)
- 马尔可夫决策过程 (Markov Decision Process, MDP)
 - MDP的定义
 - MDP求最优解
- 应用实例与扩展
 - 简单例子
 - 移动无线充电系统的应用
 - MDP的扩展

MDP简单例子：位置移动



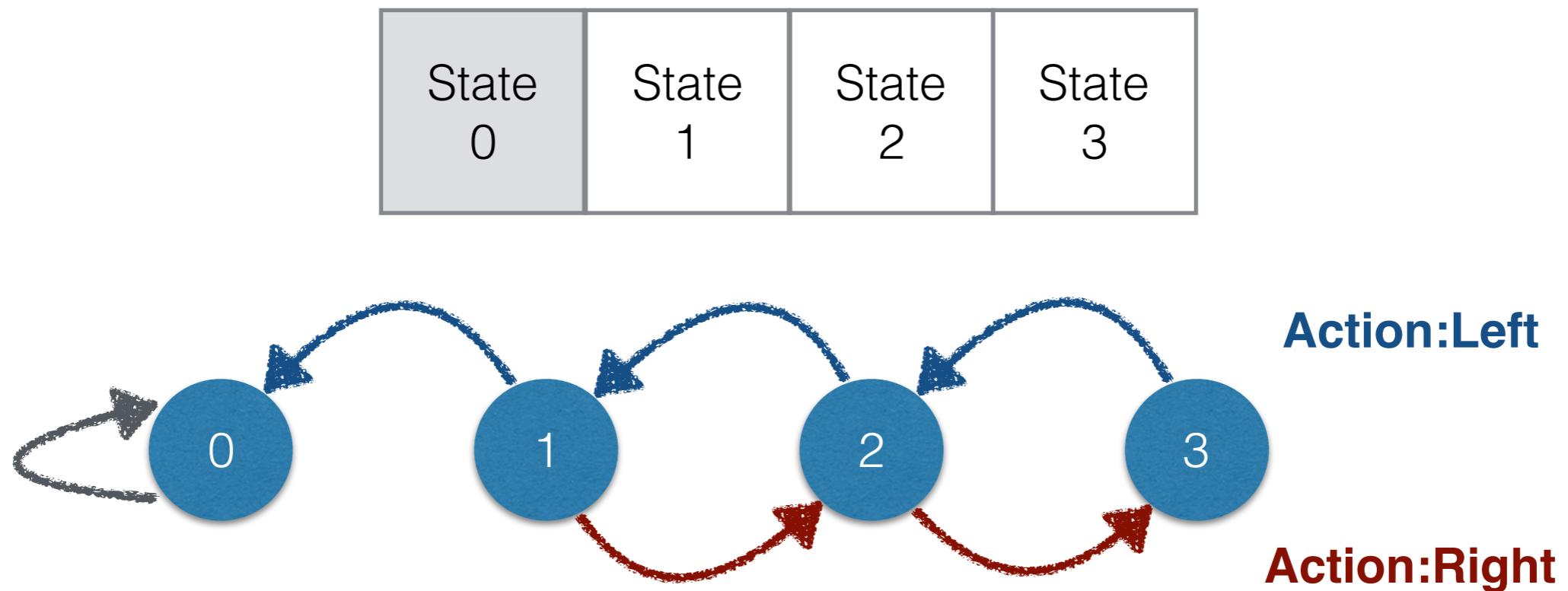
$$S = \{0, 1, 2, 3\}$$

$$A = \{\text{Left}, \text{Right}\}$$

Reward: **-1** for every step **moved**

Discount factor: 0.5

MDP简单例子：位置移动



状态转移矩阵：

$$\mathbf{P}(\mathcal{A} = \text{Left}) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$$\mathbf{P}(\mathcal{A} = \text{Right}) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Value: $H=0.0$	0.0	0.0	0.0
-------------------	-----	-----	-----

Period 1

Action:

\leftarrow / \rightarrow \leftarrow / \rightarrow \leftarrow

Value: $H=0.0$	-1.0	-1.0	-1.0
-------------------	------	------	------

Period 2

Action:

\leftarrow \leftarrow / \rightarrow \leftarrow

Value: $H=0.0$	-1.0	-1.5	-1.5
-------------------	------	------	------

Period 3

Action:

\leftarrow \leftarrow \leftarrow

...直到收敛

MDP求解中的两张表

- 由上述简单例子，可知，用对期望收益的迭代法求解MDP，即维护和更新两张表直到收敛：
 - 期望收益表：表中每个元素对应一个状态的期望收益
 - 决策行为表：表中每个元素对应当前时间的最优决策

目录

- 定义与背景介绍
 - 马尔可夫过程 (Markov Process)
 - 马尔可夫回报过程 (Markov Reward Process, MRP)
- 马尔可夫决策过程 (Markov Decision Process, MDP)
 - MDP的定义
 - MDP求最优解
- 应用实例与扩展
 - 简单例子
 - 移动无线充电系统的应用
 - MDP的扩展

移动网络中的能量管理



Copyright: Forbes

- 移动设备与应用: 传感器网络, 可穿戴部件, 移动健康管理
- 移动设备中能量: 电池/无电池 - 有限的能量供应
- 新的技术: 无线充电技术
- 基于新的技术重新思考: 移动网络中的能量管理问题

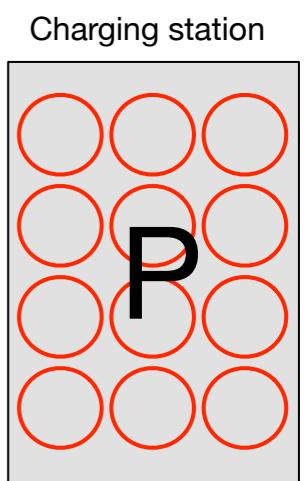
背景介绍-无线能量传输

- 无线能量传输: 通过无线连接传递能量

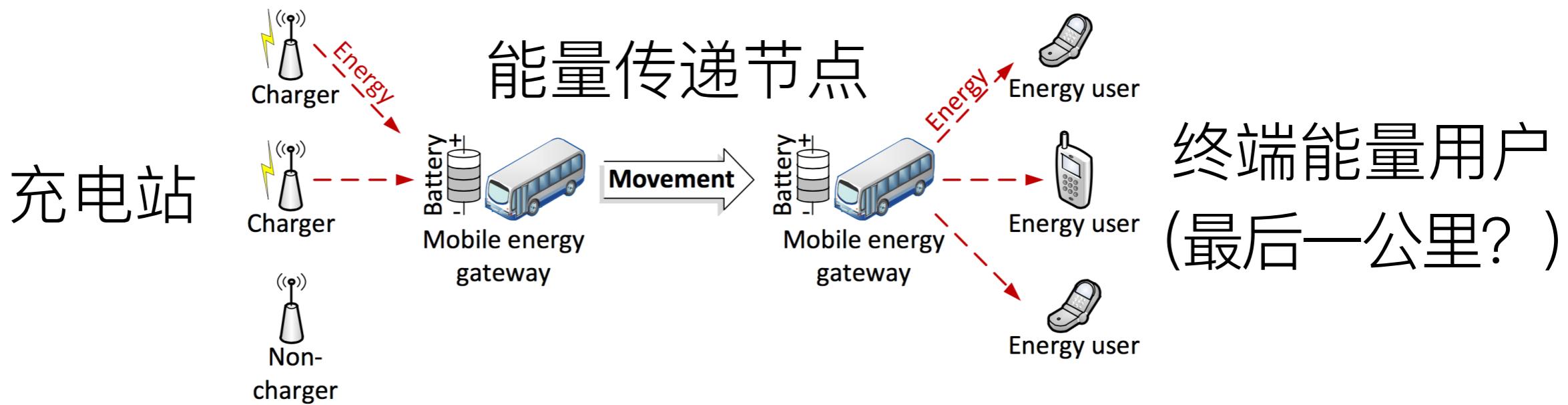
- RF energy Tx/Rx

- Friis' formula $E_R = \zeta_{RF/DC} G_t G_r \left(\frac{\lambda}{4\pi r} \right)^2 E_T,$

- Beamforming
 - 其它形式: 如非接触式电磁传输

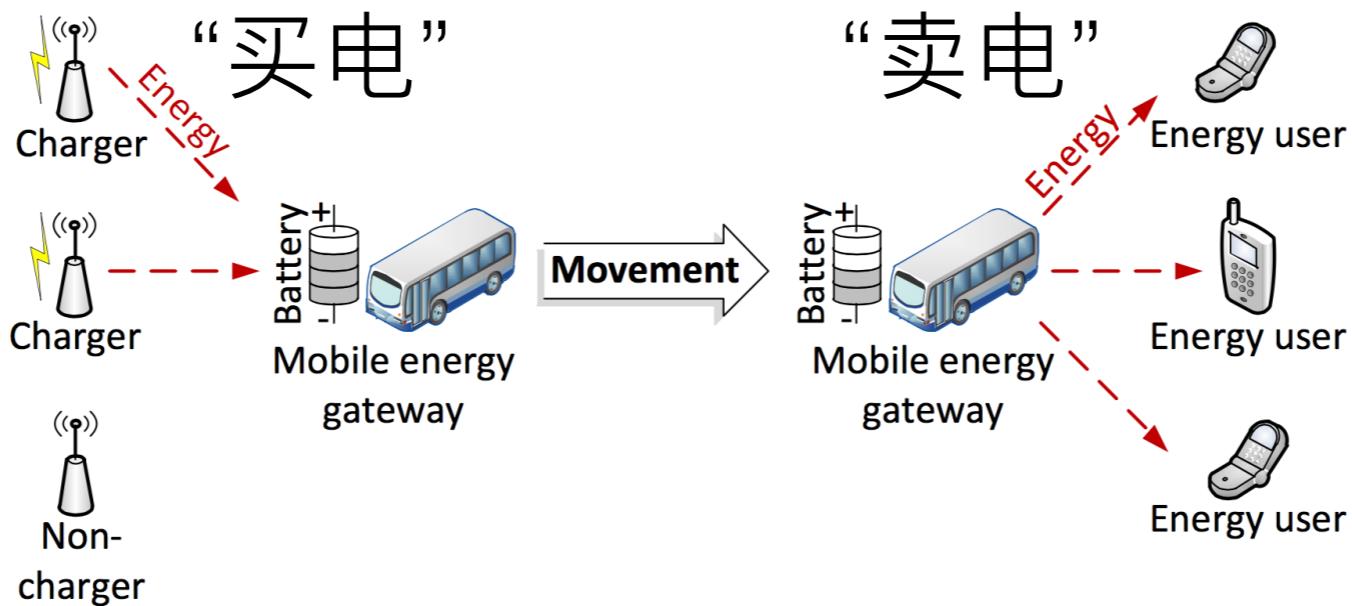


系统描述



- *Electricity chargers* 充电站: At different fixed locations, e.g., power outlets, base stations
- *End users of energy* 终端能量用户: Those who need energy, but are not covered by chargers
- *Mobile energy gateway* 能量传递节点: Moving and charge/transferring (wirelessly)

Buy/Sell energy



- Energy gateway buys from chargers (Charging)
 - Each charger asks a certain price when charging
- Energy gateway sells to end users (Transferring)
 - More users, more payments
 - Near user gets more energy, thus higher payments

能量传递节点

Mobile energy gateway

- 能量传递节点为一个移动系统用户，该用户获取能量并将能量传递给其它终端能量用户 (end user of energy)
- 扩展无线充电的充电范围 (非接触式充电如电磁线圈、RF等)
- 我们设计一个能量传递节点，并研究其**最优的能量传递策略**-什么时候传输最合算

基于MDP的系统描述建模

- 状态空间

$$\mathbb{S} = \{\mathcal{S} = (\mathcal{L}, \mathcal{E}, \mathcal{N}, \mathcal{P})\}$$

- \mathcal{L} : 位置状态; \mathcal{E} 能量状态; \mathcal{P} 充电的价格
- \mathcal{N} : 能量传输节点当前所遇到的终端用户数量

decides end user payment

- 决策空间 $\mathbb{A} = \{\mathcal{A} = 0, 1, 2\}$
- 三种决策: 空闲, 充电, 能量传输 (放电)

$\langle \mathbb{S}, \mathbb{A}, \mathbf{P}, U, \gamma \rangle$

MDP: 确定回报函数

(举例：从终端用户获得的回报)

- 终端节点到能量传输节点的距离的概率密度分布函数（距离远则接收的能量少，则给传输节点的回报少）

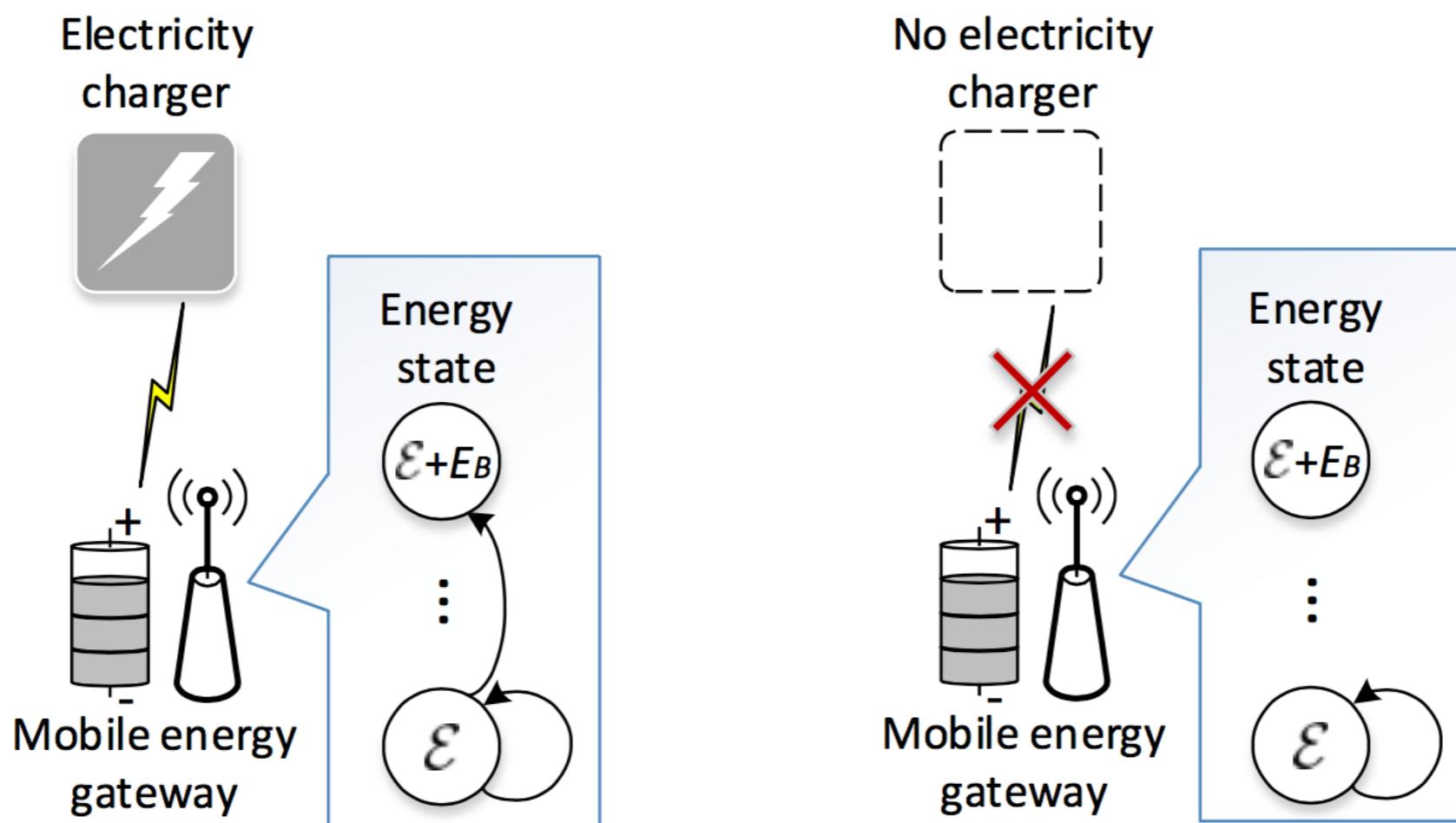
$$f(n, l|N) = \frac{3}{R} \frac{B(n + \frac{2}{3}, N - n + 1)}{B(N - n + 1, n)} \beta\left(\frac{l^3}{R^3}; n + \frac{2}{3}, N - n + 1\right)$$

- 第 n^{th} 个终端节点能给能量传输节点的期望回报

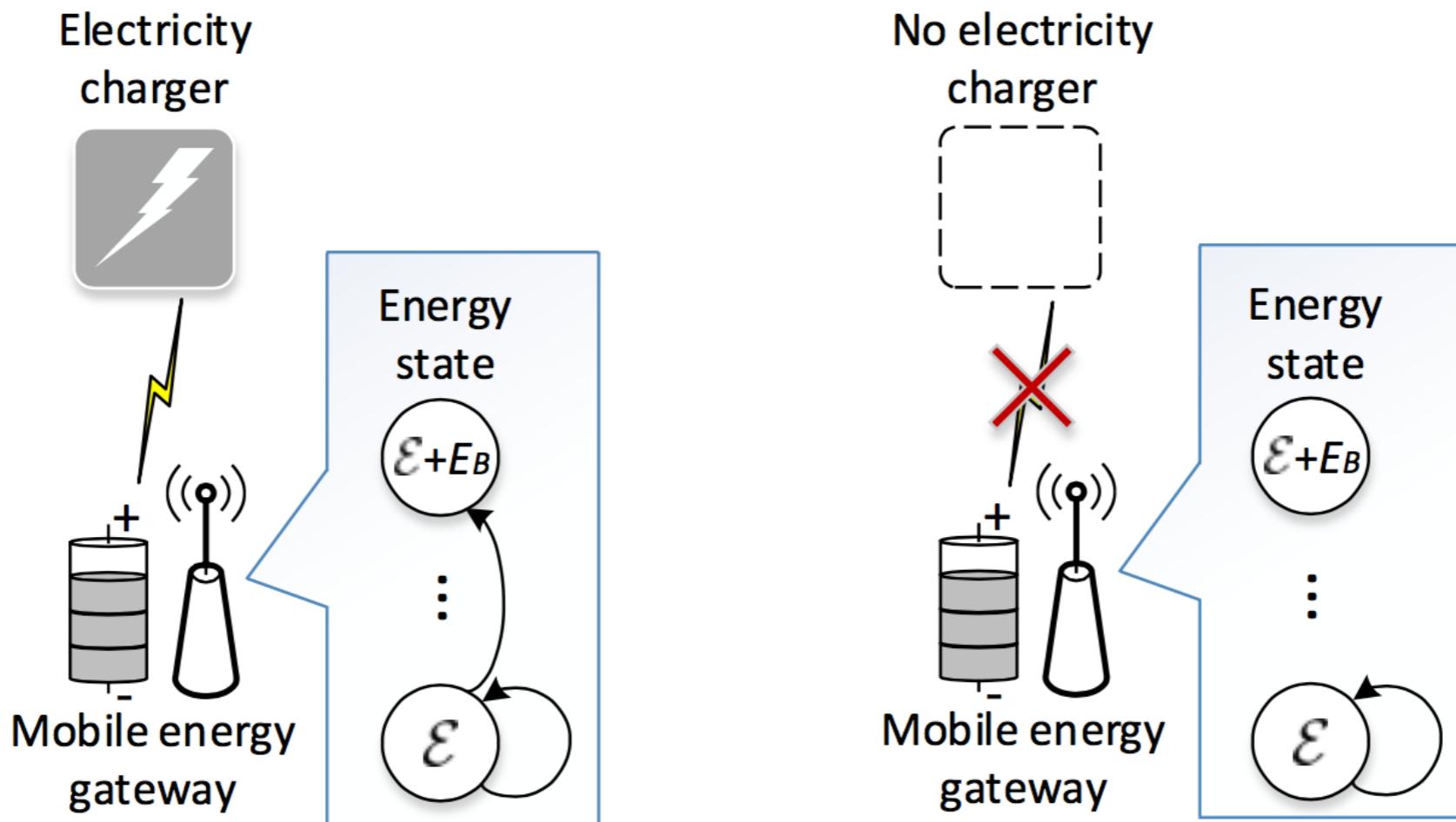
$$R(n, E_S) = \int_0^{R^\circ} f(n, l|N) r(e_n^D) dl + \int_{R^\circ}^R f(n, l|N) r(g \frac{E_S}{l^2}) dl$$

sum up to get overall payment

MDP: 系统状态转移 (例)



MDP: 系统状态转移 (例)



$$\mathbf{P}(\mathcal{A} = 1) = \begin{bmatrix} \dots & \vdots & \dots & \vdots & \dots \\ \dots & 0.3 & \dots & 0.7 & \dots \\ \vdots & & \vdots & & \end{bmatrix}$$

$$\mathbf{P}(\mathcal{A} = 0) = \begin{bmatrix} \dots & \vdots & \dots & \vdots & \dots \\ \dots & 1.0 & \dots & 0.0 & \dots \\ \vdots & & \vdots & & \end{bmatrix}$$

求解工具

- 针对Bellman equation的value iteration algorithm迭代求解
 - pymdptoolbox：输入MDP的各个参数如状态转移矩阵、回报函数等等，可提供多种解法
 - mdptoolbox，用于Matlab，同上
 - 自己构建

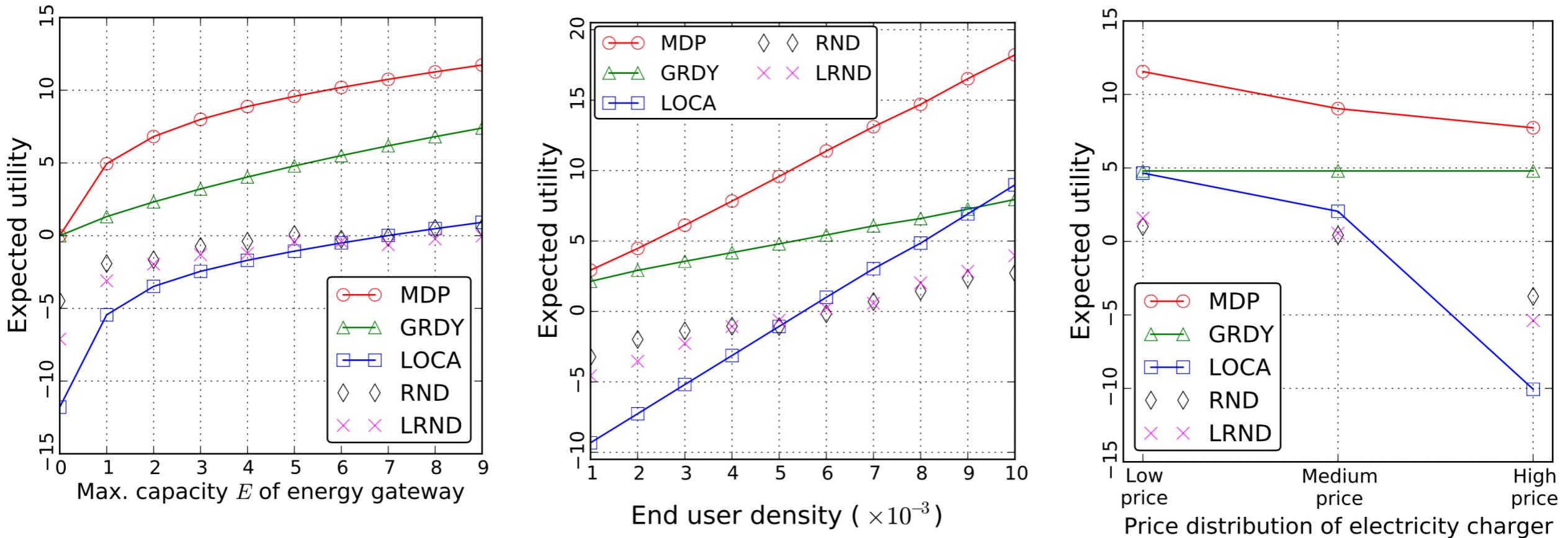
MDP优化决策和其它决策方式比较

- 对比：基于MDP的决策 / 简单决策
 - Greedy scheme [GRDY, 贪心策略]: maximizing immediate utility
 - Random scheme [RND, 随机策略]: randomly taking any action (i.e., 0,1,2) from the action set
 - Location-aware scheme [LOCA]: charging at charger, transferring at end users
 - Location-aware random scheme [LRND]: randomly taking 0 and 1 at the charger side; randomly taking 0 and 2 at the end user side

MDP优化决策和其它决策方式比较

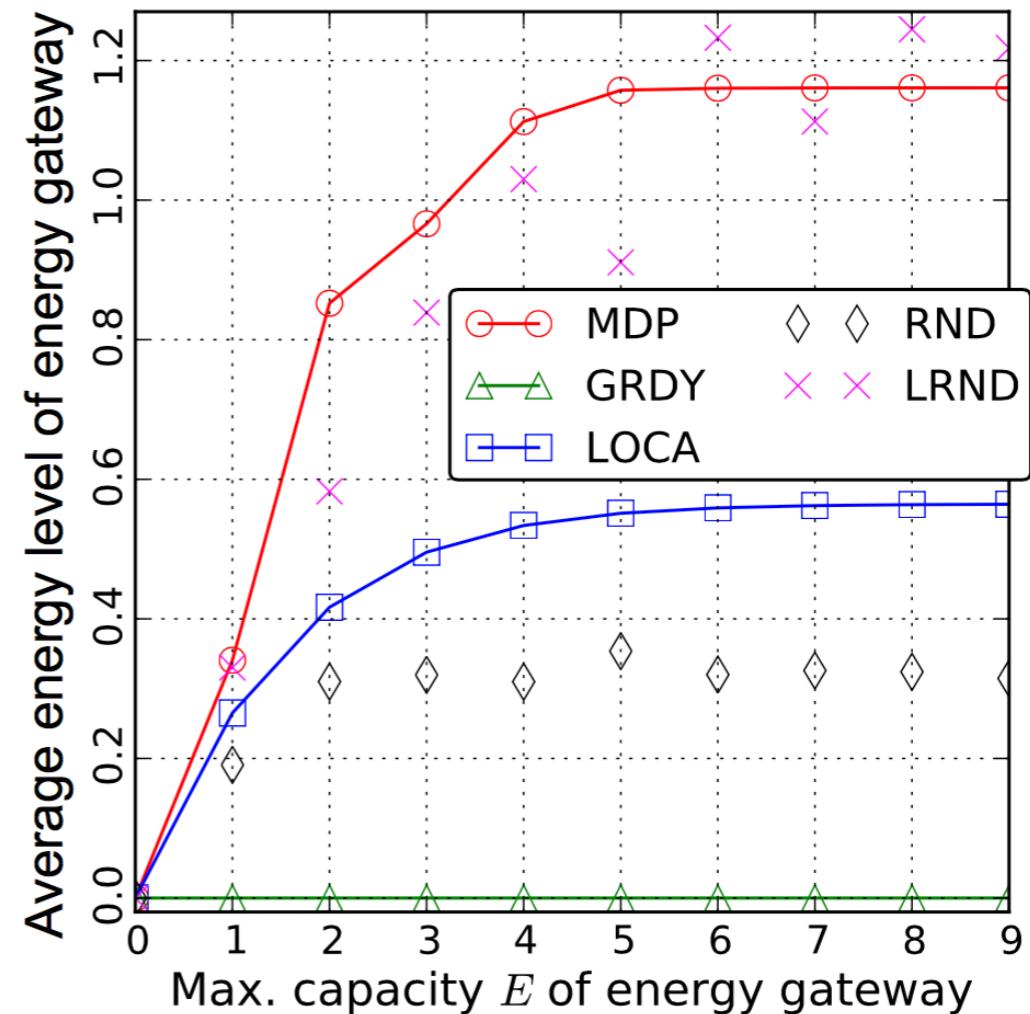
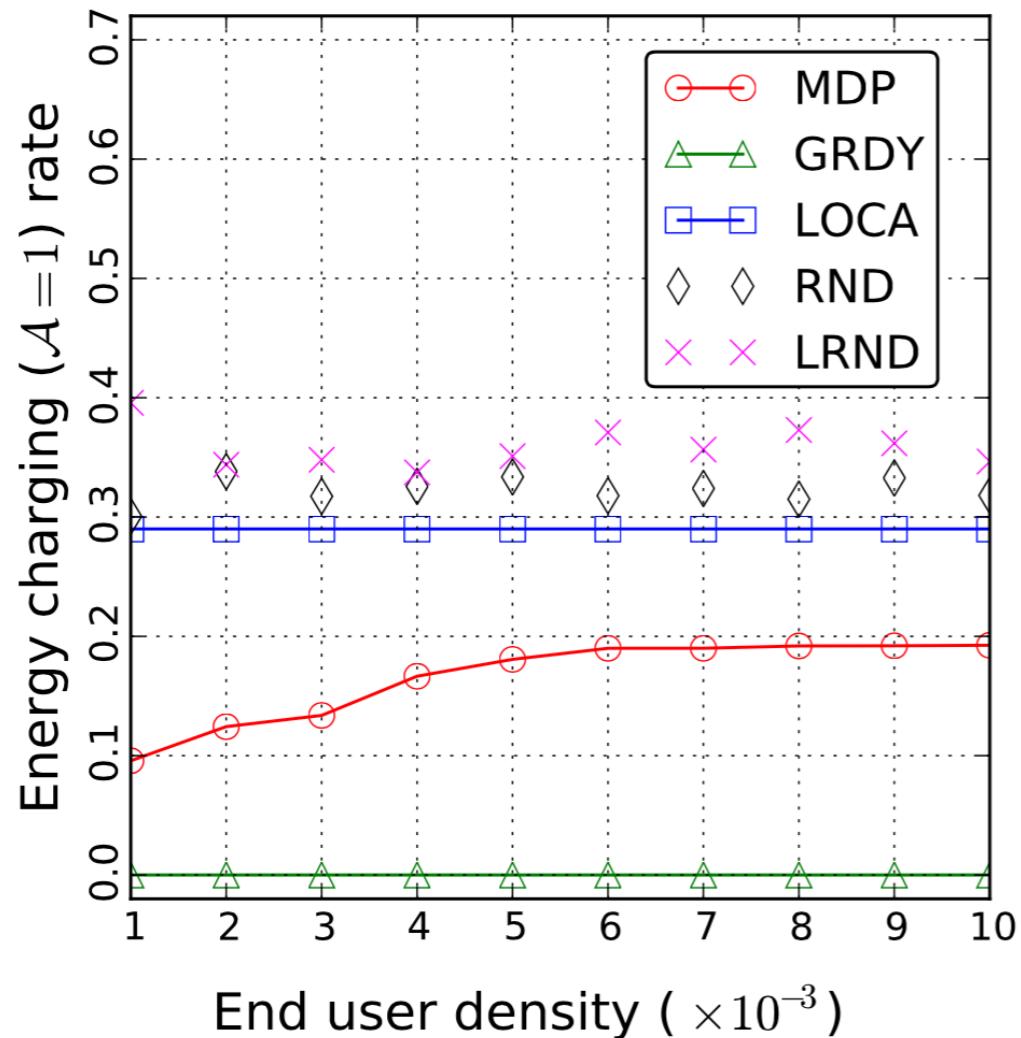
- Greedy scheme [GRDY, 贪心策略]: 只最大化当前状态下的收益
- MDP优化: 本质也是一种贪心策略, 针对于期望收益函数 (Q函数) 的贪心策略

Numerical results



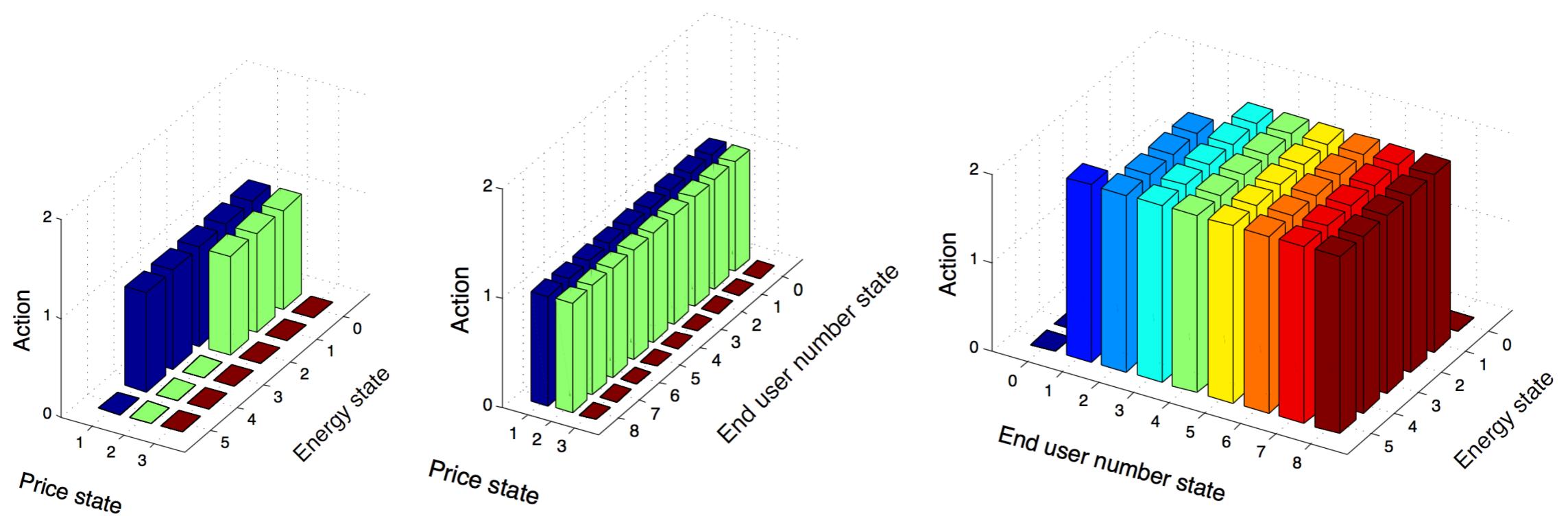
- 在多种系统场景下，基于MDP的决策可以得到较优的预期效用函数（较优的系统性能）

Numerical results



- MDP及其最优解可用于设计系统参数

MDP的解：每个状态



目 录

- 定义与背景介绍
 - 马尔可夫过程 (Markov Process)
 - 马尔可夫回报过程 (Markov Reward Process, MRP)
- 马尔可夫决策过程 (Markov Decision Process, MDP)
 - MDP的定义
 - MDP求最优解
- 应用实例与扩展
 - 简单例子
 - 移动无线充电系统的应用
 - MDP的扩展

MDP建模的一些思考与讨论

- 马尔可夫性质 (Markovian property)
 - 状态转移的马尔可夫性 - 与过去无关
 - 行为决策的马尔可夫性 - 与过去无关
 - 回报函数 (immediate reward) 的马尔可夫性

MDP建模的一些思考与讨论

- 状态之间转移的概率如何得到?
 - 理论推导与模拟仿真获取
 - 如：节点在空间中按Poisson process分布
 - 通过对已有数据的统计学习
 - 如：计算频数
 - “维数灾难” (Curse of dimensionality)
 - 系统状态过多, [Peter Marbach and John N. Tsitsiklis 2001]

MDP的扩展

- 带约束的MDP模型 (Constrained MDP, CMDP)
- 部分可观测的MDP模型 (Partially Observable MDP, POMDP)
- 最优停时问题 (Optimal stopping)
- MDP解的结构问题

参考资料

- Books
 - Martin L. Puterman, “**Markov Decision Processes: Discrete Stochastic Dynamic Programming**”
- Paper
 - M. A. Alsheikh, D. T. Hoang, D. Niyato, S. Lin, and H.-P. Tan, “**Markov Decision Processes with applications in wireless sensor networks: A survey**,” in *IEEE Communications Surveys and Tutorials*, vol. 17, no. 3, 2015. (also in arXiv:1501.00644v1)
 - P. Marbach and J. N. Tsitsiklis, “Simulation-Based Optimization of Markov Reward Processes”, in *IEEE Transactions on Automatic Control*, vol. 46, no. 2, pp. 191-209, 2001.
- Course slides
 - <http://www.cs.rice.edu/~vardi/dag01/givan1.pdf>
 - http://www0.cs.ucl.ac.uk/staff/d.silver/web/Teaching_files/MDP.pdf
 - http://castlelab.princeton.edu/ORF569papers/Powell_ADP_2ndEdition_Chapter%203.pdf
 - http://isites.harvard.edu/fs/docs/icb.topic540049.files/cs181_lec03_handout.pdf