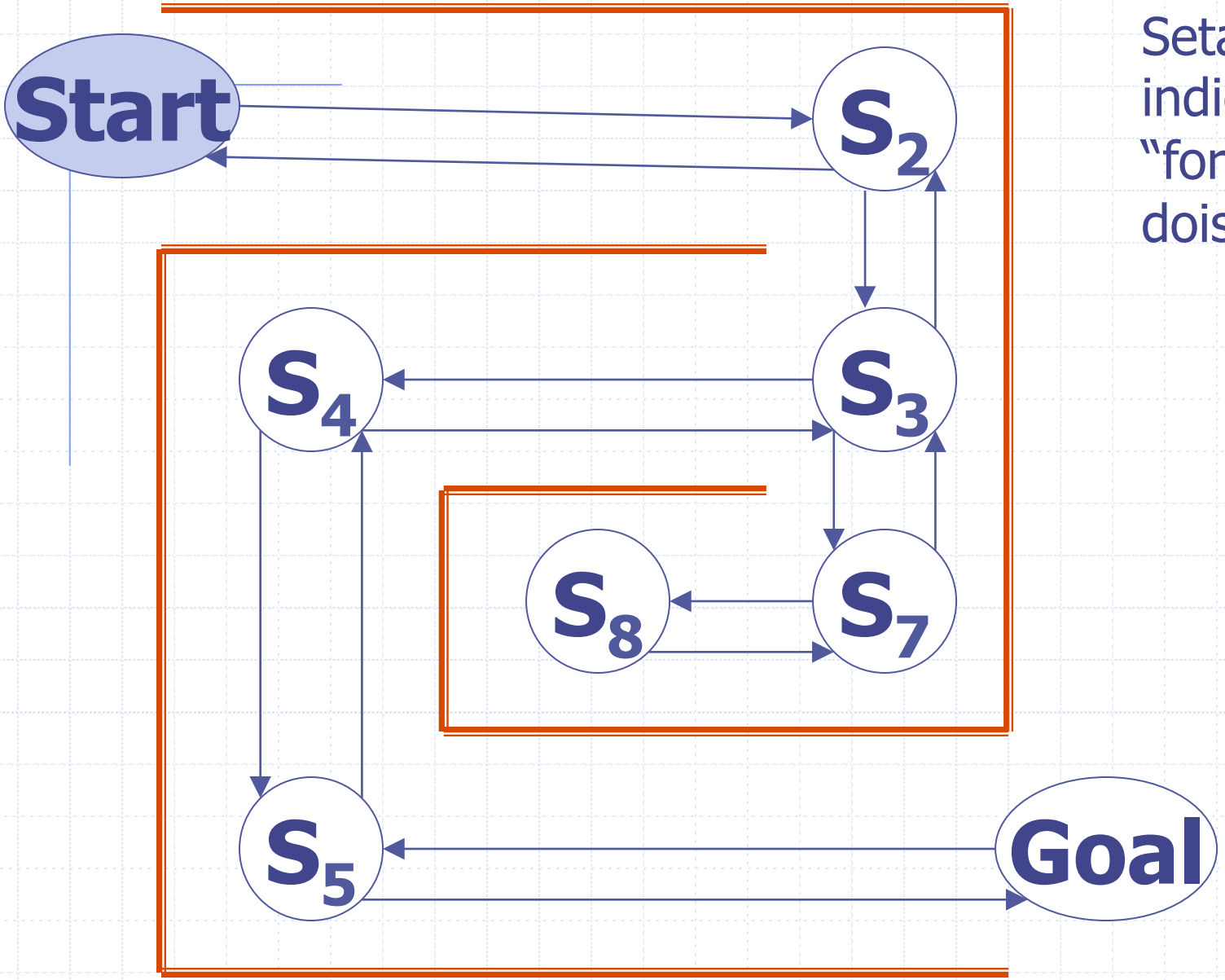
Aprendizagem por Reforço

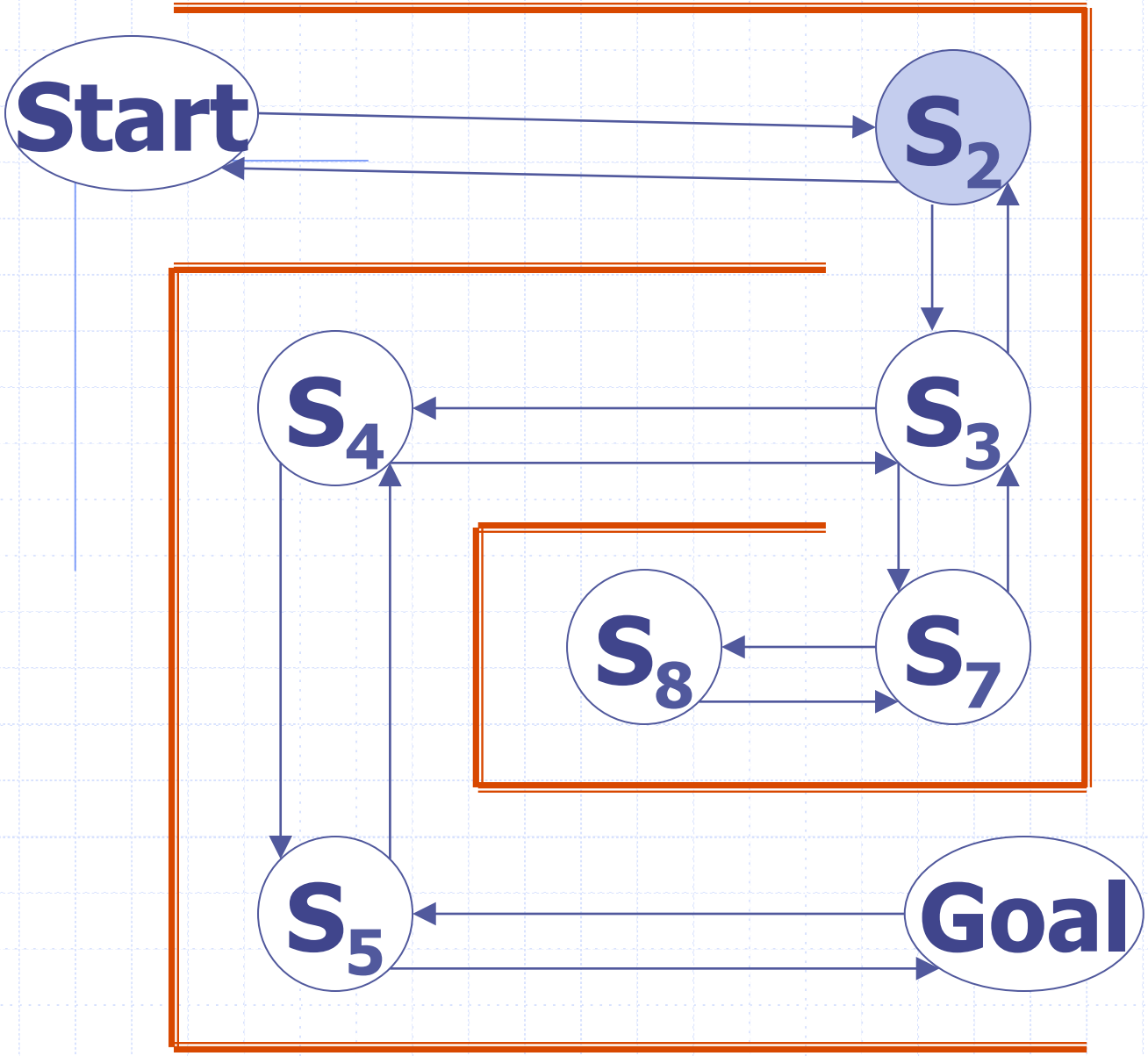
Motivação

- Como um agente aprende a escolher ações apenas interagindo com o ambiente?
  - Muitas vezes, é impraticável o uso de aprendizagem supervisionada
    - Como obter exemplos do comportamento correto e representativo para qualquer situação?
    - E se o agente for atuar em um ambiente desconhecido?
  - Exemplos:
    - Criança adquirindo coordenação motora
    - Robô interagindo com um ambiente para atingir objetivo(s)

Exemplo de aprendizado por reforço



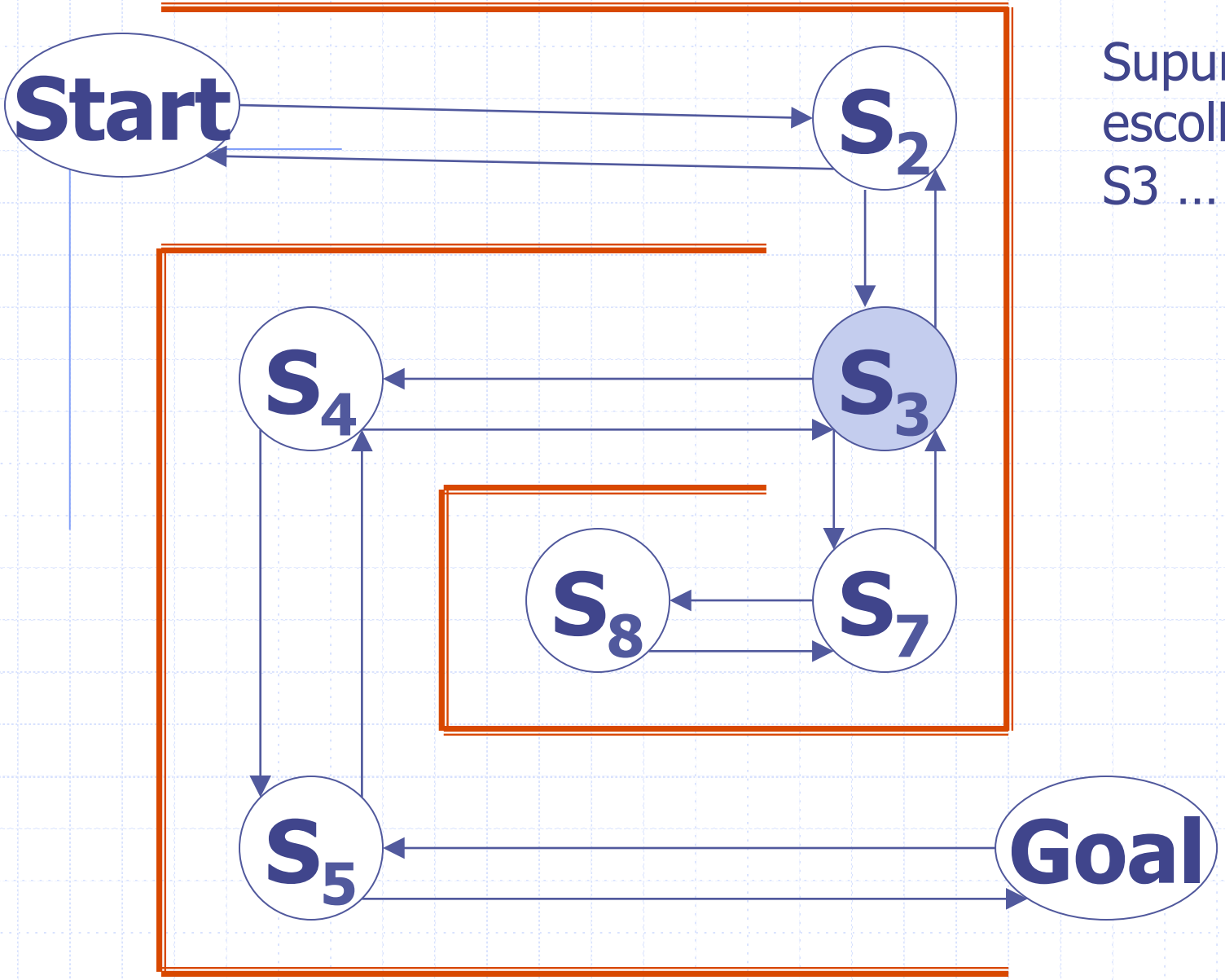
Setas indicam a "força" entre dois estados



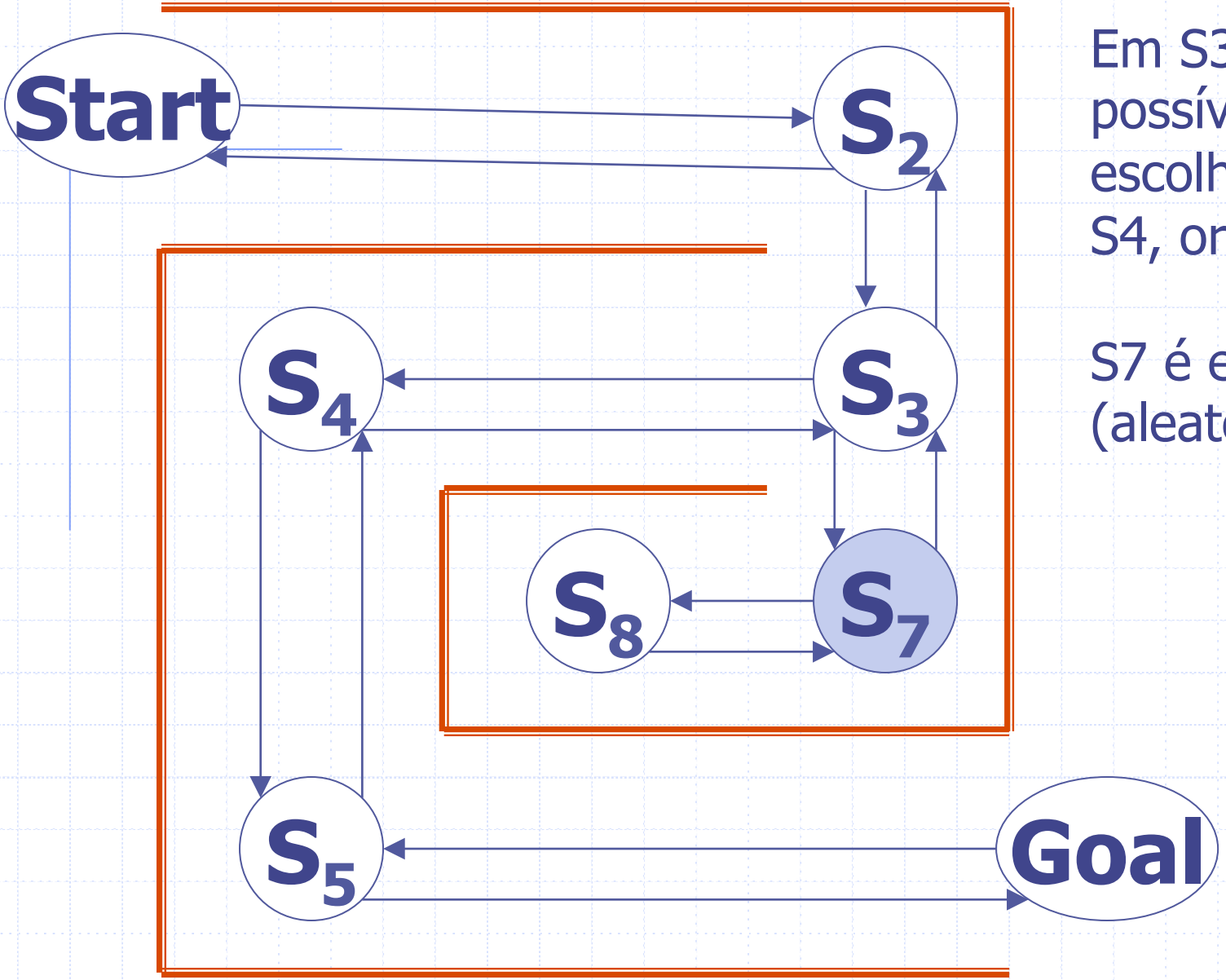
A primeira ação leva para S2 ...

Próximo estado é escolhido aleatóriamente de um dos possíveis estados, ponderado pela força da associação

Associação = largura da linha

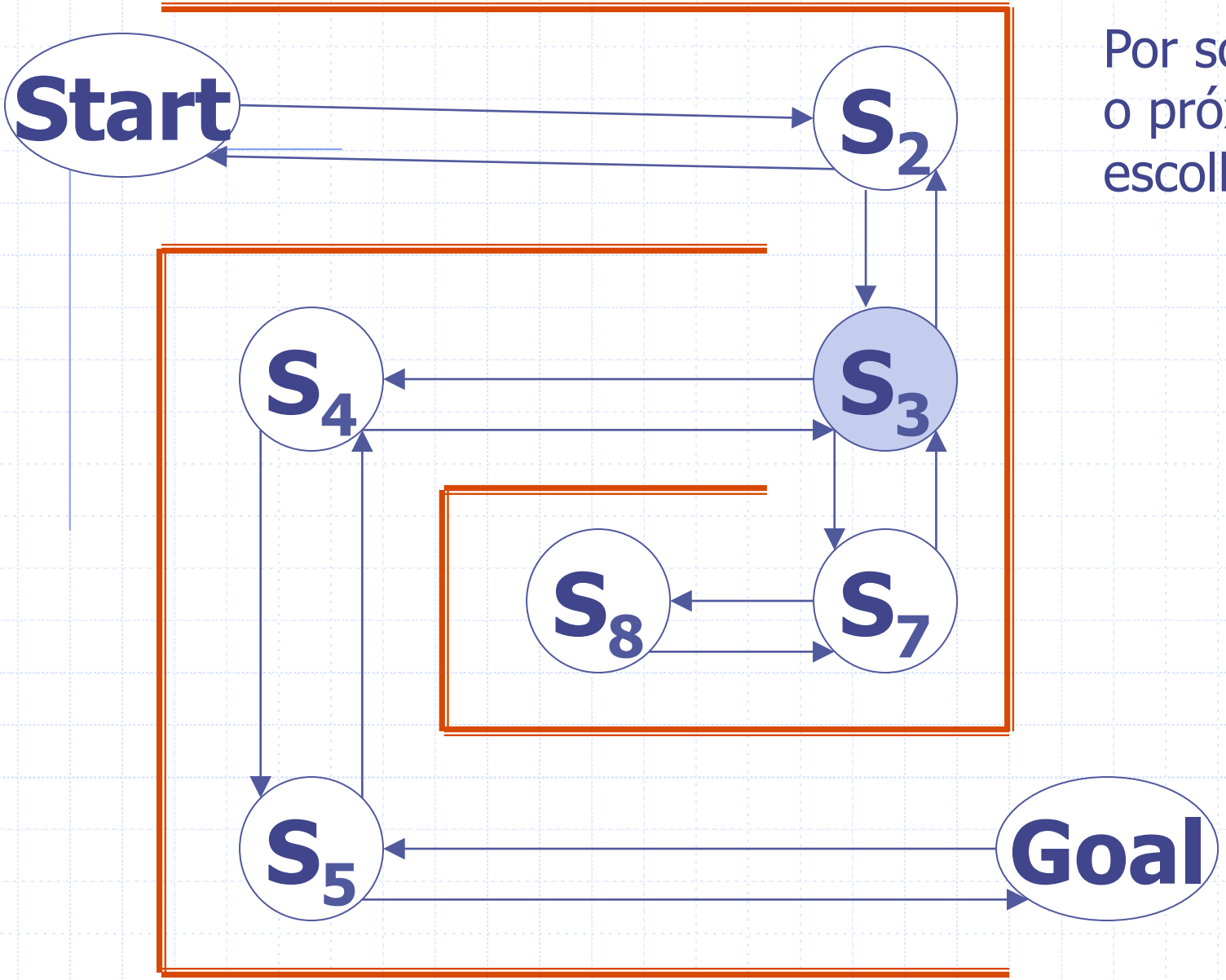


Supunha que a escolha leve a S3 ...



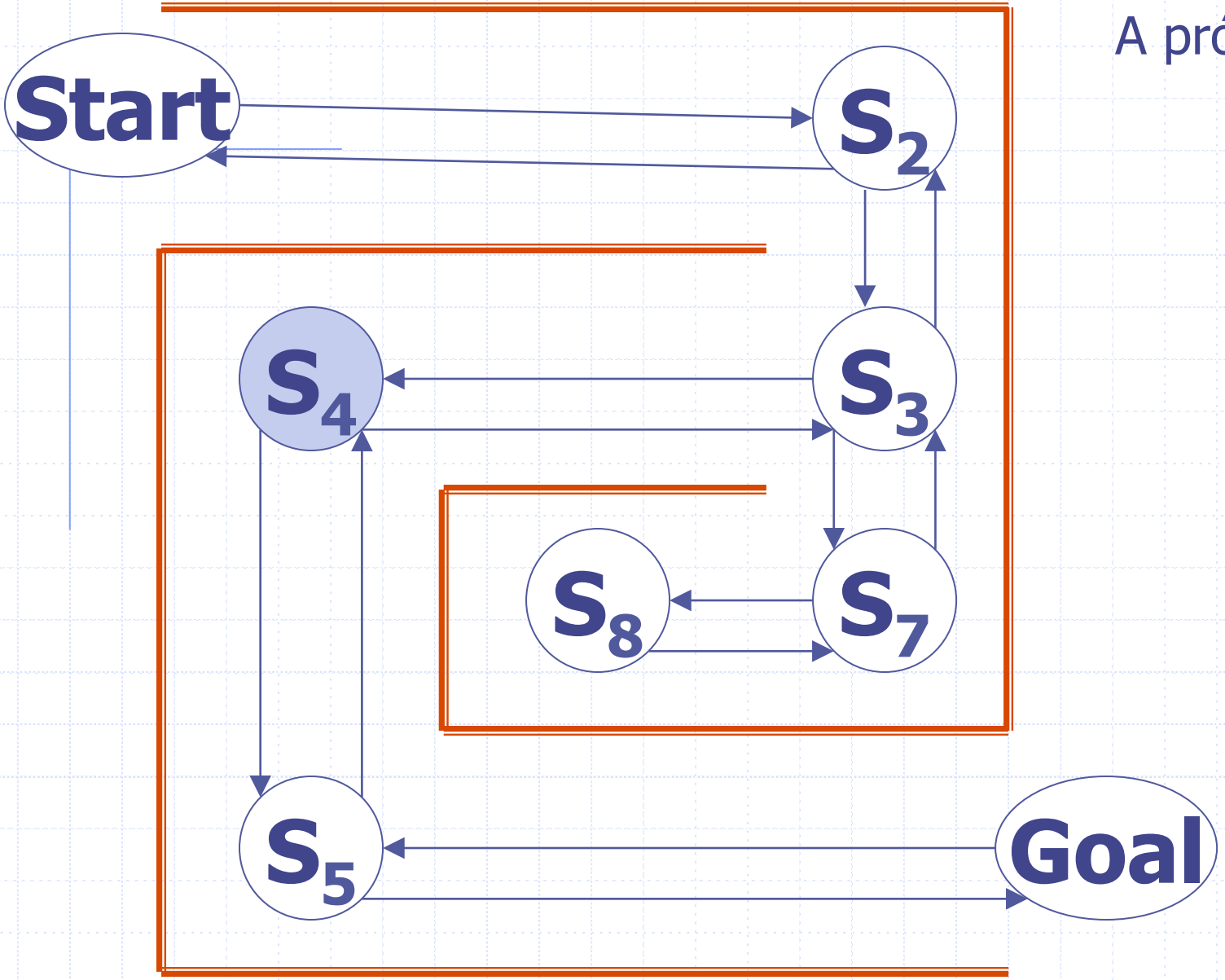
Em S3, as possíveis escolhas são S2, S4, or S7.

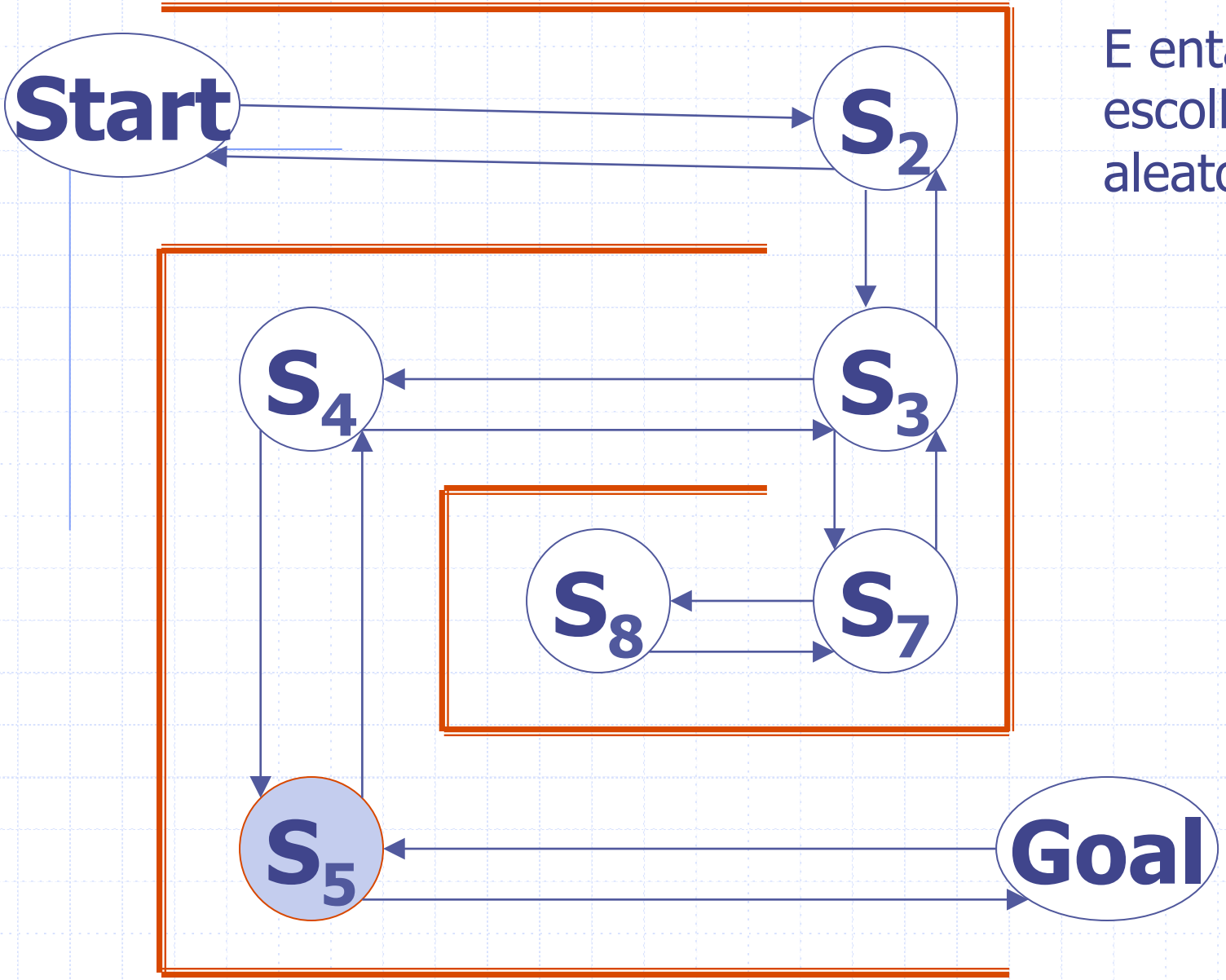
S7 é escolhido (aleatóriamente)



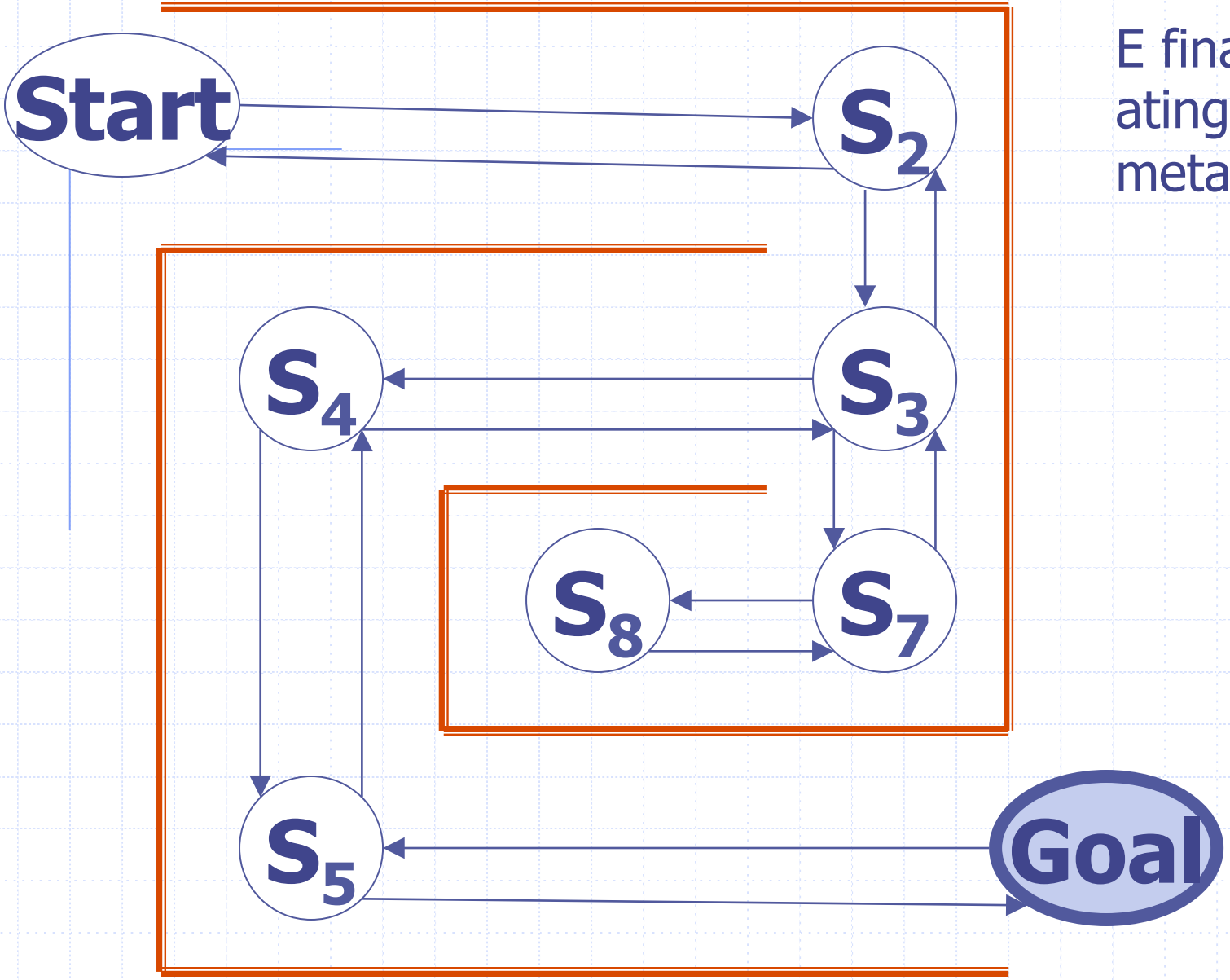
Por sorteio, S3 é o próximo escolhido...

A próxima é S4

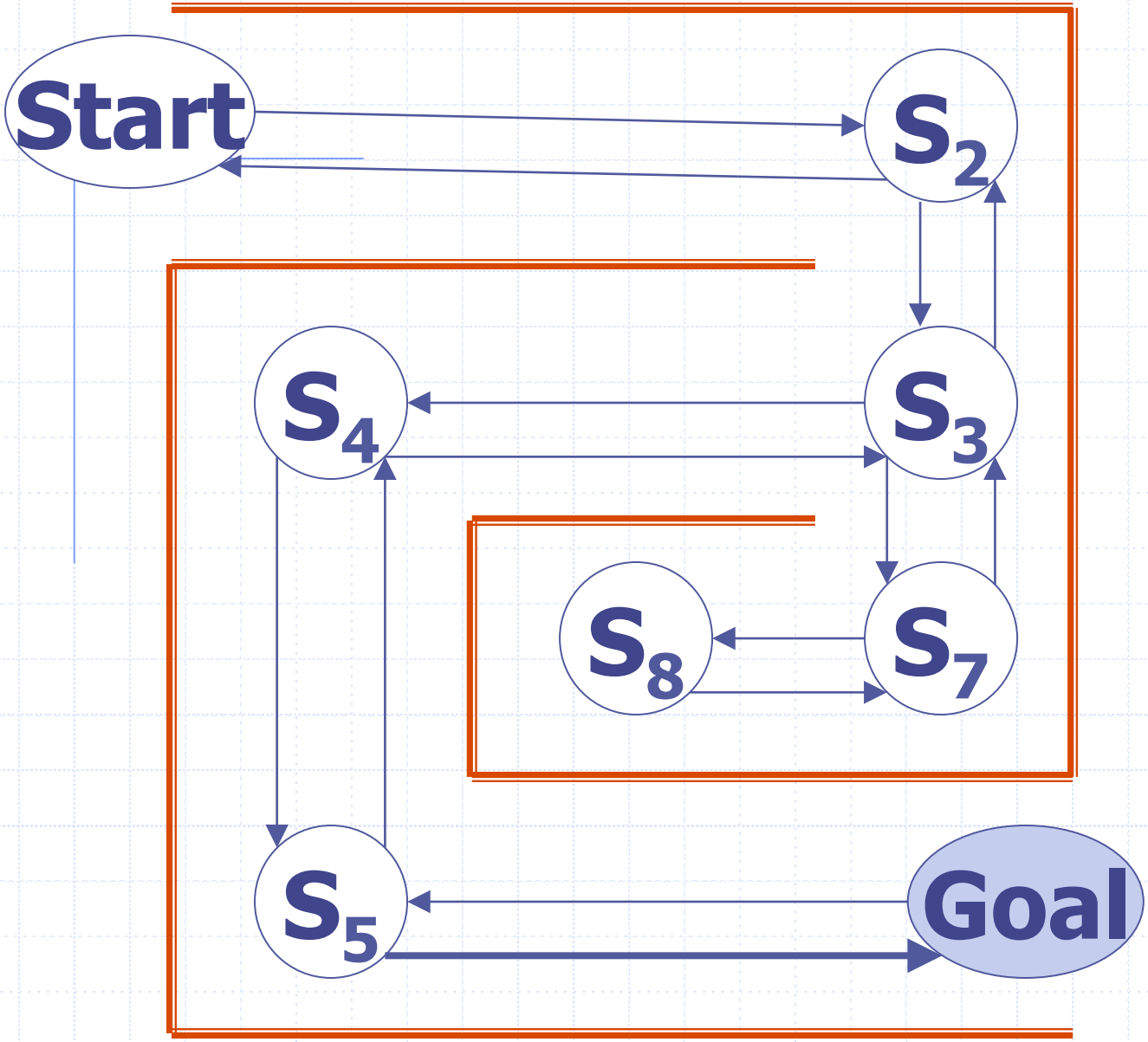




E então S5 é escolhido aleatóriamente

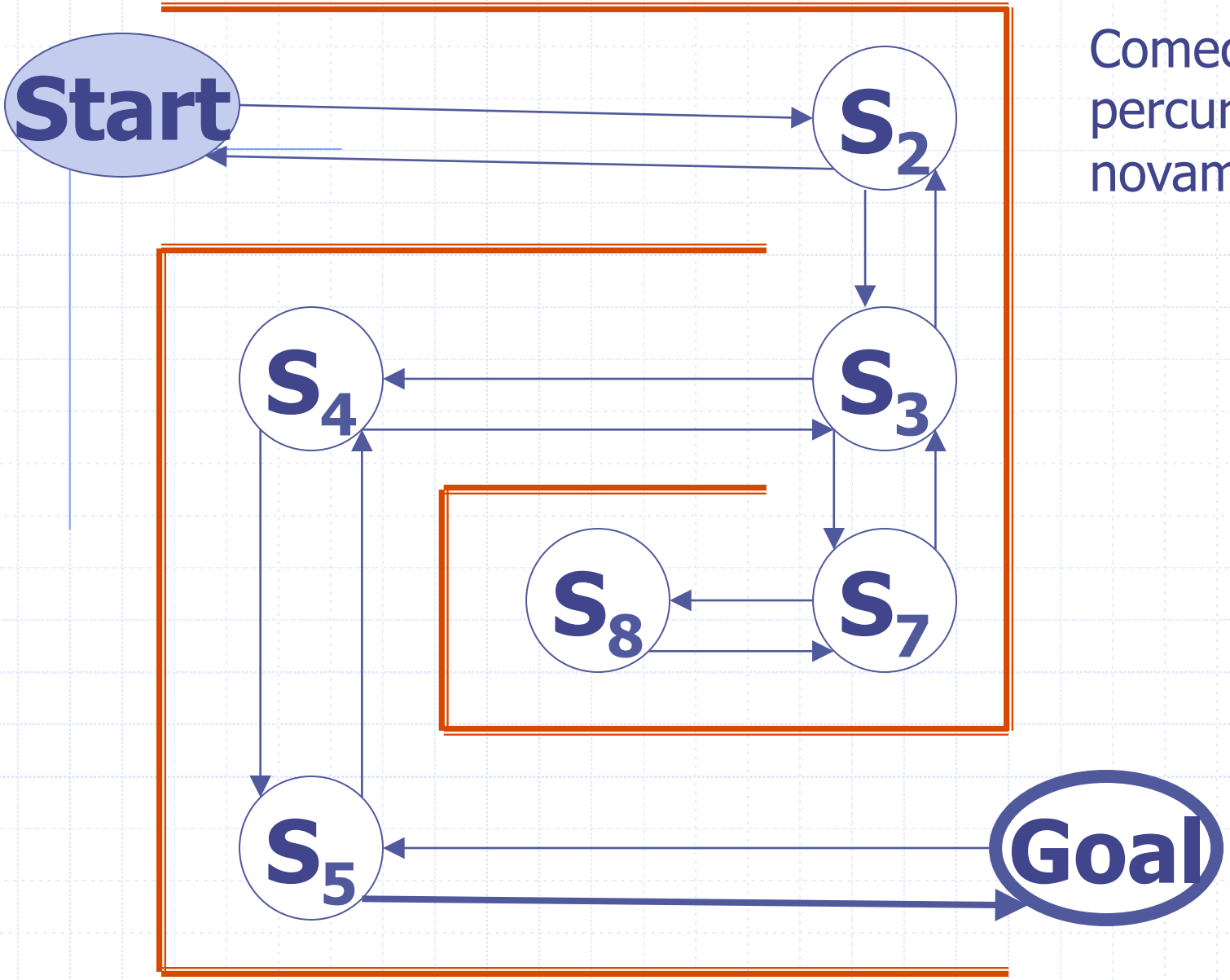


E finalmente atingimos a meta ...

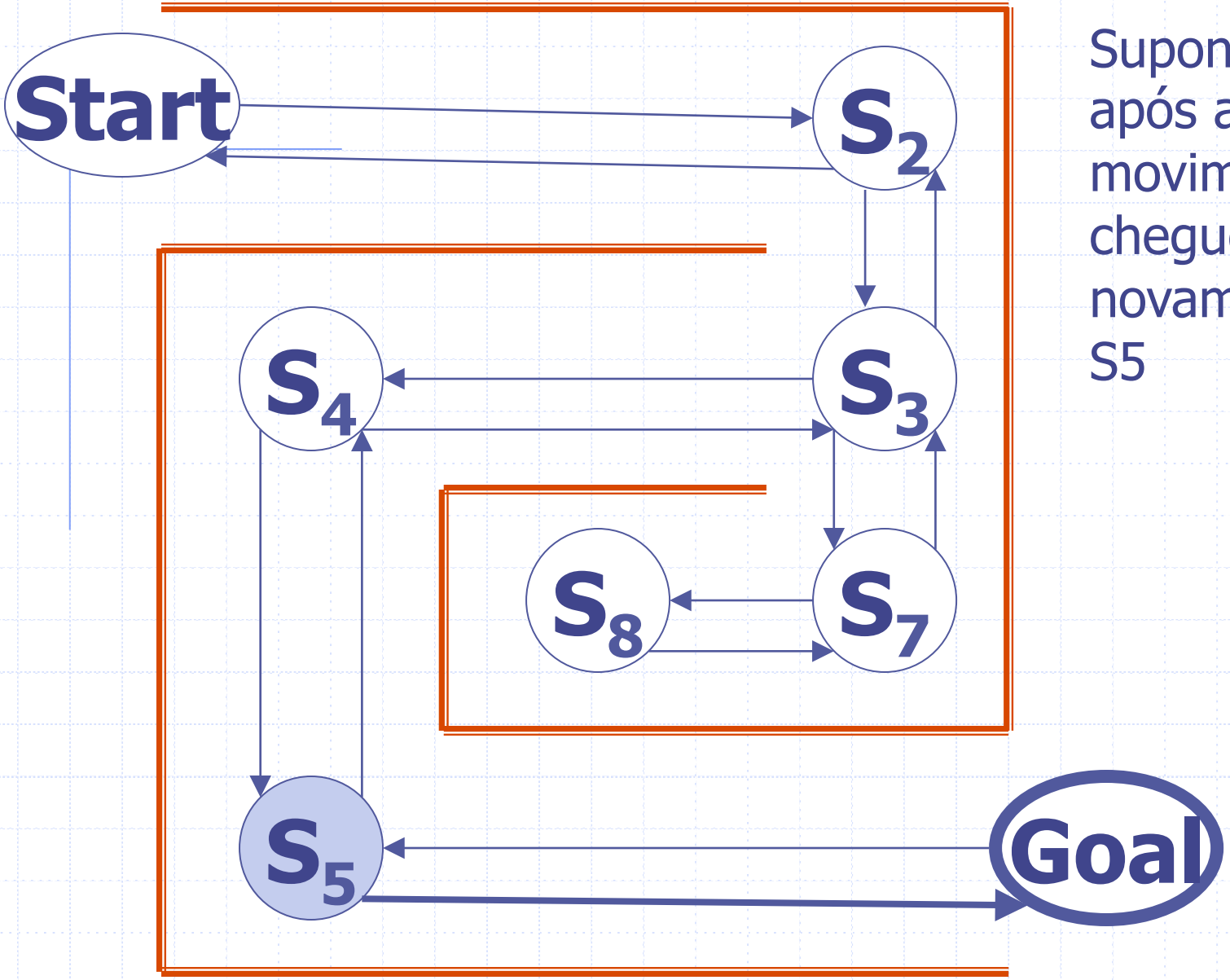


Quando a meta é atingida, reforce a conexão entre ele e o estado que levou a ele

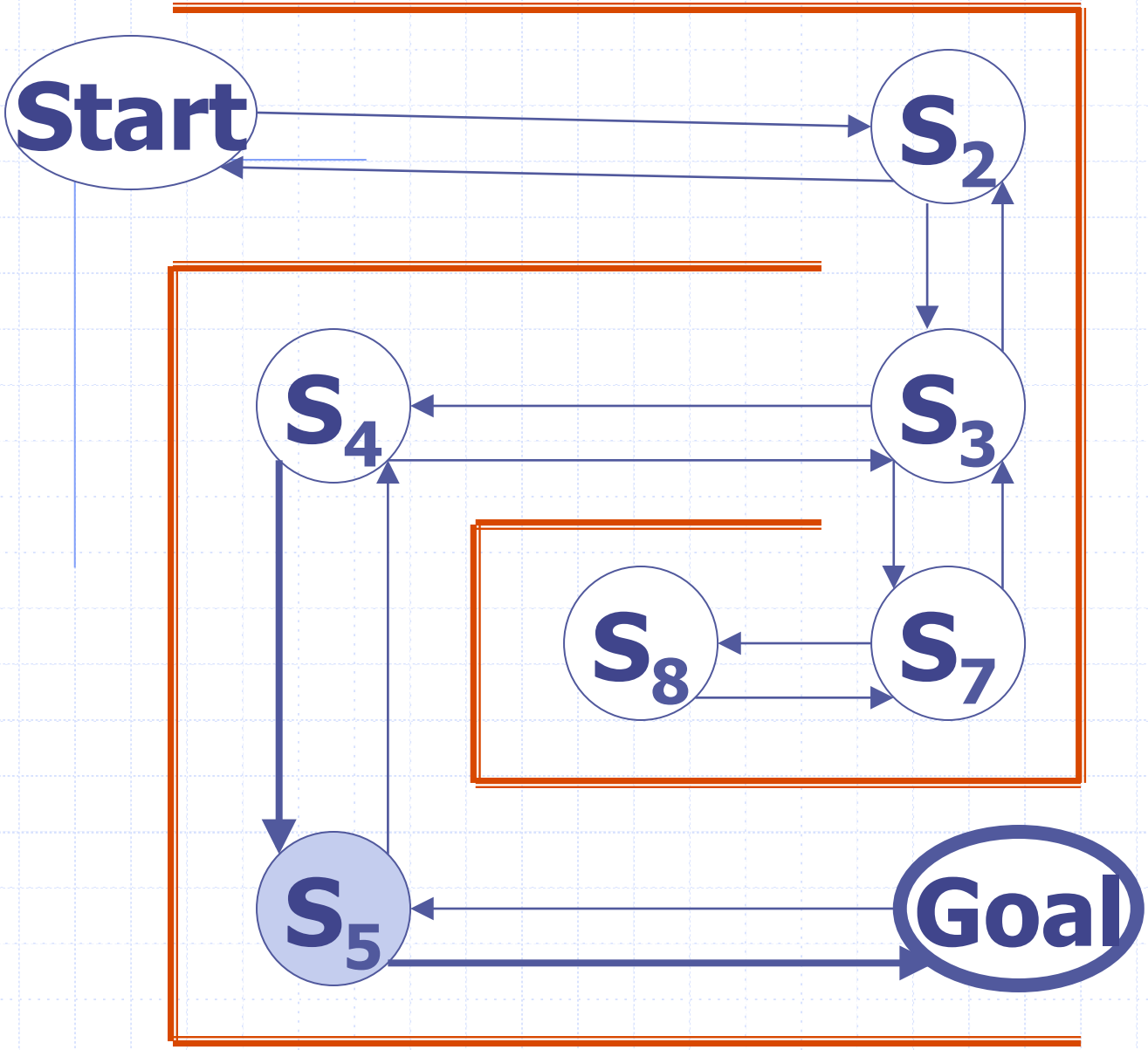
Na próxima vez que S5 for alcançado, parte da força de associação será passada para S4...



Comece o percurso novamente...



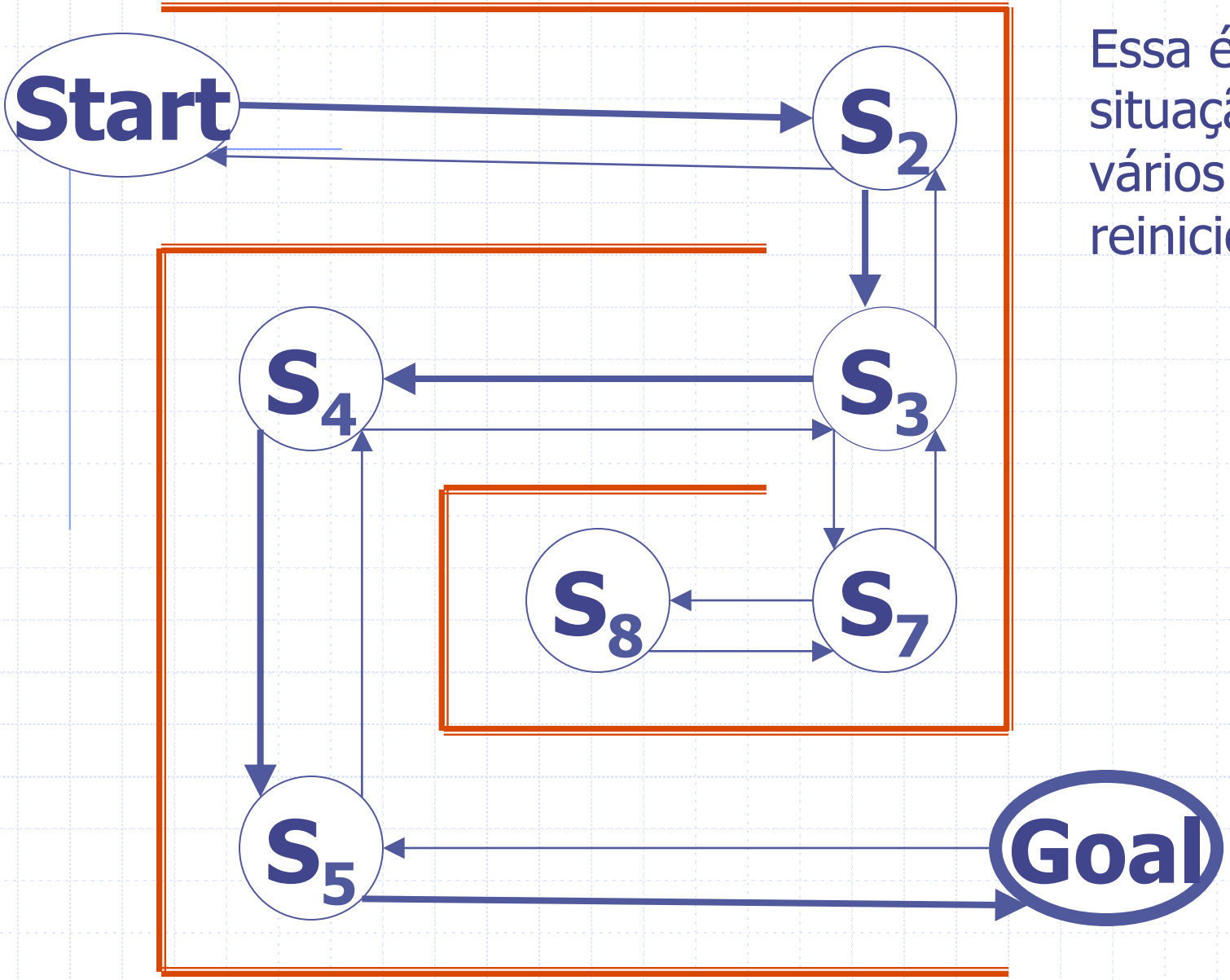
Suponha que após alguns movimentos, cheguemos novamente em S5



S5 tem grande chance de atingir a meta pela rota com mais força

Em aprendizado por reforço, a "força" é passada de volta para o estado anterior

Esse processo leva a criar um caminho entre o início e a meta

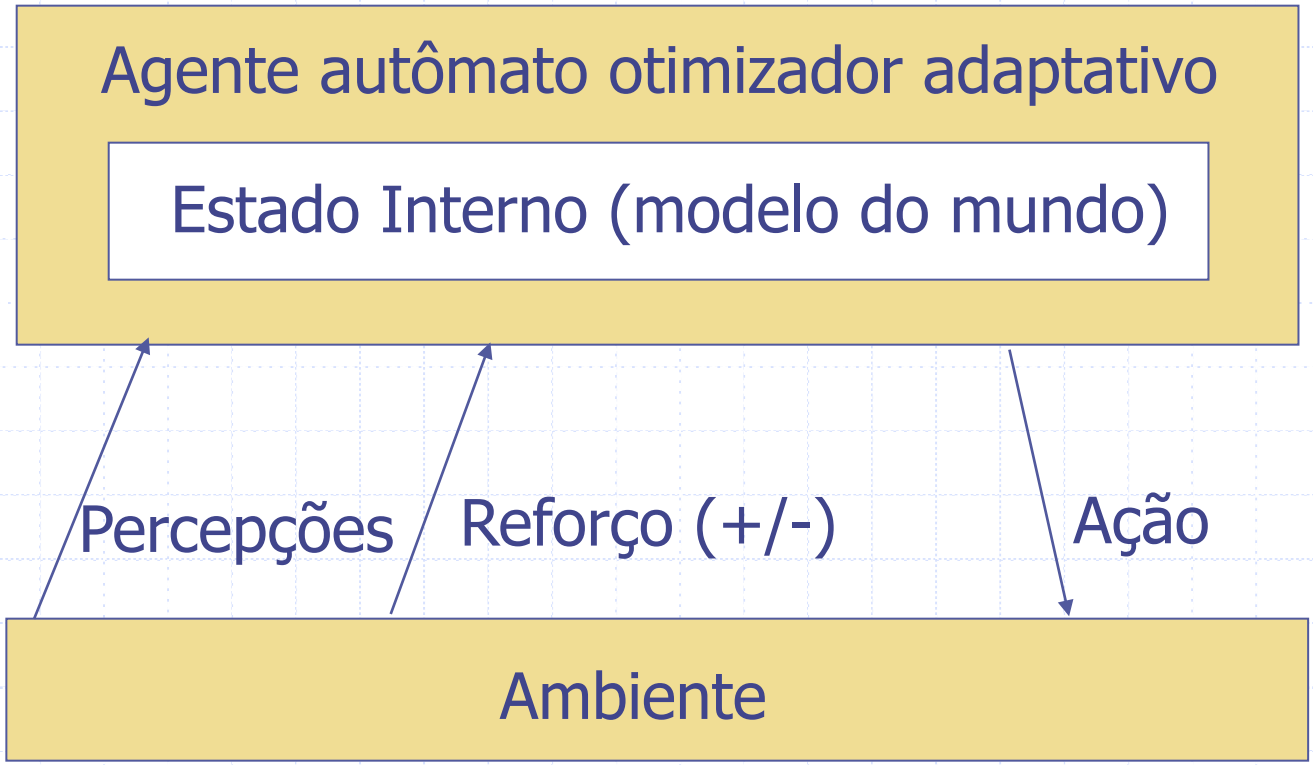


Essa é a situação após vários reinicios...

O que é aprendizagem por reforço (tradicional)?

- Problema de aprendizagem (não é uma técnica)
  - Um agente, em um ambiente
  - A cada instante de tempo **t**:
    - o agente está em um *estado* **s**
    - executa uma *ação* **a**
    - vai para um *estado* **s'**
    - recebe uma *recompensa* **r**
  - Problema da aprendizagem por reforço:
    - Como escolher uma *política de ações* que maximize o total de recompensas recebidas pelo agente

O problema da aprendizagem por reforço



Algumas aplicações

- [Tesauro, 1995] Modelagem do jogo de gamão como um problema de aprendizagem por reforço:
  - Vitória: +100
  - Derrota: – 100
  - Zero para os demais estados do jogo (*delayed reward*)
  - Após 1 milhão de partidas contra ele mesmo, joga tão bem quanto o melhor jogador humano

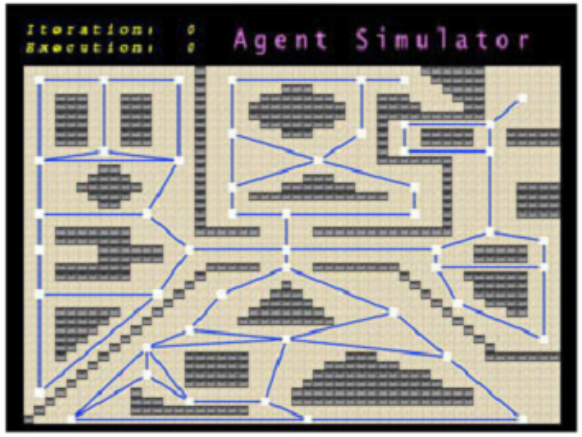
Algumas aplicações

- Time Brainstormers da Robocup (entre os 3 melhores nos 3 últimos anos)
  - Objetivo: Time cujo conhecimento é obtido 100% por técnicas de aprendizagem por reforço
  - RL em situações específicas
    - 2 atacantes contra 2 defensores
    - habilidades básicas
- Inúmeras aplicações em problemas de otimização, de controle, jogos e outros...

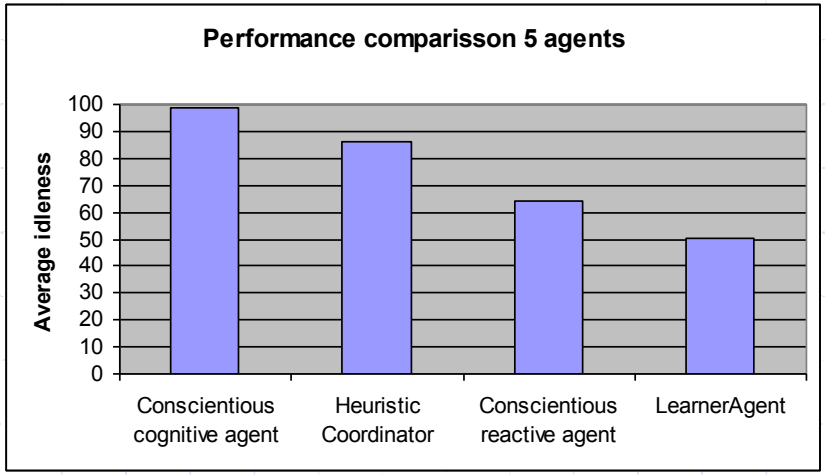
Patrulha multi-agente

- Dado um mapa, um grupo de agentes deve visitar continuamente locais específicos deste mapa de maneira a minimizar o tempo que os nós ficam sem serem visitados
- Recompensa: ociosidade dos nós visitados
- Coordenação emergente (mesmo sem comunicação explícita)

Map B

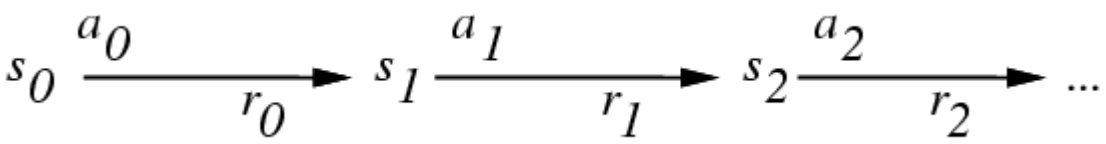


50 n, 69 a



Conceitos Básicos

- Processo de decisão de Markov (MDP)
  - Conjunto de estados **S**
  - Conjunto de ações **A**
  - Uma função de recompensa **r(s,a)**
  - Uma função de transição de estados (pode ser estocástica) **δ(s,a)**
- Política de ações **π(s)** :
  - π: **S → A**



Estados e Ações

- Estado: conjunto de características indicando como está o ambiente
  - Formado pelas percepções do agente + modelo do mundo
  - Deve prover informação para o agente de quais ações podem ser executadas
- A representação deste estado deve ser suficiente para que o agente tome suas decisões (satisfaz a propriedade de Markov)
  - A decisão de que ação tomar não pode depender da seqüência de estados anteriores
  - Ex: Um tabuleiro de dama satisfaz esta propriedade, mas de xadrez não
  - O ambiente não precisa ser episódico

A função de recompensa

- Feedback do ambiente sobre o comportamento do agente
- Indicada por $r:(S \times A) \rightarrow R$
  - r(s,a) indica a recompensa recebida quando se está no estado **s** e se executa a ação **a**
  - Pode ser determinística ou estocástica

Função de transição de estados

- δ: $(\mathbf{S} \times \mathbf{A}) \rightarrow \mathbf{S}$
- δ(s,a) indica em qual estado o agente está, dado que:
  - Estava no estado **s**
  - executou a ação **a**
- Ambientes não-determinísticos:
  - escrita como δ(s,a,s')
  - indica a probabilidade de ir para um estado **s'** dado que estava em **s** e executou **a**

Exemplos de MDPs

Problema	**Estados**	**Ações**	**Recompensas**
Agente jogador de damas	Configurações do tabuleiro	Mover uma determinada peça	#capturas – #perdas
Agente em jogo de luta	Posições/energia dos lutadores, tempo, se está sendo atacado ou não, etc...	Mover-se em uma determinada direção, lançar magia, dar porrada, etc...	(Sangue tirado – sangue perdido)
Agente patrulhador	Posição no mapa (atual e passadas), ociosidade da vizinhança, etc...	Ir para algum lugar vizinho do mapa	Ociosidade (tempo sem visitas) do lugar visitado atualmente

Política de ações (π)

- Função que modela o comportamento do agente
  - Mapeia estados em ações
- Pode ser vista como um conjunto de regras do tipo $s_n \rightarrow a_m$
  - Exemplo:
    - Se estado **s = (inimigo próximo, estou perdendo e tempo acabando)** então
      ação **a = (usar magia)**;
      Se estado **s =(outro estado)** então
      ...

Função valor dos **estados** $V\pi(s)$ $(S \rightarrow R)$

- Como saber se um determinado estado é bom ou ruim?
  - A função valor expressa esta noção, em termos das recompensas e da política de ações
  - Representa a recompensa a receber em um estado **s**, mais as recompensas futuras se seguir uma política de ações π
    - ex. tornar-se diretor, vale pelo que o cargo permite e permitirá nas próximas promoções (não interessa de onde veio - chefe de seção)
  - $V\pi(s_0) = r_0 + r_1 + r_2 + r_3 + \ldots$
    - Problema: se o tempo for infinito, a função valor do estado tende a infinito

Função Valor dos **estados**

- Para garantir convergência e diferenciar recompensas distantes do estado atual, usa-se um fator de desconto
  $0 \leq \gamma \leq 1$
- $V\pi(s_t) = r_t + \gamma\, r_{t+1} + \gamma^2\, r_{t+2} + \gamma^3\, r_{t+3} \ldots$
- $V\pi(s_t) = r_t + \gamma V\pi(s')$, onde:
  - $r_t = r(s_t, \pi(s_t))$
  - $s' = \delta(s_t, \pi(s_t))$
- Ex. Se $\gamma = 90\%$, então:
  - $V\pi(s_t) = r_t + 0.9\, r_{t+1} + 0.81\, r_{t+2} + 0.729\, r_{t+3} \ldots$

Função valor das **ações**
$Q\pi(s,a) : (S \times A) \rightarrow R$

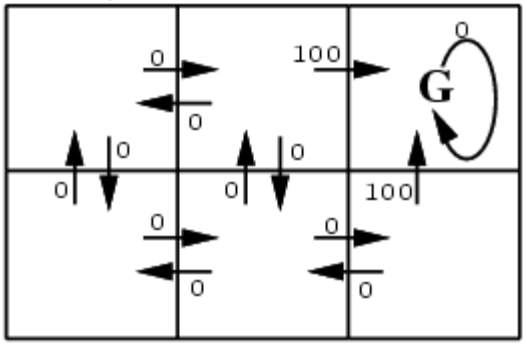
- Analogamente, ela diz a soma das recompensas a obter dado que:
  - o agente está no estado **s**
  - executou uma ação **a**
  - a partir daí, seguiu uma política de ações π
- $Q\pi(s,a) = r(s,a) + \gamma V\pi(s')$, onde:
  - $s' = \delta(s,a)$
    - o valor da ação é a recompensa da ação mais o valor do estado para onde o agente vai devido à ação

Aprendizagem por reforço

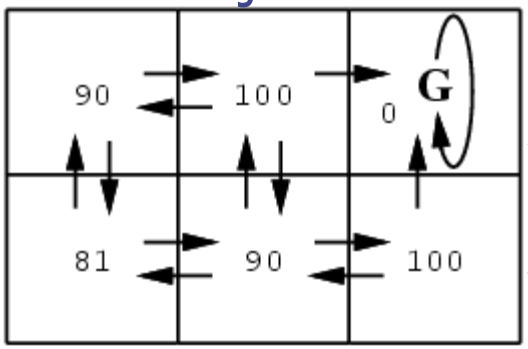
- Tarefa de aprendizagem por reforço:
  - Aprender uma política de ações π^* ótima, que maximiza a função $\mathbf{V\pi}$ **(V*)** ou a função $\mathbf{Q\pi}$ **(Q*)**
    - $\pi^* = \text{argmax}_{\pi}[V\pi(s)]$
- Em outras palavras, de que maneira o agente deve agir para maximizar as suas recompensas futuras

Exemplo: Labirinto (c/γ=0.9)

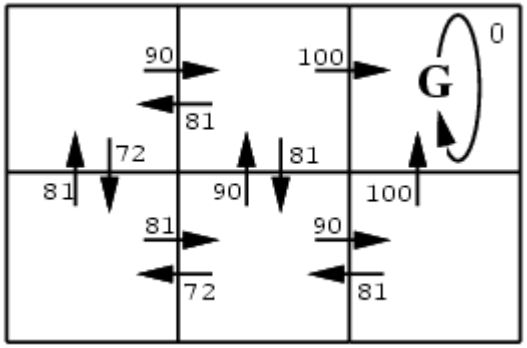
Função recompensa



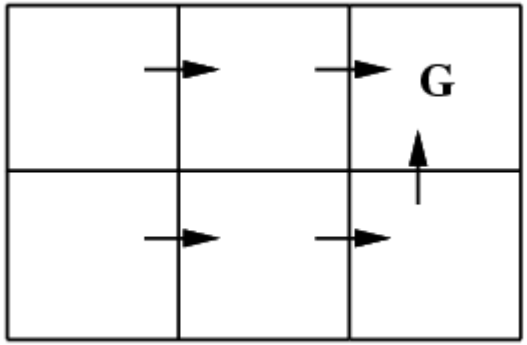
Função V*



Função Q*



Uma política de ações ótima



Aprendendo uma política ótima

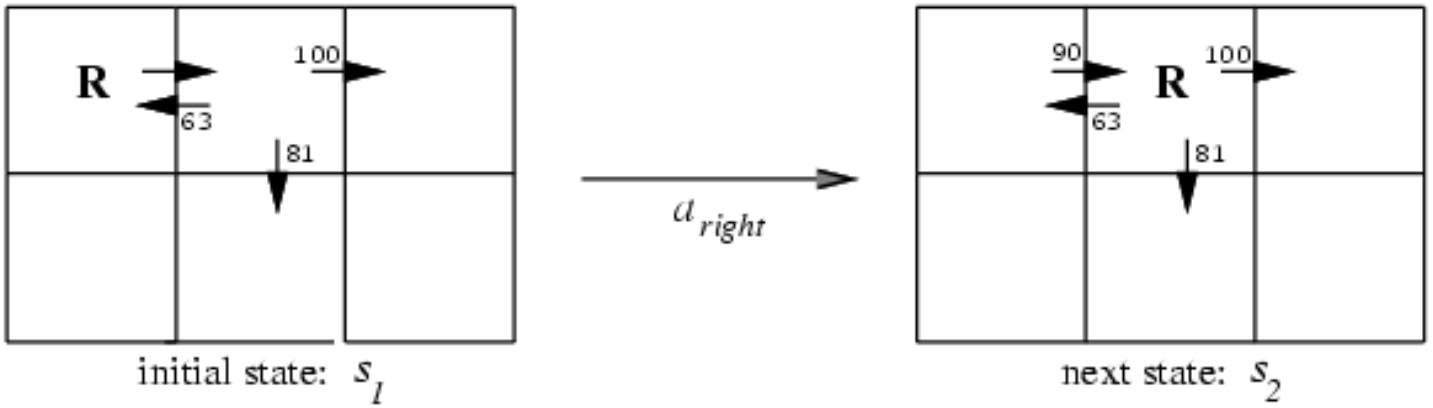
- Se o ambiente é determinístico ($\delta(s,a) = s'$ é conhecida) e r(s,a) é conhecida, a programação dinâmica computa uma política ótima :
  - $V^*(s) = \max_a [r(s,a) + \gamma V^*(\delta(s,a))]$
  - $\pi^*(s) = \text{argmax}_a [r(s,a) + \gamma V^*(\delta(s,a))]$
  - Tempo polinomial
  - Problema: se não temos conhecimento prévio das recompensas e transição de estados
- Se o ambiente é não-determinístico mas a função de probabilidade de transição de estados for conhecida, também é possível computar π^*
  - problema: É difícil estimar estas probabilidades

Q Learning

- É possível determinar π^* se eu conheço Q*
  - não precisando conhecer δ (função de transição de estados) nem **r**
  - $\pi^*(s) = \text{argmax}_a[Q(s,a)]$
    - não é função de δ nem de r
- Então, vamos aprender a função Q ótima (valor das ações) sem considerar V
  - $Q(s_t,a_t) = r(s_t,a_t) + \gamma\, V^*(\,\delta(s_t,a_t)\,)$
    $= r(s_t,a_t) + \gamma\, \max_{a'}\, [Q(s_{t+1},a')]$
    - o valor do próximo estado é o melhor Q nele
    - Como atualizar Q ?

Q-Learning

- Atualiza-se $Q(s_t)$ após observar o estado s_{t+1} e recompensa recebida



- $Q(s_1,a_{right}) = r + \gamma \max_{a'} Q(s_2,a')$
  $= 0 + 0.9 \max\{63,81,100\}$
  $= 90$

Algoritmo Q-Learning para mundos determinísticos

- Para todo estado **s** e ação **a**, inicialize a tabela Q[s][a] = 0;
- Para sempre, faça:
  - Observe o estado atual **s;**
  - Escolha uma ação **a** e execute;
  - Observe o próximo estado **s'** e recompensa **r**
  - Atualize a tabela Q:
    - $Q[s][a] = r + \gamma \max_{a'} (Q[s'][a'])$

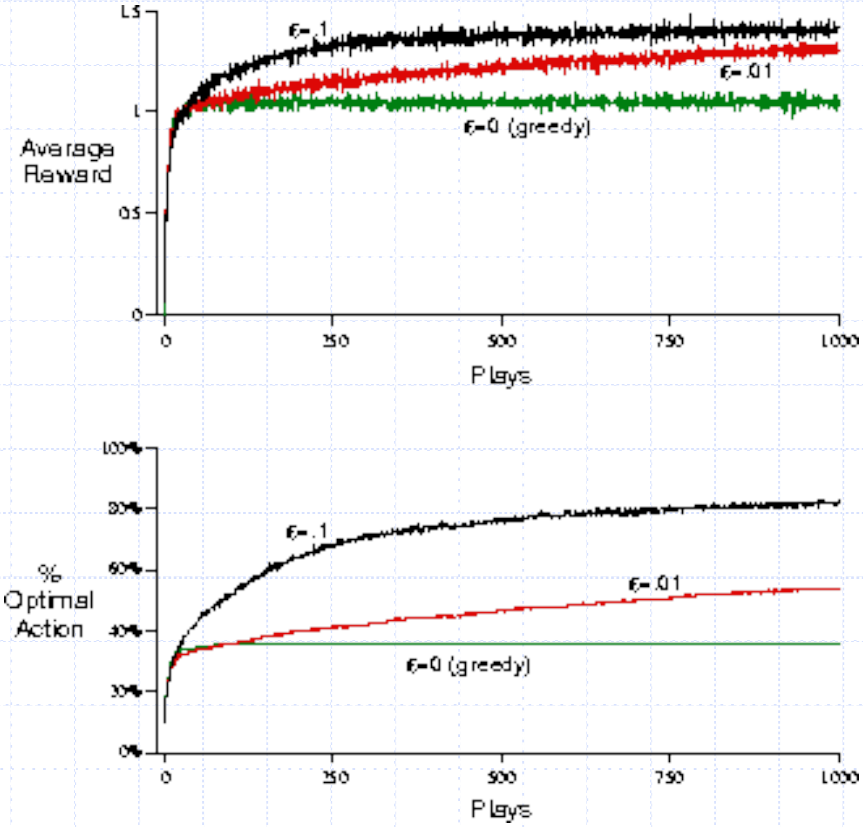
Usufruir valores conhecidos ou explorar valores não computados?

Dilema de explorar ou usufruir (exploration x exploitation)

- Usufruir
  - Escolher a ação que atualmente está com maior valor Q(s,a)
- Explorar
  - Escolher uma ação randômica, para que seu valor Q(s,a) seja atualizado
- Dilema
  - Dado que eu aprendi que Q(s,a) vale 100, vale a pena tentar executar a ação a' se Q(s,a') por enquanto vale 20 ?
    - Depende do ambiente, da quantidade de ações já tomadas e da quantidade de ações restantes

Métodos para balancear exploration e exploitation

- E-Greedy
  - A cada iteração, escolhe uma ação exploratória (randômica) com probabilidade E
  - Ex. 10-armed bandit (caça níqueis)
    - 10 caça níqueis com distribuições de prob. diferentes (desconhecidas)
    - e = 10% acha ações ótimas mais cedo, mas erra 9% do tempo
    - e = 1% obterá melhor performance no futuro!
    - e = 0 % fica preso em uma ação não ótima



Métodos para balancear exploration e exploitation

- Escolha de ações softmax
  - A probabilidade de uma ação ser escolhida varia de acordo com o valor de Q(s,a)
  - Pode-se usar o conceito de temperatura de simulated annealing:
    - T = infinito, ação totalmente exploratória
    - T = zero, ação totalmente gulosa

$$\frac{e^{Q_{t-1}(a)/\tau}}{\sum_b e^{Q_{t-1}(b)/\tau}}$$

RL em ambiente não determinístico

- Recompensas serão não determinísticas
  - $V\pi(s) = E[r_t + r_{t+1} + r_{t+2} + ...]$
  - Problema:
    - Suponha que a sequência de recompensas recebidas em um determinado estado foi:
      - 100, 98, 101, 97, 90, 103, 10
    - O valor da função Q vai refletir apenas o último valor !
- Solução: usar uma taxa de aprendizagem α
  - $Q_n(s,a) = (1-\alpha)Q_{n-1}(s,a) + \alpha[r + \max_{a'} Q_{n-1}(s',a')]$
  - A atualização de Q não "esquece" dos valores anteriores
  - Se $\alpha = 1/[1+\#visitas(s,a)]$, Q converge para Q* em tempo polinomial

Semi-MDP

- Como o agente pode levar em conta o tempo de suas ações?
  - Ex. no jogo de luta: É melhor dar vários socos fracos ou um soco forte?
    - Soco forte provavelmente traria maior recompensa
    - Demoraria mais tempo para ser executado
  - No problema da patrulha: como levar em conta o a distância entre os nós?

Semi-MDP

- O formalismo SMDP engloba este conceito
- Prova-se que a atualização de Q passa a ser dada por:
  - $Q[s][a] = r + \gamma^t \max_{a'} (Q[s'][a'])$
  - Onde **t** pode ser:
    - número de unidades de tempo que o agente executou a ação (caso discreto)
    - alguma função contínua do tempo
  - Desta maneira, as recompensas futuras passam a valer menos se o agente passar muito tempo executando uma ação

Complexidade de Q-Learning

- Escolher uma ação é barato no tempo
  - No entanto, o tempo de treinamento cresce com #S
- Em espaço: O(#S x #A)
  - Problemas
    - o número de estados possíveis cresce exponencialmente com a quantidade de características representadas
    - Como tratar estados contínuos?

Linhas de pesquisa em RL atualmente

- Substituir a tabela Q por redes neurais
  - Permite generalização
  - Tratar estados contínuos
- Tratar casos em que o estado é parcialmente observável
  - POMDP
- Aprendizagem por reforço hierárquica
- Aprendizagem por reforço multi-agente

Aprendizagem por reforço multi-agente - Cooperação

- Abordagens usando RL tradicional:
  - White box agent
    - Representação de estado global
    - Encontra a ação conjunta (a_1, a_2, ..., a_n) que maximiza uma função de reforço global (única)
    - Problemas
      - Complexidade exponencial no número de agentes
      - Como aprender as ações de maneira distribuída ?
  - Black box agent
    - O reforço é individual, mas é alguma função do bem estar global
    - O agente não toma conhecimento dos outros agentes
      - Outros agentes passam a ser ruído no ambiente

Aprendizagem por reforço multi-agente - Cooperação

- Black box agent
  - Problemas
    - Atribuição de crédito
      - Como atribuir valor a ações individuais de um agente, em termos do bem estar global?
      - Ex: Que reforço dar pra uma ação do jogador do meio de campo em um jogo de futebol?
- Gray box agent:
  - O agente toma suas decisões individualmente
  - Os agentes comunicam suas intenções

Aprendizagem por reforço multi-agente - Competição

- Min-Max Reinforcement Learning
  - RL Tradicional:
    - Executa ações com maior recompensa esperada
  - Min-Max RL
    - Modela as ações do adversário
    - Executa ações que, dado que um oponente utiliza uma política ótima, minimiza minhas perdas

Referências

- Slides de Hugo Pimentel de Santana (CIN/UFPE)
- Lecture slides do livro *Machine Learning*, do Tom Mitchell
  - http://www-2.cs.cmu.edu/~tom/mlbook-chapter-slides.html
- Livro "Reinforcement Learning: An introduction", de Sutton & Barto disponível *online*
  - http://envy.cs.umass.edu/~rich/book/the-book.html