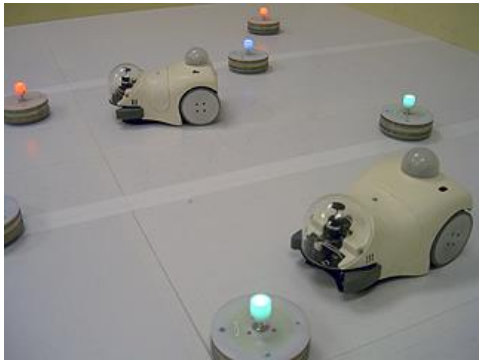
シミュレーション論 Ⅱ

第8回

強化学習

強化学習

- 強化学習：試行錯誤をくりかえして、よりよい行動方針を獲得する手法
- 状態と行動をセットにして記述し、うまくいった場合に「報酬」、失敗した場合に「罰」を与えることでよりよい行動を獲得するようになる
- 教師データが不要なため、未知の環境への応用が可能
- ロボットの行動獲得などによく利用される



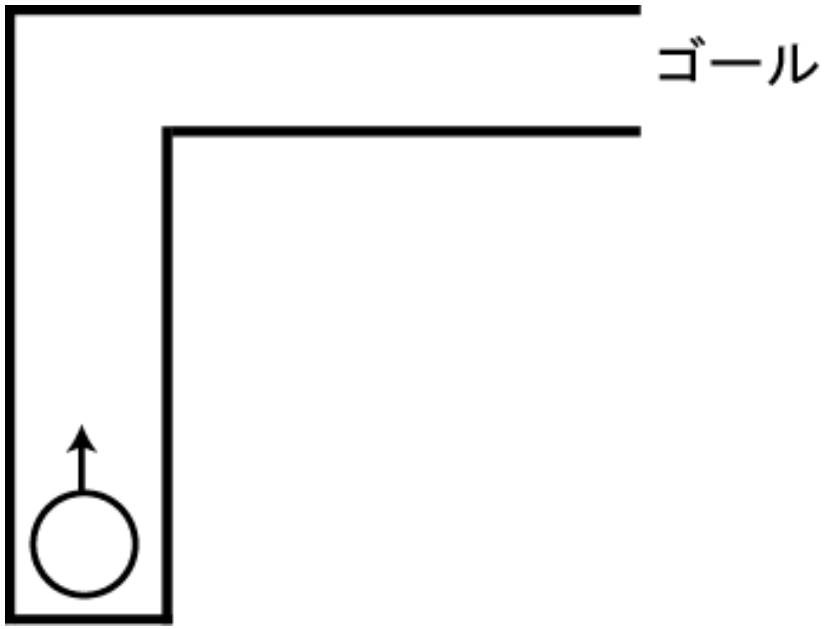
強化学習のしくみ

- 強化学習では学習をおこなう「主体」と「環境」がある
- 主体は環境の状態を観測し、行動を選択する
- 行動選択の結果として、環境から「報酬」または「罰」を得る（報酬は毎回与えられるとは限らず、特定の状況でのみ与えられる場合もある）

- 例）ロボットの行動
  - **左右と後ろが壁である環境**
  - **ロボットは周囲の状況を観察し、進む方向を決定する**
  - **無事進行できた場合→報酬**
  - **壁にぶつかった場合→罰**
- これを繰り返すことで、環境に応じた行動を選択できるようになる

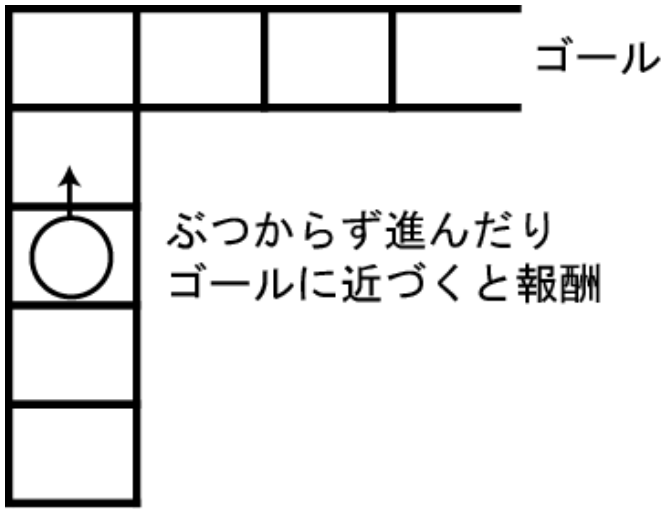
ごく単純な強化学習のモデル

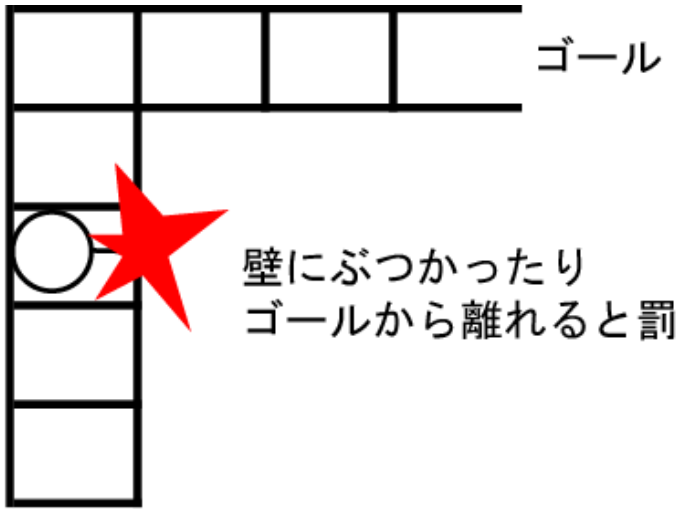
- 壁に囲まれた通路を歩いて、ゴールを目指すモデルを考えよう
- 計算式が複雑になるのでQ-leaning等の定式化は用いず、ごく簡単なモデルで強化学習のイメージをつかんでみよう



単純な強化学習のモデル(2)

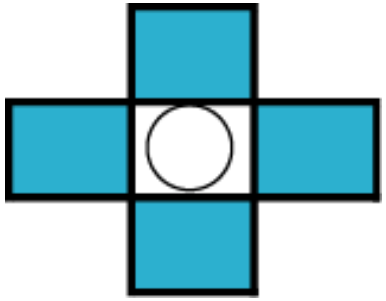
- 行動する主体(エージェント)の行動について以下のように仮定する
  - 上下左右の1マス分を観察できる
  - 1回につき1マス移動できる
  - 無事に進めたら報酬、壁にぶつかったら罰を与えられる
  - ゴールに近づいたら報酬、ゴールから離れたら罰を与えられる





単純な強化学習のモデル(3)

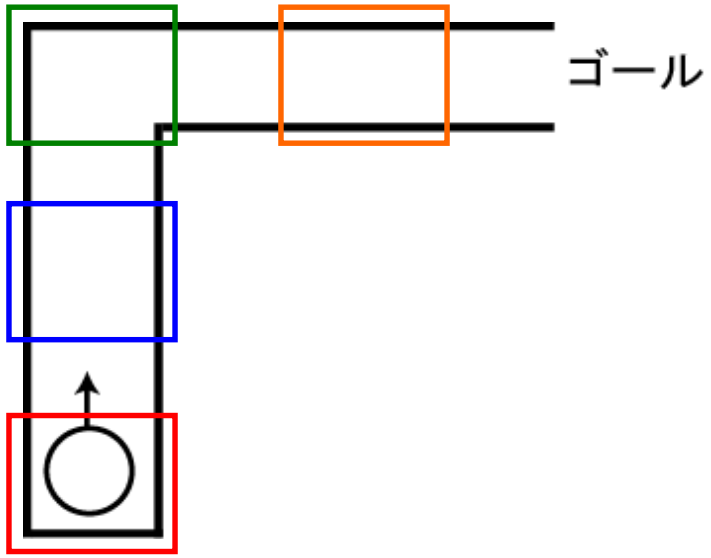
- 行動する主体(エージェント)にとっての環境は「観察できる範囲に壁があるか、ないか」で表される
- 観察できる範囲は上下左右の4マス
- また、エージェントは移動した位置がゴールに近づいたか離れたかを知ることが出来る



観察できる範囲

単純な強化学習のモデル(4)

- 先ほどの通路の例では、エージェントが観察できる状況は以下の4種類
  - ①左右と下が壁、上は空き
  - ②左右が壁、上下は空き
  - ③左と上が壁、右と下は空き
  - ④上下が壁、左右は空き



- 簡単にするためエージェントは常に上を向いていると仮定する
- ゴールに到達したら終了

単純な強化学習のモデル(5)

- エージェントの行動は上下左右いずれかに1マス移動
- 先ほどの状態に応じてそれぞれ上下左右なので、
  ４×４＝16のパターンが考えられる

- 以下の状態をそれぞれ状態１～４として、それぞれ行動との組み合わせを作成し、評価値を与える
  - 左右と下が壁、上は空き　（状態１）
  - 左右が壁、上下は空き　（状態２）
  - 左と上が壁、右と下は空き（状態３）
  - 上下が壁、左右は空き　（状態４）

単純な強化学習のモデル(6)

- 状態＋行動の組み合わせは以下のようになる
- 初期状態での各行動の評価値を5としておく

状態	行動	評価値
1	上	5
	下	5
	左	5
	右	5
2	上	5
	下	5
	左	5
	右	5
3	上	5
	下	5
	左	5
	右	5
4	上	5
	下	5
	左	5
	右	5

単純な強化学習のモデル(7)

- 報酬と罰:

  エージェントが壁にぶつからずに進んだら＋１、さらにゴールに近づいたら＋１、ゴールから離れたら－１、壁にぶつかったら－１を評価値に加える

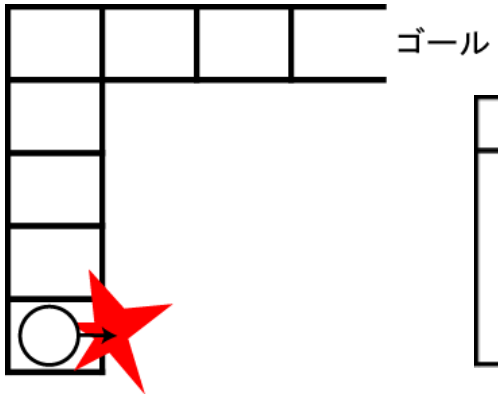
- 行動選択は「その状況において最も評価値の高いもの」を選ぶこととし、同じ評価値のものが複数ある場合はランダムに1つを選ぶ （greedy法と言われる方法）

- ゴールに到達したら終了とし、「評価値をキープしたまま」、再度スタート地点から繰り返す

実行例(1)

- スタート地点では(状態1)
- 行動の評価値は全て5なので、ランダムに行動を選択し「右」を実行したとする
- 壁にぶつかったので、(状態1－右)の組み合わせの評価値を－1とする
- 位置は変化しなかったので、同じ位置から次の行動を選択する

状態	行動	評価値
1	上	5
	下	5
	左	5
	右	5

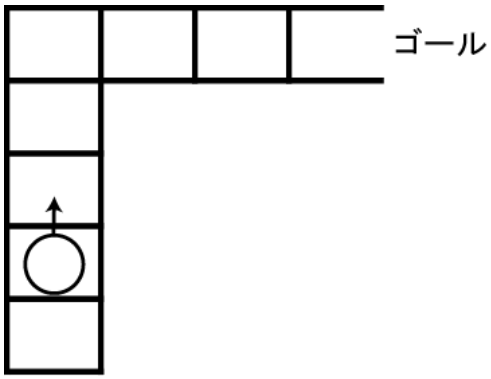


状態	行動	評価値
1	上	5
	下	5
	左	5
	右	4

実行例(2)

- 位置は変化しなかったので、同じ位置(状態1)から次の行動を選択する
- 評価値は上・下・左が5で最大なので、この中からランダムに選ぶ
- 「上」が選択されたとすると1マス進めるので評価値に＋1となる
- さらにゴールに近づいているので、評価値に＋1となる

状態	行動	評価値
1	上	5
	下	5
	左	5
	右	4

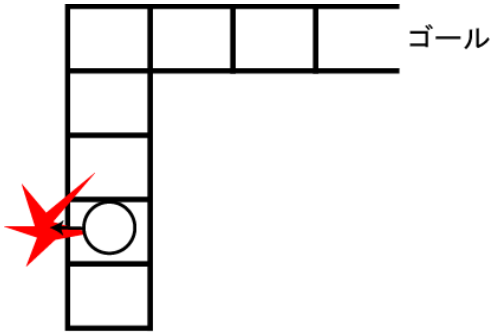


状態	行動	評価値
1	上	7
	下	5
	左	5
	右	4

実行例(3)

- 位置が1マス動いたので、(状態2)になる
- このときの行動の評価値はすべて5なので、ランダムに1つ行動を選択する
- 「左」が実行されたとすると壁にぶつかるので(状態2－左)の評価値を－1とし、位置はそのまま

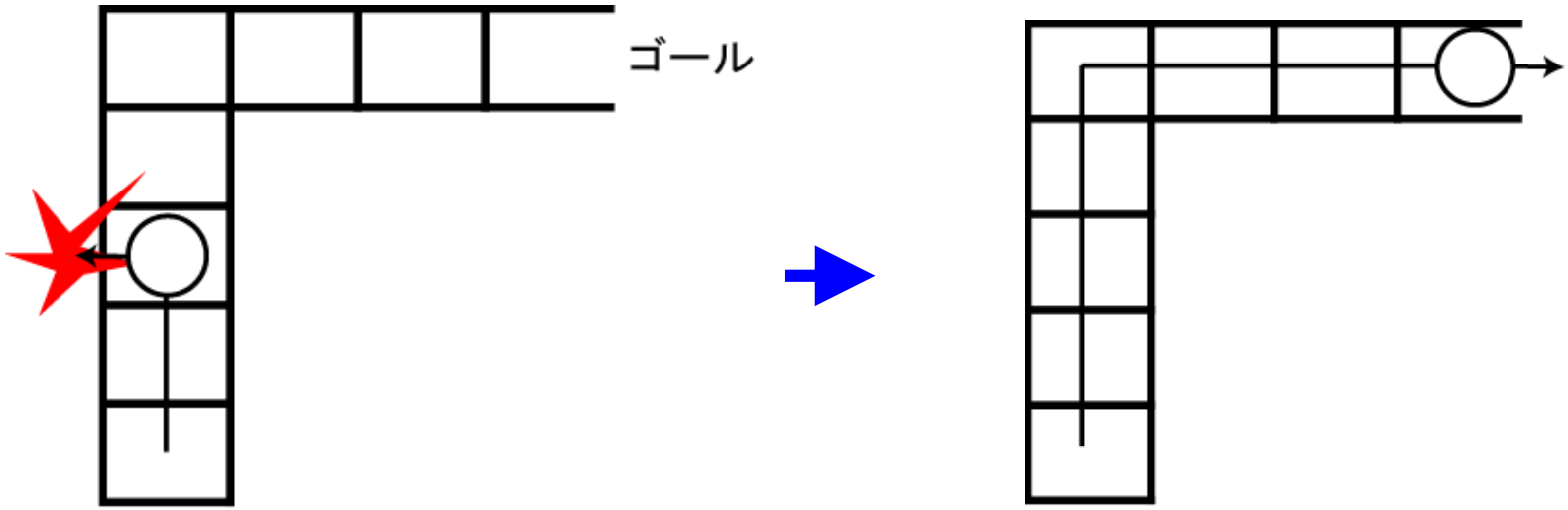
状態	行動	評価値
2	上	5
	下	5
	左	5
	右	5



状態	行動	評価値
2	上	5
	下	5
	左	4
	右	5

実行例(4)

- 以上の作業をくりかえし、評価値を変更しながら進めていく
- ゴールに到達したら終了とし、またスタート地点から繰り返す
- この作業を繰り返していくと、最終的にスムーズにゴールへ向かっていく行動が獲得できる(＝学習した)



手作業でのシミュレーション

- 先ほどの例題を手作業で試してみてください
- ゴールに到達したらその時点の評価値をキープしたまま、再度スタートからはじめ2回ゴールするまでやってみてください
  ※15回行動選択をおこなってゴールしなければ終了して次の回へ
- 同じ評価値の行動がある場合には乱数表を使用して行動を決定してください（3つある場合は1～3の乱数表を使用）

実際の強化学習アルゴリズム

- 通常の強化学習アルゴリズムでは評価値の計算方法などがもっと複雑になるが、基本は同様
- 一定期間ごとに、遺伝的アルゴリズム等を用いて行動の取捨選択などをおこなう場合もある
- 強化学習では「試行錯誤」の繰り返しで行動主体が自律的に学習するため、教師データが不要
- また、未知の環境に対しても対応できる可能性が大きい
- 学習プロセスは生物や人間の行動パターンの再現などにも利用できるのではないか？

Q-learning

- 強化学習の代表的アルゴリズム
- Q値と呼ばれる「環境と行動の組み合わせの評価値」を逐次修正してゆき、最適な行動を探す方法

（1）エージェントは環境の状態 s_t を観測する

（2）エージェントは任意の行動選択方法（探査戦略）にしたがって行動 a_t を実行する

（3）環境から報酬 r_t を受け取る

（4）状態遷移後の状態 s_{t+1} を観測する

（5）以下の更新式によりQ値を更新：

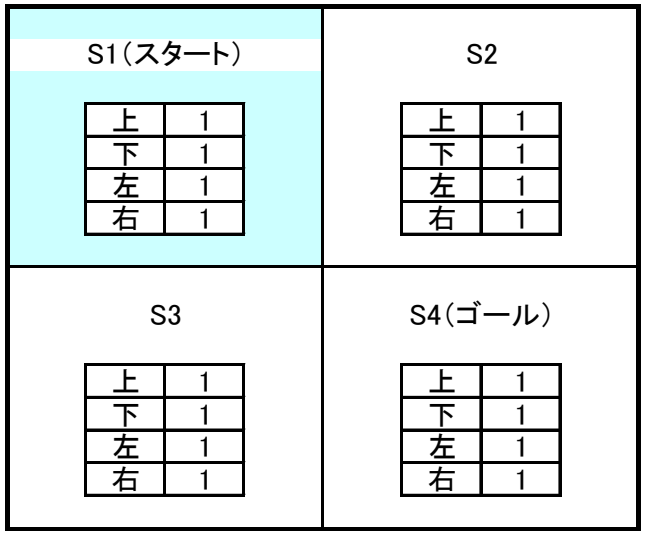
$$Q(s_t, a_t) \leftarrow (1-\alpha)Q(s_t, a_t) + \alpha[r_t + \gamma \max_a Q(s_{t+1}, a_t)]$$

ただしαは学習率（$0 < \alpha \leq 1$），γは割引率（$0 < \gamma \leq 1$）である。

（6）時間ステップ t を $t+1$ へ進めて手順（1）へ戻る

Q-learning の数値例

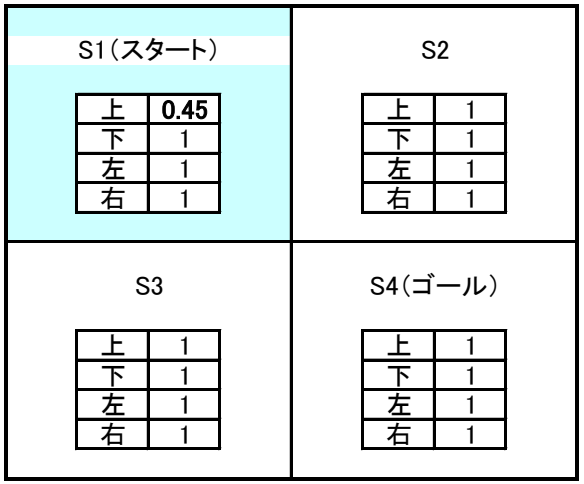
- 例）以下のような4マスの迷路を考える
- 各マスでの状態をそれぞれS1～S4とし、行動は上下左右の4種をとることができるものとする
- マスの一番外の枠は壁とし、壁方向へは移動できない（もとの場所にとどまる）
- 壁にぶつかったら報酬 －1、ゴールしたら +1、それ以外は報酬 0 とする
- 学習率 α ＝0.5、割引率 γ ＝0.9 とする



Q-learning の数値例（2）

- 各状態でのQ値の初期値を1とする
- S1からスタートし、行動「上」が選ばれたとすると
  →壁に当たるため位置はS1のまま、報酬は－1
  →よって、Q値は

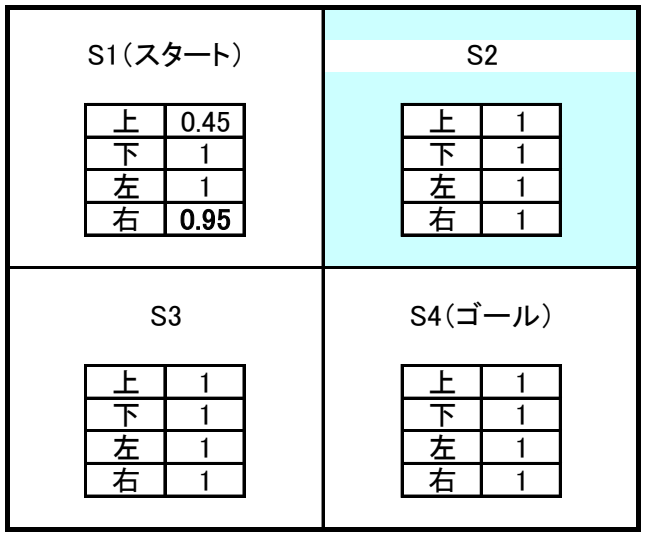
$$Q(S1, 上) \leftarrow (1-0.5)Q(S1, 上) + 0.5[-1 + 0.9 \max_a Q(S1, a)]$$
$$= 0.5 \times 1 + 0.5 \times (-1 + 0.9 \times 1)$$
$$= 0.45$$



Q-learning の数値例（3）

- 次に、S1で行動「右」が選ばれたとすると
  →状態はS2へ移動、報酬は 0
  →よって、Q値は

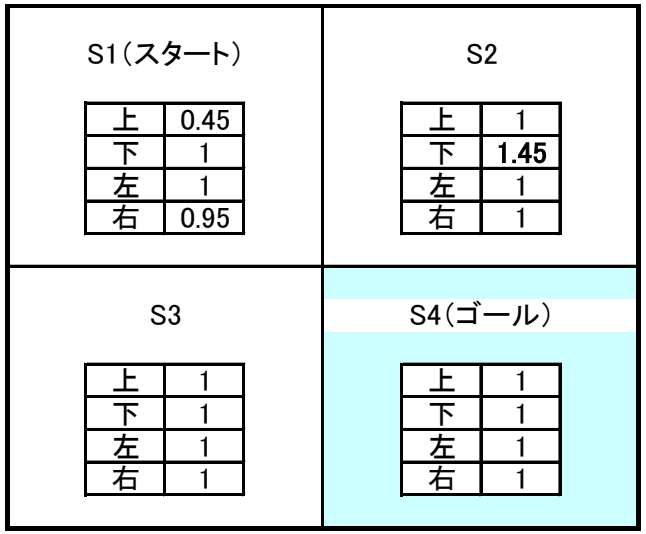
$$Q(S1,\text{右}) \leftarrow (1-0.5)Q(S1,\text{右}) + 0.5[0 + 0.9 \max_a Q(S2,a)]$$
$$= 0.5 \times 1 + 0.5 \times (0.9 \times 1)$$
$$= 0.95$$



Q-learning の数値例（4）

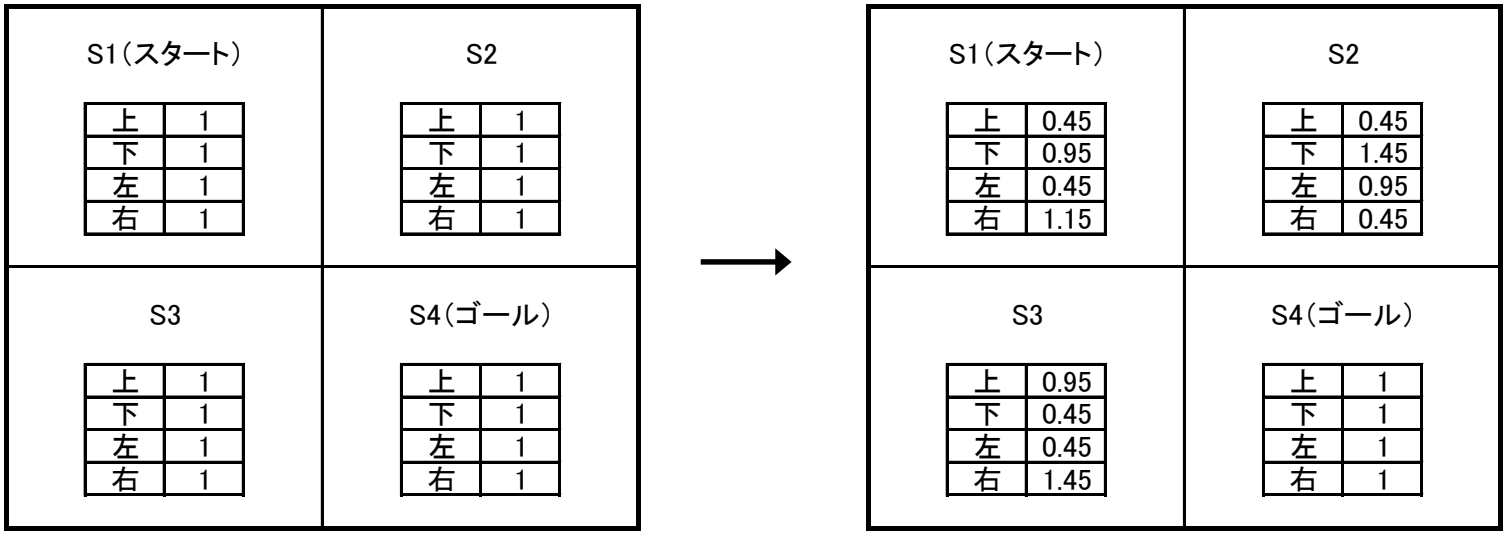
- 次に、S2で行動「下」が選ばれたとすると
  →状態はS4（ゴール）へ移動、報酬は 1
  →よって、Q値は

$$Q(S2,\text{下}) \leftarrow (1-0.5)Q(S2,\text{下}) + 0.5[1 + 0.9 \max_a Q(S4,a)]$$
$$= 0.5 \times 1 + 0.5 \times (1 + 0.9 \times 1)$$
$$= 1.45$$



Q-learning の数値例（5）

- ゴールへ到達、または一定回数繰り返してゴールに達しなければスタートへ戻り、再度同じ手順を繰り返す
- 次第にQ値が収束してゆき、「各状態でゴールへ近づく行動」の値が大きくなる
- Q値の大きい行動をたどればゴールに近づく



行動選択の方式

- Q値から行動を決定する方法には以下のようなものがある

- **ε −greedy**
  ε の確率でランダム、それ以外は最大の重みを持つルールを選択
- **ルーレット選択**
  Q(s,a)に比例した割合で行動選択
- **ボルツマン選択**
  exp(Q(s,a)/T)に比例した割合で行動選択、ただしTは時間とともに0に近付く

- ただし s は環境の状態、a は行動

Q-learning の特徴

- Q-learningは行動により状態が変わった後の「仮定の行動」を用いて評価をおこなうもので、Off-Policyの方式と言われる
- これに対し、On-Policyと呼ばれるものは厳密に「自分が行動した結果」に基づいて評価をおこなうものである
  - 代表的手法としてprofit sharing など（報酬を得た時点から過去の行動にさかのぼって報酬を与える方式）

- 強化学習には様々な方式があり、それぞれに特徴がある
- 状況や問題に応じて使い分ける

第8回のレポート

- さきほどの数値例と同じ条件で、図のS1からスタートし、「　」→「　」→「　」→「　」の順に行動が選択された場合、各状態のQ値がどうなっているか計算せよ。

  ただしQ値の初期値はすべて1とする。