自然言語処理
ー強化学習入門ー

https://satoyoshiharu.github.io/nlp/

強化学習入門
代表的なアルゴリズム

価値学習

価値の漸化式的な関係(ベルマン等式)を利用する

アルゴリズム例： DQN

ゴール:Q(s,a)を見つける

最適ポリシー:
a=argmaxQ(s,a)

応用例:Googleのレコメンド

ポリシー学習

ポリシーを直接最適化する

アルゴリズム例:PG

ゴール:π(s)を直接見つける

最適ポリシーπ(s)は確率的に a を決める

応用例:AlphaGo

Actor-Critic

価値学習とポリシー学習のいいとこどりハイブリッド

ベルマン等式

- ある時刻の選択とそれ以降の決定問題の価値との関係（後ろ向き再帰）を示すことで、単純な部分問題に分割する

$$状態価値V(s_t)=\max_a\{即時報酬r(s_t,a_t)+\gamma V(s_{t+1})\}$$

TD(時間的差分)学習

- ベルマン等式を使い、次の状態s'の価値と　r+ $\gamma V(s')$ との差を使って、学習する。

状態価値$V(s_t)$

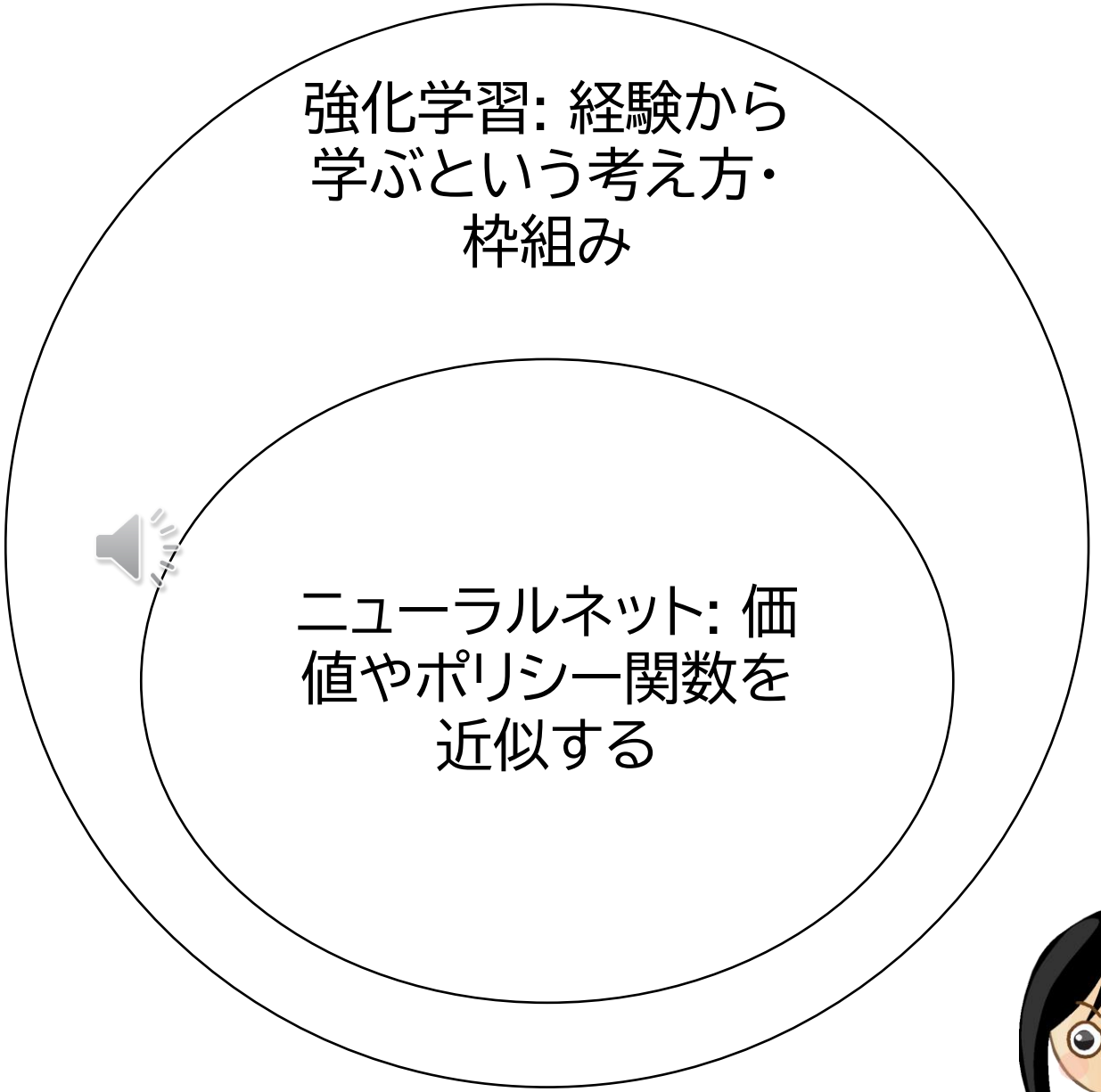
<- 状態価値$V(s_t)$

$+\alpha\{$ 即時報酬$r_t+\gamma *$次の状態の価値$V(s_{t+1})$ $-$ 状態価値$V(s_t)$ $\}$

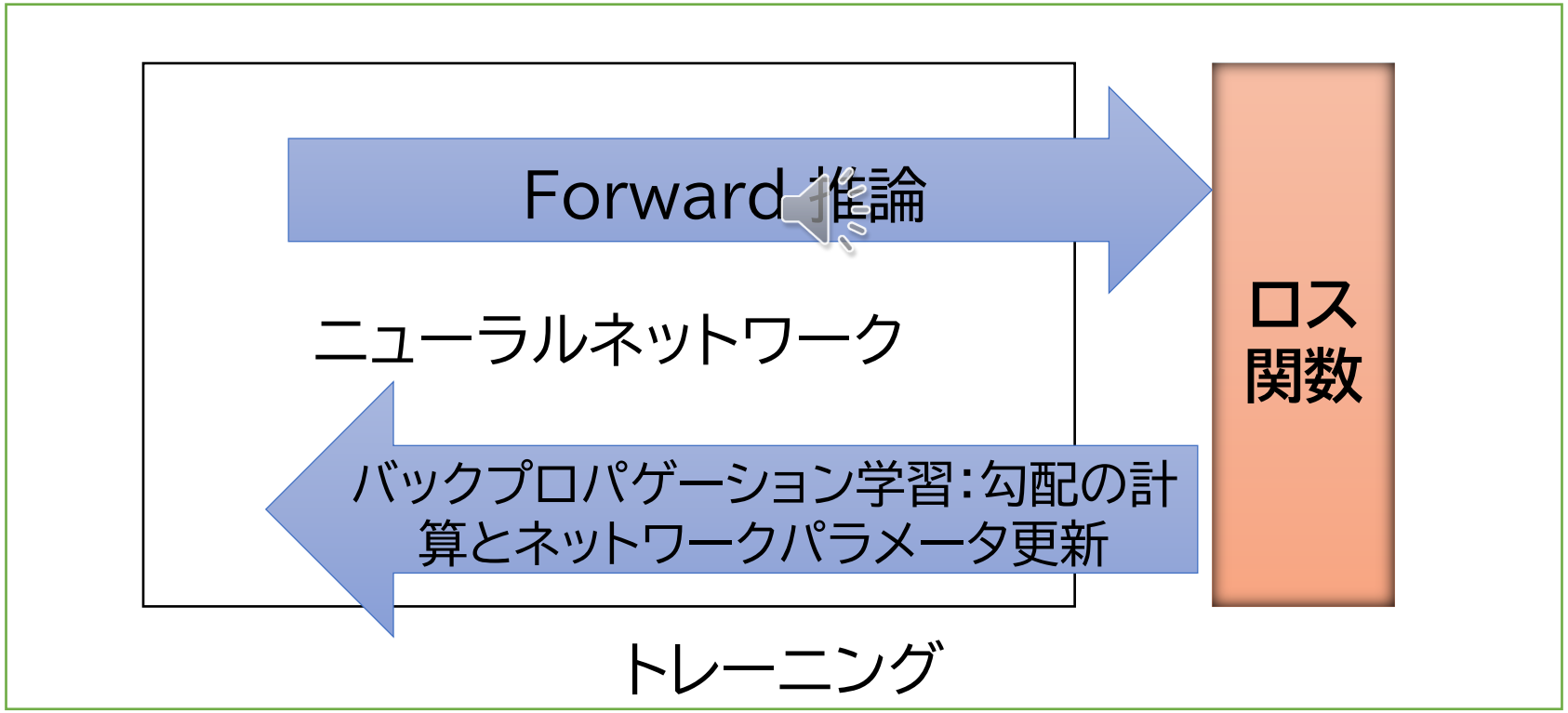
次の時刻の推定値、より正しい

現在の推定値

強化学習はニューールネットによる関数近似を利用

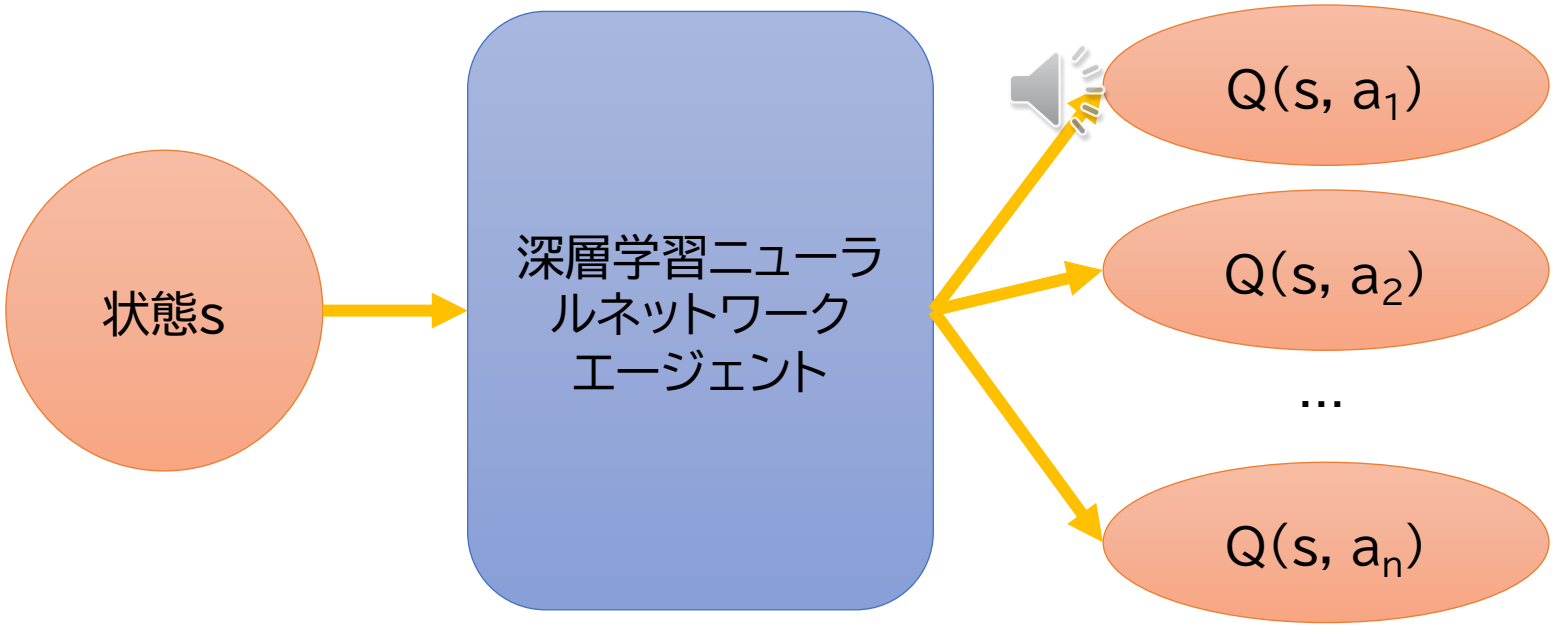


ネットワークの学習
バックプロパゲーション(逆伝播)



DQN

課題設定：入力(s,a)からQ(s,a)の関数を得たいところ、以下のように変更する。



最適ポリシー$\pi^*(s)$
＝ argmaxQ(s,a)

DQN：価値ネットワークのトレーニング

TD

- 以下のQロスを使い、バックプロパゲーション学習させる

$$\text{Qロス} = || \boxed{\text{target} - \text{prediction}} ||^2$$

$$\text{target} = r + \gamma \max Q(s_{t+1}, a_{t+1})$$

$$\text{prediction} = Q(s_t, a_t)$$

- targetとpredictionで別々のQ表を持つのがDouble DQN。targetのQ表の更新を1000回に1回などど遅らせることで、トレーニングを安定させる。

離散アクション空間

- 右左への移動など、有限個のアクションからなる空間

連続アクション空間

- 右へ3.5の力で、左へ1.2の力で、などと、実数値で表現できるアクションからなる空間。

確定的ポリシー

- アクション空間が、上下左右へのどれかへの移動、といった、有限個な世界におけるポリシー

確率的ポリシー

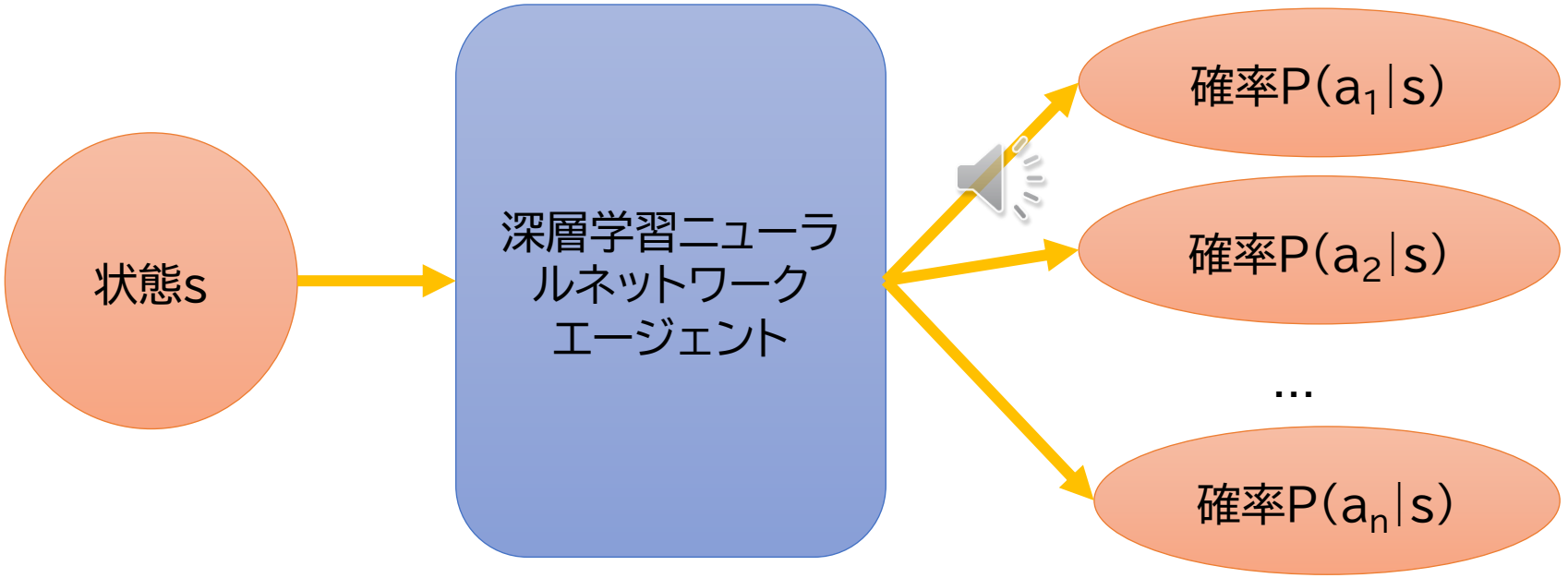
- 左へは確率0. 8, 右へは確率0. 05, などと、アクションを確率分布で持っておくポリシー。

Exploit-Explore必要

DQNの欠点

- 離散で小さなアクション空間しか対応できない。
- DQNのポリシーは、確定的なもの。

Policy Gradient（REINFORCE）

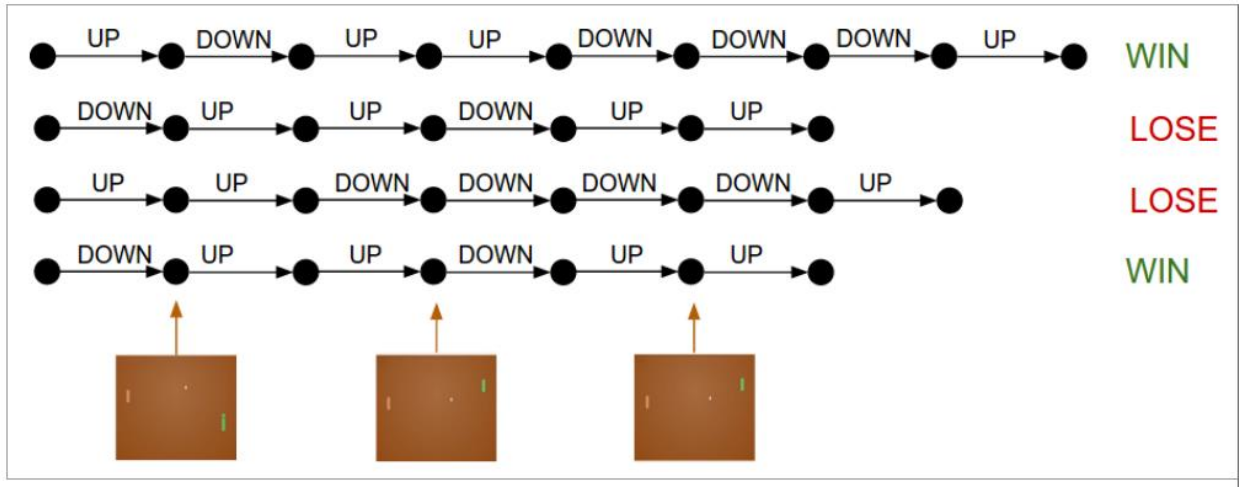


$\pi^*(s)$：P(a|s)
確率に沿ってアクションをサンプルする

$\Sigma P(a_i|s)=1$、
アクション空間は連続値でもよく、その場合 $\int P=1$

PG: ポリシーネットワークのトレーニング

- 初期化後、以下を繰り返す
- ポリシーを使い、終了状態まで実行する。
- 全(状態、アクション、報酬)を記録。
- ポリシー更新： 失敗したエピソードの全アクションの確率を低くし、成功したエピソードの全アクションの確率を高くする



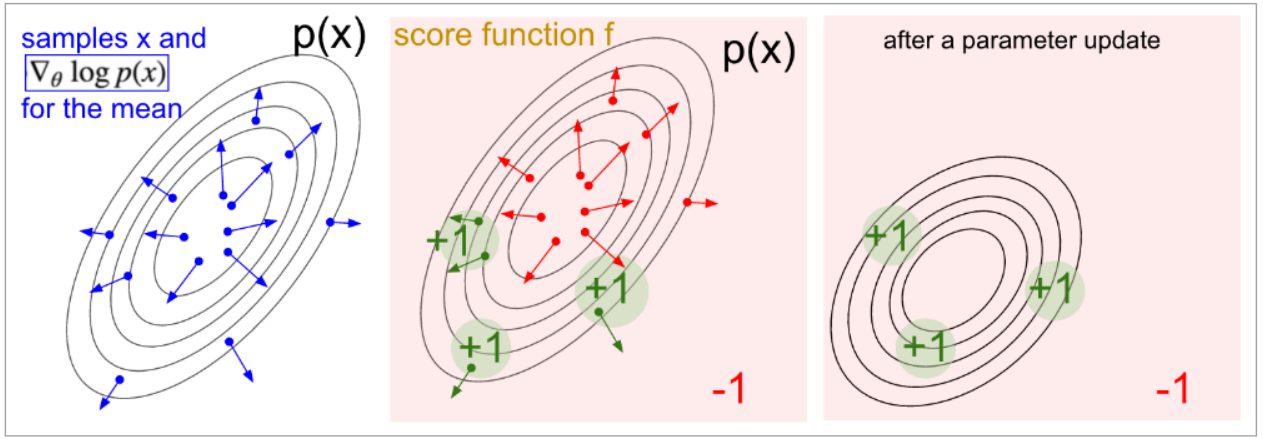
http://karpathy.github.io/2016/05/31/rl/

PG: ポリシーネットワークのトレーニング

- 報酬が大きいほど確率を大きくするために、ロス関数は、アクション確率の対数尤度 $\log P(a_t|s_t)$ にディスカウント総報酬R_tをかけた以下のものとし、バックプロパゲーション学習させる。

$$\text{ロス} = - \log P(a_t|s_t)\ R_t$$

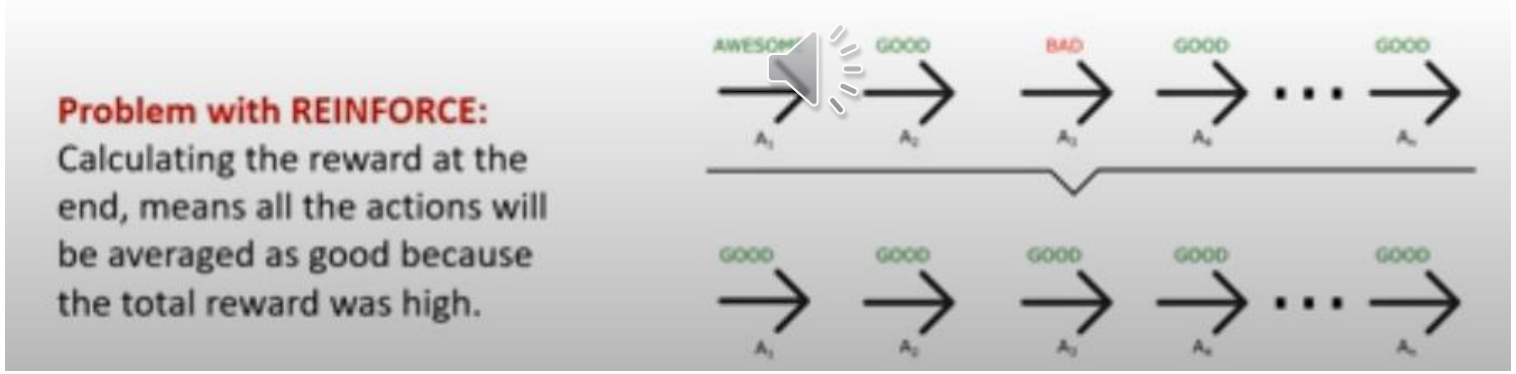
報酬の大きいほうへポリシー(とるべきアクションの確率分布)が寄っていく様子



http://karpathy.github.io/2016/05/31/rl/

PGの欠点

- それぞれのアクションがどれくらい結果に貢献したか明確でない。



MIT 6.S091: Introduction to Deep Reinforcement Learning (Deep RL)

Advantage Actor-Critic (A2C)

- 二つのニューラルネットワークを持つ
- Actor： ポリシー学習
  - w <- w + ▽logP(a|s)R で、確率勾配に重みを与えるRの代わりにQ(s,a)を使うことで、各アクションのポリシー確率に対し、最終結果にどのくらい貢献したかの個別の重みづけを与える
- Critic： Q学習

強化学習のアルゴリズムのまとめ

- 強化学習アルゴリズムは、ポリシー学習、価値学習、Actor-Criticという3つに分類されます。
- 価値学習はベルマン等式や時間的差分学習を基礎とします。
- 価値学習の代表はDQN、ポリシー学習の代表は、ポリシーグラディエントです。どちらも、ニューラルネットによる深層学習を関数近似に利用します。

https://yo-sato.com/
https://satoyoshiharu.github.io/nlp/