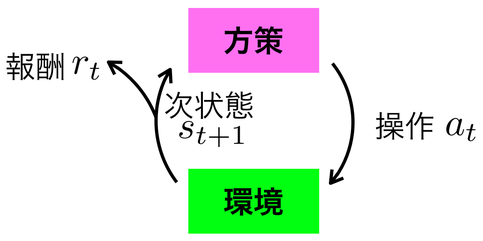
強化学習におけるゲート付き内部報酬による探索方法

早稲田大学大学院 先進理工学研究科 電気・情報生命専攻 村田研究室
修士1年　5318E090　向井滋春

背景

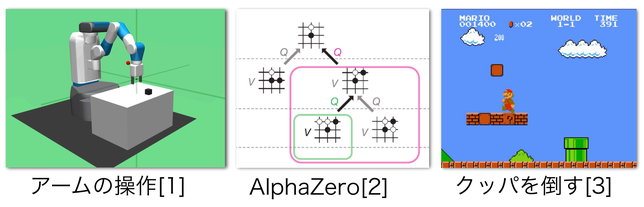
強化学習とは

・制御方策を学習する手法
・報酬の定義された環境が必要
・累積の報酬を最大化する方策を得たい



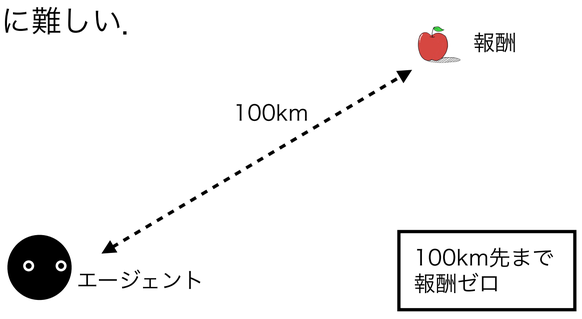
最新の応用例

・ゲームや簡単な例では成果が出てきた
・現実問題への応用はまだ少し先
　理由１：膨大な試行錯誤が必要
　理由２：報酬設計の難しさ
　→上手く報酬を与えないと全く学習できない


アームの操作[1]　AlphaZero[2]　クッパを倒す[3]

疎な報酬の問題は難解

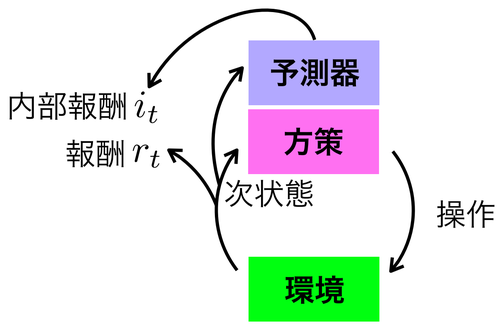
・報酬がなかなか得られない問題の学習は非常に難しい.



先行研究の問題点と新しいアイデア

【先行研究】予測誤差を報酬に加えると未知の状態へ向かう[3]

・次状態を予測する予測器を，強化学習と並行で学習
・その予測器の誤差を**内部報酬**として報酬に加算



先行研究のアイデア

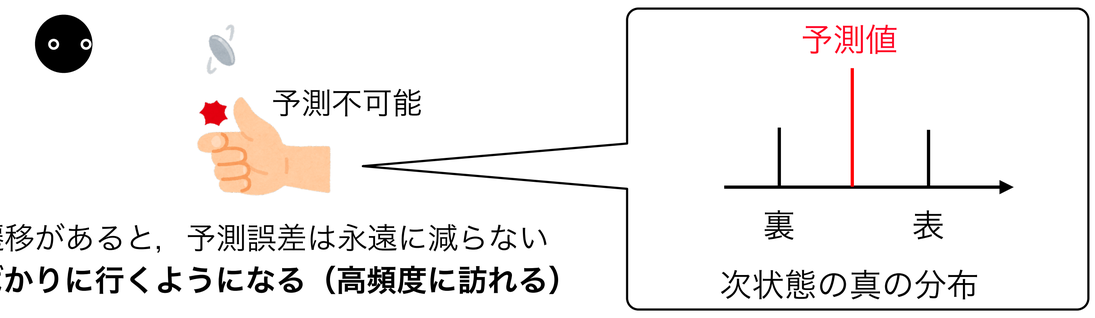
$$r_t \xrightarrow{\text{変更}} r_t + i_t$$

元々の報酬　予測誤差による内部報酬

・**狙い：未知の状態への到達　⇒　遷移の予測が難　⇒　予測誤差が大　⇒　未知の状態への誘引**
　既知の状態への到達　⇒　遷移の予測が易　⇒　予測誤差が小　⇒　何も起こらない

【問題点】根本的に予測できない場面で無限に内部報酬が出る

・根本的に予測できない例：迷路の中にコイントスを見せる障害物



・確率的な遷移があると，予測誤差は永遠に減らない
　⇒　**そこばかりに行くようになる（高頻度に訪れる）**

【新しいアイデア】高頻度に訪れる場所の内部報酬を遮断する

・ゲート関数を掛けて，高頻度に訪れるようになってしまった場所での報酬を遮断

新しいアイデア

$$r_t + i_t \xrightarrow{\text{変更}} r_t + \Phi i_t$$

先行研究の報酬　高頻度に訪れた場所での報酬を遮断するΦを追加

$$\Phi = \begin{cases} 0 & \text{訪問が高頻度} \\ 1 & \text{otherwise} \end{cases}$$

ゲート関数

定式化

内部報酬の定義と解釈

現在までに取得した状態 s_t, 行動 a_t, 次状態 s_{t+1} の組のデータ集合 D について，パラメータ θ を持つ，学習器 f_θ（ニューラルネットなど）を訓練する.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \mathbb{E}_{\{s_t, a_t, s_{t+1}\} \in D} ||s_{t+1} - f_\theta(s_t, a_t)||_2^2$$

新しい状態についての内部報酬 i_t は，トレードオフパラメータ σ を用いて

$$i_t = \frac{1}{2\sigma^2} ||s_{t+1} - f_{\hat{\theta}}(s_t, a_t)||_2^2$$

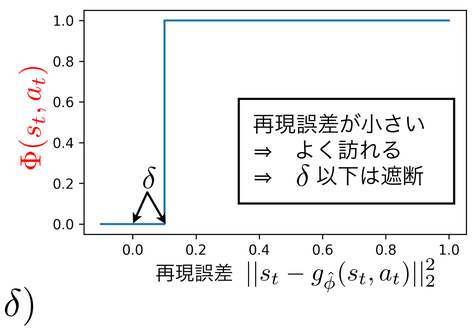
なお，この内部報酬 i_t は，これを $\Sigma = \sigma \boldsymbol{I}$ の正規分布を仮定した平均値推定と考えた場合のシャノン情報量に比例する[3].

$$-\log p(s_{t+1}) = \frac{1}{2\sigma^2} ||s_{t+1} - \hat{\mu}||_2^2 + const \propto i_t$$

ゲート関数の定義

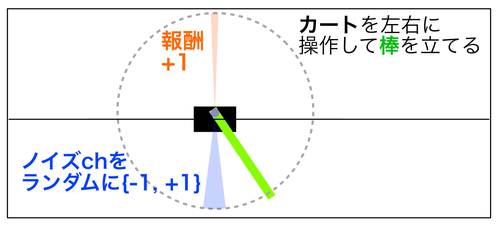
現在までに取得したデータについて，**再現誤差**が小さくなるように学習器 g_ϕ を訓練し，データによく登場する場面の誤差が小さくなることを期待して，訪問頻度を測る．ゲート関数は，ステップ関数 U，閾値 δ を用いて以下のように定義する.

$$\Phi(s_t, a_t) = U(||s_t - g_{\hat{\phi}}(s_t, a_t)||_2^2 - \delta)$$



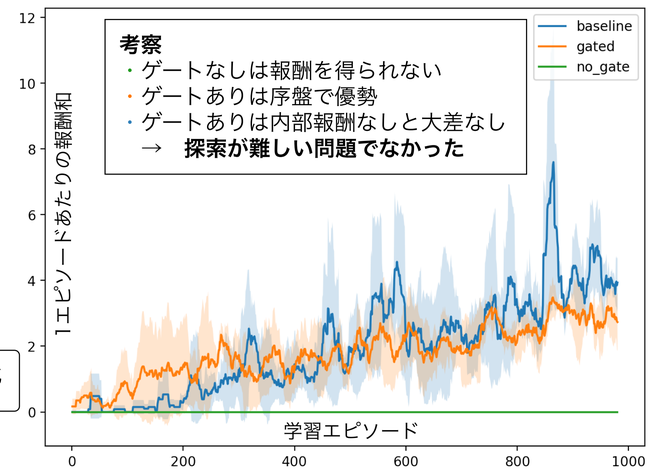
簡易実験

Noisy cartpole



・状態空間 $\mathcal{R}^5$
　位置, 速度, 角度, 角速度, ノイズch
・行動空間 $\{-1, 1\}$
　左右に力を加える
・強化学習手法
　DoubleDQN (4層MLP)

下部で確率的に変化



※それぞれ３回試行した平均，標準偏差を表示(20エピソードの移動平均)

展望

・Noisy cartpoleではなく，もう少し探索が難しい実験設定で検証する.
・問題によって誤差の大きさが変わるため，閾値 δ を適切に決める方法を考える.
・“高頻度に訪れる”の評価基準を見直す[4].

＜参考文献＞
[1] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, Wojciech Zaremba. Hindsight Experience Replay. (2017)
[2] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... & Chen, Y. Mastering the game of go without human knowledge. Nature, 550(7676), 354. (2017)
[3] Yuri Burda, Amos Storkey, Harri Edwards, Trevor Darrell, Deepak Pathak, Alexei A. Efros. Large-Scale Study of Curiosity-Driven Learning. (2018)
[4] Yuri Burda, Harrison Edwards, Amos Storkey, Oleg Klimov. Exploration by Random Network Distillation.(2018)