2021年6月18日　統計数理研究所　オープンハウス

利得がアームを引く間隔に依存する場合のバンディットアルゴリズム

谷本　悠斗　　総合研究大学院大学　統計科学専攻　博士課程(5年一貫制)3年

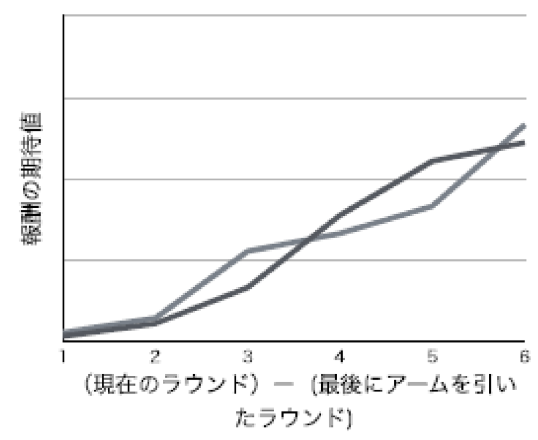
多腕バンディット問題

- 目標：リグレット最小化(<=>累積報酬最大化)
  - リグレット $= \underbrace{E[\sum_{t=1}^{T} r_t^*]}_{Oracle\ policy} - \underbrace{E[\sum_{t=1}^{T} r_t]}_{Learner}$
- 探索と活用のトレードオフ
  - 探索：各アームの報酬の情報収集
  - 活用：探索での情報を基に累積報酬を最大化

- アームが定常（報酬の期待値がroundに関わらず一定）
  ➡活用時に同じアームばかり引く
  - 満たさない例：推薦システム（同じ商品ばかり勧めてしまう）

報酬の構造

- （非定常な）報酬の構造
  - **アームiの報酬の期待値は最後にアームiを引いた時点から経過したラウンド数が長いほど大きい**
    - アームを引く ➡ そのアームの期待値は低下
    - アームを引かない ➡ そのアームの期待値は上昇

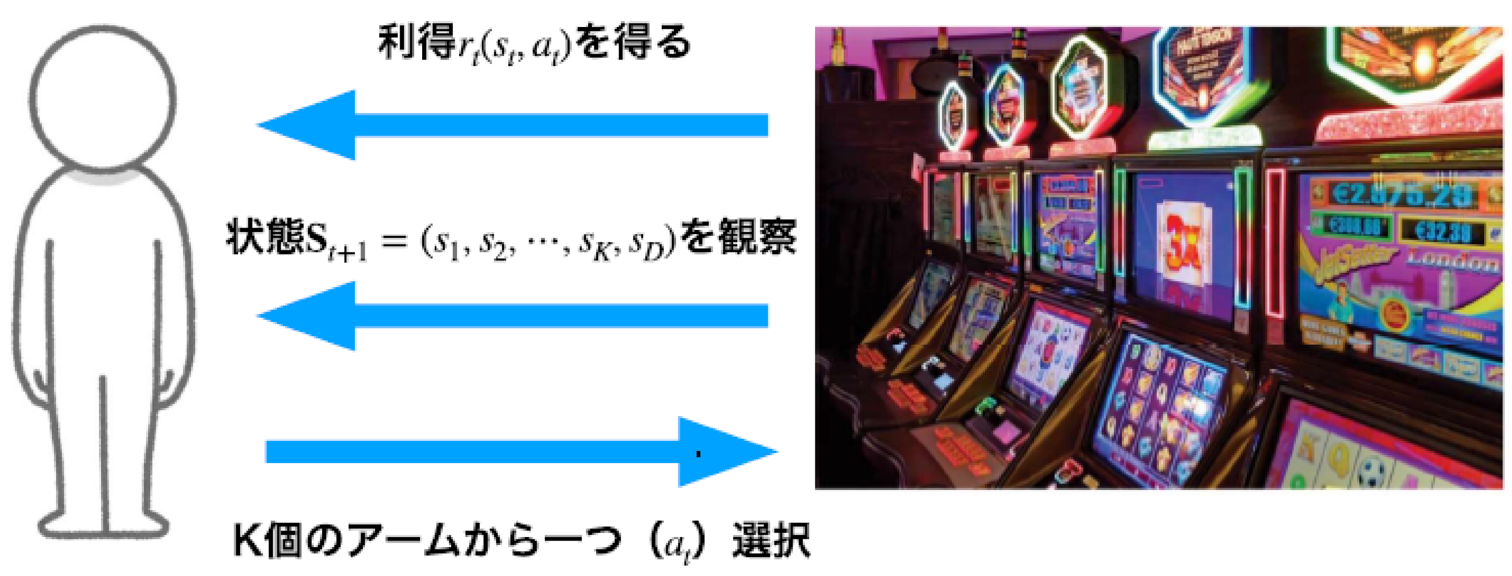


- 応用例：商品の推薦
  - 商品を購入 → 続けて同じ商品を勧められても購入しない
  - 推薦せずに時間が経過 → 再び商品を購入する可能性が上昇

アームを引かない選択肢の追加

- アームを引かない選択肢の導入
  - アームを引かないことで全てのアームの報酬の期待値が上昇
    ⇒ 累積報酬の最大化につながる可能性
- ダミーアームの導入
  - **通常のアームに加えて報酬が０の定常なアームを追加**
  - アームを引かずにスキップすることと同じ

強化学習によるアプローチ

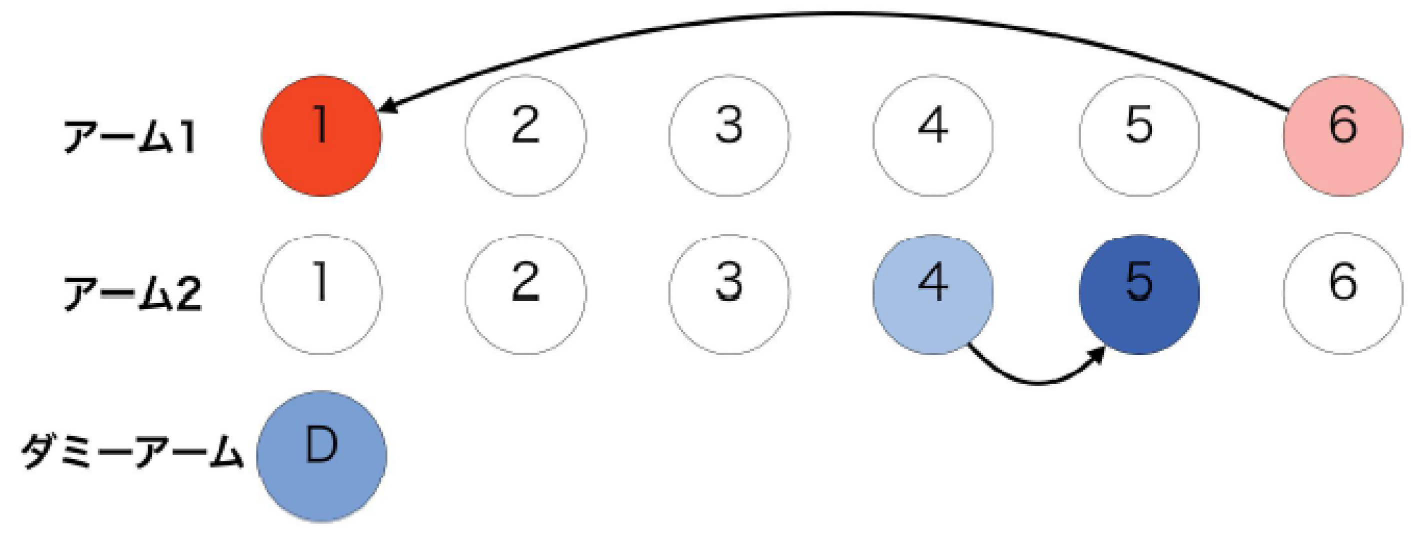


状態(state)の推移/利得

- **最後に引いた時点からの経過ラウンド数を（強化学習における）状態とみなす**
- （ダミーでない）アームがK個の時のラウンドtの状態 $\mathbf{S}_t = (s_t^1, s_t^2, \cdots, s_t^K, s_D)$
  - ダミーでないアームiの状態
    - s_t^i =min{s_{max},（現在のラウンド）ー (最後にアームを引いたラウンド)}
      - 各アームの状態の上限値をs_{max}で表す
  - ダミーアームの状態：常に一定（＝Dとおく）
- 状態の遷移（例：次ページの図）
  - ダミーでないアームaを引く→ $s_{t+1}^a = 1,\ s_{t+1}^k = min\{s_{max}, s_t^k + 1\}$ for $k \neq a$
  - ダミーアームを引く → $s_{t+1}^k = min\{s_{max}, s_t^k + 1\}$ for all k
- （有界な）利得 $r_t(s_t, a_t)$: **状態のインデックスに対して単調増加**(i. e. $r_t(s_t, a) \leq r_t(s_t + 1, a)$)

5

状態遷移の例

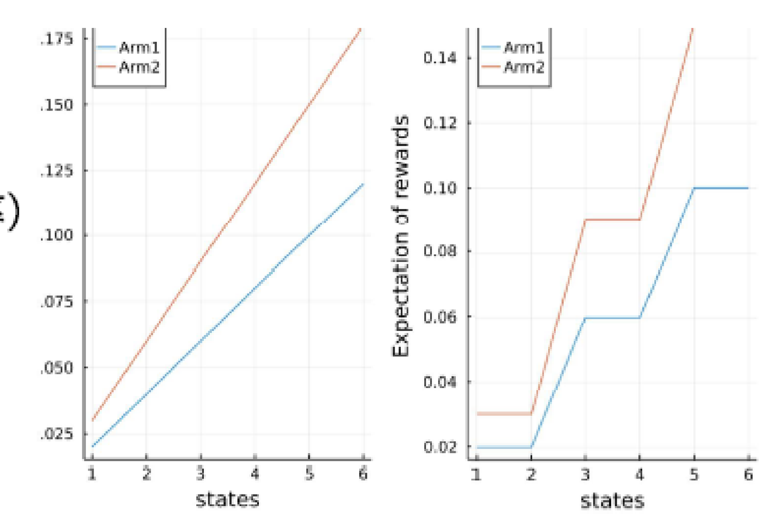


- ダミーでないアームの数：2 , $s_{max} = 6$
- $\mathbf{S}_t = (S_t^1, S_t^2, S_D) = (6,4,D) \rightarrow$ アーム$1(a_1)$を選択→ 利得$r_t(6,1)$ を得る
  $\rightarrow \mathbf{S}_{t+1} = (S_{(t+1)}^1, S_{(t+1)}^2, S_D) = (1,5,D)$

6

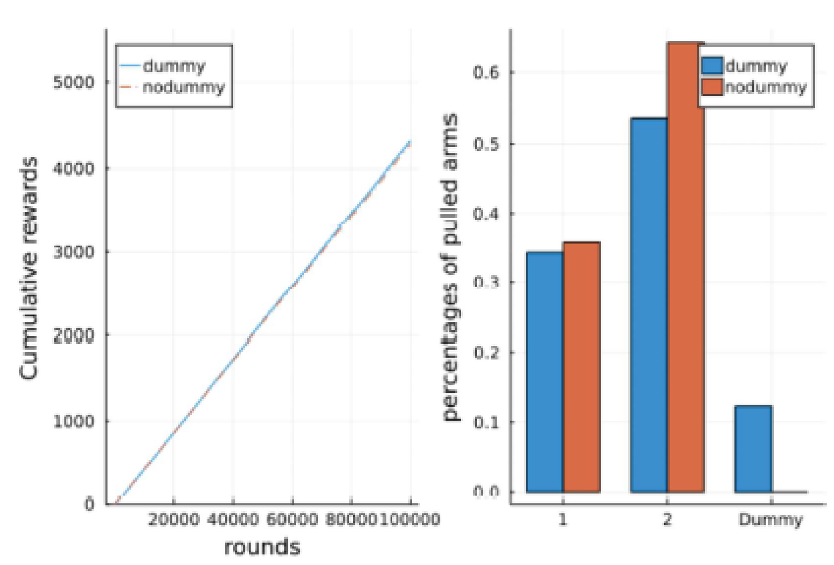
シミュレーション

- 状態数、アーム数
  - ダミーでないアームの数 : 2 ,　s_{max} = 3,6
  - ラウンド数 ＝10^5
- 報酬の期待値の構造
  - ベルヌーイ分布（例：クリック率）
    - 線形
    - 区分線形
- 強化学習アルゴリズム: Q学習
  - 方策(探索)：焼きまなしϵ-greedy

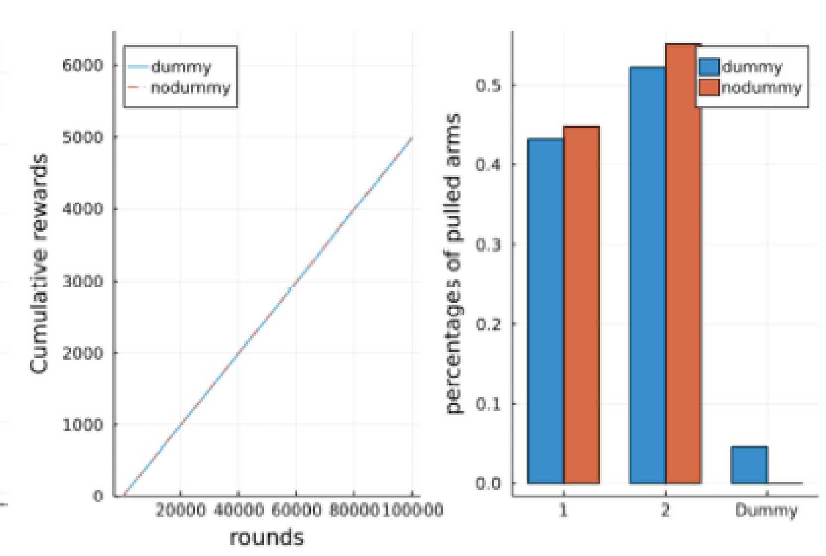


シミュレーションの結果（線形）

$s_{max} = 3$

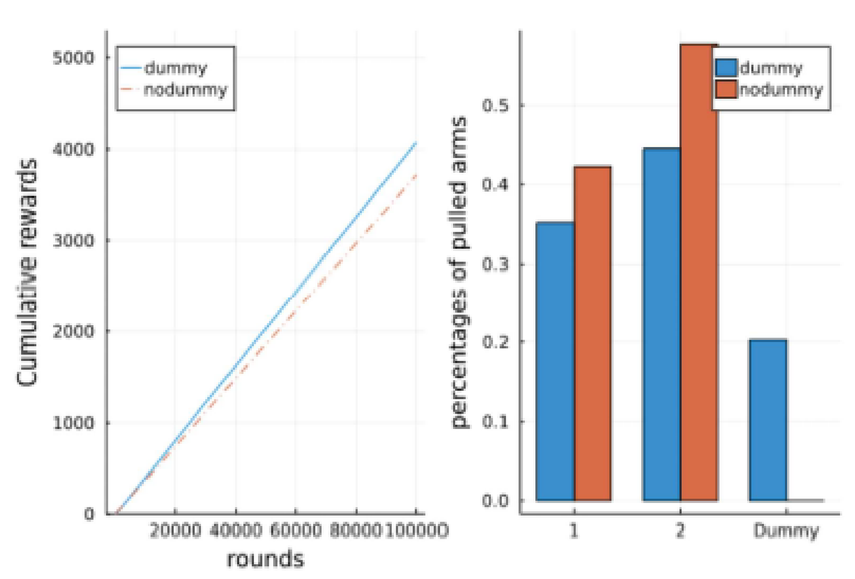


$s_{max} = 5$

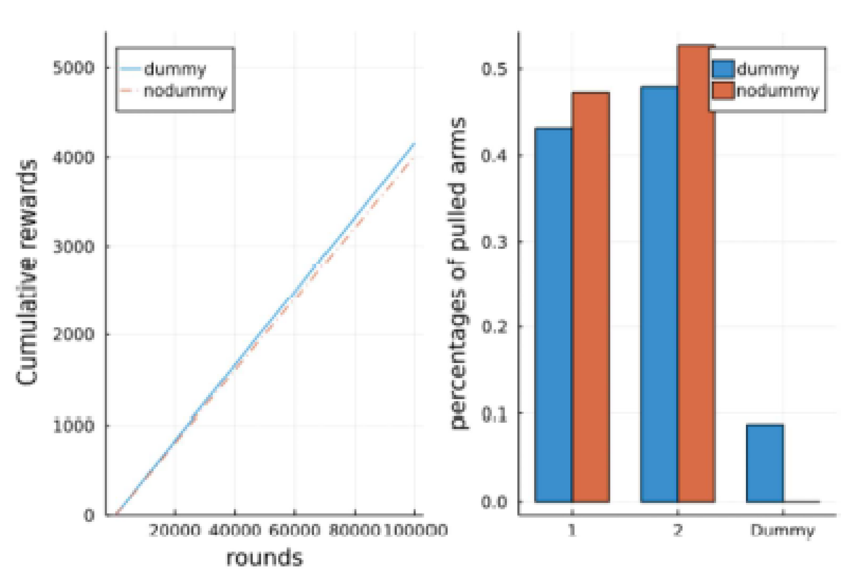


シミュレーションの結果（区分線形）

$s_{max} = 3$



$s_{max} = 5$



9

今後の課題

- 課題
  - 状態数orアーム数が多い時の価値関数の関数近似
  - 実データへの応用

10