Title	bandit問題の拡張について
Author(s)	中村, 篤祥
Citation	2010年度科学技術振興機構ERATO湊離散構造処理系プロジェクト講究録. p.295-297.
Issue Date	2011-06
Doc URL	http://hdl.handle.net/2115/48414
Type	conference presentation
Note	ERATO 湊離散構造処理系プロジェクト春のワークショップ（キックオフシンポジウム）. 2010年5月28日（金）～29日（土）. ERATO湊プロジェクト研究室.
File Information	04.nakamura_06.pdf

bandit 問題の拡張について

北海道大学情報科学研究科

中村篤祥

multi-armed bandit 問題とは

スロットマシンID 1 2 . . . K

時刻tにおける報酬(reward) $x_1(t)$ $x_2(t)$ $x_K(t)$ ← playerは知らない(選んだスロットマシンについてのみ知ることができる)

各時刻t(=1,2,…)においてplayerは以下のことを行う。

1. K台のスロットマシンから1台のスロットマシンi_tを選ぶ。
2. 選ばれたスロットマシンi_tから報酬$x_{i_t}(t)$を得る。

総利得 $\sum_t x_{i_t}(t)$ を最大化するためには各時刻においてどのようにスロットマシンを選べばよいか？

確率的なmulti-armed bandit 問題

スロットマシンID 1 2 . . . K

成功確率 θ_1 θ_2 θ_K ← playerは知らない 時刻によらず一定

報酬 $x_1(t)$ $x_2(t)$ $x_K(t)$ ← $x_i(t)=\begin{cases}1 & \text{if success}\\ 0 & \text{if fail}\end{cases}$

- $x_1(t), x_2(t),\ldots,x_K(t)$は独立
- $x_i(1), x_i(2),\ldots$は未知の成功確率θ_iのBernoulli process

目標1 expected total discounted reward $\sum_{t=1}^{\infty}\gamma^{t-1}E(x_{i_t}(t))$ の最大化 $(0<\gamma<1)$

目標2 与えられた時刻Tまでの総利得 $\sum_{t=1}^{T}x_{i_t}(t)$ の最大化

確率的なmulti-armed bandit問題の戦略

・Gittins Indexが最大のスロットマシンを選ぶ

成功、失敗回数がα、βであるようなスロットのGittins Index $G(\alpha,\beta)$

以下のような確率p：成功確率がpであるとわかっているスロットマシン1と、(成功回数,失敗回数)=(α,β)であるような成功確率未知のスロットマシン2があったとき、時刻t=1にどちらのスロットマシンを選んでも、optimal expected total discounted rewardが同じになるようなp

expected total discounted rewardを最大化する戦略 [Gittins and Jones 1974]

・ε-greedy [Sutton & Barto 1998]

1－εの確率でそれまでの平均報酬が最大のマシンを選び、εの確率でランダムに選ぶ。

・UCB1 [Auer, Cesa-Bianchi and Fisher 2002]

囲碁プログラムのモンテカルロ木探索に用いられて著名になった。

時刻tに $\bar{x}_j+\sqrt{\dfrac{2\ln t}{n_j}}$ が最大のマシンjを選ぶ。

$\bar{x}_j$:マシンjのそれまでの平均報酬

n_j :マシンjをそれまでに選んだ回数

adversarial bandit 問題 [Auer et al. 2002]

スロットマシンID 1 2 . . . K

$x_i(t)\in[0,1]$ for all i=1,2,…,K

時刻tにおける報酬(reward) $x_1(t)$ $x_2(t)$ $x_K(t)$ ← 悪魔が選ぶ。playerは知らない

各時刻t(=1,2,…,T)においてplayerは以下のことを行う。

1. K台のスロットマシンから1台のスロットマシンi_tを選ぶ。
2. 選ばれたスロットマシンi_tから報酬$x_{i_t}(t)$を得る。

$G_A(T)=\sum_{t=1}^{T}x_{i_t}(t)$ ：(乱択)アルゴリズムの時刻Tにおける総利得

期待総利得 $E(G_A(T))$ が大きなアルゴリズムAは？

乱択アルゴリズムExp3 [Auer et al. 2002]

Algorithm Exp3 // **exp**onential-weight algorithm for **exp**loration and **exp**loitation

Parameter: $\gamma\in(0,1]$

Initialization: $w_i(1)\leftarrow 1$ for i=1,2,…,K

for t=1 to T

1. $p_i(t)\leftarrow(1-\gamma)\dfrac{w_i(t)}{\sum_{j=1}^{k}w_j(t)}+\dfrac{\gamma}{K}$ for $i=1,\ldots,K$
2. i_tを$p_1(t),\ldots,p_K(t)$の分布に従ってランダムに選ぶ
3. 報酬 $x_{i_t}(t)\in[0,1]$ を得る
4. for j=1,…,k

$$\hat{x}_j(t)\leftarrow\begin{cases}x_j(t)/p_j(t) & \text{if } j=i_t\\ 0 & \text{otherwise}\end{cases}$$

$$w_j(t+1)\leftarrow w_j(t)\exp(\gamma\,\hat{x}_j(t)/K)$$

Exp3のexpected (weak) regret [Auer et al. 2002]

1つのスロットマシンを選び続けた場合の総利得の最大値 $G_{max}(T)=\max_j \sum_{t=1}^{T} x_j(t)$

と比べて平均的にどのくらい後悔するかを$G_{max}(T)-E[G_{Exp3}(T)]$で測ると

$$G_{max}(T)-E[G_{Exp3}(T)] \leqq 2.63\sqrt{TK\ln K} \quad \left(\gamma=\min\left\{1,\sqrt{\frac{K\ln K}{(e-1)T}}\right\}のとき\right)$$

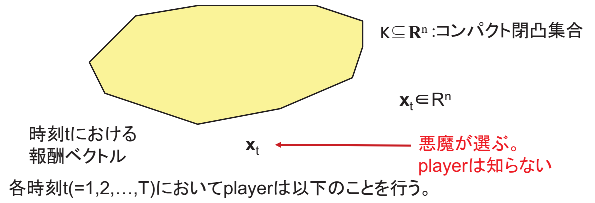
（$x_j(t)\in[0,1]$なので、$G_{max}(T)-G_{Exp3}(T)\leqq T$が成り立つことに注意。）

また、ある報酬割当分布が存在し、どのような乱択アルゴリズムAに対しても

$$E[G_{max}(T)-G_A(T)] \geq \frac{1}{20}\min\left\{\sqrt{TK},T\right\}$$

が成り立つ。ただし、期待値は報酬割当とアルゴリズムの乱択の両方に関してとるものとする。

online linear optimization 問題 [McMahan and Blum 2004]



各時刻t(=1,2,…,T)においてplayerは以下のことを行う。

1. Kから1点$\mathbf{q}_t$を選ぶ
2. 選ばれた点$\mathbf{q}_t$に対する利得 $\mathbf{x}_t\cdot\mathbf{q}_t$ を得る。

$G_A(T)=\sum_{t=1}^{T} x_t\cdot q_t$ ：(乱択)アルゴリズムの時刻Tにおける総利得

期待総利得 $E(G_A(T))$ が大きなアルゴリズムAは？

AbernethyらのアルゴリズムAbernethyらのアルゴリズム [Abernethy et al. 2008]

Algorithm BOLO (仮称) // Bandit Online Linear Optimization
Input: η>0, θ-self-concordant R
Initialization: $\mathbf{q}_1 \leftarrow \mathrm{argmin}_{\mathbf{q}\in K} R(\mathbf{q})$
for t=1 to T

1. $\{e_1,e_2,\ldots,e_n\}\leftarrow\nabla^2 R(\mathbf{q}_t)$の固有ベクトル
   $\{\lambda_1,\lambda_2,\ldots,\lambda_n\}\leftarrow\nabla^2 R(\mathbf{q}_t)$の固有値
2. i_tを{1,2,…,n}からランダムに選ぶ。
   ε_tを{−1,1}からランダムに選ぶ。
3. $\mathbf{r}_t \leftarrow \mathbf{q}_t+\varepsilon_t\lambda_{i_t}^{-1/2}e_{i_t}$
4. 報酬 $\mathbf{x}_t\cdot\mathbf{r}_t$ を得る。
5. 次のように更新する。

$$\hat{\mathbf{x}}_t \leftarrow n(\mathbf{x}_t\cdot\mathbf{r}_t)\varepsilon_t\lambda_{i_t}^{1/2}e_{i_t}$$

$$\mathbf{q}_{t+1} \leftarrow \mathrm{argmin}_{\mathbf{q}\in K} -\sum_{s=1}^{t}\hat{\mathbf{x}}_t\cdot\mathbf{q}+R(\mathbf{q})$$

BOLOのexpected regretの上界

$$\max_{\mathbf{q}\in K} E\left(\sum_{t=1}^{T}\mathbf{x}_t\cdot\mathbf{q}\right)-E(G_{BOLO}(T))=O\left(n\sqrt{\theta T\ln T}\right)$$

$$\left(\eta=\frac{\sqrt{\theta\ln T}}{4n\sqrt{T}}, T\geq 8\theta\ln Tのとき\right)$$

インターネット広告はbandit問題

スポーツページには
どの広告を出したら
T回の表示で最も多
くクリックされるか？

player=広告配信システム

multiple play設定

広告枠は複数

1回のプレイでk個選択

FExp3-k (Exp3のmultiple play 拡張版)

Algorithm FExp3
Parameter: $\gamma \in (0,1]$
Initialization: $w_i(1) \leftarrow 1$ for i=1,2,…,K

for t=1 to T

1. if $\mathrm{argmax}_{j\in\{1,2,\dots,K\}} w_j(t) \ge \left(\frac{1}{k} - \frac{\gamma}{K}\right)\sum_{i=1}^{K} w_i(t) \Big/ (1-\gamma)$ then

$$\frac{\alpha_t}{\sum_{w_i(t)\ge\alpha_t} \alpha_t + \sum_{w_i(t)<\alpha_t} w_i(t)} = \left(\frac{1}{k} - \frac{\gamma}{K}\right) \Big/ (1-\gamma)$$ を満たすα_tをもとめる。

$$S_0(t) \leftarrow \{i : w_i(t) \ge \alpha_t\}, \quad w'_i(t) \leftarrow \alpha_t \quad \text{for} \quad i \in S_0(t)$$

   else
      $S_0 \leftarrow \varphi$
2. $w'_i(t) \leftarrow w_i(t)$ for $i \in \{1,2,\dots,K\} - S_0(t)$
3. $$p_i(t) \leftarrow k\left((1-\gamma)\frac{w'_i(t)}{\sum_{j=1}^{K} w'_j(t)} + \frac{\gamma}{K}\right) \quad \text{for } i = 1,\dots,K$$

FExp3-k つづき(Exp3のmultiple play 拡張版)

4. S(t)←iが選ばれる確率が$p_i(t)$ になるように{1,2,…,K}からk個選択
5. 報酬$x_i(t) \in [0,1]$ for $i \in S(t)$を得る
6. for j=1,…,k

$$\hat{x}_j(t) \leftarrow \begin{cases} x_j(t)/p_j(t) & \text{if } j \in S(t) \\ 0 & \text{otherwise} \end{cases}$$

$$w_j(t+1) \leftarrow \begin{cases} w_j(t)\exp(k\,\gamma\,\hat{x}_j(t)/K) & \text{if} \quad j \notin S_0(t) \\ w_j(t) & \text{otherwise} \end{cases}$$

FExp3のexpected (weak) regret

同じk個のスロットマシンを選び続けた場合の総利得の最大値

$$G_{\max-k}(T) = \max_{S\subseteq\{1,2,\dots,K\},|S|=k} \sum_{j\in S}\sum_{t=1}^{T} x_j(t)$$ と比べると

$$G_{\max-k}(T) - E[G_{FExp3}(T)] \leqq 2.63\sqrt{kTK\ln\frac{K}{k}} \quad \left(\gamma = \min\left\{1, \sqrt{\frac{K\ln(K/k)}{(e-1)kT}}\right\}\text{のとき}\right)$$

また、ある報酬割当分布が存在し、どのような乱択アルゴリズムAに対しても

$$E[G_{\max-k}(T) - G_A(T)] \ge \min\left\{\frac{1}{5}\left(\frac{K-k}{K}\right)^2\sqrt{TK}, \frac{K-k}{8K}kT\right\}$$

が成り立つ。ただし、期待値は報酬割当とアルゴリズムの乱択の両方に関してとるものとする。

今後の拡張について

- 協調フィルタリングのbanditモデル
  ユーザの好みはいくつかの好みのタイプの混合として表現できると仮定して、最適な混合割合と比べた場合、リグレットが少ない方法を考案する。
- 情報検索のbanditモデル
  １回にk個ずつ結果が提示されるとしてT回までみるとした場合に最も多く欲しいものが含まれている結果提示方法は？適合度フィードバックしながら適度に良い結果を返す。低次元の学習問題にどのように落とすか。