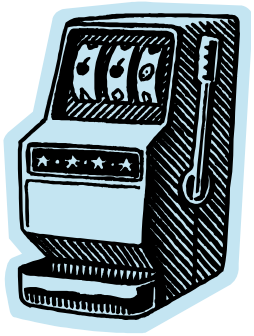
多腕バンディット問題の理論とアルゴリズム

本多淳也
東京大学 新領域創成科学研究科 助教

IBIS2014
2014/11/18

多腕バンディット問題 (multiarmed bandit problem)

- 複数台のスロットマシンをプレイするギャンブラーのモデル
- 得られる報酬の確率分布は台によって異なる
- なるべく期待値の高い台をプレイしたい



問題：

- ある程度の回数プレイしないと台の良し悪しが分からない
- 報酬の期待値の低い台を何度もプレイすると損

歴史

- 研究自体は1930年ごろからある
  - 農地への作付物の選択
  - 新薬や新療法の検証 (cf. one-armed bandit)

- 最近になって急速に応用範囲が拡大
  - 囲碁などのゲーム木探索
  - 広告・ニュース記事推薦
  - ネットワークルーティング
  - 学習アルゴリズムの選択や最適化パラメータのチューニング

概要

- バンディット問題の分類
- 報酬最大化のためのアルゴリズム
  - 理論限界
  - 決定的アルゴリズム：UCB アルゴリズム
  - ランダムアルゴリズム：Thompson sampling
- 最適腕識別との関係
- 種々の拡張

概要

- バンディット問題の分類
- 報酬最大化のためのアルゴリズム
  - 理論限界
  - 決定的アルゴリズム：UCB アルゴリズム
  - ランダムアルゴリズム：Thompson sampling
- 最適腕識別との関係
- 種々の拡張

バンディット問題の分類

- 報酬の性質
  - stochastic bandit（確率的）
  - adversarial bandit（敵対的）
- 目的関数
  - 累計報酬の最大化 (regret 最小化)
  - 期待値ベストの台の判定 (best arm identification)
- 設定の拡張
  - linear bandit
  - contextual bandit（context: 閲覧履歴・エキスパート…）
  - 連続空間上のバンディット

⋮

確率的バンディット

- K 台のスロットマシンを選んで計 n 回プレイ
- 各台からの報酬が期待値 μ_i の（例えば）ベルヌーイ分布にしたがう
- 実際にプレイした台についてのみサンプルの情報が得られる

確率的バンディット

- K 台のスロットマシンを選んで計 n 回プレイ
- 各台からの報酬が期待値 μ_i のベルヌーイ分布にしたがう

$\mu^* = \max_i \mu_i$

reward 最大化 (= regret 最小化):

- 期待値最大の台を知っていた場合の報酬期待値: $n\mu^*$
- それぞれ台 i を $T_i(n)$ 回選んだ場合の期待値: $\sum_i \mu_i T_i(n)$
- 目標:期待値の損失 (regret) の最小化

$$\mathrm{regret}(n) = \sum_i (\mu^* - \mu_i) T_i(n)$$

確率的バンディット

- K 台のスロットマシンを選んで計 n 回プレイ
- 各台からの報酬が期待値 μ_i のベルヌーイ分布にしたがう

最適腕識別 (best arm identification):

- n 回のプレイの後で期待値最大の台をなるべく高確率で判定 (fixed budget)
  - 確率 δ 以上で正解するための試行回数を最小化するという設定も（fixed confidence）
- ある種の能動学習の問題
- 理論解析の道具は regret 最小化と似ているが，実際にとるべき戦略は多少異なる

敵対的バンディット

- 各マシンからの報酬 $x_{i,t} \in [0,1]$ が adversary により敵対的に決められる

	1回目	2回目	3回目	4回目
台1				
台2				
台3				

敵対的バンディット

- 各マシンからの報酬 $x_{i,t} \in [0,1]$ が adversary により敵対的に決められる

	1回目	2回目	3回目	4回目
台1	0.8			
台2	0.3			
台3	0.4			

敵対的バンディット

- 各マシンからの報酬 $x_{i,t} \in [0,1]$ が adversary により敵対的に決められる

	1回目	2回目	3回目	4回目
台1	0.8			
台2	0.3			
台3	0.4			

敵対的バンディット

- 各マシンからの報酬 $x_{i,t} \in [0,1]$ が adversary により敵対的に決められる

	1回目	2回目	3回目	4回目
台1	0.8	0.5		
台2	0.3	0.9		
台3	0.4	0.1		

敵対的バンディット

- 各マシンからの報酬 $x_{i,t} \in [0,1]$ が adversary により敵対的に決められる

	1回目	2回目	3回目	4回目
台1	0.8	0.5		
台2	0.3	0.9		
台3	0.4	0.1		

敵対的バンディット

- 各マシンからの報酬 $x_{i,t} \in [0,1]$ が adversary により敵対的に決められる

	1回目	2回目	3回目	4回目
台1	0.8	0.5	0.2	
台2	0.3	0.9	0.1	
台3	0.4	0.1	0.4	

敵対的バンディット

- 各マシンからの報酬 $x_{i,t} \in [0,1]$ が adversary により敵対的に決められる

	1回目	2回目	3回目	4回目
台1	0.8	0.5	0.2	
台2	0.3	0.9	0.1	
台3	0.4	0.1	0.4	

敵対的バンディット

- 各マシンからの報酬 $x_{i,t} \in [0,1]$ が adversary により敵対的に決められる

	1回目	2回目	3回目	4回目
台1	0.8	0.5	0.2	0.3
台2	0.3	0.9	0.1	0.8
台3	0.4	0.1	0.4	0.6

敵対的バンディット

- 各マシンからの報酬 $x_{i,t} \in [0,1]$ が adversary により敵対的に決められる

	1回目	2回目	3回目	4回目
台1	0.8	0.5	0.2	0.3
台2	0.3	0.9	0.1	0.8
台3	0.4	0.1	0.4	0.6

敵対的バンディット

- 各マシンからの報酬 $x_{i,t} \in [0,1]$ が adversary により敵対的に決められる

	1回目	2回目	3回目	4回目	累計
台1	0.8	0.5	0.2	0.3	1.8
台2	0.3	0.9	0.1	0.8	2.1
台3	0.4	0.1	0.4	0.6	1.5

- Regret ＝ 同じ台を選び続けた場合の最大報酬 － 実際の報酬
  ＝ 2.1 － 1.5 ＝ 0.6
- 確率分布を仮定しなくていい

概要

- バンディット問題の分類
- 報酬最大化のためのアルゴリズム
  - 理論限界
  - 決定的アルゴリズム：UCB アルゴリズム
  - ランダムアルゴリズム：Thompson sampling
- 最適腕識別との関係
- 種々の拡張

理論限界

- ある台に決め打ちしたりしない「真っ当な」戦略のみ考えたい
  - 真っ当な戦略：ベストでない台のプレイ回数が $\mathrm{o}(\mathrm{Poly}(n))$
- このような戦略を用いた場合には

$$\mathrm{E}[T_i(n)] \geq \frac{\log n}{D(\mu_i \| \mu^*)} + \mathrm{o}(\log n)$$

ベルヌーイ分布の KL ダイバージェンス

(Lai & Robbins, 1985)

  - 分布 $\mathrm{Ber}(\mu_i)$ と $\mathrm{Ber}(\mu^*)$ の「紛らわしさ」でプレイ回数が決まる
- 逆にこれより良い性能を達成する（場合がある）戦略は設定次第で regret が多項式オーダーに

漸近最適戦略の構成

- 理論限界: $T_i(n) \geq \mathrm{O}(\log n)$
  - ベストでない台 i を引いてしまう1ラウンドごとの確率は $\mathrm{O}(1/n)$
  - 逆に $\mathrm{O}(1/n)$ 程度の確率でダメっぽい台を選べばよさそう
- UCB 戦略: 有意水準 $1/n$ での信頼区間の上限 (Upper Confidence Bound) が最大になる台をプレイ
  - Hoeffding の不等式: $\Pr[\hat{\mu}_i \leq x] \leq \mathrm{e}^{-2T_i(n)(\mu_i - x)^2}$
  - Hoeffding の不等式から導かれる UCB (Auer et al., 2002):

$$\mathrm{UCB}_i = \max\{\mu : \mathrm{e}^{-2T_i(n)(\mu - \hat{\mu}_i)^2} \geq 1/n\}$$

$$= \hat{\mu}_i + \sqrt{\frac{\log n}{2T_i(n)}}$$

漸近最適戦略の構成

- 理論限界: $T_i(n) \geq \mathrm{O}(\log n)$
  - ベストでない台 i を引いてしまう1ラウンドごとの確率は $\mathrm{O}(1/n)$
  - 逆に $\mathrm{O}(1/n)$ 程度の確率でダメっぽい台を選べばよさそう
- UCB 戦略: 有意水準 $1/n$ での信頼区間の上限 (Upper Confidence Bound) が最大になる台をプレイ
  - Chebycheff の不等式: $\Pr[\hat{\mu}_i \leq x] \leq \mathrm{e}^{-T_i(n)D(x\|\mu_i)}$
  - Chebycheff の不等式から導かれる UCB (KL-UCB) (Burnetas+, 1996, Garivier+, 2011):

$$\begin{aligned}\mathrm{UCB}_i &= \max\{\mu : \mathrm{e}^{-T_i(n)D(\hat{\mu}_i\|\mu)} \geq 1/n\} \\ &= \mu > \hat{\mu}_i \text{ s.t. } D(\hat{\mu}_i\|\mu) = (\log n)/T_i(n)\end{aligned}$$

UCB の性質

- KL-UCB は理論限界を達成
  - ベルヌーイ分布・分散既知の正規分布 (Burnetas+, 1996)
  - 1パラメーター指数型分布族の一部 (Garivier+, 2011)

問題点

- KL ダイバージェンスの逆関数を計算するのが若干面倒
  - ある種の双対アルゴリズムを導入すると逆関数の計算は不要 (DMED・IMEDアルゴリズム, Honda+, 2010, 2012)
    - $[0, 1]$ あるいは $(-\infty, 1]$ 上のノンパラメトリックモデルも可
- 信頼区間の有意水準 $1/n$ とかいう謎のパラメーター
  - 理論限界を知らないと出てこない
  - 完全に漸近論に依拠

Thompson sampling

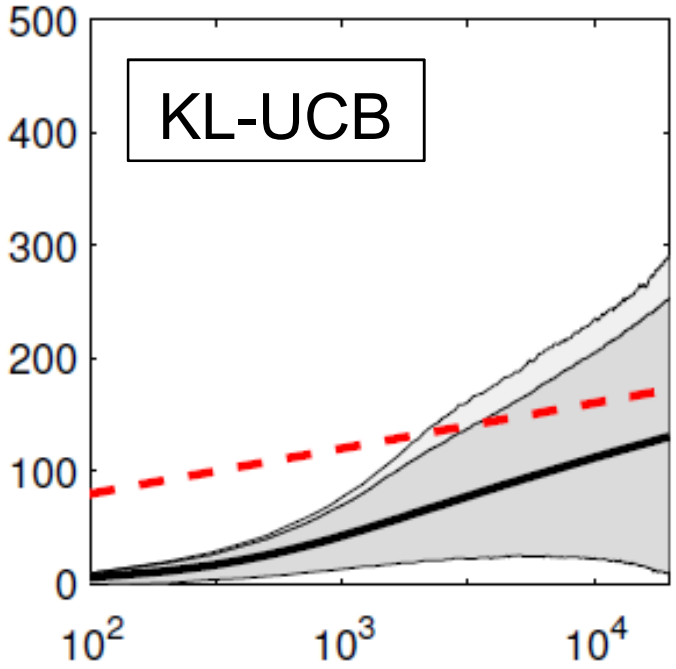
- ベイズ戦略の一種 (Thompson (1933) にちなむ)
- その台が最適である事後確率に従って台をランダムに選択

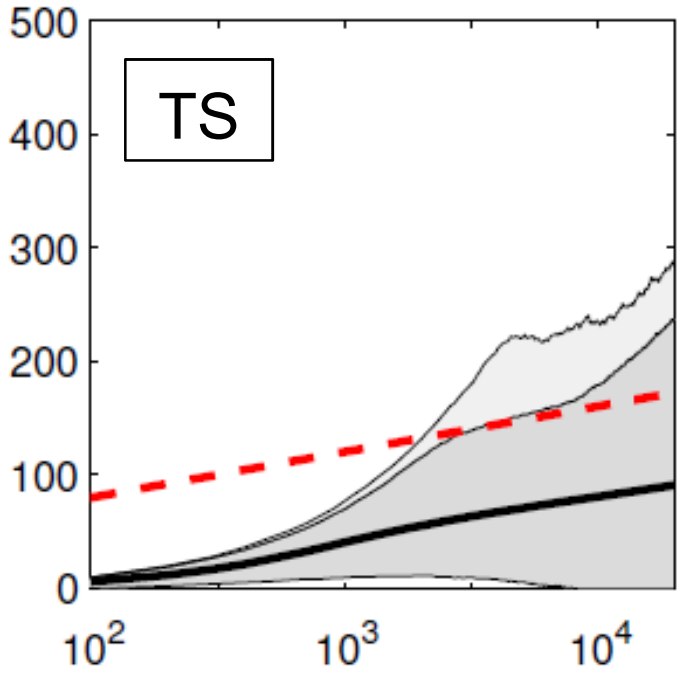
- 経験的には非常に高性能
- あまり事前分布の取り方にも依存しない
- (linear や contextual, 非定常の場合含めて) 事後分布が計算可能なモデルでさえあればとりあえず動かせる

- 最近になって理論解析が整備
  - ベルヌーイ分布モデル (Agrawal & Goyal, 2013 ほか)
  - 1変数指数型分布族 (Korda, Kaufmann & Munos, 2013)
  - 平均・分散未知の正規分布 (Honda & Takemura, 2014)

Thompson sampling の性能

- 大抵の場合にUCBよりも高性能
- 例：10個のスロットマシン，
  クリック率 10%×1台 ・ 5%×3台 ・ 2%×3台・1%×3台





[Kaufmann et al., 2012]

ランダムアルゴリズムの利点

- TS ではそれぞれの台を「その台がベストである事後確率」にしたがってランダムに選択
  - この確率は計算困難

$$\pi(\mu_i \geq \max_{j \neq i} \mu_j | \{\boldsymbol{x}_i\})$$

$$= \int_0^1 \pi(\theta_i | \boldsymbol{x}_i) \left(\prod_{j \neq i} \int_0^{\theta_i} \pi(\theta_j | \boldsymbol{x}_j) \mathrm{d}\theta_j \right) \mathrm{d}\theta_i$$

$\mu_i = \theta_i$ となる確率　　他の台で $\mu_j \leq \theta_i$ となる確率

  - 実際にはこの確率の計算は不要：
    各台から乱数 θ_i を事後分布 $\pi(\mu_i | \boldsymbol{x}_i)$ にしたがって生成して θ_i が最大となった台 i をプレイすれば同等な動作となる

TS の実装

- 「事後分布のもとでの各台が期待値ベストな確率」は実際には計算が困難
- 単に事後分布 $\pi(\mu_i|\boldsymbol{x}_i)$ にしたがって θ_i をそれぞれランダムにサンプリングし，それが最大になった台を選択すればよい
  - 例：事前分布 $\pi(\mu_i) = \mathrm{unif}(0,1)$ の場合

	台1	台2	台3
これまでの結果	1,0,1,1	0,0	1,0,1
μ_i の事後分布	$\mathrm{Be}(4,2)$	$\mathrm{Be}(1,3)$	$\mathrm{Be}(3,2)$
事後分布からのサンプル			

TS の実装

- 「事後分布のもとでの各台が期待値ベストな確率」は実際には計算が困難
- 単に事後分布 $\pi(\mu_i|\boldsymbol{x}_i)$ にしたがって θ_i をそれぞれランダムにサンプリングし，それが最大になった台を選択すればよい
  - 例：事前分布 $\pi(\mu_i) = \mathrm{unif}(0,1)$ の場合

	台1	台2	台3
これまでの結果	1,0,1,1	0,0	1,0,1
μ_i の事後分布	$\mathrm{Be}(4,2)$	$\mathrm{Be}(1,3)$	$\mathrm{Be}(3,2)$
事後分布からのサンプル	0.78	0.16	0.68

TS の実装

- 「事後分布のもとでの各台が期待値ベストな確率」は実際には計算が困難
- 単に事後分布 $\pi(\mu_i|\boldsymbol{x}_i)$ にしたがって θ_i をそれぞれランダムにサンプリングし，それが最大になった台を選択すればよい
  - 例：事前分布 $\pi(\mu_i) = \mathrm{unif}(0,1)$ の場合

	台1	台2	台3
これまでの結果	1,0,1,1,0	0,0	1,0,1
μ_i の事後分布	$\mathrm{Be}(4,3)$	$\mathrm{Be}(1,3)$	$\mathrm{Be}(3,2)$
事後分布からのサンプル			

TS の実装

- 「事後分布のもとでの各台が期待値ベストな確率」は実際には計算が困難
- 単に事後分布 $\pi(\mu_i|\boldsymbol{x}_i)$ にしたがって θ_i をそれぞれランダムにサンプリングし，それが最大になった台を選択すればよい
  - 例：事前分布 $\pi(\mu_i) = \mathrm{unif}(0,1)$ の場合

	台1	台2	台3
これまでの結果	1,0,1,1,0	0,0	1,0,1
μ_i の事後分布	$\mathrm{Be}(4,3)$	$\mathrm{Be}(1,3)$	$\mathrm{Be}(3,2)$
事後分布からのサンプル	0.67	0.19	0.81

TS の実装

- 「事後分布のもとでの各台が期待値ベストな確率」は実際には計算が困難
- 単に事後分布 $\pi(\mu_i|\boldsymbol{x}_i)$にしたがって θ_i をそれぞれランダムにサンプリングし，それが最大になった台を選択すればよい
  - 例：事前分布 $\pi(\mu_i) = \mathrm{unif}(0,1)$ の場合

	台1	台2	台3
これまでの結果	1,0,1,1,0	0,0	1,0,1,1
μ_i の事後分布	$\mathrm{Be}(4,3)$	$\mathrm{Be}(1,3)$	$\mathrm{Be}(4,2)$
事後分布からのサンプル			

TS の実装

- 「事後分布のもとでの各台が期待値ベストな確率」は実際には計算が困難
- 単に事後分布 $\pi(\mu_i|\boldsymbol{x}_i)$ にしたがって θ_i をそれぞれランダムにサンプリングし，それが最大になった台を選択すればよい
  - 例：事前分布 $\pi(\mu_i) = \mathrm{unif}(0,1)$ の場合

	台1	台2	台3
これまでの結果	1,0,1,1,0	0,0	1,0,1,1
μ_i の事後分布	$\mathrm{Be}(4,3)$	$\mathrm{Be}(1,3)$	$\mathrm{Be}(4,2)$
事後分布からのサンプル	0.58	0.61	0.77

TS の意味付け

- 期待値ベストでない台を選んでしまった回数を最小化したい
- 台 i の損失への寄与：
  台 i がベストな台である事後確率 p_i に対して

台 i を選んだ回数

$$\mathrm{E}[(\text{誤選択の回数})|\{\boldsymbol{x}_i\}] = (1-p_i)\times n_i + p_i \times (n-n_i)$$

台 i がベストでない場合　台 i がベストな場合

- 2つの項がバランスするのは
  $(1-p_i)\times n_i = p_i \times (n-n_i) \Leftrightarrow n_i = p_i n$ のとき
  - 台 i を計 n 回中 $p_i n$ 回プレイしているのが理想的な状態
  - 確率 p_i で台 i をプレイすればよい → Thompson sampling

UCB との関係 (1/2)

Thompson sampling

- 台 i が選ばれる確率＝台 i がベストである事後確率 p_i
- 台 i が次にプレイされるまでの平均待ち時間 $1/p_i$
- TS ≒ 前回のプレイから $1/p_i$ ラウンド経過したらその台をプレイ

(1回プレイ後の待ち時間≪2回プレイ後の待ち時間≪ … なので)

≒ 累計で $1/p_i$ ラウンド経過したらその台をプレイ

＝ 期待値ベストである事後確率が $p_i \geq 1/n$ ならプレイ

UCB 戦略

- UCB ＝ 有意水準 $1/n$ で見積もった期待値が大きい台をプレイ

≒ 期待値ベストである尤度が $1/n$ 以上の台をプレイ

UCB との関係 (2/2)

- Thompson sampling も UCB も(大雑把には)期待値ベストである尤度(あるいは事後確率)が $1/n$ 以上の台を選択
  - KL-UCB ではこの尤度を Chebycheff の不等式

$$\Pr[\hat{\mu}_i \leq x] \leq \mathrm{e}^{-T_i(n)D(x\|\mu_i)}$$

    で大きめに近似
  - TSでは近似でなく正確な確率がベースになっているので余分な探索がない
    - 実は漸近展開の知識を使うと

$$\Pr[\hat{\mu}_i \leq x] \approx \frac{c}{\sqrt{T_i(n)}}\mathrm{e}^{-T_i(n)D(x\|\mu_i)}$$

    - だったらUCBでも同種の改良ができる？

UCB のチューニング (1/2)

- UCBでは「有意水準を $1/n$ から $T_i(n)/n$ に置き換えると(**何が何だかよく分からないけど**)性能が上がるらしい」というテクニックが昔からまことしやかに伝わっている
  (Audibert+, 2010, Garivier+, 2011)
  - 元のKL-UCBスコア:

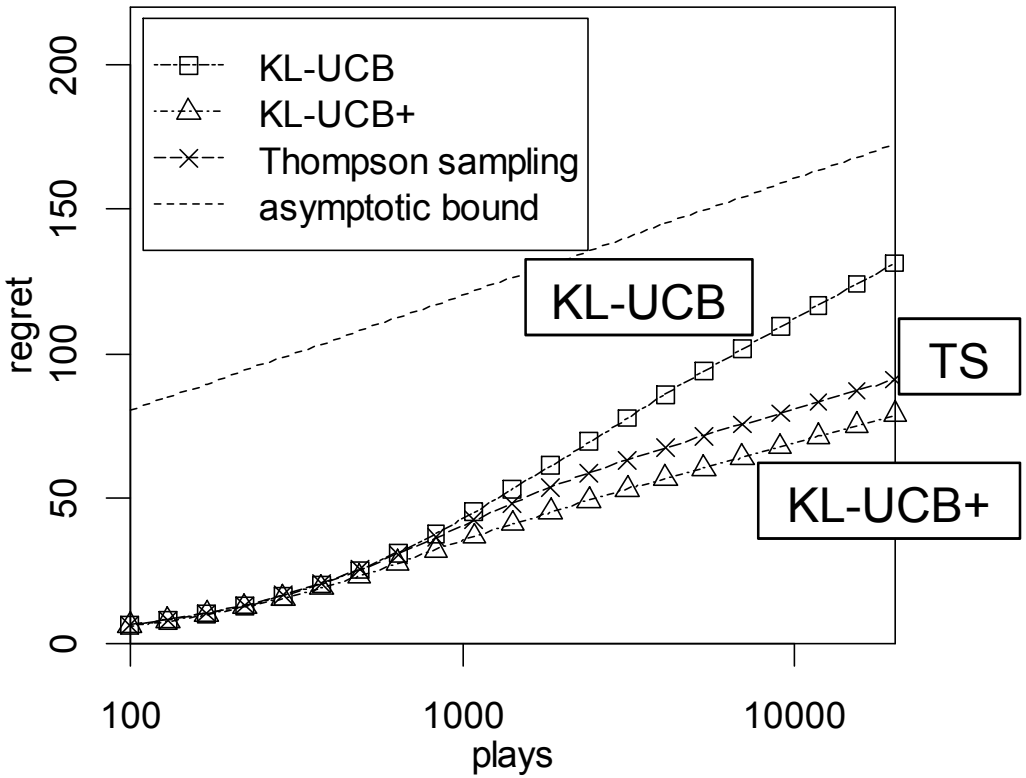
$$\mathrm{UCB}_i = \max\{\mu : \mathrm{e}^{-T_i(n)D(\hat{\mu}_i\|\mu)} \geq 1/n\}$$

  - チューニングを行った場合(KL-UCB+):

$$\mathrm{UCB}_i = \max\{\mu : \mathrm{e}^{-T_i(n)D(\hat{\mu}_i\|\mu)} \geq T_i(n)/n\}$$
$$= \max\{\mu : (T_i(n))^{-1}\mathrm{e}^{-T_i(n)D(\hat{\mu}_i\|\mu)} \geq 1/n\}$$

UCB のチューニング (2/2)

- TS は KL-UCB+ と KL-UCB の中間くらいの性質に
- KL-UCB+ も理論限界を達成 (Kaufmann, 2014)



- 一方で TS の事前分布の調整は確率の指数部を

$\exp(-(T_i(n)+\alpha)D(\hat{\mu}_i\|\mu))$ とすることにほぼ対応

（→大きな差は出にくい）

チューニングと最適性の関係

- KL-UCB では真の期待値 μ の尤度をパラメーター c を用いて

$$(T_i(n))^{-c} \exp(-T_i(n) D(\hat{\mu}_i \| \mu))$$

と見積もることで性能をチューニング可

- Thompson sampling は $c = 1/2$ に近い

- 疑問:
  - TS では事前分布は全く気にしないでも大丈夫？
  - むしろ UCB のチューニングもいつでもOK？
- 答え:どちらもダメな場合がある
  - ベルヌーイ分布はパラメータが1個かつ有界な特殊ケース

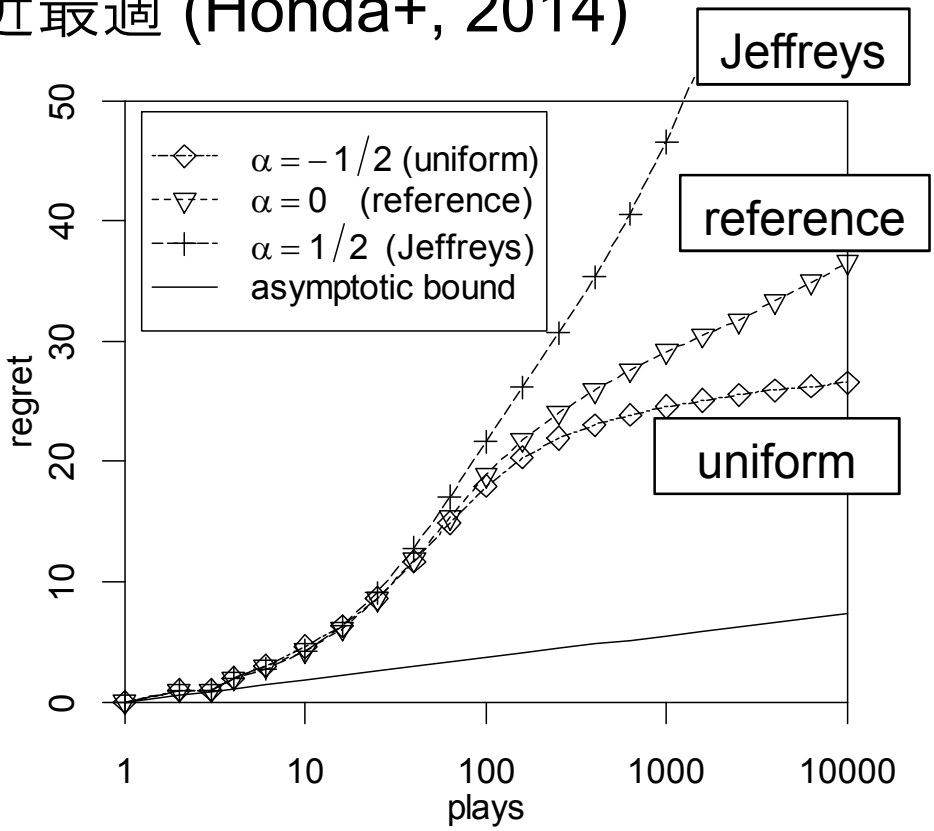
正規分布モデルの Thompson sampling

- 分散既知：Jeffreys 事前分布で漸近最適 (Korda+, 2013)
- 分散未知：reference・Jeffreys 事前分布ともにダメ
  一様事前分布なら漸近最適 (Honda+, 2014)

- 例：$(\mu_1, \mu_2) = (1, 0)$

  $(\sigma_1, \sigma_2) = (3.0, 0.3)$

  事前分布 $\pi(\mu_i, \sigma_i) \propto \sigma^{-1-2\alpha}$
  - α が大きいと探索減
- 少ない試行回数では
  逆の結果になりやすい



概要

- バンディット問題の分類
- 報酬最大化のためのアルゴリズム
  - 理論限界
  - 決定的アルゴリズム：UCB アルゴリズム
  - ランダムアルゴリズム：Thompson sampling
- 最適腕識別との関係
- 種々の拡張

regret 最小化との違い

- トータル N 回のプレイで期待値最大の台を高確率で判定したい

- 基本的な方針：ダメそうな台のプレイは早めに打ち切って
  良さそうな台の比較に専念
- よくある間違い：深く考えず regret 最小化のための戦略を
  そのまま使ってしまう
  - 通常の UCB や Thompson sampling は期待値最大で
    ない台を $\mathrm{O}(\log N)$ 回程度しかプレイしない
    → プレイ数の偏りが大きすぎる
  - 例えば $K = 2$, 分散既知の正規分布の場合には明らかに
    両者を同じ回数プレイするのがベスト

最適腕識別のための戦略

- UCB 等のアルゴリズムは有意水準を $1/n$ から $\mathrm{e}^{-\mathrm{O}(N)}$ にすることで容易に修正可能 (Audibert+, 2010 ほか)
- 現時点でベストな台については UCB よりも LUB (Lower Confidence Bound) を基準にするとうまくいきやすい
  （候補数が多い場合には枝刈りの基準としても有効）

- TS は有意水準に対応するものを自由に設定できない
  - 「期待値最大である事後確率 p_i が $\mathrm{e}^{-\mathrm{O}(N)}$ より大きい台を選択」等の修正は考えられるが，p_i の計算は大変
  - ただし，モデルが複雑になるとそもそも UCB でも信頼区間の構成が難しい場合があり，そのような場合には TS が相対的に高性能となる場合も

概要

- バンディット問題の分類
- 報酬最大化のためのアルゴリズム
  - 理論限界
  - 決定的アルゴリズム：UCB アルゴリズム
  - ランダムアルゴリズム：Thompson sampling
- 最適腕識別との関係
- 種々の拡張

時間発展のある場合

- 時刻 t とともに各台の報酬分布が変化する（ニュース推薦など）
- 時間発展のない場合の regret :

$$\mathrm{regret}(n) = \sum_{t=1}^{n} \left(\max_{i \in \{1, \cdots, K\}} \{\mu_i\} - \mu_{I_t} \right)$$

- 時間発展のある場合:

$$\mathrm{regret}(n) = \sum_{t=1}^{n} \left(\max_{i \in \{1, \cdots, K\}} \{\mu_{i,t}\} - \mu_{I_t,t} \right)$$

  - 時間発展のない場合に比べて本質的に難しい
  - 例) 確率分布が途中で m 回だけ変化する場合:
    regret が $\mathrm{O}(\sqrt{n})$ に (Garivier & Moulines, 2011)

Linear bandit

- 取り得る行動(マシン): $a \in D \subset \mathbb{R}^d$ (D: 有界集合)
  - プレイごとに d 次元のベクトルをスロットマシンとして選択
- 報酬: $Y_t = \theta^T a_t + \epsilon_t$ ($\theta \in \mathbb{R}^d$: 未知パラメータ)
- 例: Web デザイン:

$$a_t = \begin{pmatrix} \text{広告の大きさ} \\ \text{フォントサイズ} \\ \text{レイアウト} \\ \vdots \end{pmatrix}$$

Linear bandit

- 取り得る行動(マシン): $a \in D \subset \mathbb{R}^d$ (D : 有界集合)
  - プレイごとに d 次元のベクトルをスロットマシンとして選択
- 報酬: $Y_t = \theta^T a_t + \epsilon_t$ ($\theta \in \mathbb{R}^d$: 未知パラメータ)

- UCB に似たアルゴリズムにより $\mathrm{regret} = \mathrm{O}(d \log n)$ を高確率で達成可能 (Dani, Hayes & Kakade, 2008)
  - 行動の候補数 $|D|$ が大きくても(あるいは無限大でも)パラメータの次元数のみで regret が抑えられる

Contextual bandit

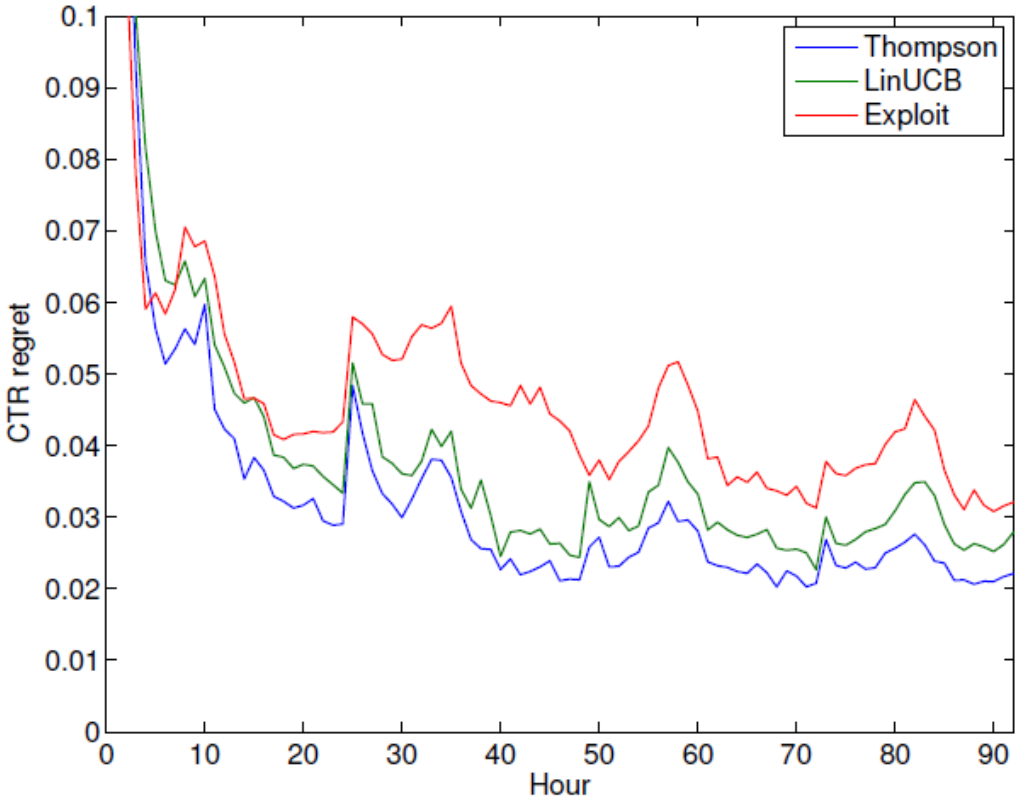
- 各試行ごとに補助情報 (context) x_t が与えられる
  - 性別・年齢・位置情報・閲覧履歴……，またはエキスパートの助言等
- 台 i を選んだ場合の報酬: $Y_i \sim f(\cdot|\theta_i, x_t)$
  - 例: ロジスティック回帰モデル

$$\Pr[Y_i = 1|x_t] = \frac{1}{1 + \mathrm{e}^{\theta_i^{\mathsf{T}} x_t}}$$

- bound が保証されるようなものはあまりないが，Thompson Sampling が高性能

Contextual bandit における TS

- Contextual bandit
  （ユーザ情報・閲覧履歴）
- クリック率 (CTR: Click Through Rate) の比較
- ロジスティック回帰モデル



[Chapelle & Li, 2011]

その他の定式化

- Combinatorial bandit
  - 単位時間ごとに複数本の腕(広告等)を選択
  - 単に期待値の和を最大化する･多様性を考慮する，等の目的の設定
- Dueling bandit
  - 単位時間ごとに2つの候補を選択して優劣のみが提示される(映画の好み等)
- Gaussian process 上の bandit
  - 腕の候補が大量 or 無限個ある（パラメータの最適化等）といった場合で最近成功している

まとめ

- UCB :
  - 有意水準をモデルに応じて適切に設計する必要
  - ある程度性質の分かっているモデルならチューニング次第で非常に高性能に
  - 最適腕識別にも自然に適用可能
- Thompson sampling:
  - 理論限界を介さずにある程度自然な意味付けが可能
  - (事後分布からのサンプリングが出来れば)簡単に実装可能
  - 事前分布の選択にも左右されにくい(が，未知の部分も多い)
  - 最適腕識別には使いにくい

References

概説

- S. Scott, “A modern Bayesian look at the multi-armed bandit,” Applied Stochastic Models in Business and Industry, 2010.

- S. Bubeck and Nicolo Cesa-Bianchi, “Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems,” Fundations and Trends in machine Learning, 2012.

Thompson samplimg (理論)

- E. Kaufmann, N. Korda and R. Munos, “Thompson sampling: an asymptotically optimal finite-time analysis”, ALT2012.

最適腕識別

- J. Audibert, S. Bubeck and R. Munos, “ Best Arm Identification in Multi-Armed Bandits”, COLT2010.

References

Linear bandit

- V. Dani, T. P. Hayes and S. M. Kakade, “Stochastic Linear Optimization under Bandit Feedback”, COLT 2008.

Combinatorial bandit

- Y. Gai, B. Krishnamachari and R. Jain, “Learning Multiuser Channel Allocations in Cognitive Radio Networks: A Combinatorial Multi-Armed Bandit Formulation”, IEEE DySPAN 2010.

Gaussian process 上のバンディット

- N. Srinivas, A. Krause, S. Kakade, and M. Seeger, “Information-Theoretic Regret Bounds for Gaussian Process Optimization in the Bandit Setting”, IEEE Transactions on Information Theory, 2012.