2H1-4

多腕バンディットにおけるリグレットの非線形拡張

Multi-armed Bandit with Nonlinear Regret

梁 曽漢 *1 Zenghan Liang　小宮山 純平 *1 Junpei Komiyama　大岩 秀和 *1 Hidekazu Oiwa　佐藤 一誠 *2 Issei Sato　中川 裕志 *2 Hiroshi Nakagawa

東京大学大学院情報理工学研究科 *1
Graduate School of Information and Technology, The University of Tokyo

東京大学情報基盤センター *2
Information Technology Center, The University of Tokyo

Traditional algorithms for Multi-armed Bandit Problem (MAB) has been designed to minimize the *regret*, the difference of expected reward between the globally optimal policy and a sequence of rewards derived from algorithms. However, minimizing *regret* is not a desirable approach to deploy it to real applications. In this work, we designed p-UCB, a new variant of UCB algorithms for maximizing nonlinear utility function. We investigate the theoretical guarantees of p-UCB, and report empirical superiority when the trial number is small.

1. 序論

多腕バンディット問題 (Multi-armed Bandit Problem, MAB) とは, 異なる見込み配当率が設定された複数のスロットマシンから「いかに期待報酬が最も高いスロットマシンを少ない探索により見つけるか」という課題を定式化したものである. スロットマシン (以降アームと呼ぶ) から得られる報酬は適当な確率分布に従うと仮定する. 多腕バンディット問題は, プレイヤーが各ラウンドで自ら選択したアームに関する報酬の情報しか得られないという点において, 一般的な逐次学習の設定に比べ, 限定されたフィードバックのみから学習を行うことに特徴がある [Cesa-Bianchi 06]. 多腕バンディット問題の起源は古く, [Thompson 33] による治験計画に関するものが最初であった. これは複数の新薬を開発した際に, どの新薬の効用が高いかを調べるために治験の被験者をどのように逐次的に分配するか, という問題に関する考察であった. 現在では, 情報通信技術の発展に伴い, インターネット広告の逐次配置 [Chakrabarti 08, Babaioff 09, Graepel 10, Xu 13] での応用が数多くみられる. インターネット広告の掲載により広告主が得る報酬は成果ベースが一般的であり, 表示された広告がクリックされた場合のみに広告主に報酬が支払われる. 広告ごとにクリックされる確率 (報酬が与えられる確率) や報酬額が異なるため, 限られた広告枠へ期待報酬を最大化するように広告を割り当てることは重要な課題である. 多腕バンディット問題の本質的な難しさは「探索と利用のトレードオフ」である. 探索 (Exploration) とは, 試行回数が少ないアームを引き, そのアームの経験報酬の分散を減らすことをいう. 一方で, 利用 (Exploitation) とは, 累積効用を最大化するために経験報酬が最大のアームを引き続けることをいう. プレイヤーの目的は, 獲得する累計効用を最大化することである. そのため, これまでの試行結果から得た見込み配当率が最大のアームをプレイし続けるか, あるいは, さらに配当率が高いアームが存在すると想定し, 別のアームにも試行回数を割くか, というジレンマが生じる. 多腕バンディット問題におけるアーム選択の戦略の「良さ」を評価する普遍的な指標としてリグレットがある. リグレットは, 期待報酬が最大のアームを選び続けた時に得られる期待効用と自ら設定した戦略によって得られる期待効用の差として定義される. 「良い」戦略とは, 試行回数 n に対してリグレットが低いオーダーで抑えられる戦略を指してきた. 特に, 任意のアーム群および定数 $a > 0$ に対して $o(n^a)$ で抑えられる戦略は強一致性を持つという. [Lai 85] では強一致性をもつ戦略のリグレットの漸近的な下限が $\mathcal{O}(\ln n)$ で与えられることを示した. さらに, [Auer 02] ではリグレットのオーダーが $\mathcal{O}(\ln n)$ で与えられる実用的な戦略である UCB1 が提案され, 多腕バンディット問題におけるベンチマーク手法となっている. このように, UCB1 を始めとする多くの戦略は, リグレットという期待損失に対して線形な効用を最小化する目的のもので設計されてきた. しかしながら, 真に最小化すべき効用が期待損失に線形である必然性はなく, 目的に応じて設計すべきである.

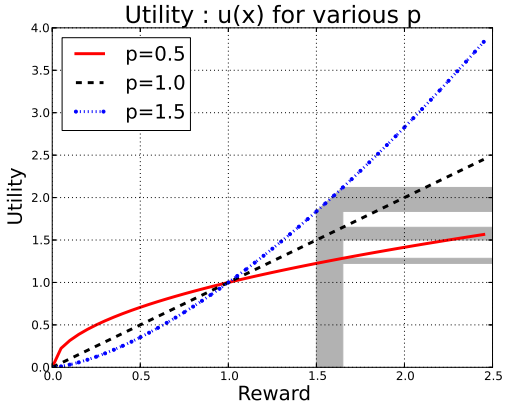


図 1: 効用関数 $u(x) = x^p$. 横軸の報酬の増分 δ_x に対する効用の増分 $u(x+\delta_x) - u(x)$ がパラメタ p によって制御される.

広告ごとのクリック確率及び報酬額が異なる状況において, 期待報酬は高いがクリック確率が低い「リスクの高い商品」の広告ばかり配置するよりも, 期待報酬は低いがクリック確率が高い「リスクの低い商品」の配置にも適切な回数を割くほうが経営戦略上好ましいであろう. 本研究では, このような状況に対するリスク回避的な戦略を設計する. まず, UCB1 の自然

連絡先: 梁 曽漢, zenghan_liang@mist.i.u-tokyo.ac.jp

な拡張である g-UCB を比較手法として設計する．ここでは報酬額は**既知**, 報酬確率は**未知**として, 線形期待損失であるリグレットを最小化するように構築する．次に, 同様な状況設定において非線形効用 $u(x) = x^p$ $(0 \leq p)$ を最大化する p-UCB を初めて提案する．非線形効用を導入するのは, 試行可能回数が少ない時によりリスク回避的な戦略をとることで, 報酬の期待値を最大化するのみならず, 報酬の安定化も図るためである．最後に, 人工データに対し p-UCB $(p = 0.5)$ は g-UCB よりも試行回数が少ない時に得られる効用がより多く, かつ, 探索の傾向が強いことを実験的に検証する．

2. 問題設定

本論文にて用いる記法ならびに問題設定の定式化を行う．選択できるアームは K 本与えられているとし, 総試行数 n に対し第 t 回目 $(1 \leq t \leq n)$ の試行でプレイヤーがアーム i $(i = 1, \ldots, K)$ から得る報酬を $X_{i,t} \in \{0, c_i\}$ とする．c_i はアーム i から得られる**既知**の報酬であり, μ_i はアーム i から報酬が得られる**未知**の確率とする．すなわち, アームが与える報酬に関する確率分布は:

$$p(X_{i,t} \,|\, \mu_i) := \mu_i^{\mathbb{I}\{X_{i,t}=c_i\}}(1-\mu_i)^{1-\mathbb{I}\{X_{i,t}=c_i\}} \qquad (1)$$

で与えられる．ここで, $\mathbb{I}\{\cdot\}$ は以下で定義される指示関数である:

$$\mathbb{I}\{A\} = \begin{cases} 1 & \text{if } A \text{ is true,} \\ 0 & \text{otherwise.} \end{cases} \qquad (2)$$

過去の試行結果から次の引くアームを引く戦略を $\mathcal{A}$ とし, 第 t 回目の試行で戦略 $\mathcal{A}$ に基づき選択したアームのインデックスを $\mathcal{A}_t$ で表す．また, 戦略 $\mathcal{A}$ により, t 回試行を行ったときにアーム i を選択した回数を $T_{i,t}(\mathcal{A})$ とする．すなわち, $T_{i,t}(\mathcal{A}) := \sum_{s=1}^{t} \mathbb{I}\{i = \mathcal{A}_s\}$ である (文脈から戦略 $\mathcal{A}$ が明らかである場合, 単に $T_{i,t}$ と書く)．多腕バンディット問題における普遍的な指標としてのリグレット $\mathcal{R}_n(\mathcal{A})$ は次のように定義される:

$$\mathcal{R}_n(\mathcal{A}) := R_n^\star - R_{\mathcal{A},n}. \qquad (3)$$

ここで, $R_n^\star$ は期待効用を最大化する最適戦略であり, 線形効用を最大化する従来の設定においては最適なアーム $i^\star := \arg\max_i c_i\mu_i$ を n ラウンド選び続けた時に得られる期待効用である．$R_{\mathcal{A},n}$ は戦略 $\mathcal{A}$ に従い n ラウンド選び続けた時に得られる期待効用である．以後の議論のために, $c^\star := c_{i^\star}, \mu^\star := \mu_{i^\star}$, $\Delta_i := c^\star\mu^\star - c_i\mu_i$ と定義しておく．また, $\bar{X}_{i,t}$ は t ラウンド目までにアーム i から得た報酬の経験平均, すなわち:

$$\bar{X}_{i,t} := \frac{1}{T_{i,t}(\mathcal{A})} \sum_{s=1}^{t} X_{i,s} \cdot \mathbb{I}\{i = \mathcal{A}_s\} \qquad (4)$$

とする．また, $\mathbb{E}[\cdot]$ は期待値を表す．我々の目的は, 各アームの報酬額が与えられた時に, 全試行を通して得られる累積効用を最大化する性質のよい戦略を構築することである．

3. 提案手法

3.1 g-UCB

[Auer 02] ではアームからの報酬が Bernoulli 分布で与えられる時の戦略 UCB1 を提案した．UCB1 は線形期待損失であるリグレットを最小化するように構築された戦略であった．UCB1 を各アームの報酬の集合が $\{0, c_i\}$ で与えられる設定に対し拡張した戦略 g-UCB を以下とする:

$$\mathcal{A}_{t+1}^{g-\mathrm{UCB}} = \arg\max_i \ \bar{X}_{i,t} + c_i\sqrt{\frac{2\ln t}{T_{i,t}}}. \qquad (5)$$

これは, $\forall i, c_i = 1$ とすると UCB1 に一致するため, UCB1 の自然な拡張になっていることがわかる．g-UCB のリグレット上界は以下に示すように直ちに求まる:

定理 1 ($\mathcal{A}_t^{g-\mathrm{UCB}}$ **のリグレット上界**)**.** 任意のアーム集合および試行回数 n に対し, 戦略 $\mathcal{A}_t^{g-\mathrm{UCB}}$ のリグレットは高々

$$8\sum_{j\neq i^\star}\frac{c_j^2\ln n}{\Delta_j} + \left(1+\frac{\pi^2}{3}\right)\sum_{j=1}^{K}\Delta_j. \qquad (6)$$

Proof. リグレットは期待損失に対する線形な効用であったから, $R_n^\star := \mathbb{E}\left[\sum_{t=1}^n X_{i^\star,t}\right] = nc^\star\mu^\star$, $R_{\mathcal{A},n} := \mathbb{E}\left[\sum_{t=1}^n X_{\mathcal{A}_t,t}\right] = \sum_{i=1}^K c_i\mu_i\,\mathbb{E}[T_{i,n}(\mathcal{A})]$ と書ける．以降, [Auer 02] の Theorem1 同様に $\mathbb{E}[T_{i,n}(\mathcal{A})]$ の上界を評価すればよい. □

3.2 p-UCB

報酬 $x \geq 0$ に対し, 以下の非線形な効用関数を考えよう:

$$u(x) = x^p \quad (0 \leq p). \qquad (7)$$

図 1 に示すように, 非線形な効用の増分はこれまでに獲得した報酬及び p の値に依存して定まるのが特徴であり, 特に $0 \leq p < 1$ では累計報酬が多いほど効用が増加しにくくなる．このような効用をベースにした期待効用の項 $R_n^\star, R_{\mathcal{A},n}$ について考える．期待効用を最大化する最適戦略は, これまで同様に $i^\star$ を選び続けることであることを示そう:

定理 2 (**効用に対する最適戦略**)**.** 期待報酬最大のアーム $i^\star$ を選び続けた時に得られる効用がそれ以外のアーム $j \neq i^\star$ を選び続けた時に得られる効用よりも高くなる確率は, ラウンド数 n が大きくなるにつれ 1 に近づく．すなわち:

$$\lim_{n\to\infty} P\left\{u\left(\sum_{t=1}^n X_{i^\star,t}\right) \leq\ u\left(\sum_{t=1}^n X_{j,t}\right)\right\} = 0. \qquad (8)$$

Proof. 報酬は正, かつ, 効用関数 (7) は単調増加であるから $P\{u(\sum X_{i^\star,t}) \leq u(\sum X_{j,t})\} = P\{\sum X_{i^\star,t} \leq \sum X_{j,t}\}$. $\delta_{j,n} = n\Delta_j$ に対して Hoeffding の不等式より $P\{\sum X_{i^\star,t} \leq \sum X_{j,t}\} \leq P\{\sum X_{i^\star,t} - nc^\star\mu^\star \leq -\frac{\delta_{j,n}}{2}\} + P\{\sum X_{j,t} - nc_j\mu_j \geq \frac{\delta_{j,n}}{2}\} \leq 2\exp\left(\frac{-\delta_{j,n}^2}{2n\max\{c^\star,c_j\}^2}\right) \overset{n\to\infty}{\to} 0$. □

よって, $R_n^\star, R_{\mathcal{A},n}$ は次のように書き下せる:

$$R_n^\star = \mathbb{E}\left[u\left(\sum_{t=1}^n X_{i^\star,t}\right)\right],\ R_{\mathcal{A},n} = \mathbb{E}\left[u\left(\sum_{t=1}^n X_{\mathcal{A}_t,t}\right)\right]. \qquad (9)$$

特に, $R_n^\star$ に関しては具体的な形を簡単に求めることができる:

定理 3 (**期待効用**)**.** 確率変数の列 $X_1, X_2, \ldots, X_n \in \{0, c\}$ が $p(X_t \,|\, \mu) = \mu^{\mathbb{I}\{X_t=c\}}(1-\mu)^{1-\mathbb{I}\{X_t=c\}}$ として独立同一分布で与えられるとする．この時, 効用関数 (7) に基づく期待効用は以下のように求まる:

$$\mathbb{E}\left[u\left(\sum_{t=1}^n X_t\right)\right] = \sum_{t=0}^{n}(tc)^p \cdot {}_n\mathrm{C}_t \cdot \mu^t(1-\mu)^{n-t}. \qquad (10)$$

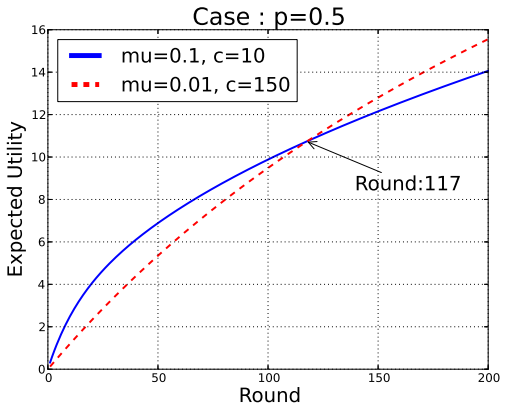


図 2: アーム 1, 2 に対する期待効用. 線形効用 $u(x) = x$ の元ではアーム 1 はアーム 2 より低くなるが, 非線形効用 $u(x) = \sqrt{x}$ の元では 116 試行目までアーム 1 はアーム 2 より高い.

Proof. 省略する. □

効用関数 (7) に基づく期待報酬の具体例を考え, 非線形効用が持つ意味合いについて説明する. 例えば, $p = 1/2$ の時に $K = 2$ アームから選択を行う状況を想定する:

$$\text{アーム } 1 : \mu_1 = 0.1,\ c_1 = 10,$$
$$\text{アーム } 2 : \mu_2 = 0.01,\ c_2 = 150$$

効用関数 (7) に基づく 1 試行目の期待効用は

$$\text{アーム } 1 : 0.9 \times \sqrt{0} + 0.1 \times \sqrt{10} = \sqrt{10}/10$$
$$\text{アーム } 2 : 0.99 \times \sqrt{0} + 0.01 \times \sqrt{150} = \sqrt{6}/20$$

である. 単純な期待値ではアーム 2 を選択した方が良いにもかかわらず, 1 試行目までの期待効用に関しては効用関数を適用したがためにアーム 1 を選択した方が良い結果になっている. アーム 1, 2 をそれぞれ引き続けた時の期待効用を定理 3 より計算し, 200 試行目まで期待効用の変化を図 2 に示した. 非線形な効用により, 116 試行まではアーム 1 の効用がアーム 2 の効用よりも高くなっているとわかる. このように非線形効用を最大化する戦略を設計することで, よりリスクの低いアームを選択する性質を戦略に持たせることができる, と考えられる.

効用関数 (7) を最大化する戦略 p-UCB を以下とする:

$$\mathcal{A}_{t+1}^{p\text{-UCB}_j} = \arg\max_i \ \bar{Y}_{i,t} \cdot \delta_{i,t}^p + d_j \sqrt{\frac{2 \ln t}{T_{i,t}}}, \tag{11}$$

$$d_1 := \delta_{i,t}^p,\ d_2 := p\delta_{i,t}^p,\ d_3 := c_i^p,\ d_4 := pc_i^p.$$

ここで

$$\bar{Y}_{i,t} := \frac{1}{T_{i,t}} \sum_{s=1}^{t} \mathbb{I}\{i = \mathcal{A}_s \wedge X_{i,s} = c_i\}, \tag{12}$$

$$\delta_{i,t}^p := u\left(\sum_{s=1}^{T_{i,t}} X_{i,s} + c_i\right) - u\left(\sum_{s=1}^{T_{i,t}} X_{i,s}\right). \tag{13}$$

である. $\bar{Y}_{i,t}$ は μ_i に関する推定量であり, $\delta_{i,t}^p$ はアーム i を選んだ時の効用の増加分である. $p = 1$ とすると g-UCB に一致するため, g-UCB の自然な拡張になっていることがわかる. また, $p = 0$ では各アームに対するスコアはすべて同じとなり, ランダムにアームを選択する戦略となる. よって, p は探索をコントロールするパラメタとして働き, p が小さい時は探索が強く, p が大きい時は利用が強くなる, と考えられる.

表 1: 各 p-UCB$_j$ におけるアーム選択回数の標準偏差.

	p-UCB$_1$	p-UCB$_2$	p-UCB$_3$	p-UCB$_4$
$p = 0.5$	147.97	175.03	177.44	163.89
$p = 1.0$	241.36	267.85	258.69	247.64
$p = 1.5$	303.93	310.91	304.08	295.95

4. 数値実験

アームごとの報酬額と報酬が得られる確率が異なる状況での p-UCB と g-UCB の挙動を人工データを用いて検証を行った. $K = 6$ のアームセット E を以下のように作成する:

1. 各アームの価値 (期待報酬) $v_i \sim \text{Beta}(6, 6)$,
2. 各アームの報酬確率 $\mu_i \sim \text{Beta}(1, 20)$,
3. 各アームの報酬 $c_i = v_i/\mu_i$.

これは実際の各広告の価値の分散が小さく, かつ, クリック率が低い状況を想定した設定である. 実験に用いたコードは [Cappe 12] をもとに作成した.

4.1 実験 1. 効用損失の比較

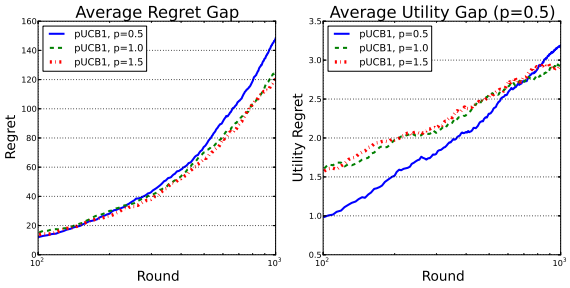
効用損失 $\mathcal{R}_n^p$ を式 (3)(9) より定義し, $p = 0.5$ に固定する. これに対し, ハイパーパラメタ $p = 0.5$ とした時の p-UCB 戦略 0.5-UCB$_j$ の効用損失を g-UCB (1.0-UCB$_j$ と同値) 及び 1.5-UCB$_j$ のそれと比較する. 異なるアーム環境 E を 20 回生成し, 各環境に対し $n = 1000$ ラウンドからなる試行を 100 回繰り返したリグレット及び効用損失の平均を取り, 得た結果を図 3 に示す. 0.5-UCB$_j$ の各手法 (青線) ともリグレットの基準においては最適ではないが, 効用損失の基準においては, 試行回数が少ない時には他の戦略に比べ効用損失が最も少ないことがわかる.

4.2 実験 2. アーム選択回数の標準偏差比較

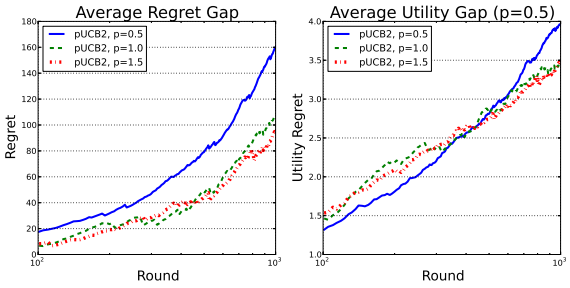
各 p-UCB$_j$ 戦略におけるアーム選択回数のばらつきを調べ, p の値によってアーム選択結果から異なるかどうか検証した. 設定は実験 1 と同じであり, 各環境に対し 100 回の試行でのアーム選択回数の平均値を取り, さらに各環境での平均アーム選択数の標準偏差を取り, それらを平均したものを表 1 に示した. 表 1 によれば, 0.5-UCB は g-UCB に比べアーム選択のばらつきが小さい, 故に探索が多く, 一方で 1.5-UCB は g-UCB に比べアーム選択のばらつきが大きい, 故に利用が多いことがわかる. このことから, p は探索と利用のバランスを制御する性質を持つことがわかる.

5. 結論

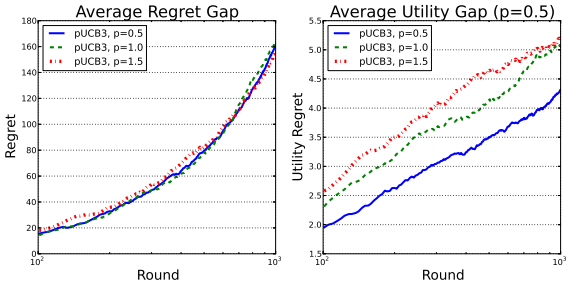
広告ごとのクリック確率及び報酬額が異なる状況において, 多腕バンディット問題に基づく広告配置戦略を提案した. 非線形な効用関数 $u(x) = x^p$ を最大化するように戦略 p-UCB を設計し, パラメタ p を調整することで探索と利用をコントロー



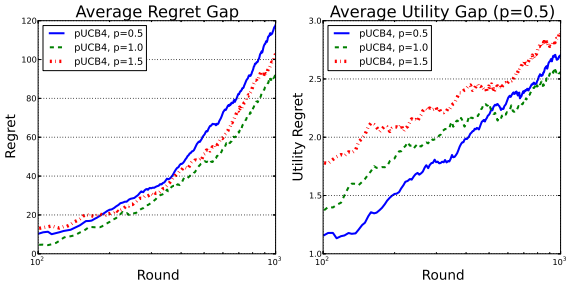
(a) p-UCB$_1$, $d_1 = \delta_{i,t}^p$ のケース



(b) p-UCB$_2$, $d_2 = p\delta_{i,t}^p$ のケース



(c) p-UCB$_3$, $d_3 = c_i^p$ のケース



(d) p-UCB$_1$, $d_4 = pc_i^p$ のケース

図 3: p-UCB$_j$($j = 1, 2, 3, 4$) に対する線形効用損失 (リグレット) 及び非線形効用損失 ($p = 0.5$) の比較. 0.5-UCB$_j$ の各手法 (青線) ともリグレットの基準においては最適ではないが, 非線形効用損失の基準においては, 試行回数が少ない時には効用損失が小さい.

ルできることを考察した. 特に, 効用関数 $u(x) = \sqrt{x}$ に基づく効用損失を最小化するのに, g-UCB よりも 0.5-UCB が優れることを検証し, 0.5-UCB がリスク回避的なアーム選択を行うことを確認した.

参考文献

[Auer 02] Auer, P., Cesa-Bianchi, N., and Fischer, P.: Finite-time analysis of the multiarmed bandit problem, *Machine learning*, Vol. 47, No. 2-3, pp. 235–256 (2002)

[Babaioff 09] Babaioff, M., Sharma, Y., and Slivkins, A.: Characterizing truthful multi-armed bandit mechanisms, in *Proceedings of the 10th ACM conference on Electronic commerce*, pp. 79–88ACM (2009)

[Cappe 12] Cappe, O., Garivier, A., and Kaufmann, E.: pymaBandits (2012), `http://mloss.org/software/view/415/`

[Cesa-Bianchi 06] Cesa-Bianchi, N.: *Prediction, learning, and games*, Cambridge University Press (2006)

[Chakrabarti 08] Chakrabarti, D., Kumar, R., Radlinski, F., and Upfal, E.: Mortal multi-armed bandits, in *Advances in Neural Information Processing Systems*, pp. 273–280 (2008)

[Graepel 10] Graepel, T., Candela, J. Q., Borchert, T., and Herbrich, R.: Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine, in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 13–20 (2010)

[Lai 85] Lai, T. L. and Robbins, H.: Asymptotically efficient adaptive allocation rules, *Advances in applied mathematics*, Vol. 6, No. 1, pp. 4–22 (1985)

[Thompson 33] Thompson, W. R.: On the likelihood that one unknown probability exceeds another in view of the evidence of two samples, *Biometrika*, Vol. 25, No. 3/4, pp. 285–294 (1933)

[Xu 13] Xu, M., Qin, T., and Liu, T.-Y.: Estimation Bias in Multi-Armed Bandit Algorithms for Search Advertising, in *Advances in Neural Information Processing Systems*, pp. 2400–2408 (2013)