WIKIPEDIA

Multi-armed bandit

A row of slot machines in Las Vegas

In probability theory, the **multi-armed bandit problem** (sometimes called the ***K*-[1] or *N*-armed bandit problem**[2]) is a problem in which a fixed limited set of resources must be allocated between competing (alternative) choices in a way that maximizes their expected gain, when each choice's properties are only partially known at the time of allocation, and may become better understood as time passes or by allocating resources to the choice.[3][4] This is a classic reinforcement learning problem that exemplifies the exploration–exploitation tradeoff dilemma. The name comes from imagining a gambler at a row of slot machines (sometimes known as "one-armed bandits"), who has to decide which machines to play, how many times to play each machine and in which order to play them, and whether to continue with the current machine or try a different machine.[5] The multi-armed bandit problem also falls into the broad category of stochastic scheduling.

In the problem, each machine provides a random reward from a probability distribution specific to that machine. The objective of the gambler is to maximize the sum of rewards earned through a sequence of lever pulls.[3][4] The crucial tradeoff the gambler faces at each trial is between "exploitation" of the machine that has the highest expected payoff and "exploration" to get more information about the expected payoffs of the other machines. The trade-off between exploration and exploitation is also faced in machine learning. In practice, multi-armed bandits have been used to model problems such as managing research projects in a large organization like a science foundation or a pharmaceutical company.[3][4] In early versions of the problem, the gambler begins with no initial knowledge about the machines.

Herbert Robbins in 1952, realizing the importance of the problem, constructed convergent population selection strategies in "some aspects of the sequential design of experiments".[6] A theorem, the Gittins index, first published by John C. Gittins, gives an optimal policy for maximizing the expected discounted reward.[7]

Contents

Empirical motivation

The multi-armed bandit problem models an agent that simultaneously attempts to acquire new knowledge (called "exploration") and optimize their decisions based on existing knowledge (called "exploitation"). The agent attempts to balance these competing tasks in order to maximize their total value over the period of time considered. There are many practical applications of the bandit model, for example:

- clinical trials investigating the effects of different experimental treatments while minimizing patient losses,[3][4][8][9]

- *adaptive routing efforts for minimizing delays in a network,*
- *financial portfolio design*[10][11]

How must a given budget be distributed among these research departments to maximize results?

In these practical examples, the problem requires balancing reward maximization based on the knowledge already acquired with attempting new actions to further increase knowledge. This is known as the *exploitation vs. exploration tradeoff* in machine learning.

The model has also been used to control dynamic allocation of resources to different projects, answering the question of which project to work on, given uncertainty about the difficulty and payoff of each possibility.[12]

Originally considered by Allied scientists in World War II, it proved so intractable that, according to Peter Whittle, the problem was proposed to be dropped over Germany so that German scientists could also waste their time on it.[13]

The version of the problem now commonly analyzed was formulated by Herbert Robbins in 1952.

The multi-armed bandit model

The multi-armed bandit (short: *bandit* or MAB) can be seen as a set of real distributions $B = \{R_1, \dots, R_K\}$, each distribution being associated with the rewards delivered by one of the $K \in \mathbb{N}^+$ levers. Let $\mu_1, \dots, \mu_K$ be the mean values associated with these reward distributions. The gambler iteratively plays one lever per round and observes the associated reward. The objective is to maximize the sum of the collected rewards. The horizon H is the number of rounds that remain to be played. The bandit problem is formally equivalent to a one-state Markov decision process. The regret ρ after T rounds is defined as the expected difference between the reward sum associated with an optimal strategy and the sum of the collected rewards:

$$\rho = T\mu^* - \sum_{t=1}^{T} \hat{r}_t,$$

where μ^* is the maximal reward mean, $\mu^* = \max_k \{\mu_k\}$, and $\hat{r}_t$ is the reward in round t.

A *zero-regret strategy* is a strategy whose average regret per round ρ/T tends to zero with probability 1 when the number of played rounds tends to infinity.[14] Intuitively, zero-regret strategies are guaranteed to converge to a (not necessarily unique) optimal strategy if enough rounds are played.

Variations

A common formulation is the *Binary multi-armed bandit* or *Bernoulli multi-armed bandit,* which issues a reward of one with probability p, and otherwise a reward of zero.

Another formulation of the multi-armed bandit has each arm representing an independent Markov machine. Each time a particular arm is played, the state of that machine advances to a new one, chosen according to the Markov state evolution probabilities. There is a reward depending on the

current state of the machine. In a generalization called the "restless bandit problem", the states of non-played arms can also evolve over time.[15] There has also been discussion of systems where the number of choices (about which arm to play) increases over time.[16]

Computer science researchers have studied multi-armed bandits under worst-case assumptions, obtaining algorithms to minimize regret in both finite and infinite (asymptotic) time horizons for both stochastic[1] and non-stochastic[17] arm payoffs.

Bandit strategies

A major breakthrough was the construction of optimal population selection strategies, or policies (that possess uniformly maximum convergence rate to the population with highest mean) in the work described below.

Optimal solutions

In the paper "Asymptotically efficient adaptive allocation rules", Lai and Robbins[18] (following papers of Robbins and his co-workers going back to Robbins in the year 1952) constructed convergent population selection policies that possess the fastest rate of convergence (to the population with highest mean) for the case that the population reward distributions are the one-parameter exponential family. Then, in Katehakis and Robbins[19] simplifications of the policy and the main proof were given for the case of normal populations with known variances. The next notable progress was obtained by Burnetas and Katehakis in the paper "Optimal adaptive policies for sequential allocation problems",[20] where index based policies with uniformly maximum convergence rate were constructed, under more general conditions that include the case in which the distributions of outcomes from each population depend on a vector of unknown parameters. Burnetas and Katehakis (1996) also provided an explicit solution for the important case in which the distributions of outcomes follow arbitrary (i.e., non-parametric) discrete, univariate distributions.

Later in "Optimal adaptive policies for Markov decision processes"[21] Burnetas and Katehakis studied the much larger model of Markov Decision Processes under partial information, where the transition law and/or the expected one period rewards may depend on unknown parameters. In this work the explicit form for a class of adaptive policies that possess uniformly maximum convergence rate properties for the total expected finite horizon reward, were constructed under sufficient assumptions of finite state-action spaces and irreducibility of the transition law. A main feature of these policies is that the choice of actions, at each state and time period, is based on indices that are inflations of the right-hand side of the estimated average reward optimality equations. These inflations have recently been called the optimistic approach in the work of Tewari and Bartlett,[22] Ortner[23] Filippi, Cappé, and Garivier,[24] and Honda and Takemura.[25]

When optimal solutions to multi-arm bandit tasks [26] are used to derive the value of animals' choices, the activity of neurons in the amygdala and ventral striatum encodes the values derived from these policies, and can be used to decode when the animals make exploratory versus exploitative choices. Moreover, optimal policies better predict animals' choice behavior than alternative strategies (described below). This suggests that the optimal solutions to multi-arm bandit problems are biologically plausible, despite being computationally demanding. [27]

- *UCBC (Historical Upper Confidence Bounds with clusters): [28] The algorithm adapts UCB for a new setting such that it can incorporate both clustering and historical information. The algorithm incorporates the historical observations by utilizing both in the computation of the observed mean rewards and the uncertainty term. The algorithm incorporates the clustering information by playing*

at two levels: first picking a cluster using a UCB-like strategy at each time step, and subsequently picking an arm within the cluster, again using a UCB-like strategy.

Approximate solutions

Many strategies exist which provide an approximate solution to the bandit problem, and can be put into the four broad categories detailed below.

Semi-uniform strategies

Semi-uniform strategies were the earliest (and simplest) strategies discovered to approximately solve the bandit problem. All those strategies have in common a greedy behavior where the *best* lever (based on previous observations) is always pulled except when a (uniformly) random action is taken.

- Epsilon-greedy strategy:[29] The best lever is selected for a proportion $1-\epsilon$ of the trials, and a lever is selected at random (with uniform probability) for a proportion ϵ. A typical parameter value might be $\epsilon = 0.1$, but this can vary widely depending on circumstances and predilections.
- Epsilon-first strategy: A pure exploration phase is followed by a pure exploitation phase. For N trials in total, the exploration phase occupies ϵN trials and the exploitation phase $(1-\epsilon)N$ trials. During the exploration phase, a lever is randomly selected (with uniform probability); during the exploitation phase, the best lever is always selected.
- Epsilon-decreasing strategy: Similar to the epsilon-greedy strategy, except that the value of ϵ decreases as the experiment progresses, resulting in highly explorative behaviour at the start and highly exploitative behaviour at the finish.
- Adaptive epsilon-greedy strategy based on value differences (VDBE): Similar to the epsilon-decreasing strategy, except that epsilon is reduced on basis of the learning progress instead of manual tuning (Tokic, 2010).[30] High fluctuations in the value estimates lead to a high epsilon (high exploration, low exploitation); low fluctuations to a low epsilon (low exploration, high exploitation). Further improvements can be achieved by a softmax-weighted action selection in case of exploratory actions (Tokic & Palm, 2011).[31]
- Adaptive epsilon-greedy strategy based on Bayesian ensembles (Epsilon-BMC): An adaptive epsilon adaptation strategy for reinforcement learning similar to VBDE, with monotone convergence guarantees. In this framework, the epsilon parameter is viewed as the expectation of a posterior distribution weighting a greedy agent (that fully trusts the learned reward) and uniform learning agent (that distrusts the learned reward). This posterior is approximated using a suitable Beta distribution under the assumption of normality of observed rewards. In order to address the possible risk of decreasing epsilon too quickly, uncertainty in the variance of the learned reward is also modeled and updated using a normal-gamma model. (Gimelfarb et al., 2019).[32]
- Contextual-Epsilon-greedy strategy: Similar to the epsilon-greedy strategy, except that the value of ϵ is computed regarding the situation in experiment processes, which lets the algorithm be Context-Aware. It is based on dynamic exploration/exploitation and can adaptively balance the two aspects by deciding which situation is most relevant for exploration or exploitation, resulting in highly

explorative behavior when the situation is not critical and highly exploitative behavior at critical situation.[33]

Probability matching strategies

Probability matching strategies reflect the idea that the number of pulls for a given lever should *match* its actual probability of being the optimal lever. Probability matching strategies are also known as Thompson sampling or Bayesian Bandits,[34][35] and are surprisingly easy to implement if you can sample from the posterior for the mean value of each alternative.

Probability matching strategies also admit solutions to so-called contextual bandit problems.

Pricing strategies

Pricing strategies establish a *price* for each lever. For example, as illustrated with the POKER algorithm,[14] the price can be the sum of the expected reward plus an estimation of extra future rewards that will gain through the additional knowledge. The lever of highest price is always pulled.

Strategies with ethical constraints

- Behavior Constrained Thompson Sampling (BCTS) [36]: In this paper the authors detail a novel online agent that learns a set of behavioral constraints by observation and uses these learned constraints as a guide when making decisions in an online setting while still being reactive to reward feedback. To define this agent, the solution was to adopt a novel extension to the classical contextual multi-armed bandit setting and provide a new algorithm called Behavior Constrained Thompson Sampling (BCTS) that allows for online learning while obeying exogenous constraints. The agent learns a constrained policy that implements the observed behavioral constraints demonstrated by a teacher agent, and then uses this constrained policy to guide the reward-based online exploration and exploitation.

These strategies minimize the assignment of any patient to an inferior arm ("physician's duty"). In a typical case, they minimize expected successes lost (ESL), that is, the expected number of favorable outcomes that were missed because of assignment to an arm later proved to be inferior. Another version minimizes resources wasted on any inferior, more expensive, treatment.[8]

Contextual bandit

A particularly useful version of the multi-armed bandit is the contextual multi-armed bandit problem. In this problem, in each iteration an agent has to choose between arms. Before making the choice, the agent sees a d-dimensional feature vector (context vector), associated with the current iteration. The learner uses these context vectors along with the rewards of the arms played in the past to make the choice of the arm to play in the current iteration. Over time, the learner's aim is to collect enough information about how the context vectors and rewards relate to each other, so that it can predict the next best arm to play by looking at the feature vectors.[37]

Approximate solutions for contextual bandit

Many strategies exist that provide an approximate solution to the contextual bandit problem, and can be put into two broad categories detailed below.

Online linear bandits

- *LinUCB (Upper Confidence Bound) algorithm: the authors assume a linear dependency between the expected reward of an action and its context and model the representation space using a set of linear predictors.*[38][39]
- *LinRel (Linear Associative Reinforcement Learning) algorithm: Similar to LinUCB, but utilizes Singular-value decomposition rather than Ridge regression to obtain an estimate of confidence.*[40][41]
- *HLINUCBC (Historic LINUCB with clusters): proposed in the paper* [42]*, extends the LinUCB idea with both historical and clustering information.*[43]

Online non-linear bandits

- *UCBogram algorithm: The nonlinear reward functions are estimated using a piecewise constant estimator called a regressogram in nonparametric regression. Then, UCB is employed on each constant piece. Successive refinements of the partition of the context space are scheduled or chosen adaptively.*[44][45][46]
- *Generalized linear algorithms: The reward distribution follows a generalized linear model, an extension to linear bandits.*[47][48][49][50]
- *NeuralBandit algorithm: In this algorithm several neural networks are trained to modelize the value of rewards knowing the context, and it uses a multi-experts approach to choose online the parameters of multi-layer perceptrons.*[51]
- *KernelUCB algorithm: a kernelized non-linear version of linearUCB, with efficient implementation and finite-time analysis.*[52]
- *Bandit Forest algorithm: a random forest is built and analyzed w.r.t the random forest built knowing the joint distribution of contexts and rewards.*[53]
- *Oracle-based algorithm: The algorithm reduces the contextual bandit problem into a series of supervised learning problem, and does not rely on typical realizability assumption on the reward function.*[54]

Constrained contextual bandit

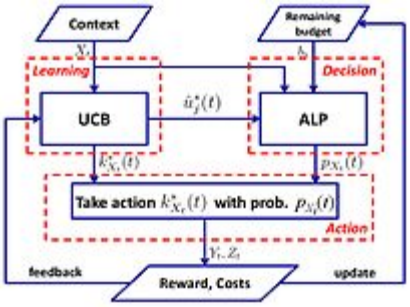
- *Context Attentive Bandits or Contextual Bandit with Restricted Context* [55]*: The authors consider a novel formulation of the multi-armed bandit model, which is called contextual bandit with restricted context, where only a limited number of features can be accessed by the learner at every iteration. This novel formulation is motivated by different online problems arising in clinical trials, recommender systems and attention modeling. Herein, they adapt the standard multi-armed bandit algorithm known as Thompson Sampling to take advantage of the restricted context setting, and propose two novel algorithms, called the Thompson Sampling with Restricted Context (TSRC) and*

the Windows Thompson Sampling with Restricted Context (WTSRC),for handling stationary and nonstationary environments, respectively..

In practice, there is usually a cost associated with the resource consumed by each action and the total cost is limited by a budget in many applications such as crowdsourcing and clinical trials. Constrained contextual bandit (CCB) is such a model that considers both the time and budget constraints in a multi-armed bandit setting. A. Badanidiyuru et al.[56] first studied contextual bandits with budget constraints, also referred to as Resourceful Contextual Bandits, and show that a $O(\sqrt{T})$ regret is achievable. However, their work focuses on a finite set of policies, and the algorithm is computationally inefficient.

A simple algorithm with logarithmic regret is proposed in:[57]

- *UCB-ALP algorithm: The framework of UCB-ALP is shown in the right figure. UCB-ALP is a simple algorithm that combines the UCB method with an Adaptive Linear Programming (ALP) algorithm, and can be easily deployed in practical systems. It is the first work that show how to achieve logarithmic regret in constrained contextual bandits. Although[57] is devoted to a special case with single budget constraint and fixed cost, the results shed light on the design and analysis of algorithms for more general CCB problems.*



Framework of UCB-ALP for constrained contextual bandits

Adversarial bandit

Another variant of the multi-armed bandit problem is called the adversarial bandit, first introduced by Auer and Cesa-Bianchi (1998). In this variant, at each iteration, an agent chooses an arm and an adversary simultaneously chooses the payoff structure for each arm. This is one of the strongest generalizations of the bandit problem[58] as it removes all assumptions of the distribution and a solution to the adversarial bandit problem is a generalized solution to the more specific bandit problems.

Example: Iterated prisoner's dilemma

An example often considered for adversarial bandits is the iterated prisoner's dilemma. In this example, each adversary has two arms to pull. They can either Deny or Confess. Standard stochastic bandit algorithms don't work very well with these iterations. For example, if the opponent cooperates in the first 100 rounds, defects for the next 200, then cooperate in the following 300, etc. Then algorithms such as UCB won't be able to react very quickly to these changes. This is because after a certain point sub-optimal arms are rarely pulled to limit exploration and focus on exploitation. When the environment changes the algorithm is unable to adapt or may not even detect the change.

Approximate solutions

Exp3[59]

Algorithm

Parameters: Real $\gamma \in (0,1]$

Initialisation: $\omega_i(1) = 1$ *for* $i = 1, \dots, K$

For each t = 1, 2, ..., T

1. *Set* $p_i(t) = (1-\gamma)\frac{\omega_i(t)}{\sum_{j=1}^{K}\omega_j(t)} + \frac{\gamma}{K} \qquad i = 1, \dots, K$
2. *Draw* i_t *randomly according to the probabilities* $p_1(t), \dots, p_K(t)$
3. *Receive reward* $x_{i_t}(t) \in [0,1]$
4. *For* $j = 1, \dots, K$ *set:*

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

$$\omega_j(t+1) = \omega_j(t) exp(\gamma \hat{x}_j(t)/K)$$

Explanation

Exp3 chooses an arm at random with probability $(1-\gamma)$ it prefers arms with higher weights (exploit), it chooses with probability γ to uniformly randomly explore. After receiving the rewards the weights are updated. The exponential growth significantly increases the weight of good arms.

Regret analysis

The (external) regret of the Exp3 algorithm is at most $O(\sqrt{KTlog(K)})$

Follow the perturbed leader (FPL) algorithm

Algorithm

Parameters: Real η

Initialisation: $\forall i : R_i(1) = 0$

For each t = 1,2,...,T

1. *For each arm generate a random noise from an exponential distribution* $\forall i : Z_i(t) \sim Exp(\eta)$
2. *Pull arm* $I(t)$*:* $I(t) = arg\max_i \{R_i(t) + Z_i(t)\}$
   Add noise to each arm and pull the one with the highest value
3. *Update value:* $R_{I(t)}(t+1) = R_{I(t)}(t) + x_{I(t)}(t)$
   The rest remains the same

Explanation

We follow the arm that we think has the best performance so far adding exponential noise to it to provide exploration.[60]

Exp3 vs FPL

Exp3	FPL
Maintains weights for each arm to calculate pulling probability	Doesn't need to know the pulling probability per arm
Has efficient theoretical guarantees	The standard FPL does not have good theoretical guarantees
Might be computationally expensive (calculating the exponential terms)	Computationally quite efficient

Infinite-armed bandit

In the original specification and in the above variants, the bandit problem is specified with a discrete and finite number of arms, often indicated by the variable K. In the infinite armed case, introduced by Agarwal (1995), the "arms" are a continuous variable in K dimensions.

Non-stationary bandit

Garivier and Moulines derive some of the first results with respect to bandit problems where the underlying model can change during play. A number of algorithms were presented to deal with this case, including Discounted UCB[61] and Sliding-Window UCB.[62]

Another work by Burtini et al. introduces a weighted least squares Thompson sampling approach (WLS-TS), which proves beneficial in both the known and unknown non-stationary cases.[63] In the known non-stationary case, the authors in [64] produce an alternative solution, a variant of UCB named Adjusted Upper Confidence Bound (A-UCB) which assumes a stochastic model and provide upper-bounds of the regret.

Other variants

Many variants of the problem have been proposed in recent years.

Dueling bandit

The dueling bandit variant was introduced by Yue et al. (2012)[65] to model the exploration-versus-exploitation tradeoff for relative feedback. In this variant the gambler is allowed to pull two levers at the same time, but they only get a binary feedback telling which lever provided the best reward. The difficulty of this problem stems from the fact that the gambler has no way of directly observing the reward of their actions. The earliest algorithms for this problem are InterleaveFiltering,[65] Beat-The-Mean.[66] The relative feedback of dueling bandits can also lead to voting paradoxes. A solution is to take the Condorcet winner as a reference.[67]

More recently, researchers have generalized algorithms from traditional MAB to dueling bandits: Relative Upper Confidence Bounds (RUCB),[68] Relative EXponential weighing (REX3),[69] Copeland Confidence Bounds (CCB),[70] Relative Minimum Empirical Divergence (RMED),[71] and Double Thompson Sampling (DTS).[72]

Collaborative bandit

The collaborative filtering bandits (i.e., COFIBA) was introduced by Li and Karatzoglou and Gentile (SIGIR 2016),[73] where the classical collaborative filtering, and content-based filtering methods try to learn a static recommendation model given training data. These approaches are far from ideal in highly dynamic recommendation domains such as news recommendation and computational advertisement, where the set of items and users is very fluid. In this work, they investigate an adaptive clustering technique for content recommendation based on exploration-exploitation strategies in contextual multi-armed bandit settings.[74] Their algorithm (COFIBA, pronounced as "Coffee Bar") takes into account the collaborative effects[73] that arise due to the interaction of the users with the items, by dynamically grouping users based on the items under consideration and, at the same time, grouping items based on the similarity of the clusterings induced over the users. The resulting algorithm thus takes advantage of preference patterns in the data in a way akin to collaborative filtering methods. They provide an empirical analysis on medium-size real-world datasets, showing scalability and increased prediction performance (as measured by click-through rate) over state-of-the-art methods for clustering bandits. They also provide a regret analysis within a standard linear stochastic noise setting.

Combinatorial bandit

The Combinatorial Multiarmed Bandit (CMAB) problem[75][76][77] arises when instead of a single discrete variable to choose from, an agent needs to choose values for a set of variables. Assuming each variable is discrete, the number of possible choices per iteration is exponential in the number of variables. Several CMAB settings have been studied in the literature, from settings where the variables are binary[76] to more general setting where each variable can take an arbitrary set of values.[77]

See also

- Gittins index – a powerful, general strategy for analyzing bandit problems.
- Greedy algorithm
- Optimal stopping
- Search theory
- Stochastic scheduling

References

1. Auer, P.; Cesa-Bianchi, N.; Fischer, P. (2002). "Finite-time Analysis of the Multiarmed Bandit Problem" (https://doi.org/10.1023%2FA%3A1013689704352). Machine Learning. 47 (2/3): 235–256. doi:10.1023/A:1013689704352 (https://doi.org/10.1023%2FA%3A1013689704352).
2. Katehakis, M. N.; Veinott, A. F. (1987). "The Multi-Armed Bandit Problem: Decomposition and Computation" (https://semanticscholar.org/paper/e4fe28113fed71999a0db30a930e0b42d3ce55f1). Mathematics of Operations Research. 12 (2): 262–268. doi:10.1287/moor.12.2.262 (https://doi.org/10.1287%2Fmoor.12.2.262). S2CID 656323 (https://api.semanticscholar.org/CorpusID:656323).
3. Gittins, J. C. (1989), Multi-armed bandit allocation indices, Wiley-Interscience Series in Systems and Optimization., Chichester: John Wiley & Sons, Ltd., ISBN 978-0-471-92059-5
4. Berry, Donald A.; Fristedt, Bert (1985), Bandit problems: Sequential allocation of experiments, Monographs on Statistics and Applied Probability, London: Chapman & Hall, ISBN 978-0-412-24810-8
5. Weber, Richard (1992), "On the Gittins index for multiarmed bandits", Annals of Applied Probability, 2 (4): 1024–1033, doi:10.1214/aoap/1177005588 (https://doi.org/10.1214%2Faoap%2F1177005588), JSTOR 2959678 (https://www.jstor.org/stable/2959678)

6. Robbins, H. (1952). "Some aspects of the sequential design of experiments" (https://doi.org/10.1090%2FS0002-9904-1952-09620-8). Bulletin of the American Mathematical Society. 58 (5): 527–535. doi:10.1090/S0002-9904-1952-09620-8 (https://doi.org/10.1090%2FS0002-9904-1952-09620-8).
7. J. C. Gittins (1979). "Bandit Processes and Dynamic Allocation Indices". Journal of the Royal Statistical Society. Series B (Methodological). 41 (2): 148–177. doi:10.1111/j.2517-6161.1979.tb01068.x (https://doi.org/10.1111%2Fj.2517-6161.1979.tb01068.x). JSTOR 2985029 (https://www.jstor.org/stable/2985029).
8. Press, William H. (2009), "Bandit solutions provide unified ethical models for randomized clinical trials and comparative effectiveness research", Proceedings of the National Academy of Sciences, 106 (52): 22387–22392, Bibcode:2009PNAS..10622387P (https://ui.adsabs.harvard.edu/abs/2009PNAS..106622387P), doi:10.1073/pnas.0912378106 (https://doi.org/10.1073%2Fpnas.0912378106), PMC 2793317 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2793317), PMID 20018711 (https://pubmed.ncbi.nlm.nih.gov/20018711).
9. Press (1986)
10. Brochu, Eric; Hoffman, Matthew W.; de Freitas, Nando (September 2010), Portfolio Allocation for Bayesian Optimization, arXiv:1009.5419 (https://arxiv.org/abs/1009.5419), Bibcode:2010arXiv1009.5419B (https://ui.adsabs.harvard.edu/abs/2010arXiv1009.5419B)
11. Shen, Weiwei; Wang, Jun; Jiang, Yu-Gang; Zha, Hongyuan (2015), "Portfolio Choices with Orthogonal Bandit Learning" (http://www.aaai.org/ocs/index.php/IJCAI/IJCAI15/paper/viewPDFInterstitial/10972/10798), Proceedings of International Joint Conferences on Artificial Intelligence (IJCAI2015)
12. Farias, Vivek F; Ritesh, Madan (2011), "The irrevocable multiarmed bandit problem", Operations Research, 59 (2): 383–399, CiteSeerX 10.1.1.380.6983 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.380.6983), doi:10.1287/opre.1100.0891 (https://doi.org/10.1287%2Fopre.1100.0891)
13. Whittle, Peter (1979), "Discussion of Dr Gittins' paper", Journal of the Royal Statistical Society, Series B, 41 (2): 148–177, doi:10.1111/j.2517-6161.1979.tb01069.x (https://doi.org/10.1111%2Fj.2517-6161.1979.tb01069.x)
14. Vermorel, Joannes; Mohri, Mehryar (2005), Multi-armed bandit algorithms and empirical evaluation (http://bandit.sourceforge.net/Vermorel200Spoker.pdf) (PDF), In European Conference on Machine Learning, Springer, pp. 437–448
15. Whittle, Peter (1988), "Restless bandits: Activity allocation in a changing world", Journal of Applied Probability, 25A: 287–298, doi:10.2307/3214163 (https://doi.org/10.2307%2F3214163), JSTOR 3214163 (https://www.jstor.org/stable/3214163), MR 0974588 (https://www.ams.org/mathscinet-getitem?mr=0974588)
16. Whittle, Peter (1981), "Arm-acquiring bandits", Annals of Probability, 9 (2): 284–292, doi:10.1214/aop/1176994469 (https://doi.org/10.1214%2Faop%2F1176994469)
17. Auer, P.; Cesa-Bianchi, N.; Freund, Y.; Schapire, R. E. (2002). "The Nonstochastic Multiarmed Bandit Problem". SIAM J. Comput. 32 (1): 48–77. CiteSeerX 10.1.1.130.158 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.130.158). doi:10.1137/S0097539701398375 (https://doi.org/10.1137%2FS0097539701398375).
18. Lai, T.L.; Robbins, H. (1985). "Asymptotically efficient adaptive allocation rules". Advances in Applied Mathematics. 6 (1): 4–22. doi:10.1016/0196-8858(85)90002-8 (https://doi.org/10.1016%2F0196-8858%2885%2990002-8).
19. Katehakis, M.N.; Robbins, H. (1995). "Sequential choice from several populations" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC41010). Proceedings of the National Academy of Sciences of the United States of America. 92 (19): 8584–5. Bibcode:1995PNAS...92.8584K (https://ui.adsabs.harvard.edu/abs/1995PNAS...92.8584K). doi:10.1073/pnas.92.19.8584 (https://doi.org/10.1073%2Fpnas.92.19.8584). PMC 41010 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC41010). PMID 11607577 (https://pubmed.ncbi.nlm.nih.gov/11607577).
20. Burnetas, A.N.; Katehakis, M.N. (1996). "Optimal adaptive policies for sequential allocation problems". Advances in Applied Mathematics. 17 (2): 122–142. doi:10.1006/aama.1996.0007 (https://doi.org/10.1006%2Faama.1996.0007).
21. Burnetas, A.N.; Katehakis, M.N. (1997). "Optimal adaptive policies for Markov decision processes". Math. Oper. Res. 22 (1): 222–255. doi:10.1287/moor.22.1.222 (https://doi.org/10.1287%2Fmoor.22.1.222).

22. Tewari, A.; Bartlett, P.L. (2008). "Optimistic linear programming gives logarithmic regret for irreducible MDPs" (http://books.nips.cc/papers/files/nips20/NIPS2007_0673.pdf) (PDF). Advances in Neural Information Processing Systems. 20. CiteSeerX 10.1.1.69.5482 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.69.5482).
23. Ortner, R. (2010). "Online regret bounds for Markov decision processes with deterministic transitions". Theoretical Computer Science. 411 (29): 2684–2695. doi:10.1016/j.tcs.2010.04.005 (https://doi.org/10.1016%2Fj.tcs.2010.04.005).
24. Filippi, S. and Cappé, O. and Garivier, A. (2010). "Online regret bounds for Markov decision processes with deterministic transitions", Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on, pp. 115–122
25. Honda, J.; Takemura, A. (2011). "An asymptotically optimal policy for finite support models in the multi-armed bandit problem". Machine Learning. 85 (3): 361–391. arXiv:0905.2776 (https://arxiv.org/abs/0905.2776). doi:10.1007/s10994-011-5257-4 (https://doi.org/10.1007%2Fs10994-011-5257-4). S2CID 821462 (https://api.semanticscholar.org/CorpusID:821462).
26. Averbeck, B.B. (2015). "Theory of choice in bandit, information sampling, and foraging tasks" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4376795). PLOS Computational Biology. 11 (3): e1004164. Bibcode:2015PLSCB..11E4164A (https://ui.adsabs.harvard.edu/abs/2015PLSCB..11E4164A). doi:10.1371/journal.pcbi.1004164 (https://doi.org/10.1371%2Fjournal.pcbi.1004164). PMC 4376795 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4376795). PMID 25815510 (https://pubmed.ncbi.nlm.nih.gov/25815510).
27. Costa, V.D.; Averbeck, B.B. (2019). "Subcortical Substrates of Explore-Exploit Decisions in Primates" (https://www.cell.com/neuron/pdfExtended/S0896-6273(19)30442-8#secsectitle0010). Neuron. 103 (3): 533–535. doi:10.1016/j.neuron.2019.05.017 (https://doi.org/10.1016%2Fj.neuron.2019.05.017). PMC 6687547 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6687547). PMID 31196672 (https://pubmed.ncbi.nlm.nih.gov/31196672).
28. Bouneffouf, D. (2019). Optimal Exploitation of Clustering and History Information in Multi-Armed Bandit. Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligen. AAAI. Soc. pp. 270–279. doi:10.1109/sfcs.2000.892116 (https://doi.org/10.1109%2Fsfcs.2000.892116). ISBN 978-0769508504. S2CID 28713091 (https://api.semanticscholar.org/CorpusID:28713091).
29. Sutton, R. S. & Barto, A. G. 1998 Reinforcement learning: an introduction. Cambridge, MA: MIT Press.
30. Tokic, Michel (2010), "Adaptive ε-greedy exploration in reinforcement learning based on value differences" (http://www.tokic.com/www/tokicm/publikationen/papers/AdaptiveEpsilonGreedyExploration.pdf) (PDF), KI 2010: Advances in Artificial Intelligence, Lecture Notes in Computer Science, 6359, Springer-Verlag, pp. 203–210, CiteSeerX 10.1.1.458.464 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.458.464), doi:10.1007/978-3-642-16111-7_23 (https://doi.org/10.1007%2F978-3-642-16111-7_23), ISBN 978-3-642-16110-0.
31. Tokic, Michel; Palm, Günther (2011), "Value-Difference Based Exploration: Adaptive Control Between Epsilon-Greedy and Softmax" (http://www.tokic.com/www/tokicm/publikationen/papers/KI2011.pdf) (PDF), KI 2011: Advances in Artificial Intelligence, Lecture Notes in Computer Science, 7006, Springer-Verlag, pp. 335–346, ISBN 978-3-642-24455-1.
32. Gimelfarb, Michel; Sanner, Scott; Lee, Chi-Guhn (2019), "ε-BMC: A Bayesian Ensemble Approach to Epsilon-Greedy Exploration in Model-Free Reinforcement Learning" (http://auai.org/uai2019/proceedings/papers/162.pdf) (PDF), Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, AUAI Press, p. 162.
33. Bouneffouf, D.; Bouzeghoub, A.; Gançarski, A. L. (2012). "A Contextual-Bandit Algorithm for Mobile Context-Aware Recommender System". Neural Information Processing. Lecture Notes in Computer Science. 7665. p. 324. doi:10.1007/978-3-642-34487-9_40 (https://doi.org/10.1007%2F978-3-642-34487-9_40). ISBN 978-3-642-34486-2.
34. Scott, S.L. (2010), "A modern Bayesian look at the multi-armed bandit", Applied Stochastic Models in Business and Industry, 26 (2): 639–658, doi:10.1002/asmb.874 (https://doi.org/10.1002%2Fasmb.874)
35. Olivier Chapelle, Lihong Li (2011), "An empirical evaluation of Thompson sampling" (http://papers.nips.cc/paper/4321-an-empirical-evaluation-of-thompson-sampling), Advances in Neural Information Processing Systems 24 (NIPS), Curran Associates: 2249–2257

36. Bouneffouf, D. (2018). "Incorporating Behavioral Constraints in Online AI Systems". The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19). AAAI.: 270–279. arXiv:1809.05720 (https://arxiv.org/abs/1809.05720). https://arxiv.org/abs/1809.05720%7Cyear=2019
37. Langford, John; Zhang, Tong (2008), "The Epoch-Greedy Algorithm for Contextual Multi-armed Bandits" (http://papers.nips.cc/paper/3178-the-epoch-greedy-algorithm-for-multi-armed-bandits-with-side-information), Advances in Neural Information Processing Systems 20, Curran Associates, Inc., pp. 817–824
38. Lihong Li, Wei Chu, John Langford, Robert E. Schapire (2010), "A contextual-bandit approach to personalized news article recommendation", Proceedings of the 19th International Conference on World Wide Web (WWW 2010): 661–670, arXiv:1003.0146 (https://arxiv.org/abs/1003.0146), Bibcode:2010arXiv1003.0146L (https://ui.adsabs.harvard.edu/abs/2010arXiv1003.0146L), doi:10.1145/1772690.1772758 (https://doi.org/10.1145%2F1772690.1772758), ISBN 9781605587998, S2CID 207178795 (https://api.semanticscholar.org/CorpusID:207178795)
39. Wei Chu, Lihong Li, Lev Reyzin, Robert E. Schapire (2011), "Contextual bandits with linear payoff functions" (http://proceedings.mlr.press/v15/chu11a/chu11a.pdf) (PDF), Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS): 208–214
40. Auer, P. (2000). "Using upper confidence bounds for online learning". Proceedings 41st Annual Symposium on Foundations of Computer Science. IEEE Comput. Soc. pp. 270–279. doi:10.1109/sfcs.2000.892116 (https://doi.org/10.1109%2Fsfcs.2000.892116). ISBN 978-0769508504. S2CID 28713091 (https://api.semanticscholar.org/CorpusID:28713091). Missing or empty |title= (help)
41. Hong, Tzung-Pei; Song, Wei-Ping; Chiu, Chu-Tien (November 2011). Evolutionary Composite Attribute Clustering. 2011 International Conference on Technologies and Applications of Artificial Intelligence. IEEE. doi:10.1109/taai.2011.59 (https://doi.org/10.1109%2Ftaai.2011.59). ISBN 9781457721748. S2CID 14125100 (https://api.semanticscholar.org/CorpusID:14125100).
42. Optimal Exploitation of Clustering and History Information in Multi-Armed Bandit.
43. Bouneffouf, D. (2019). Optimal Exploitation of Clustering and History Information in Multi-Armed Bandit. Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligen. AAAI. Soc. pp. 270–279. doi:10.1109/sfcs.2000.892116 (https://doi.org/10.1109%2Fsfcs.2000.892116). ISBN 978-0769508504. S2CID 28713091 (https://api.semanticscholar.org/CorpusID:28713091).
44. Rigollet, Philippe; Zeevi, Assaf (2010), Nonparametric Bandits with Covariates, Conference on Learning Theory, COLT 2010, arXiv:1003.1630 (https://arxiv.org/abs/1003.1630), Bibcode:2010arXiv1003.1630R (https://ui.adsabs.harvard.edu/abs/2010arXiv1003.1630R)
45. Slivkins, Aleksandrs (2011), Contextual bandits with similarity information. (http://www.jmlr.org/papers/volume15/slivkins14a/slivkins14a.pdf) (PDF), Conference on Learning Theory, COLT 2011
46. Perchet, Vianney; Rigollet, Philippe (2013), "The multi-armed bandit problem with covariates", Annals of Statistics, 41 (2): 693–721, arXiv:1110.6084 (https://arxiv.org/abs/1110.6084), doi:10.1214/13-aos1101 (https://doi.org/10.1214%2F13-aos1101), S2CID 14258665 (https://api.semanticscholar.org/CorpusID:14258665)
47. Sarah Filippi, Olivier Cappé, Aurélien Garivier, Csaba Szepesvári (2010), "Parametric Bandits: The Generalized Linear Case" (http://papers.nips.cc/paper/4166-parametric-bandits-the-generalized-linear-case), Advances in Neural Information Processing Systems 23 (NIPS), Curran Associates: 586–594
48. Lihong Li, Yu Lu, Dengyong Zhou (2017), "Provably optimal algorithms for generalized linear contextual bandits" (http://proceedings.mlr.press/v70/li17c.html), Proceedings of the 34th International Conference on Machine Learning (ICML): 2071–2080, arXiv:1703.00048 (https://arxiv.org/abs/1703.00048), Bibcode:2017arXiv170300048L (https://ui.adsabs.harvard.edu/abs/2017arXiv170300048L)
49. Kwang-Sung Jun, Aniruddha Bhargava, Robert D. Nowak, Rebecca Willett (2017), "Scalable generalized linear bandits: Online computation and hashing" (http://papers.nips.cc/paper/6615-scalable-generalized-linear-bandits-online-computation-and-hashing), Advances in Neural Information Processing Systems 30 (NIPS), Curran Associates: 99–109, arXiv:1706.00136 (https://arxiv.org/abs/1706.00136), Bibcode:2017arXiv170600136J (https://ui.adsabs.harvard.edu/abs/2017arXiv170600136J)

50. Branislav Kveton, Manzil Zaheer, Csaba Szepesvári, Lihong Li, Mohammad Ghavamzadeh, Craig Boutilier (2020), "Randomized exploration in generalized linear bandits", Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS), arXiv:1906.08947 (https://arxiv.org/abs/1906.08947), Bibcode:2019arXiv190608947K (https://ui.adsabs.harvard.edu/abs/2019arXiv190608947K)
51. Allesiardo, Robin; Féraud, Raphaël; Djallel, Bouneffouf (2014), "A Neural Networks Committee for the Contextual Bandit Problem", Neural Information Processing – 21st International Conference, ICONIP 2014, Malaisia, November 03-06,2014, Proceedings, Lecture Notes in Computer Science, 8834, Springer, pp. 374–381, arXiv:1409.8191 (https://arxiv.org/abs/1409.8191), doi:10.1007/978-3-319-12637-1_47 (https://doi.org/10.1007%2F978-3-319-12637-1_47), ISBN 978-3-319-12636-4, S2CID 14155718 (https://api.semanticscholar.org/CorpusID:14155718)
52. Michal Valko; Nathan Korda; Rémi Munos; Ilias Flaounas; Nello Cristianini (2013), Finite-Time Analysis of Kernelised Contextual Bandits, 29th Conference on Uncertainty in Artificial Intelligence (UAI 2013) and (JFPDA 2013)., arXiv:1309.6869 (https://arxiv.org/abs/1309.6869), Bibcode:2013arXiv1309.6869V (https://ui.adsabs.harvard.edu/abs/2013arXiv1309.6869V)
53. Féraud, Raphaël; Allesiardo, Robin; Urvoy, Tanguy; Clérot, Fabrice (2016). "Random Forest for the Contextual Bandit Problem" (http://jmlr.org/proceedings/papers/v51/feraud16.html). Aistats: 93–101.
54. Alekh Agarwal, Daniel J. Hsu, Satyen Kale, John Langford, Lihong Li, Robert E. Schapire (2014), "Taming the monster: A fast and simple algorithm for contextual bandits" (http://proceedings.mlr.press/v32/agarwalb14.html), Proceedings of the 31st International Conference on Machine Learning (ICML): 1638–1646, arXiv:1402.0555 (https://arxiv.org/abs/1402.0555), Bibcode:2014arXiv1402.0555A (https://ui.adsabs.harvard.edu/abs/2014arXiv1402.0555A)
55. Contextual Bandit with Restricted Context, Djallel Bouneffouf, 2017 <https://www.ijcai.org/Proceedings/2017/0203.pdf>
56. Badanidiyuru, A.; Langford, J.; Slivkins, A. (2014), "Resourceful contextual bandits" (http://www.jmlr.org/proceedings/papers/v35/badanidiyuru14.pdf) (PDF), Proceeding of Conference on Learning Theory (COLT)
57. Wu, Huasen; Srikant, R.; Liu, Xin; Jiang, Chong (2015), "Algorithms with Logarithmic or Sublinear Regret for Constrained Contextual Bandits" (https://papers.nips.cc/paper/6008-algorithms-with-logarithmic-or-sublinear-regret-for-constrained-contextual-bandits), The 29th Annual Conference on Neural Information Processing Systems (NIPS), Curran Associates: 433–441, arXiv:1504.06937 (https://arxiv.org/abs/1504.06937), Bibcode:2015arXiv150406937W (https://ui.adsabs.harvard.edu/abs/2015arXiv150406937W)
58. Burtini, Giuseppe, Jason Loeppky, and Ramon Lawrence. "A survey of online experiment design with the stochastic multi-armed bandit." arXiv preprint arXiv:1510.00757 (https://arxiv.org/abs/1510.00757) (2015).
59. Seldin, Y., Szepesvári, C., Auer, P. and Abbasi-Yadkori, Y., 2012, December. Evaluation and Analysis of the Performance of the EXP3 Algorithm in Stochastic Environments. In EWRL (pp. 103–116).
60. Hutter, M. and Poland, J., 2005. Adaptive online prediction by following the perturbed leader (http://www.jmlr.org/papers/volume6/hutter05a/hutter05a.pdf). Journal of Machine Learning Research, 6(Apr), pp.639–660.
61. Discounted UCB, Levente Kocsis, Csaba Szepesvári, 2006
62. On Upper-Confidence Bound Policies for Non-Stationary Bandit Problems, Garivier and Moulines, 2008 <https://arxiv.org/abs/0805.3415>
63. Improving Online Marketing Experiments with Drifting Multi-armed Bandits, Giuseppe Burtini, Jason Loeppky, Ramon Lawrence, 2015 <http://www.scitepress.org/DigitalLibrary/PublicationsDetail.aspx?ID=Dx2xXEB0PJE=&t=1>
64. Bouneffouf, Djallel; Feraud, Raphael (2016), "Multi-armed bandit problem with known trend", Neurocomputing
65. Yue, Yisong; Broder, Josef; Kleinberg, Robert; Joachims, Thorsten (2012), "The K-armed dueling bandits problem", Journal of Computer and System Sciences, 78 (5): 1538–1556, CiteSeerX 10.1.1.162.2764 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.162.2764), doi:10.1016/j.jcss.2011.12.028 (https://doi.org/10.1016%2Fj.jcss.2011.12.028)
66. Yue, Yisong; Joachims, Thorsten (2011), "Beat the Mean Bandit", Proceedings of ICML'11

67. Urvoy, Tanguy; Clérot, Fabrice; Féraud, Raphaël; Naamane, Sami (2013), "Generic Exploration and K-armed Voting Bandits" (http://www.jmlr.org/proceedings/papers/v28/urvoy13.pdf) (PDF), Proceedings of the 30th International Conference on Machine Learning (ICML-13)
68. Zoghi, Masrour; Whiteson, Shimon; Munos, Remi; Rijke, Maarten D (2014), "Relative Upper Confidence Bound for the K-Armed Dueling Bandit Problem" (http://www.jmlr.org/proceedings/papers/v32/zoghi14.pdf) (PDF), Proceedings of the 31st International Conference on Machine Learning (ICML-14)
69. Gajane, Pratik; Urvoy, Tanguy; Clérot, Fabrice (2015), "A Relative Exponential Weighing Algorithm for Adversarial Utility-based Dueling Bandits" (http://jmlr.org/proceedings/papers/v37/gajane15.pdf) (PDF), Proceedings of the 32nd International Conference on Machine Learning (ICML-15)
70. Zoghi, Masrour; Karnin, Zohar S; Whiteson, Shimon; Rijke, Maarten D (2015), "Copeland Dueling Bandits", Advances in Neural Information Processing Systems, NIPS'15, arXiv:1506.00312 (https://arxiv.org/abs/1506.00312), Bibcode:2015arXiv150600312Z (https://ui.adsabs.harvard.edu/abs/2015arXiv150600312Z)
71. Komiyama, Junpei; Honda, Junya; Kashima, Hisashi; Nakagawa, Hiroshi (2015), "Regret Lower Bound and Optimal Algorithm in Dueling Bandit Problem" (http://jmlr.org/proceedings/papers/v40/Komiyama15.pdf) (PDF), Proceedings of the 28th Conference on Learning Theory
72. Wu, Huasen; Liu, Xin (2016), "Double Thompson Sampling for Dueling Bandits", The 30th Annual Conference on Neural Information Processing Systems (NIPS), arXiv:1604.07101 (https://arxiv.org/abs/1604.07101), Bibcode:2016arXiv160407101W (https://ui.adsabs.harvard.edu/abs/2016arXiv160407101W)
73. Li, Shuai; Alexandros, Karatzoglou; Gentile, Claudio (2016), "Collaborative Filtering Bandits", The 39th International ACM SIGIR Conference on Information Retrieval (SIGIR 2016), arXiv:1502.03473 (https://arxiv.org/abs/1502.03473), Bibcode:2015arXiv150203473L (https://ui.adsabs.harvard.edu/abs/2015arXiv150203473L)
74. Gentile, Claudio; Li, Shuai; Zappella, Giovanni (2014), "Online Clustering of Bandits", The 31st International Conference on Machine Learning, Journal of Machine Learning Research (ICML 2014), arXiv:1401.8257 (https://arxiv.org/abs/1401.8257), Bibcode:2014arXiv1401.8257G (https://ui.adsabs.harvard.edu/abs/2014arXiv1401.8257G)
75. Gai, Y. and Krishnamachari, B. and Jain, R. (2010), Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation (http://www.academia.edu/download/30758682/DySPAN2010.pdf) (PDF), pp. 1–9
76. Chen, Wei and Wang, Yajun and Yuan, Yang (2013), Combinatorial multi-armed bandit: General framework and applications (http://www.jmlr.org/proceedings/papers/v28/chen13a.pdf) (PDF), pp. 151–159
77. Santiago Ontañón (2017), "Combinatorial Multi-armed Bandits for Real-Time Strategy Games" (https://www.jair.org/index.php/jair/article/download/11053/26230), Journal of Artificial Intelligence Research, 58: 665–702, arXiv:1710.04805 (https://arxiv.org/abs/1710.04805), Bibcode:2017arXiv171004805O (https://ui.adsabs.harvard.edu/abs/2017arXiv171004805O), doi:10.1613/jair.5398 (https://doi.org/10.1613%2Fjair.5398), S2CID 8517525 (https://api.semanticscholar.org/CorpusID:8517525)

Further reading

- Guha, S.; Munagala, K.; Shi, P. (2010). "Approximation algorithms for restless bandit problems". Journal of the ACM. 58: 1–50. arXiv:0711.3861 (https://arxiv.org/abs/0711.3861). doi:10.1145/1870103.1870106 (https://doi.org/10.1145%2F1870103.1870106). S2CID 1654066 (https://api.semanticscholar.org/CorpusID:1654066).
- Dayanik, S.; Powell, W.; Yamazaki, K. (2008), "Index policies for discounted bandit problems with availability constraints", Advances in Applied Probability, 40 (2): 377–400, doi:10.1239/aap/1214950209 (https://doi.org/10.1239%2Faap%2F1214950209).

- Powell, Warren B. (2007), "Chapter 10", Approximate Dynamic Programming: Solving the Curses of Dimensionality, New York: John Wiley and Sons, ISBN 978-0-470-17155-4.
- Robbins, H. (1952), "Some aspects of the sequential design of experiments", Bulletin of the American Mathematical Society, 58 (5): 527–535, doi:10.1090/S0002-9904-1952-09620-8 (https://doi.org/10.1090%2FS0002-9904-1952-09620-8).
- Sutton, Richard; Barto, Andrew (1998), Reinforcement Learning (https://web.archive.org/web/20131211192714/http://webdocs.cs.ualberta.ca/~sutton/book/the-book.html), MIT Press, ISBN 978-0-262-19398-6, archived from the original (http://webdocs.cs.ualberta.ca/~sutton/book/the-book.html) on 2013-12-11.
- Allesiardo, Robin (2014), "A Neural Networks Committee for the Contextual Bandit Problem", Neural Information Processing – 21st International Conference, ICONIP 2014, Malaisia, November 03-06,2014, Proceedings, Lecture Notes in Computer Science, 8834, Springer, pp. 374–381, arXiv:1409.8191 (https://arxiv.org/abs/1409.8191), doi:10.1007/978-3-319-12637-1_47 (https://doi.org/10.1007%2F978-3-319-12637-1_47), ISBN 978-3-319-12636-4, S2CID 14155718 (https://api.semanticscholar.org/CorpusID:14155718).
- Weber, Richard (1992), "On the Gittins index for multiarmed bandits", Annals of Applied Probability, 2 (4): 1024–1033, doi:10.1214/aoap/1177005588 (https://doi.org/10.1214%2Faoap%2F1177005588), JSTOR 2959678 (https://www.jstor.org/stable/2959678).
- Katehakis, M. and C. Derman (1986), "Computing optimal sequential allocation rules in clinical trials", Adaptive statistical procedures and related topics, Institute of Mathematical Statistics Lecture Notes - Monograph Series, 8, pp. 29–39, doi:10.1214/lnms/1215540286 (https://doi.org/10.1214%2Flnms%2F1215540286), ISBN 978-0-940600-09-6, JSTOR 4355518 (https://www.jstor.org/stable/4355518).
- Katehakis, M. and A. F. Veinott, Jr. (1987), "The multi-armed bandit problem: decomposition and computation" (https://semanticscholar.org/paper/e4fe28113fed71999a0db30a930e0b42d3ce55f1), Mathematics of Operations Research, 12 (2): 262–268, doi:10.1287/moor.12.2.262 (https://doi.org/10.1287%2Fmoor.12.2.262), JSTOR 3689689 (https://www.jstor.org/stable/3689689), S2CID 656323 (https://api.semanticscholar.org/CorpusID:656323).

External links

- MABWiser (https://github.com/fmr-llc/mabwiser), open source Python implementation of bandit strategies that supports context-free, parametric and non-parametric contextual policies with built-in parallelization and simulation capability.
- PyMaBandits (http://mloss.org/software/view/415/), open source implementation of bandit strategies in Python and Matlab.
- Contextual (https://github.com/Nth-iteration-labs/contextual), open source R package facilitating the simulation and evaluation of both context-free and contextual Multi-Armed Bandit policies.

- bandit.sourceforge.net Bandit project (http://bandit.sourceforge.net), open source implementation of bandit strategies.
- Banditlib (https://github.com/jkomiyama/banditlib), Open-Source implementation of bandit strategies in C++.
- Leslie Pack Kaelbling and Michael L. Littman (1996). Exploitation versus Exploration: The Single-State Case (https://archive.is/20121212095047/http://www.cs.washington.edu/research/jair/volume4/kaelbling96a-html/node6.html).
- Tutorial: Introduction to Bandits: Algorithms and Theory. Part1 (http://techtalks.tv/talks/54451/). Part2 (http://techtalks.tv/talks/54455/).
- Feynman's restaurant problem (https://www.feynmanlectures.caltech.edu/info/exercises/Feynmans_restaurant_problem.html), a classic example (with known answer) of the exploitation vs. exploration tradeoff.
- Bandit algorithms vs. A-B testing (http://www.chrisstucchio.com/blog/2012/bandit_algorithms_vs_ab.html).
- S. Bubeck and N. Cesa-Bianchi A Survey on Bandits (http://homes.di.unimi.it/~cesabian/Pubblicazioni/banditSurvey.pdf).
- A Survey on Contextual Multi-armed Bandits (https://arxiv.org/abs/1508.03326), a survey/tutorial for Contextual Bandits.
- Blog post on multi-armed bandit strategies, with Python code (https://mpatacchiola.github.io/blog/2017/08/14/dissecting-reinforcement-learning-6.html).
- Animated, interactive plots (https://pavlov.tech/2019/03/02/animated-multi-armed-bandit-policies/) illustrating Epsilon-greedy, Thompson sampling, and Upper Confidence Bound exploration/exploitation balancing strategies.

Retrieved from "https://en.wikipedia.org/w/index.php?title=Multi-armed_bandit&oldid=984830525"

This page was last edited on 22 October 2020, at 10:49 (UTC).

Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.