

Multi-armed bandit

From Wikipedia, the free encyclopedia

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 gambler at a row of slot machines (sometimes known as "one-armed bandits") has to decide which machines to play, how many times to play each machine and in which order to play them.^[3] When played, 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.^{[4][5]}

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]

In practice, multi-armed bandits have been used to model the problem of managing research projects in a large organization, like a science foundation or a pharmaceutical company. Given a fixed budget, the problem is to allocate resources among the competing projects, whose properties are only partially known at the time of allocation, but which may become better understood as time passes.^{[4][5]}

In early versions of the multi-armed bandit problem, the gambler has no initial knowledge about the machines. 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 reinforcement learning.



A row of slot machines in Las Vegas.

Contents

- 1 Empirical motivation
- 2 The multi-armed bandit model
- 3 Variations
- 4 Bandit strategies
 - 4.1 Optimal solutions
 - 4.2 Approximate solutions
 - 4.2.1 Semi-uniform strategies
 - 4.2.2 Probability matching strategies
 - 4.2.3 Pricing strategies
 - 4.2.4 Strategies with ethical constraints
- 5 Contextual bandit

- 5.1 Approximate solutions for contextual bandit
 - 5.1.1 Online linear classifier
 - 5.1.2 Online non-linear classifier
- 5.2 Constrained contextual bandit
- 6 Adversarial bandit
- 7 With known trend
- 8 Infinite armed bandit
- 9 Dueling bandit
- 10 Non-stationary bandit
- 11 Clustering bandit
- 12 Distributed bandit
- 13 Collaborative bandit
- 14 Spatially correlated bandit
- 15 See also
- 16 References
- 17 Further reading
- 18 External links

Empirical motivation

The multi-armed bandit problem models an agent that simultaneously attempts to acquire new knowledge (called "exploration") and optimize his or her decisions based on existing knowledge (called "exploitation"). The agent attempts to balance these competing tasks in order to maximize his 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,^{[4][5][8][9]}
- adaptive routing efforts for minimizing delays in a network,
- financial portfolio design^{[10][11]}

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 reinforcement 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]



How to distribute a given budget among these research departments to maximize results?

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*) can be seen as a set of real distributions $\mathcal{B} = \{R_1, \dots, R_K\}$, each distribution being associated with the rewards delivered by one of the $K \in \mathbb{N}^+$ levers. Let μ_1, \dots, μ_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 at time 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 generalisation 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]

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:**^[26] 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).^[27] 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).^[28]
- **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 let 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.^[29]

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,^[30] 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

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.^[31]

Approximate solutions for contextual bandit

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

Online linear classifier

- **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.

Online non-linear classifier

- **UCBogram algorithm:** The nonlinear reward functions are estimated using piecewise constant over a functions using a piecewise constant estimator called *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.^{[32][33][34]}
- **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.^[35]
- **KernelUCB algorithm:** a kernelized non-linear version of linearUCB, with efficient implementation and finite-time analysis.^[36]
- **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.^[37]

Constrained contextual bandit

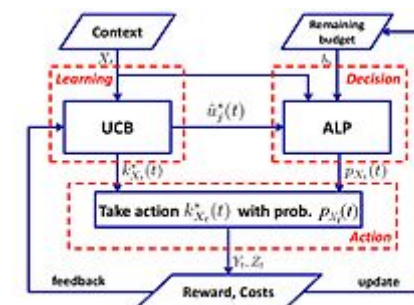
In practice, there is usually a cost associated with the resource consumed by each action and the total cost is limited by a budget in many applications such as crowdsourcing and clinical trials. Constrained contextual bandit (CCB) is such a model that consider both the time and budget constraints in multi-armed bandit setting. A. Badanidiyuru et al.^[38] first studies the contextual bandits with budget constraints, also referred to as Resourceful Contextual Bandits, and show that a $O(\sqrt{T})$ regret is achievable. However,^[38] focuses on a finite set of policies and the algorithm is computationally inefficient.

A simple algorithm with logarithmic regret is proposed in:^[39]

- **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^[39] 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.

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^[40] as it removes all



Framework of UCB-ALP for constrained contextual bandits

assumptions of the distribution and a solution to the adversarial bandit problem is a generalized solution to the more specific bandit problems.

With known trend

Multi-armed bandit problem with known trend is a variant of the multi-armed bandit model, where the gambler knows the shape of the reward function of each arm but not its distribution. This new problem is motivated by different on-line problems like active learning, music and interface recommendation applications, where when an arm is sampled by the model the received reward change according to a known trend. By adapting the standard multi-armed bandit algorithm UCB1 to take advantage of this setting, the authors in^[41] propose the new algorithm named Adjusted Upper Confidence Bound (A-UCB) that assumes a stochastic model and provide upper bounds of the regret which compare favorably with the ones of UCB1.

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.

Dueling bandit

The dueling bandit variant was introduced by Yue et al. (2012)^[42] 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,^[42] Beat-The-Mean.^[43] The relative feedback of dueling bandits can also lead to voting paradoxes. A solution is to take the Condorcet winner as a reference.^[44]

More recently, researchers have generalized algorithms from traditional MAB to dueling bandits: Relative Upper Confidence Bounds (RUCB),^[45] Relative EXponential weighing (REX3),^[46] Copeland Confidence Bounds (CCB),^[47] Relative Minimum Empirical Divergence (RMED),^[48] and Double Thompson Sampling (DTS).^[49]

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^[50] and Sliding-Window UCB.^[51]

Another work by Burtini et al. introduces a weighted least squares Thompson sampling approach, which proves beneficial in both the known and unknown non-stationary cases.^[52]

Clustering bandit

The clustering of bandits (i.e., CLUB) was introduced by Gentile and Li and Zappela (ICML 2014),^[53] with a novel algorithmic approach to content recommender systems based on adaptive clustering of exploration-exploitation ("bandit") strategies. They provide a sharp regret analysis of this algorithm in a standard stochastic noise setting, demonstrate its scalability properties, and prove its effectiveness on a number of artificial and real-world datasets. Their experiments show a significant increase in prediction performance over state-of-the-art methods for bandit problems.

Distributed bandit

The distributed clustering of bandits (i.e., DCCB) was introduced by Korda and Szorenyi and Li (ICML 2016),^[54] they provide two distributed confidence ball algorithms for solving linear bandit problems in peer to peer networks with limited communication capabilities. For the first, they assume that all the peers are solving the same linear bandit problem, and prove that their algorithm achieves the optimal asymptotic regret rate of any centralised algorithm that can instantly communicate information between the peers. For the second, they assume that there are clusters of peers solving the same bandit problem within each cluster as in,^[53] and they prove that their algorithm discovers these clusters, while achieving the optimal asymptotic regret rate within each one. Through experiments on several real-world datasets, they demonstrate the performance of proposed algorithms compared to the state-of-the-art.

Collaborative bandit

The collaborative filtering bandits (i.e., COFIBA) was introduced by Li and Karatzoglou and Gentile (SIGIR 2016),^[55] 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.^[53] Their algorithm (COFIBA, pronounced as "Coffee Bar") takes into account the collaborative effects^[55] 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.



Spatially correlated bandit

The spatially correlated multi-armed bandit (SCMAB) was introduced by Wu, Schulz, Speekenbrink, Nelson, and Meder (2017)^[56], to study how humans generalize from observed to unobserved outcomes in limited horizon search. An underlying reward function was used to map the spatial location of each playable arm of the bandit to a mean reward, where the correlation between rewards decreased as an exponential function of the distance between two arms (generated using a radial basis function kernel). So far, SCMABs have been used to study human behavior, with the finding that human choices are best predicted by combining Gaussian process regression as a function learning mechanism with Upper confidence bound sampling^[56].

See also

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

References

1. Auer, P.; Cesa-Bianchi, N.; Fischer, P. (2002). "Finite-time Analysis of the Multiarmed Bandit Problem". *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". *Mathematics of Operations Research*. **12** (2): 262–268. doi:10.1287/moor.12.2.262 (https://doi.org/10.1287%2Fmoor.12.2.262).
3. Weber, Richard (1992), "On the Gittins index for multiarmed bandits", *Annals of Applied Probability*, **2** (4): 1024–1033, JSTOR 2959678 (https://www.jstor.org/stable/2959678), doi:10.1214/aoap/1177005588 (https://doi.org/10.1214%2Faoap%2F1177005588)
4. Gittins, J. C. (1989), *Multi-armed bandit allocation indices*, Wiley-Interscience Series in Systems and Optimization., Chichester: John Wiley & Sons, Ltd., ISBN 0-471-92059-2
5. Berry, Donald A.; Fristedt, Bert (1985), *Bandit problems: Sequential allocation of experiments*, Monographs on Statistics and Applied Probability, London: Chapman & Hall, ISBN 0-412-24810-7
6. 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).
7. J. C. Gittins (1979). "Bandit Processes and Dynamic Allocation Indices". *Journal of the Royal Statistical Society. Series B (Methodological)*. **41** (2): 148–177. 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" (http://www.pnas.org/content/106/52/22387), *Proceedings of the National Academy of Sciences*, **106** (52): 22387–22392, PMC 2793317 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2793317) , PMID 20018711 (https://www.ncbi.nlm.nih.gov/pubmed/20018711), doi:10.1073/pnas.0912378106 (https://doi.org/10.1073%2Fpnas.0912378106).
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) 
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" (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.380.6983&rep=rep1&type=pdf), *Operations Research*, **59** (2): 383–399
13. Whittle, Peter (1979), "Discussion of Dr Gittins' paper", *Journal of the Royal Statistical Society, Series B*, **41** (2): 165, JSTOR 2985029 (https://www.jstor.org/stable/2985029)
14. Vermorel, Joannes; Mohri, Mehryar (2005), *Multi-armed bandit algorithms and empirical evaluation* (http://bandit.sourceforge.net/Vermorel2005poker.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, MR 974588 (<https://www.ams.org/mathscinet-getitem?mr=974588>), doi:10.2307/3214163 (<https://doi.org/10.2307%2F3214163>)
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. 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. PMC 41010 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC41010>). PMID 11607577 (<https://www.ncbi.nlm.nih.gov/pubmed/11607577>). doi:10.1073/pnas.92.19.8584 (<https://doi.org/10.1073%2Fpnas.92.19.8584>).
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>).
26. Sutton, R. S. & Barto, A. G. 1998 Reinforcement learning: an introduction. Cambridge, MA: MIT Press.
27. Tokic, Michel (2010), "Adaptive ϵ -greedy exploration in reinforcement learning based on value differences", *KI 2010: Advances in Artificial Intelligence* (<http://www.tokic.com/www/tokicm/publikationen/papers/AdaptiveEpsilonGreedyExploration.pdf>) (PDF), Lecture Notes in Computer Science, **6359**, Springer-Verlag, pp. 203–210, ISBN 978-3-642-16110-0, doi:10.1007/978-3-642-16111-7_23 (https://doi.org/10.1007%2F978-3-642-16111-7_23).
28. Tokic, Michel; Palm, Günther (2011), "Value-Difference Based Exploration: Adaptive Control Between Epsilon-Greedy and Softmax", *KI 2011: Advances in Artificial Intelligence* (<http://www.tokic.com/www/tokicm/publikationen/papers/KI2011.pdf>) (PDF), Lecture Notes in Computer Science, **7006**, Springer-Verlag, pp. 335–346, ISBN 978-3-642-24455-1.
29. 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. ISBN 978-3-642-34486-2. doi:10.1007/978-3-642-34487-9_40 (https://doi.org/10.1007%2F978-3-642-34487-9_40).
30. 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>)
31. Langford, John; Zhang, Tong (2008), "The Epoch-Greedy Algorithm for Contextual Multi-armed Bandits", *Advances in Neural Information Processing Systems 20* (<http://papers.nips.cc/paper/3178-the-epoch-greedy-algorithm-for-multi-armed-bandits-with-side-information>), Curran Associates, Inc., pp. 817–824

32. Rigollet, Philippe; Zeevi, Assaf (2010), *Nonparametric Bandits with Covariates*, Conference on Learning Theory, COLT 2010
33. Slivkins, Aleksandrs (2011), *Contextual bandits with similarity information.*, Conference on Learning Theory, COLT 2011
34. Perchet, Vianney; Rigollet, Philippe (2013), "The multi-armed bandit problem with covariates", *Annals of Statistics*, **41** (2): 693–721, doi:10.1214/13-aos1101 (<https://doi.org/10.1214%2F13-aos1101>)
35. Allesiardo, Robin; Féraud, Raphaël; Djallel, Bouneffouf (2014), "A Neural Networks Committee for the Contextual Bandit Problem", *Neural Information Processing - 21st International Conference, ICONIP 2014, Malaysia, November 03-06, 2014, Proceedings*, Lecture Notes in Computer Science, **8834**, Springer, pp. 374–381, ISBN 978-3-319-12636-4, doi:10.1007/978-3-319-12637-1_47 (https://doi.org/10.1007%2F978-3-319-12637-1_47)
36. 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>) 
37. 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*.
38. Badanidiyuru, A.; Langford, J.; Slivkins, A. (2014), "Resourceful contextual bandits", *Proceeding of Conference on Learning Theory (COLT)*
39. 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)*
40. Burtini (2015)
41. Bouneffouf, Djallel; Feraud, Raphael (2016), "Multi-armed bandit problem with known trend", *Neurocomputing*
42. Yue, Yisong; Broder, Josef; Kleinberg, Robert; Joachims, Thorsten (2012), "The K-armed Dueling Bandits Problem", *Journal of Computer and System Sciences* (<http://www.sciencedirect.com/science/article/pii/S0022000012000281>), **78** (5), pp. 1538–1556, doi:10.1016/j.jcss.2011.12.028 (<https://doi.org/10.1016%2Fj.jcss.2011.12.028>)
43. Yue, Yisong; Joachims, Thorsten (2011), "Beat the Mean Bandit", *Proceedings of ICML'11*
44. Urvoy, Tanguy; Clérot, Fabrice; Féraud, Raphaël; Naamane, Sami (2013), "Generic Exploration and K-armed Voting Bandits", *Proceedings of the 30th International Conference on Machine Learning (ICML-13)* (<http://www.jmlr.org/proceedings/papers/v28/urvoy13.pdf>) (PDF)
45. Zoghi, Masrour; Whiteson, Shimon; Munos, Remi; Rijke, Maarten D (2014), "Relative Upper Confidence Bound for the $\$K\$$ -Armed Dueling Bandit Problem", *Proceedings of the 31st International Conference on Machine Learning (ICML-14)* (<http://www.jmlr.org/proceedings/papers/v32/zoghi14.pdf>) (PDF)
46. Gajane, Pratik; Urvoy, Tanguy; Clérot, Fabrice (2015), "A Relative Exponential Weighing Algorithm for Adversarial Utility-based Dueling Bandits", *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)* (<http://jmlr.org/proceedings/papers/v37/gajane15.pdf>) (PDF)
47. 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>) 
48. Komiyama, Junpei; Honda, Junya; Kashima, Hisashi; Nakagawa, Hiroshi (2015), "Regret Lower Bound and Optimal Algorithm in Dueling Bandit Problem", *Proceedings of The 28th Conference on Learning Theory* (<http://jmlr.org/proceedings/papers/v40/Komiyama15.pdf>) (PDF)
49. 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>) 
50. Discounted UCB, Levente Kocsis, Csaba Szepesvári, 2006
51. On Upper-Confidence Bound Policies for Non-Stationary Bandit Problems, Garivier and Moulines, 2008 <<http://arxiv.org/abs/0805.3415>>
52. 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>>

53. 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>) 
54. Korda, Nathan Korda; Szorenyi, Balazs; Li, Shuai (2016), "Distributed Clustering of Linear Bandits in Peer to Peer Networks", *The 33rd International Conference on Machine Learning, Journal of Machine Learning Research (ICML 2016)*, arXiv:1604.07706 (<https://arxiv.org/abs/1604.07706>) 
55. 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>) 
56. Wu, Charley M; Schulz, Eric; Speekenbrink, Maarten; Nelson, Jonathan D; Meder, Björn (2017), "Mapping the unknown: The spatially correlated multi-armed bandit", *Proceedings of the 39th Annual Cognitive Science Society* (<http://biorxiv.org/content/biorxiv/early/2017/04/28/106286.full.pdf>) (PDF)

Further reading

- Guha, S.; Munagala, K.; Shi, P. (2010). "Approximation algorithms for restless bandit problems". *Journal of the ACM*. **58**: 1–50. doi:10.1145/1870103.1870106 (<https://doi.org/10.1145%2F1870103.1870106>).
- 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 0-470-17155-3.
- 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* (<http://webdocs.cs.ualberta.ca/~sutton/book/the-book.html>), MIT Press, ISBN 0-262-19398-1.
- Allesiardo, Robin (2014), "A Neural Networks Committee for the Contextual Bandit Problem", *Neural Information Processing - 21st International Conference, ICONIP 2014, Malaysia, November 03-06,2014, Proceedings*, Lecture Notes in Computer Science, **8834**, Springer, pp. 374–381, ISBN 978-3-319-12636-4, doi:10.1007/978-3-319-12637-1_47 (https://doi.org/10.1007%2F978-3-319-12637-1_47).
- Bouneffouf, Djallel (2012), "A Contextual-Bandit Algorithm for Mobile Context-Aware Recommender System", *Neural Information Processing - 19th International Conference, ICONIP 2012, Doha, Qatar, November 12-15,2012, Proceedings, Part III* (http://link.springer.com/chapter/10.1007%2F978-3-642-34487-9_40), Lecture Notes in Computer Science, **7665**, Springer, pp. 324–331, ISBN 978-3-642-34486-2, doi:10.1007/978-3-642-34487-9_40 (https://doi.org/10.1007%2F978-3-642-34487-9_40).
- Weber, Richard (1992), "On the Gittins index for multiarmed bandits", *Annals of Applied Probability*, **2** (4): 1024–1033, JSTOR 2959678 (<https://www.jstor.org/stable/2959678>), doi:10.1214/aoap/1177005588 (<https://doi.org/10.1214%2Faoap%2F1177005588>).
- Katehakis, M. and C. Derman (1986), "Computing Optimal Sequential Allocation Rules in Clinical Trials", *IMS Lecture Notes-Monograph Series*, **8**: 29–39, JSTOR 4355518 (<https://www.jstor.org/stable/4355518>), doi:10.1214/lnms/1215540286 (<https://doi.org/10.1214%2Flnms%2F1215540286>).
- Katehakis, M. and A. F. Veinott, Jr. (1987), "The multi-armed bandit problem: decomposition and computation", *Mathematics of Operations Research*, **12** (2): 262–268, JSTOR 3689689 (<https://www.jstor.org/stable/3689689>), doi:10.1287/moor.12.2.262 (<https://doi.org/10.1287%2Fmoor.12.2.262>).

External links

- PyMaBandits (<http://mloss.org/software/view/415/>), Open-Source implementation of bandit strategies in Python and Matlab
- 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 (<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 (http://www.feynmanlectures.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.

Retrieved from "https://en.wikipedia.org/w/index.php?title=Multi-armed_bandit&oldid=791415239"

-
- This page was last edited on 20 July 2017, at 03:39.
 - 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.