Statistical Learning for Decision*

Ben Lowery

January 20, 2022

Abstract

By establishing key concepts within the field, reviewing primarily Bayesian based methods, and stating some relevant theoretical guarantees, this report aims to give insight into the field of multi-armed bandits. The foray into the field focuses on Thompson sampling, a method with rather interesting origins and resurgence, before branching into other possible stochastic bandit policies and a culminating computational comparison of such.

1 Introduction

When it comes to decisions under uncertainty, little to nothing is known about the outcomes and these decisions need to be made using underlying assumptions of the situation at hand. Decisions under uncertainty is a problem focused on in regards to: (1) human choices and rational decision making (see expected utility and Von Neumann & Morgenstern (1953) theory) and (2) statistical learning (machine learning for example).

Within sequential decision making, are multi-armed bandit problems. These can be summarised as allocation of resources that maximise an expected gain for some partially known events. The idea is to find a strategy that balances two key areas, exploration and exploitation. With this, it is seen as a method for decision making under uncertainty viewed as one of the simplest problem set ups within the wider field of Reinforcement Learning.

2 Expected Utility

Expected utility concerns the choices made by individuals in complex situations for which factors such as risk appetite need to be taken into account. Through the definitions of some pre-defined axiomatic behaviour taken by a rational decision maker, they should aim to maximise the expected value of some function that models the potential outcomes of an event.

Foundations of expected utility are often traced to the St. Petersburg paradox with the question posed by N. Bernoulli in 1713 (Peterson 2020) and a corresponding solution by his cousin D. Bernoulli (1738). The paradox concerned itself with assuming infinite expected utility in a simple game of chance, when in reality the utility from participation could be modelled in such a way that a player would posses a 'risk-averse' strategy and model utility such that expected utility, and subsequently a stake in which is appropriate to play the game, could be given as a finite value. For a more broad explanation into the problem, alongside a solution that takes a non-utility approach, see Peters (2011).

Progressing from the work of the Bernoulli's, was a more formalised theory into utility by John von Neumann and Oskar Morgenstern (VNM) in their seminal work in the field *Theory of Games and Economic Behavior* (Von Neumann & Morgenstern 1953). The book primarily serves as the foundations for game theory but within contains pivotal work regarding risk attitudes and expected utility theory. The most pertinent of these are the four axioms essential in the application of VNM theory. Binmore (2008) presents an approachable and anecdotal explanation of the axioms, with an amalgamation of the two definitions defined in the following postulates.

*Special extended edition.

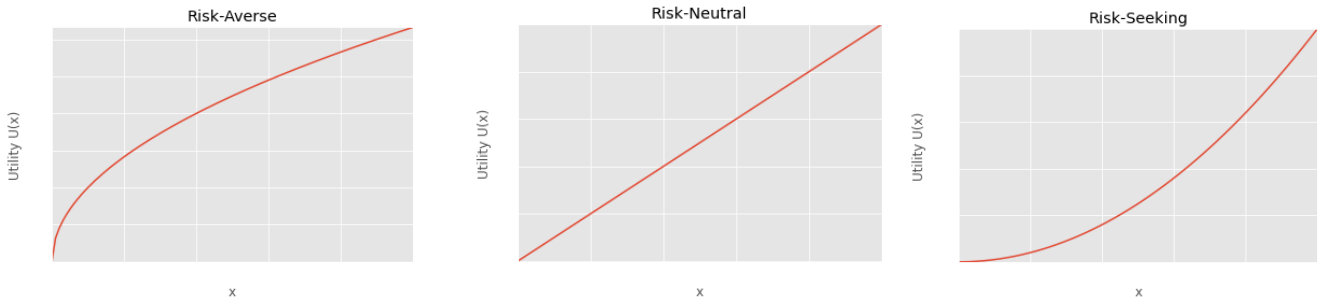
Postulate 1 (Completeness)**.** *There is a well defined set of preferences and the individual can clearly decide between two alternatives.*

Postulate 2 (Transitivity)**.** *Follows from the completeness criterion and states preferences are chosen consistently.*

Postulate 3 (Independence)**.** *An individual does not care about a new independent outcome if they are indifferent about itself and the one it is replacing.*

Postulate 4 (Continuity)**.** *Small changes in outcomes only lead to small changes in preference.*

A key subset of the theory, utility functions, models a preference relation between utility and some measurable quantity (wealth, food, etc.), the incorrect assumption of a utility function is something that often occurs and was behind the problems that plagued the St. Petersburg paradox earlier (Hansson 2005). Utility functions are required to have satisfied the four axioms from VNM Theory, and can be utilised to measure and model risk appetite. Under VNM theory, these attitudes to risk are consistent throughout, with three categories being risk-averse, risk-neutral, and risk-seeking behaviours. The first of these is represented by a concave function and provides a preference for low variance outcomes. A risk-neutral individual is indifferent towards an outcome and can be represented by an affine function. Whereas, risk-seeking individuals will have more utility in achieving a higher level of outcome, all of which can referred to in Figure 1. An important rule of the utility function is that utility is ordinal and that we objectively choose the higher number and **not** claim that one value is preferred a number of times more than another. As an example, if for events x_1 and x_2, $U(x_1) = 42$ and $U(x_2) = 3$, x_1 is not 14 times more preferred than x_2.



(a) Modelled by $U(x) = \ln(x)$ (b) Modelled by $U(x) = x$ (c) Modelled by $U(x) = x^2$

Figure 1: Utility functions modelling various risk appetites

Mathematically, these risk-attitudes can also be reduced to the relationship between the expected utility of some event, and the utility of an expected event. Let some quantity, x, represent food that a bird scavenges for over an arbitrary period of time. The bird has the choice as to whether go beyond what they usually are able to gather and risk dying in the process and obviously returning with nothing as they're dead, or settle for an amount they know is safe to be collected rather quickly. Depending on the scenario the bird is in, different risk-appetites may become apparent:

Scenario 1 (The bird is starving)**.** *In this scenario, a risk-seeking approach would need to be made to attain enough food for survival. To make sure this is reflected in their utility function, it would be required to see the bird value the expected utility from the amount of food it finds over the average amount of food it's expected to find, and thus forage for food even if it's more than should be expected. The inequality to represent this is given in the following Equation 1:*

$$\mathbb{E}[U(x)] \geq U(\mathbb{E}[x]). \tag{1}$$

Scenario 2 (The bird likely has the means to attain ample amounts of food under both scenarios)**.** *Here, they are indifferent to the amount of food they're expected to get, and the utility it attains from this. Therefore, it does not value taking the risk and potentially gaining a greater payoff, than it does going for a safe amount. This is represented by an equivalence and a risk-neutral approach seen below,*

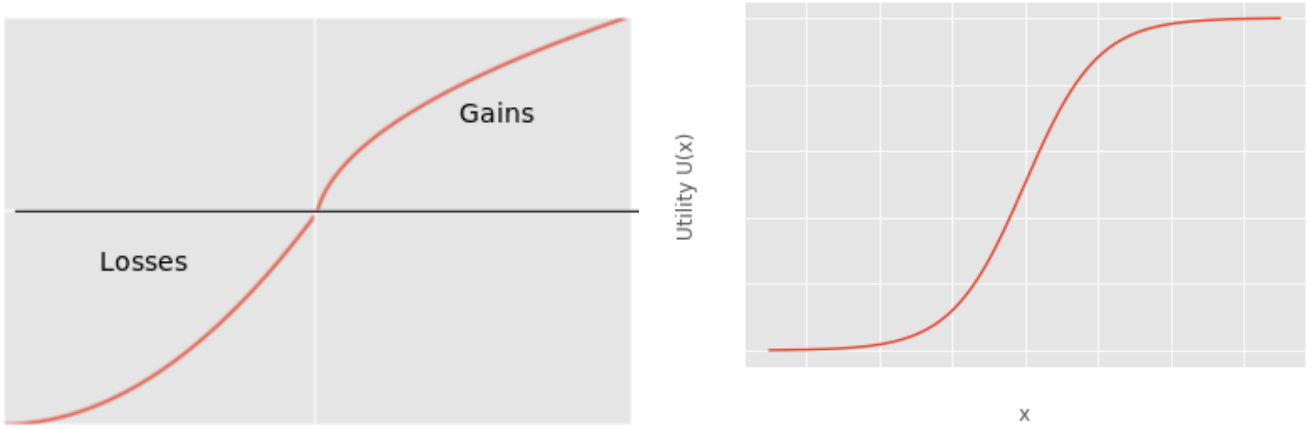
$$\mathbb{E}[U(x)] = U(\mathbb{E}[x]). \tag{2}$$

Scenario 3 (The bird likely only needs a small amount of extra food)**.** *Finally, here the bird only requires a small amount of extra food and is willing to take a safe payoff for this, even if it's less than the amount they usually expect to get. Here the inequality is a reversal of the risk-seeking approach and is a risk-averse strategy given as follows,*

$$\mathbb{E}[U(x)] \leq U(\mathbb{E}[x]). \tag{3}$$

These can be checked with the utility functions seen in Figure 1 to verify Equations (1)-(3) are satisfied.

VNM theory can be extended towards having shifting attitudes towards utility as the situation changes. Consider now the bird looking for a certain quantity of food to survive, except there is now a threshold as to which the bird will turn from not having enough food (a scarcity) to having enough food (a surplus) to survive an arbitrary amount of time. It could be reasonably assumed, based on the literature explored, the bird would be risk seeking for scarcity (or losses) and risk averse for when food becomes a surplus (or gains). The culminates in a proposal utility function compromising of a piece-wise function and is seen in Figure 2a, and has functions for risk-seeking as $U(x) = x^2$ and risk-aversion by $U(x) = log(x)$. Note, this assumes individuals aren't as risk averse as they are risk seeking, but this need not be the case as seen in Figure 2b, using the logistic function to model utility, the bird is as risk averse as they are risk seeking.



(a) Piecewise function with losses ($U(x) = x^2$) and gains ($U(x) = \ln(x)$).

(b) Logistic function for both losses and gains given by $U(x) = (1 + e^{-x})^{-1}$.

Figure 2: Example utility functions for the bird problem.

Expected utility theory is a diverse, interdisciplinary and often abstract concept, allowing decision makers to better assess different risk appetites individuals might have towards a problem setting. Ultimately, expected utility theory serves as to describe the rationale of decision making under risk and there is much more work in the field. Particular points of further interest include a more thorough exposition of expected utility and a Bayesian approach of expected utility given by Savage's representation theorem, contained within his work *The Foundations of Statistics* (Savage 1954). Alongside this is the Friedman-Savage utility function; a model of utility widely seen in economics in which there are multiple points of inflection leading to concave and convex regions (Friedman & Savage 1948). This has been used to explore pattern behaviours such as why people are prone to be risk averse for insurance but risk

seeking for gambling. Utility theory is not the only school of thought when it comes to choices under uncertainty. In the field of Economics lies *Prospect Theory*, introduced by Kahneman & Tversky (1979), which formulates its description of behaviour without as much emphasis on the axiomatic predispositions that are assumed under VNM theory.

3 Multi-armed bandits

Progressing on from the idea of making decisions under uncertainty, we consider a framework for which algorithms can be formulated to make decisions over time for a set of competing choices. Multi-armed bandits is a far reaching and dense field of work, stemming from initial beginnings by Thompson (1933). In the simplest sense, an algorithm encompassed within bandit problems involve choosing from K possible actions (or arms), and doing so for T rounds. Here, both K and T are known. This setup is specifically applicable to the case for models with independent and identically distributed rewards, known as stochastic bandits. Rewards in this sense pertain to only be collected by the arm selected by the algorithm, and are drawn independently from a fixed distribution which is not known by the algorithm (Slivkins 2019).

The *K-armed Bernoulli bandit* is a simple yet well studied model, with many pertinent applications. In this model the initial set up follows on from that of the aforementioned stochastic bandit problem, and for K actions, each individual action returns a binary reward, with success probability $\theta_k \in [0, 1]$. In the bandit setting, these success probabilities are fixed but unknown to the decision maker, with the goals of maximising the cumulative number of successes.

An important issue for bandit strategies to tackle is the exploitation/exploration trade-off. This is a key component in bandit strategies which tries to find parity between exploring new areas in hopes of better rewards, as well as persisting in areas it knows has had favourable rewards before. Russo et al. (2017) notes that the Bernoulli bandit problem lends itself well to being able to "crystallise" understanding of the exploitation/exploration trade-off.

3.1 Regret

To evaluate these algorithms, plots of cumulative regret are often employed. Given the array of possible strategies available, it would be useful to be able to evaluate the performance and determine a preference for which strategy is performing the best. Regret can be quantified as the gap between the optimal behaviour and the behaviour chosen up until that point.

Given K arms, T rounds and a set of actions a_t, we obtain a stochastic reward by selecting an action at round t, denote this as $X_{k,t}$. By defining the mean reward $\mu_k = \mathbb{E}[X_{k,t}]$ and the 'best' of these by $\mu^* = \max_{k \in \{1,\dots,K\}} \mu_k$, the expected cumulative regret is then given in Equation 4,

$$R(T) = \mu^* \cdot T - \sum_{t=1}^{T} \mathbb{E}\left[\mu_{a_t}\right]. \tag{4}$$

Note, the minimization of cumulative regret is equivalent to maximizing the cumulative reward.

With the exploration/exploitation trade-off in Bandit strategies, it is clear that a policy with zero regret is unattainable, as exploration would lead to some regret being apparent. Instead, there can be a focus on theoretical *regret bounds*, with much work in online optimisation focusing on finding these performance guarantees (Russo & Roy 2016). The next step is to consider popular strategies in the field and compare the regret these algorithms produce, as well as stating some theoretical guarantees on regret that have been found.

3.2 Thompson sampling

Despite being one of the earliest strategies, the Bayesian based Thompson (1933) sampling algorithm did not get much attention until more contemporary work, such as that by Chapelle & Li (2011), showcased

its strong empirical performance and effectiveness at addressing the exploitation/exploration trade-off. Initially derived for a two-armed bandit setting for clinical trials, since its resurgence their have been use cases for the algorithm particularly in large tech companies, in areas ranging from website optimisation by Amazon (Hill et al. 2017) to Ad selection by LinkedIn (Tang et al. 2013). Being a Bayesian method, Thompson sampling requires a prior belief, a set past observations and then a likelihood function; with the posterior distribution subsequently recovered from Bayes rule.

Considering Thompson sampling in the context of the Bernoulli-bandit for K-arms, a sensible prior to select would be its conjugate prior, the Beta distribution. This subsequently models the mean reward for each arm with parameters $\alpha = (\alpha_1, ..., \alpha_K)$ and $\beta = (\beta_1, ..., \beta_k)$. The posterior distribution is then updated by observing a reward (r_t) when an action (x_t) is taken by the following rule,

$$(\alpha_k, \beta_k) \leftarrow \begin{cases} (\alpha_k, \beta_k) + (r_t, 1 - r_t) & \text{if} \quad x_t = k \\ (\alpha_k, \beta_k) & \text{otherwise} \end{cases}. \tag{5}$$

Algorithm 1 provides the procedure for implementing Thompson Sampling in the context for the Beta-Bernoulli bandit.

Algorithm 1: K-armed Bernoulli bandit Thompson Sampling

Input: K, α, β
for *t=1,...,T* **do**
  for *k=1,...,K* **do**
    Draw sample $\hat{\theta}_k \sim \text{Beta}(\alpha_k, \beta_k)$
  end
  $x_t \leftarrow \text{argmax}_k \hat{\theta}_k$ and observe reward r_t.
  $(\alpha_{x_t}, \beta_{x_t}) \leftarrow (\alpha_{x_t} + r_t, \beta_{x_t} + (1 - r_t))$
end

The algorithm can also be generalised to any Bandit algorithm with an unknown reward distribution, as long as it has support $[0, 1]$. As before, a reward (now denoted $\tilde{r}_t$) is observed, however a Bernoulli trial is then performed on this new observation with success probability $\tilde{r}_t$. The results of the Bernoulli trials can then be used to update the distributions. To see why this works, consider the probability of observing a success, with f denoting the pdf of the unknown reward distribution with support $[0, 1]$:

$$\mathbb{P}[r_t = 1] = \int_0^1 \tilde{r} f(\tilde{r}) d\tilde{r} = \mu. \tag{6}$$

Given Equation 6, it can be seen that the probability of observing a success is the same as the mean reward, a fundamental property of the recently derived Bernoulli bandit and thus success and failures are updated in the same way even when the reward distribution is not necessarily Bernoulli. Algorithm 2 provides this new approach, alongside a change in notation to denote success and failure by S and F respectively, as opposed to the parameters α and β to signify a generalisation of distribution choice.

Algorithm 2: K-armed stochastic bandit Thompson Sampling

Input: K, S, F
for *t=1,...,T* **do**
  for *k=1,...,K* **do**
    Draw sample $\hat{\theta}_k \sim \text{Beta}(S_k + 1, F_k + 1)$
  end
  $x_t \leftarrow \text{argmax}_k \hat{\theta}_k$ and observe reward $\tilde{r}_t$.
  $r_t \leftarrow 1$ **if** (Random$[0, 1] < \tilde{r}_t$) **else** 0 # Bernoulli Trial
  $(S_{x_t}, F_{x_t}) \leftarrow (S_{x_t} + r_t, R_{x_t} + (1 - r_t))$
end

Agrawal & Goyal (2012), which derived some key regret properties of Thompson Sampling, states and then utilises this generalised version as to be able to use Bernoulli bandits for analysis that can be applied to a more general class of stochastic bandits as long as it possesses same mean. General Thompson sampling from distributions with any support is also possible yet omitted for brevity and can be found in Section 4 of Russo et al. (2017).

3.2.1 Regret bounds

The bounds of expected regret for Thompson sampling have been studied extensively in the last decade. Upper bounds were found for the K-armed bandit by Agrawal & Goyal (2012), followed by Kaufmann, Korda & Munos (2012) providing an improved upper bound, in the specific case of the Bernoulli bandit with uniform priors. This regret bound was found to be the same as that of another class of algorithms known as *upper confidence bound* (UCB) policies. As opposed to Thompson, UCB is a frequentist approach to mutli-armed bandits, however it can be adapted to the Bayesian setting (see Section 3.3).

Lower bounds for any bandit algorithm had been established as far back as Lai & Robbins (1985), shown to be $\Omega(log(T))$. For the specific case of Thompson sampling, Liu & Li (2016) establishes lower and then matching upper bounds for both a good and poor prior choices, finding differences that emphasises how an informative prior can improve best and worse case algorithmic performance. While Russo & Roy (2016) noted that most of these theoretical bounds relied what they refer to as *hard* knowledge which pertains to the prior knowledge mapping an action to an outcome distribution. With little in the way of establishing bounds for *soft* knowledge, which tries to match these mappings to reality. This subsequently lead to a bound of order $\sqrt{\log(|\mathcal{A}|)dT}$, for an action set $\mathcal{A}$, in the context of online linear optimization problems under bandit feedback.

3.3 Other Algorithms

ϵ-greedy

A simple heuristic and a baseline for methods that balance exploration and exploitation, the ϵ-greedy algorithm will, at any time, choose to either explore with probability ϵ or exploit with probability $1 - \epsilon$. The choice of ϵ can be fixed from the start, or subject to annealing in which it is progressively lowered over time. In terms of regret bounds, for a constant ϵ, this method is linear (Kuleshov & Precup 2014). While for ϵ subjected to annealing, it has been found to possess $O(\log(T))$ expected regret, however there does not appear to be much practical use for this bound (Auer et al. 2002, Burtini et al. 2015).

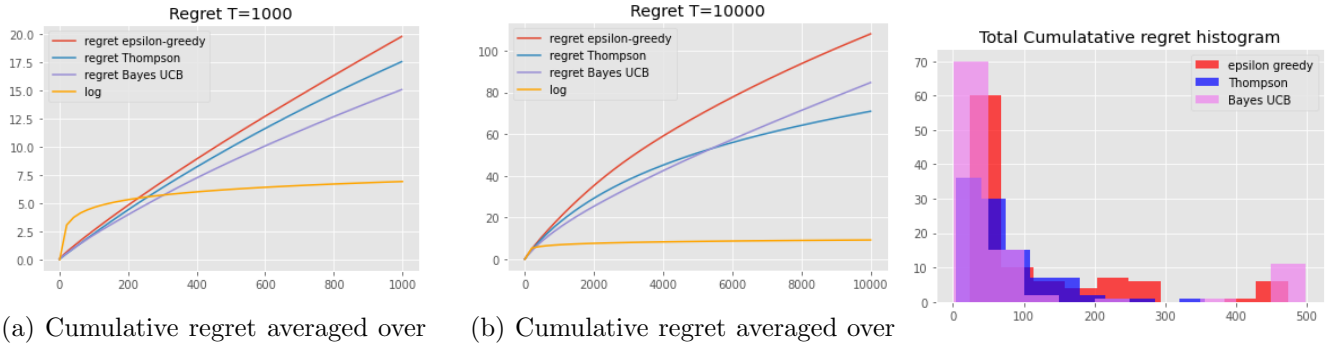
Bayes-UCB

UCB based policies provide an optimistic approach to mutli-armed bandit problems, aiming to reduce the chance of overlooking the best arm. Bayes-UCB (Kaufmann, Cappe & Garivier 2012) provides a similar approach to that of Thompson sampling, but maintaining the UCBs focus on not underestimating the possible best arm. It achieves this by maximising quantiles rather than the posterior mean. Bayes-UCB is considered an asymptotically efficient algorithm, outperforming its contemporary frequentist based counterpart KL-UCB (Burtini et al. 2015); a method derived from the Kullback Leibler divergence. In terms of regret bounds, Ghavamzadeh et al. (2016) states that the frequentist based upper bound match that of the lower bound of Lai & Robbins (1985) when considering a Bernoulli bandit, and possesses a general non-asymptotic finite-time regret of $O(T)$.

3.4 Simulation of Methods

Thompson, ϵ-greedy ($\epsilon = 0.1$) and Bayes-UCB are now implemented for the 2-armed Bernoulli bandit problem with a $Beta(1, 1)$ prior. For this simulation setup, each method was simulated 100 times, for

$T = 10,000$ rounds and arm probabilities for each arm being 0.5 and 0.55 respectively. The average cumulative regret over these simulations are plotted in Figures 3a and 3b, with the full number of rounds and a subset of $T \leq 1000$ on show. Alongside this is a log plot to visualise if the functions are providing the "log-like" regret behaviour the theoretical bounds of Bayes-UCB and Thompson claim. It can be seen that the results maintain what was expected, with ϵ-greedy algorithm suffering from worse than the other two methods, who themselves possess log behaviour in their cumulative regret when viewed over the full range of T rounds. In addition to this, Figure 3c highlights the distribution of the total cumulative regret for each of the simulations, Bayes-UCB appears more likely to produce low regret results, whereas Thompson shows consistency skewed towards the low regret end, with less chance of anomalously high cumulative regret that has been seen with the other two methods.



(a) Cumulative regret averaged over 1000 rounds

(b) Cumulative regret averaged over 1000 rounds

(c) Regret distribution.

Figure 3: Plots of cumulative regret and the distribution of regret over 100 simulations.

This study has obvious caveats with the most prominent being the lack of simulations due to limited computational power. It would also be worth in future studies to show the interval of trajectories each method had, alongside tests with different priors, arm lengths and other methodologies. Pilarski et al. (2021) and May et al. (2012) are some examples of more thorough computational studies of bandit algorithms in the literature.

3.5 Algorithm Limitations

A publication by Russo & Roy (2014) which ties itself to their paper on soft and hard knwledge regret bounds (Russo & Roy 2016), notes issues with Thompson, as well as UCB methods. These include... Therefore, the authors proposed a new algorithm, *information-directed sampling* (IDS). However, this also comes with some performance caveats and another potential improvement with the TS-UCB algorithm by Baek & Farias (2020), albeit currently in pre-print.

4 Conclusions and Further Work

Despite its dormancy for much of the 20th and early 21st century, Thompson Sampling provides a empirically strong and incredibly simple method to solving Bandit problems. However mulit-armed bandits extend greatly from this method with other popular Bayesian, and alternatively Frequentist, families of algorithms, each containing policies that give high performing and state of the art results that can be actively compared through a regret metric and readily applicable to a multitude of fields.

A natural extension to the work outlined would be implementation and comparison of further methodologies in the area, both Frequentist and Bayesian. Delving further into the theoretical work of establishing regret bounds for various algorithms would also be a worthwhile endeavour. As would an expanded explanation on why Thompson works and some more mathematical underpinnings that create the Algorithms

provided in Section 3.2. Further reading to expand on these topics are found in Russo et al. (2017), Ghavamzadeh et al. (2016), and Bubeck & Cesa-Bianchi (2012).

5 Data Availability

Code used in the Multi-armed bandit section is available at the following GitHub Link: `https://github.com/BenSLowery/MResCode/tree/main/608_report`.

References

Agrawal, S. & Goyal, N. (2012), Analysis of thompson sampling for the multi-armed bandit problem, *in* S. Mannor, N. Srebro & R. C. Williamson, eds, 'Proceedings of the 25th Annual Conference on Learning Theory', Vol. 23 of *Proceedings of Machine Learning Research*, PMLR, Edinburgh, Scotland, pp. 39.1–39.26.
URL: *https://proceedings.mlr.press/v23/agrawal12.html*

Auer, P., Cesa-Bianchi, N. & Fischer, P. (2002), *Machine Learning* **47**(2/3), 235–256.
URL: *https://doi.org/10.1023/a:1013689704352*

Baek, J. & Farias, V. F. (2020), 'Ts-ucb: Improving on thompson sampling with little to no additional computation', *arXiv preprint arXiv:2006.06372* .

Bernoulli, D. (1738), 'Exposition of a new theory on the measurement of risk', *Econometrica* **22**(1).

Binmore, K. (2008), *Rational Decisions*, Princeton University Press.

Bubeck, S. & Cesa-Bianchi, N. (2012), 'Regret analysis of stochastic and nonstochastic multi-armed bandit problems', *arXiv preprint arXiv:1204.5721* .

Burtini, G., Loeppky, J. L. & Lawrence, R. (2015), 'A survey of online experiment design with the stochastic multi-armed bandit', *ArXiv* **abs/1510.00757**.

Chapelle, O. & Li, L. (2011), An empirical evaluation of thompson sampling, *in* 'Proceedings of the 24th International Conference on Neural Information Processing Systems', NIPS'11, Curran Associates Inc., Red Hook, NY, USA, p. 2249–2257.

Friedman, M. & Savage, L. J. (1948), 'The utility analysis of choices involving risk', *Journal of Political Economy* **56**(4), 279–304.

Ghavamzadeh, M., Mannor, S., Pineau, J. & Tamar, A. (2016), 'Bayesian reinforcement learning: A survey', *arXiv preprint arXiv:1609.04436* .

Hansson, S. O. (2005), *Decision Theory: A Brief Introduction*.

Hill, D. N., Nassif, H., Liu, Y., Iyer, A. & Vishwanathan, S. (2017), An efficient bandit algorithm for realtime multivariate optimization, *in* 'Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', KDD '17, Association for Computing Machinery, New York, NY, USA, p. 1813–1821.
URL: *https://doi.org/10.1145/3097983.3098184*

Kahneman, D. & Tversky, A. (1979), 'Prospect theory: An analysis of decision under risk', *Econometrica* **47**(2), 263–291.

Kaufmann, E., Cappe, O. & Garivier, A. (2012), On bayesian upper confidence bounds for bandit problems, *in* N. D. Lawrence & M. Girolami, eds, 'Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics', Vol. 22 of *Proceedings of Machine Learning Research*, PMLR, La Palma, Canary Islands, pp. 592–600.
URL: *https://proceedings.mlr.press/v22/kaufmann12.html*

Kaufmann, E., Korda, N. & Munos, R. (2012), Thompson sampling: An asymptotically optimal finite-time analysis, *in* N. H. Bshouty, G. Stoltz, N. Vayatis & T. Zeugmann, eds, 'Algorithmic Learning Theory', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 199–213.

Kuleshov, V. & Precup, D. (2014), 'Algorithms for multi-armed bandit problems', *Journal of Machine Learning Research* **1**.

Lai, T. & Robbins, H. (1985), 'Asymptotically efficient adaptive allocation rules', *Advances in Applied Mathematics* **6**(1), 4–22.
URL: *https://www.sciencedirect.com/science/article/pii/0196885885900028*

Liu, C.-Y. & Li, L. (2016), On the prior sensitivity of thompson sampling, *in* 'International Conference on Algorithmic Learning Theory', Springer, pp. 321–336.

May, B. C., Korda, N., Lee, A. & Leslie, D. S. (2012), 'Optimistic bayesian sampling in contextual-bandit problems', *Journal of Machine Learning Research* **13**(67), 2069–2106.
URL: *http://jmlr.org/papers/v13/may12a.html*

Peters, O. (2011), 'The time resolution of the st petersburg paradox', *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **369**(1956), 4913–4931.

Peterson, M. (2020), The St. Petersburg Paradox, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', Fall 2020 edn, Metaphysics Research Lab, Stanford University.

Pilarski, S., Pilarski, S. & Varró, D. (2021), 'Optimal policy for bernoulli bandits: Computation and algorithm gauge', *IEEE Transactions on Artificial Intelligence* **2**(1), 2–17.

Russo, D. & Roy, B. V. (2014), 'Learning to optimize via information directed sampling', *CoRR* **abs/1403.5556**.
URL: *http://arxiv.org/abs/1403.5556*

Russo, D. & Roy, B. V. (2016), 'An information-theoretic analysis of thompson sampling', *Journal of Machine Learning Research* **17**(68), 1–30.
URL: *http://jmlr.org/papers/v17/14-087.html*

Russo, D., Van Roy, B., Kazerouni, A., Osband, I. & Wen, Z. (2017), 'A tutorial on thompson sampling', *arXiv preprint arXiv:1707.02038* .

Savage, L. J. (1954), *The foundations of statistics*, The foundations of statistics., John Wiley & Sons, Oxford, England.

Slivkins, A. (2019), 'Introduction to multi-armed bandits', *arXiv preprint arXiv:1904.07272* .

Tang, L., Rosales, R., Singh, A. & Agarwal, D. (2013), Automatic ad format selection via contextual bandits, CIKM '13, Association for Computing Machinery, New York, NY, USA, p. 1587–1594.
URL: *https://doi.org/10.1145/2505515.2514700*

Thompson, W. R. (1933), 'On the likelihood that one unknown probability exceeds another in view of the evidence of two samples', *Biometrika* **25**(3/4), 285–294.
URL: *http://www.jstor.org/stable/2332286*

Von Neumann, J. & Morgenstern, O. (1953), *Theory of games and economic behavior*.