CS 221 Project Progress Report: Exploring Bandit Algorithms

David Chen
Stanford University
dchen11@stanford.edu

***Abstract*—This progress report outlines an ongoing project on exploring bandit algorithms within the classic stochastic multi-armed bandit framework. The project's goal is to provide a comprehensive comparison of algorithms through reproduction of existing empirical analysis of regret and runtime, qualitative observations of behavior, and theoretical comparison of proven regret bounds. To date, a simulation framework has been developed and populated with algorithms such as random, ε-greedy, explore-then-commit, Bayes-UCB, and Thompson sampling, with preliminary evaluations conducted in the beta-Bernoulli bandit setting. Future work will expand the analysis to include additional settings with more complex information structure, alongside the implementation and evaluation of more advanced algorithms such as information-directed sampling (IDS).**

I. Introduction

In the broad context of online decision-making under uncertainty, the class of problems known as multi-armed bandits (MABs) has provided a rich environment for insightful theoretical analysis and applicable algorithms. Multi-armed bandit problems have been studied in a wide variety of fields, ranging from computer science and statistics, to operations research and economics. Although closely related to the more general setting of reinforcement learning, the study of MABs typically has a strong focus on the classic dilemma of exploration-exploitation. Still, fundamental insights from theoretical developments have been further developed for more complex reinforcement learning settings, and simple yet effective bandit algorithms have been deployed for many practical use-cases in the industry with great success.

A. Project goals and current progress.

The primary focus of this course project is to provide a comprehensive comparison of a collection of algorithms for the classic stochastic multi-armed bandit setting. Namely, I do not focus on settings such as those of contextual or adversarial bandits. I aim to accomplish this task by reproducing simulation results such as those presented in Russo and Van Roy [1]. I also intend to include a selection of derivations on important theoretical results. In summary, comparisons will take the form of empirical analysis of regret, runtime, and qualitative observations of algorithm behavior, as well as theoretical comparison of proven regret bounds.

At the time of writing, progress includes a functioning framework for simulation of various bandit algorithms. Implemented algorithms include: random, ε-greedy, explore-then-commit (ETC), Bayes-UCB, and Thompson sampling. Preliminary results entail simulation on the independent beta-Bernoulli bandit setting.

B. Brief overview of the multi-armed bandit problem.

I examine the stochastic multi-armed bandit problem. At each time period, an agent is allowed to choose an action (an arm) to execute, and subsequently observes a random outcome, often in the form of a scalar reward. Outcomes are associated with the specific arm, and can either be *independent or dependent* with respect to the other arms. The true distribution of the outcomes is unknown, and thus exploration of arms is necessary in order to gather knowledge about rewards. In this project, I restrict analysis to settings with stationary outcome distributions over time, as well as restricting the class of eligible actions to be fixed finite sets.

The objective of MAB problems is to maximize the average cumulative reward over time. Thus, a central issue that arises is that of *exploration-exploitation*, where a tradeoff is necessary in order to discover actions associated with higher rewards, while still leveraging high reward actions over the time horizon. In general, the time horizon can be infinite, but I only analyze and implement problems in a finite-time setting for this project.

A key difference between MABs and the more general reinforcement learning framework is the lack of "state". In the setting of Markov decision processes, outcomes are associated with a changing state as well as the selected action, whereas in the restricted bandit setting, any given action is assumed to produce i.i.d. outcomes when chosen in different time periods.

Typical theoretical analysis of MABs often involves the notion of *regret*, which intuitively is the expected difference in the sum of rewards between a strategy that chooses the optimal action at every round, and the actual strategy. There is also the notion of *per-period regret*, which is specific to a single round. Upper and lower bounds on regret are usually of interest for various algorithms, and much of the literature is dedicated to deriving and improving these bounds.

II. Related Works

The first formulation of the multi-armed bandit problem is most commonly attributed to a paper from Robbins in 1952 [2]. Since then, numerous techniques and settings have appeared in the literature. The introduction of "upper confidence bound" strategies as an approach to more efficient exploration appeared in Lai and Robbins [3].

Many early approaches were more aligned with the frequentist perspective, and extensions of the idea of upper confidence bounds resulted in algorithms such as UCB1 [4], which also proved upper bounds on the cumulative regret that scaled logarithmically with time. Over time, analysis for the Bayesian approach also gained popularity, such as the Bayes-UCB approach introduced by Kaufmann et al. [5].

Around the same time, an approach known as Thompson sampling started gaining recognition in the context of MABs. Thompson sampling itself pre-dated bandits, first introduced by Thompson in 1933 [6]. In the last couple of decades, theoretical and experimental analysis demonstrated impressive performance in the context of bandits [7].

Eventually, this culminated in an elegant approach to deriving upper bounds on regret for Thompson sampling using concepts from information theory. Russo and Van Roy introduced the concept of the *information ratio*, which was used to prove general bounds that depended on the entropy of the prior distribution of the optimal action [8]. The information ratio turned out to be quite useful beyond a one-time analysis of Thompson sampling; Russo and Van Roy eventually developed a novel algorithm that explicitly minimized the information ratio, and provided theoretical and experimental results that demonstrated its superiority over Thompson sampling in various settings [1]. It is this paper that I take inspiration in terms of reproduction of simulation results.

Finally, there are multiple other resources that have gathered results and techniques across the field of bandits as a whole, including extensions such as contextual, adversarial, and many other related settings [9], [10]. I aim to provide worked proofs of some results presented in these comprehensive texts from Slivkins [10], and Lattimore and Szepesvári [9].

III. Methodology

A. Problem formulation

I leave a more detailed formulation for the final paper. However, I'll briefly summarize the main terminology and setting I have been working with.

We work with a probability space $(\Omega, \mathbb{F}, \mathbb{P})$, and all random variables are defined with respect to this space, including the random variables that model prior uncertainty as described commonly in the Bayesian formulation.

The agent chooses actions $(A_t)_{t\in\mathbb{N}}$ from a finite set $\mathcal{A}$, and subsequently observes the outcomes $\left(Y_{t,A_t}\right)_{t\in\mathbb{N}}$, where each $Y_{t,a} \in \mathcal{Y}$. We assume, according to the Bayesian perspective, that there is a random element θ that describes the true distribution of outcomes, such that conditioned on θ, the sequence $(Y_t)_{t\in\mathbb{N}} = \left((Y_t, a)_{a\in\mathcal{A}}\right)_{t\in\mathbb{N}}$ is independent and identically distributed.

Furthermore, the agent observes a reward associated with the outcome. In many cases, the reward and outcome are equivalent, but generally, reward can be a known function $R : \mathcal{Y} \to \mathbb{R}$. For convenience, we can denote $R_{t,a} := R\big(Y_{t,a}\big)$.

Once we have the notion of reward, we can consider maximization of that reward, and we can define the optimal action(s) to be A^* such that $A^* \in \arg\max_{a\in\mathcal{A}} \mathbb{E}\big[R_{t,a} \,\big|\, \theta\big]$. Building on top of this, we can finally define the T-period *regret* of a strategy of choosing actions π to be:

$$\operatorname{Regret}(T, \pi) = \sum_{t=1}^{T} \Big(R_{t,A^*} - R_{t,A_t}\Big), \tag{1}$$

where the sequence of actions is understood to be chosen by π. We can take an expectation on both sides, with respect to randomness in the choice of actions, outcomes, and over the prior distribution of θ, which leaves us with the *expected regret*.

Further concepts which are useful but I do not define in detail in the progress report are fundamental concepts in information theory, such as the *Shannon entropy* $H(P)$ of a probability distribution P, the *Kullback-Leibler divergence* $D_{\mathrm{KL}}(P \,\|\, Q)$, the *mutual information* $I(X\,;Y)$ with respect to random variables X and Y, and finally, the *information ratio* as first described in [8] and utilized in IDS [1].

B. Dataset (bandit simulation)

All relevant "data" is generated online (ad-hoc) during simulations. Outcomes associated with specific distributions are generated randomly using existing libraries, such as numpy.random and scipy.stats.

For example, in a Bernoulli bandit instance, the initial parameters θ_k across K arms are generated independently from a continuous uniform distribution on $[0, 1]$ using np.random.uniform. The subsequent rewards are then generated using an indicator function implemented using np.random.rand. These are generated as needed during simulation.

C. Baseline (simple algorithms)

The baseline involves implementation of simple non-adaptive exploration algorithms for the beta-Bernoulli bandit setting. The "non-adaptive exploration" terminology is borrowed from Slivkins [10]. These include a random strategy, various ε-greedy strategies, and explore-then-commit.

The random strategy simply chooses an action uniformly at random from the set of available actions, at each time-step. This is mainly chosen to demonstrate a worst-case upper bound on regret for all subsequent algorithms.

The ε-greedy algorithms involve choosing a uniformly random action at each time period with probability ε_t, and otherwise choosing the action with the maximum point-estimate of the mean reward. Notably ε_t can vary over time, but overall this class of algorithms is still classified into the non-adaptive exploration category, given it does not change its exploration strategy based on the realized history.

Some examples of valid choices of ε_t are:

- Constant (ex. $\varepsilon_t = \varepsilon \in [0, 1)$),
- Decaying (ex. $\varepsilon_t = \varepsilon(t) = t^{-1/3}$),
- Explore-then-commit (ex. $\varepsilon_t = \varepsilon(t) = \mathbb{1}_{t<200}$).

These approaches are chosen as the baseline of algorithms that do not incorporate any additional notion of uncertainty into the exploration strategy, which leads to provably worse

regret-bounds and demonstrably worse realized regret in simulation.

D. Main approach (advanced algorithms)

More interesting algorithms arise when we attempt to balance exploration-exploitation through use of confidence intervals, probability matching, or explicitly minimizing the information ratio.

A major family of algorithms in this area are the UCB algorithms, which range from frequentist algorithms such as UCB1 to the Bayesian Bayes-UCB.

On the other hand, Thompson sampling selects an action based off of the statistical possibility that it is optimal under the posterior distributions.

Finally, I examine information-directed sampling, where the action is chosen to minimize an information ratio based on the expected regret and mutual information.

These algorithms have all been shown to have improved upper-bounds for regret compared to the baseline algorithms, and have also demonstrated better performance in simulation.

E. Additional settings

While the beta-Bernoulli bandit setting provides useful insights by itself, extension of analysis to a wider variety of settings may provide a more complete picture of the capabilities of all the algorithms, as well as distinguish algorithms that are capable of taking advantage of settings where there exists a richer information structure.

To that end, I believe it will be worthwhile to implement the following additional settings:

- *Independent Gaussian*, where the reward for each arm is sampled from a Gaussian distribution with a fixed known variance, and the mean parameters are assumed to be independent samples from a Gaussian prior.
- *Independent Poisson*, where the reward for each arm is sampled from a Poisson distribution, and the rate parameters are assumed to be independent samples from a Gamma prior.
- *Linear Gaussian*, where actions $a \in \mathbb{R}^d$ are known d-dimensional vectors. The rewards correspond to $a^\top \theta + \epsilon_t$, where θ is drawn from a multivariate Gaussian prior, and ϵ_t is independent Gaussian noise.

F. Evaluation

I evaluate all algorithms over 2000 simulations (trials), each running for $T = 2000$. For each trial, we calculate the cumulutative sum over time, and that sequence is then averaged over all trials.

At the time of writing, this is the extent of the quantitative evaluation. A more qualitative comparison of the regret between each algorithm is discussed.

In addition, at the time of writing, only the beta-Bernoulli bandit setting is evaluated. The other settings mentioned previously have yet to be implemented.

For the final discussion, I believe it would also be worthwhile to compare the runtimes of the different algorithms. Some algorithms, such as ε-greedy, only rely on a few elementary operations each iteration, while some, like IDS, involve more intensive numerical methods to approximate integrals.

A final point of comparison can be done theoretically through best known regret bounds in similar settings. Derivations will be provided for selected results, and comparison between algorithms as well as their empirical results will be shown.

IV. Results and Discussion

Through a round of preliminary simulations, I am able to successfully reproduce results similar to existing experimental results from Russo and Van Roy [1]. For the algorithms not included in prior simulations, namely the non-adaptive exploration algorithms, I observe noticably higher regret for this horizon. Interestingly, within this interval, ε-greedy with exponential decay seems to perform worse than constant ε-greedy. Indeed, checking the ε_t factor, I find that ε_t decays to 0.1 only after the 1000th time-step, at which point the slope of the regret curve seems to become more shallow than constant ε-greedy with $\varepsilon = 0.1$. Past $T = 2000$, I believe that ε-greedy with decay should have lower cumulative regret, and further analysis to characterize this trend may be interesting.

Explore-then-commit exhibits a very different regret curve, as it is essentially fully random for 200 time-steps before switching to fully greedy. The slope of the regret curve is initially much shallower than the other algorithms at $t = 200$, and further analysis of the exact behavior and slope of the regret curve would also be interesting to investigate. For any instance of the problem, the cumulative regret is equivalent to the probability that the optimal action is correctly identified after the exploration period, and if not, the average difference in reward between the actual chosen action and the optimal one.

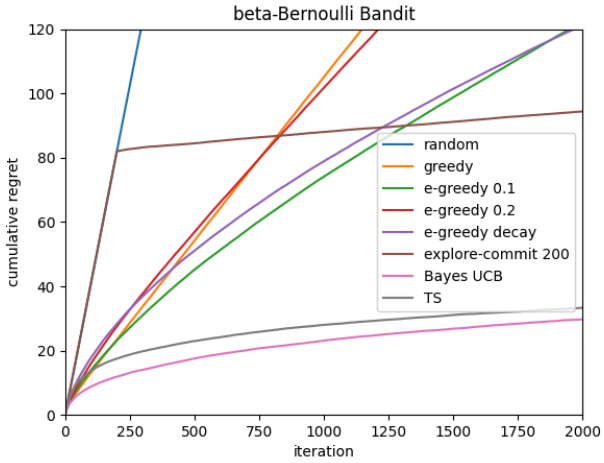

Fig. 1: The average cumulative regret over $T = 2000$ for beta-Bernoulli problems with 10 arms. # trials $= N = 2000$.

V. Future Work

To summarize concisely the planned work to be done after the submission of this progress report:

- Additional settings
  - Independent Gaussian
  - Independent Poisson
  - Linear Gaussian
- Additional algorithms
  - IDS
  - Variance-based IDS
  - Other UCB algorithms
- Additional analysis
  - Selection and derivation of key regret results
  - Comparison with simulation
  - Runtime analysis

References

[1] D. Russo and B. Van Roy, "Learning to Optimize via Information-Directed Sampling," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., Curran Associates, Inc., 2014, p. . [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2014/file/90720a2fcc41f9332e6a1558da327089-Paper.pdf

[2] H. Robbins, "Some aspects of the sequential design of experiments," 1952.

[3] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4–22, 1985, doi: https://doi.org/10.1016/0196-8858(85)90002-8.

[4] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time Analysis of the Multiarmed Bandit Problem," *Mach. Learn.*, vol. 47, no. 2–3, pp. 235–256, May 2002, doi: 10.1023/A:1013689704352.

[5] E. Kaufmann, O. Cappe, and A. Garivier, "On Bayesian Upper Confidence Bounds for Bandit Problems," in *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, N. D. Lawrence and M. Girolami, Eds., in Proceedings of Machine Learning Research, vol. 22. La Palma, Canary Islands: PMLR, 2012, pp. 592–600. [Online]. Available: https://proceedings.mlr.press/v22/kaufmann12.html

[6] W. R. Thompson, "On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples," *Biometrika*, vol. 25, no. 3/4, pp. 285–294, 1933, Accessed: May 22, 2025. [Online]. Available: http://www.jstor.org/stable/2332286

[7] O. Chapelle and L. Li, "An Empirical Evaluation of Thompson Sampling," in *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds., Curran Associates, Inc., 2011, p. . [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2011/file/e53a0a2978c28872a4505bdb51db06dc-Paper.pdf

[8] D. Russo and B. V. Roy, "An Information-Theoretic Analysis of Thompson Sampling," *Journal of Machine Learning Research*, vol. 17, no. 68, pp. 1–30, 2016, [Online]. Available: http://jmlr.org/papers/v17/14-087.html

[9] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge University Press, 2020. [Online]. Available: https://books.google.com/books?id=bydXzAEACAAJ

[10] A. Slivkins, "Introduction to Multi-Armed Bandits." [Online]. Available: https://arxiv.org/abs/1904.07272