New Paradigms for Adaptive Decision Making under Bandit Feedback

Kirthevasan Kandasamy
University of California, Berkeley

Machine learning has made great strides in the recent past, achieving human or super-human level performance in tasks requiring learning from data. However, intelligent systems should go beyond simply learning; they should be able to make decisions in unknown environments so as to achieve a certain goal. Many such problems fall within the framework of adaptive decision-making under *bandit feedback*. In typical use cases, a decision-maker sequentially chooses an action(s) and observes feedback. She then uses the information collected to make inferences about the environment and plan future actions to fulfil a given goal. Below are a few of the many applications for the bandit framework which have inspired my work.

1. *Fair resource allocation in multi-tenant clusters:* In shared data-centres, we wish to allocate computational resources to different users fairly. Typically, users have difficulty estimating their resource requirements if their jobs are complex or changing, especially in the presence of heterogeneous resources. Bandit methods can be used to estimate the resource requirements of users' jobs online, while simultaneously allocating the resources in a manner that is efficient and fair for all users [BKB+21][RFS+20].

2. *Black-box optimisation & configuration tuning:* Many scientific and engineering problems can be framed as configuration tuning problems, where we need to find the optimum configuration of an unknown system via repeated experimentation. Examples include hyperparameter tuning of statistical models [DKM+21, KVN+20][SLA12, LJDT17], materials discovery [DMK+20, KXK+20][BT96], and optimising simulation-based astrophysical experiments [KSP15b][Dav07, XEN+19].

Main themes: The majority of my past work has been in bandit settings. My research questions have usually arisen via my collaborations with practitioners in different fields, primarily in systems, and additionally in materials science and astrophysics. Not only has this led to practically grounded problem settings, but it has also resulted in novel methods and theoretical insights. I have learned that while the bandit framework has great potential for scientific, social, and economic impact, we also need new problem formulations to tackle modern challenges and opportunities. Below are some of the central themes in my past work.

1. **Multiple rational and strategic stakeholders:** Many real world systems are economic and multi-agent in nature, and decisions taken by or for one agent should be weighed against the considerations of others, especially when they have competing goals and/or when there is scarcity. Additionally, individual agents may be strategic and try to manipulate the system to obtain outcomes that are favourable to them, but might not be socially desirable. For instance, in the above multi-tenant resource allocation example, we may wish to learn user (agent) resource requirements via feedback so that we can allocate resources efficiently. At the same time, we also wish to treat users fairly for their long term happiness, and account for strategic users who may misreport their feedback with the intention of obtaining a large number of resources. Therefore, in addition to efficiency, bandit methods should also satisfy fairness and strategy-proofness (incentivise good behaviour from users).

2. **Real world time and resource constraints:** Most theoretical formulations of bandit problems study the sample complexity (number of observations collected from the environment) to fulfil a given goal. Historically, this has been because the number of samples usually translates to the amount of time taken and/or the amount of resources consumed. However, formulating problems directly in terms such of time and resource constraints give rise to new algorithms and theoretical results. For example, in some use cases, we may not be constrained to a finite set of resources to execute our actions; instead, we may have flexibility to scale up or down the amount of resources we use, such as when performing computer-

based experiments on the cloud, or using high throughput platforms for drug and materials discovery. In a different example, when experiments are conducted on distributed computing platforms, we may wish to account for the sub-linear scaling characteristics when planning actions under time constraints.

3. **Multi-fidelity decision-making:** Conventional wisdom in the bandit literature suggests that we need to design better algorithms in order to improve the theoretical and empirical performance of algorithms. However, in many applications, cheap approximations to an expensive experiment may be available. By leveraging these approximations, we may be able to fulfil our goal at lower overall cost. For example, in simulation-based scientific studies, the simulations can be carried out at varying levels of granularity to obtain cheaper simulations. Similarly, when tuning hyperparameters of machine learning models, we may train using less data or for fewer iterations to obtain cheap approximations of its final performance.

Past work

1. Mechanism design with bandit feedback. Mechanism design, a core area of research in the economics and game theory literature, deals with settings where there are multiple rational and strategic stakeholders. It finds applications in fair division (page 1), auction design, kidney exchange [RSÜ04], matching markets [Rot86], and several more. The goal of a *mechanism* is to truthfully elicit the preferences of each agent and obtain a socially desirable outcome. However, the majority of the literature in mechanism design assume that agents know their preferences. In the following work, we study repeated mechanisms in a variety of settings where agents do not know their preferences; on each round, a mechanism first chooses an outcome; at the end of the round, the agents provide bandit feedback based on their experience of this outcome. The goal of the mechanism is to learn agent preferences and while finding a socially desirable outcome. Simultaneously, it needs to account for the rationality and strategic considerations of the agents.

1. *In fair division [KSG+20, GKG+21].* In fair division, multiple agents share a common resource. Each agent prefers to have more resources for herself; i.e. her utility increases with the amount of resources she receives. While fair division mechanisms typically assume that users know their utilities, this may not be true in many applications of interest, such as when allocating resources in multi-tenant clusters (page 1). In [KSG+20], we studied learning the utilities under bandit feedback when there is a single resource type. We formalise this task via asymptotic notions of Pareto-efficiency (PE), fairness, and strategy-proofness. Under nonparametric assumptions on the utilities, we describe an algorithm with $\mathcal{O}(T^{2/3})$ rates for all three desiderata, and a second algorithm with $\mathcal{O}(T^{1/2})$ rates for PE and fairness but with no strategy-proofness guarantees. In [GKG+21], we studied fair division with multiple resource types. Unlike in the single resource case, here, agents might find it mutually beneficial to exchange resources. We propose a randomised algorithm which achieves $\mathcal{O}(T^{1/2})$ rates for PE and fairness.

2. *In auction-like settings [KGJS20].* We studied a multi-round auction design setting when buyers do not know their values for the different outcomes that may be chosen by a mechanism. We defined three notions of regret for the welfare, the individual utilities of each buyer and that of the seller. We showed that these three terms were interdependent via an $\Omega(T^{2/3})$ lower bound for the maximum of these three regret terms, and described an algorithm which achieves this rate. Additionally, we provided asymptotic rates which bound the violations of strategy-proofness and individual rationality.

3. *Cilantro: SLO-based resource allocation in multi-tenant clusters [BKB+21].* Micro-services deployed in real-time systems are evaluated on how well they perform relative to the given service level objective (SLO). While applications track these performance metrics, prior fair allocation mechanisms are more or less oblivious to SLOs [DKS89, GZH+11]. We designed Cilantro, a Kubernetes-based system for resource allocation which learns the resource requirements of jobs via feedback on its performance. We

employ several bandit methods, including those designed above for fair division [KSG+20, GKG+21]. Across a wide variety of workloads, Cilantro improves the utility for more than half of the users by $1.5 - 2.5\times$,while satisfying fairness and strategy-proofness empirically.

2. Best arm identification (BAI) under real world time and resource constraints. BAI is one of the most common abstractions for configuration tuning in the bandit literature. Here, we repeatedly and adaptively pull arms (draw samples/rewards) from one of several arms, with the aim of finding the "best arm", i.e. the arm with the highest expected reward. With a few exceptions, the literature on this topic is restricted to settings where we draw one sample at a time, or a fixed number of multiple samples at a time. However, in the real world, we need to perform BAI under a variety of time and resource constraints. Formulating BAI directly in terms such constraints can give rise to new algorithms and theoretical results.

1. *Overcoming sub-linear scaling in parallel BAI [TKS+21a].* We studied BAI when distributed computing resources can be allocated in varying amounts to execute arm pulls. By allocating more resources per pull, we obtain results faster and can make more informed decisions subsequently, but might have reduced throughput due to sub-linear scaling. For example, an astrophysical simulation can be sped up by running it on multiple cores; but this speed-up is partly offset due to communication and synchronisation costs. We proposed an algorithm which optimally manages this trade-off and show that the time taken can be upper bounded by the solution of a dynamic program whose inputs are the gaps between the optimal and sub-optimal arms. We complement this with a matching lower bound.
2. *BAI with elastic resources [TKS+21b].* We studied BAI under a time deadline of T rounds. In typical use cases, T is small, i.e. we have limited adaptivity, but can execute multiple arm pulls per round. For instance, when executing computer-based experiments on the public cloud, such as in hyperparameter tuning and scientific simulations, we can scale up or down the resources we use as we wish. However, we need to pay for the total resource-time used, which corresponds to the number of arm pulls. Similarly, in materials discovery, high-throughput experimental platforms can be used to simultaneously conduct as many experiments as we wish in parallel, but the number of experiments (samples) needs to be minimised to reduce costs. We proved two hardness results on the sample complexity for this problem, and designed an algorithm which was optimal with respect to both results.
3. *Hyperparameter tuning in the cloud [MLD+21, DKM+21].* We designed Rubberband, a system for hyperparameter tuning using elastic resources. Our methods, which were designed using the intuitions developed in the above work, outperform many baselines which use a fixed amount of resources.

3. Multi-fidelity optimisation. In [KDSP16, KDSP17, KDO+19], we studied methods for optimising an expensive function, when we have access to cheaper approximations. Both theoretically and empirically, and under a variety of assumptions, our methods, which use upper confidence bound techniques, are more cost-efficient than naive strategies which only query the most expensive function.

4. Other work. In addition to the above, my other work in this space include design of experiments under application-specific sub-modular objectives [KNZ+19], executing multiple actions in parallel [KKSP18], and optimising in graph-structured [KXK+20, KNS+18] and high-dimensional [KSP15a] domains.

Future vision

There are two research agendas that I am excited to pursue in the near future. First, in many practical use cases, we need to rely on data/feedback from multiple agents in order to be able to learn, and I wish to study how the strategic considerations of such agents can affect learning. As one example, we might wish to develop a system that mutually benefits multiple agents, but if the agents are competitive, they might be

reluctant to share their data; e.g. banks, who are otherwise competitive, might wish to collaborate on fraud detection. The theory of public goods [BBV86] can help us deal with such scenarios, but requires new ideas to quantify the utility each agent gains from shared data, dealing with nefarious agents who might poison their data, and finding the Nash equilibrium of this game. In a different example, myopic buyers are usually tempted to purchase products that they are familiar with or those with the most positive reviews. This makes it difficult to learn buyer preferences for new products. In such cases, either the sellers or an e-commerce platform should decide how to best incentivise such buyers to try new products.

Second, I wish to study theoretical formulations for adaptive decision-making that capture the constraints and opportunities that arise in practice. For instance, the systems literature is ubiquitous with problems where we wish to optimise for different performance characteristics, such as run time, energy usage, SLO-violation, fairness, and other application-specific criteria [RCMC20, VAPGZ17, RFS+20]. Moreover, since these systems cannot be modelled analytically, this needs to be done via repeated experimentation, making it suitable for bandit methods. However, practitioners often simply treat this as black-box optimisation problems by designing heuristic rewards based on these criteria. As we demonstrated in [KSG+20, GKG+21, BKB+21], accounting for various desiderata in a principled manner often leads to better empirical and theoretical properties. Going forward, I wish to pursue such problems in systems and other application domains.

References

[BBV86] Theodore Bergstrom, Lawrence Blume, and Hal Varian. On the Private Provision of Public Goods. *Journal of public economics*, 1986.

[BKB+21] Romil Bhardwaj, Kirthevasan Kandasamy, Asim Biswal, Wenshuo Guo, Ben Hindman, Joseph E Gonzalez, Michael I Jordan, and Ion Stoica. Cilantro: SLO-based Resource Allocation in Multi-tenant Clusters. unpublished, 2021.

[BT96] James R Broach and Jeremy Thorner. High-throughput Screening for Drug Discovery. *Nature*, 1996.

[Dav07] T. M. Davis et al. Scrutinizing Exotic Cosmological Models Using ESSENCE Supernova Data Combined with Other Cosmological Probes. *Astrophysical Journal*, 2007.

[DKM+21] Lisa Dunlap, Kirthevasan Kandasamy, Ujval Misra, Richard Liaw, Joseph E Gonzalez, Ion Stoica, and Michael I Jordan. SEER: Hyperparameter Tuning on the Cloud. In *Proceedings of the Symposium on Cloud Computing (SoCC)*, 2021.

[DKS89] Alan Demers, Srinivasan Keshav, and Scott Shenker. Analysis and Simulation of a Fair Queueing Algorithm. *ACM SIGCOMM Computer Communication Review*, 1989.

[DMK+20] Adarsh Dave, Jared Mitchell, Kirthevasan Kandasamy, Han Wang, Sven Burke, Biswajit Paria, Barnabás Póczos, Jay Whitacre, and Venkatasubramanian Viswanathan. Autonomous Discovery of Battery Electrolytes with Robotic Experimentation and Machine Learning. *Cell Reports Physical Science*, 2020.

[GKG+21] Wenshuo Guo, Kirthevasan Kandasamy, Joseph E Gonzalez, Michael I Jordan, and Ion Stoica. Online Learning of Competitive Equilibria in Exchange Economies. *arXiv preprint arXiv:2106.06616*, 2021.

[GZH+11] Ali Ghodsi, Matei Zaharia, Benjamin Hindman, Andy Konwinski, Scott Shenker, and Ion Stoica. Dominant Resource Fairness: Fair Allocation of Multiple Resource Types. In *NSDI*, 2011.

[KDO+19] Kirthevasan Kandasamy, Gautam Dasarathy, Junier Oliva, Jeff Schneider, and Barnabas Poczos. Multi-fidelity Gaussian Process Bandit Optimisation. *Journal of Artificial Intelligence Research*, 2019.

[KDSP16] Kirthevasan Kandasamy, Gautam Dasarathy, Jeff Schneider, and Barnabas Poczos. The Multi-fidelity Multi-armed Bandit. In *Advances in Neural Information Processing Systems*, 2016.

[KDSP17] Kirthevasan Kandasamy, Gautam Dasarathy, Jeff Schneider, and Barnabás Póczos. Multi-fidelity Bayesian Optimisation with Continuous Approximations. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

[KGJS20] Kirthevasan Kandasamy, Joseph E Gonzalez, Michael I Jordan, and Ion Stoica. Mechanism Design with Bandit Feedback. *arXiv preprint arXiv:2004.08924*, 2020.

[KKSP18] Kirthevasan Kandasamy, Akshay Krishnamurthy, Jeff Schneider, and Barnabás Póczos. Parallelised

Bayesian Optimisation via Thompson Sampling. In *International Conference on Artificial Intelligence and Statistics*, 2018.

[KNS+18] Kirthevasan Kandasamy, Willie Neiswanger, Jeff Schneider, Barnabas Poczos, and Eric Xing. Neural Architecture Search with Bayesian Optimisation and Optimal Transport. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

[KNZ+19] Kirthevasan Kandasamy, Willie Neiswanger, Reed Zhang, Akshay Krishnamurthy, Jeff Schneider, and Barnabas Poczos. Myopic Posterior Sampling for Adaptive Goal Oriented Design of Experiments. In *International Conference on Machine Learning*, 2019.

[KSG+20] Kirthevasan Kandasamy, Gur-Eyal Sela, Joseph E Gonzalez, Michael I Jordan, and Ion Stoica. Online Learning Demands in Max-min Fairness. *arXiv preprint arXiv:2012.08648*, 2020.

[KSP15a] Kirthevasan Kandasamy, Jeff Schenider, and Barnabás Póczos. High Dimensional Bayesian Optimisation and Bandits via Additive Models. In *International Conference on Machine Learning*, 2015.

[KSP15b] Kirthevasan Kandasamy, Jeff Schneider, and Barnabás Póczos. Bayesian Active Learning for Posterior Estimation. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[KVN+20] Kirthevasan Kandasamy, Karun Raju Vysyaraju, Willie Neiswanger, Biswajit Paria, Christopher R Collins, Jeff Schneider, Barnabas Poczos, and Eric P Xing. Tuning Hyperparameters without Grad Students: Scalable and Robust Bayesian Optimisation with Dragonfly. *Journal of Machine Learning Research*, 2020.

[KXK+20] Ksenia Korovina, Sailun Xu, Kirthevasan Kandasamy, Willie Neiswanger, Barnabas Poczos, Jeff Schneider, and Eric Xing. ChemBO: Bayesian Optimization of Small Organic Molecules with Synthesizable Recommendations. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020.

[LJDT17] Lisha Li, Kevin Jamieson, Giulia DeSalvo, and Ameet Talwalkar. Hyperband: A Novel Bandit-based Approach to Hyperparameter Optimization. *Journal of Machine Learning Research*, 2017.

[MLD+21] Ujval Misra, Richard Liaw, Lisa Dunlap, Romil Bhardwaj, Kirthevasan Kandasamy, Joseph E Gonzalez, Ion Stoica, and Alexey Tumanov. RubberBand: Cloud-based Hyperparameter Tuning. In *Proceedings of the Sixteenth European Conference on Computer Systems*, 2021.

[RCMC20] Alex Renda, Yishen Chen, Charith Mendis, and Michael Carbin. Difftune: Optimizing CPU Simulator Parameters with Learned Differentiable Surrogates. In *2020 53rd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)*. IEEE, 2020.

[RFS+20] Krzysztof Rzadca, Pawel Findeisen, Jacek Swiderski, Przemyslaw Zych, Przemyslaw Broniek, Jarek Kusmierek, Pawel Nowak, Beata Strack, Piotr Witusowski, Steven Hand, et al. Autopilot: Workload Autoscaling at Google. In *Proceedings of the European Conference on Computer Systems*, 2020.

[Rot86] Alvin E Roth. On the Allocation of Residents to Rural Hospitals: A General Property of Two-sided Matching Markets. *Econometrica: Journal of the Econometric Society*, pages 425–427, 1986.

[RSÜ04] Alvin E Roth, Tayfun Sönmez, and M Utku Ünver. Kidney Exchange. *The Quarterly Journal of Economics*, 2004.

[SLA12] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems*, 2012.

[TKS+21a] Brijen Thananjeyan, Kirthevasan Kandasamy, Ion Stoica, Michael Jordan, Ken Goldberg, and Joseph Gonzalez. Resource Allocation in Multi-armed Bandit Exploration: Overcoming Sublinear Scaling with Adaptive Parallelism. In *International Conference on Machine Learning*, 2021.

[TKS+21b] Brijen Thananjeyan, Kirthevasan Kandasamy, Ion Stoica, Michael I Jordan, Ken Goldberg, and Joseph E Gonzalez. PAC Best Arm Identification Under a Deadline. *arXiv preprint arXiv:2106.03221*, 2021.

[VAPGZ17] Dana Van Aken, Andrew Pavlo, Geoffrey J Gordon, and Bohan Zhang. Automatic Database Management System Tuning Through Large-scale Machine Learning. In *Proceedings of the 2017 ACM International Conference on Management of Data*, 2017.

[XEN+19] ZA Xing, D Eldon, AO Nelson, WJ Eggert, MA Roelofs, O Izacard, AS Glasser, NC Logan, R Nazikian, DA Humphreys, et al. Automating Kinetic Equilibrium Reconstruction for Tokamak Stability Analysis. *Bulletin of the American Physical Society*, 2019.