Kelly W. Zhang: Research Statement

My research interests lie at the intersection of **adaptive experimentation**, **sequential decision making**, and **statistical inference**. Specifically, my research has focused on developing methods for statistical inference on data collected with bandit and reinforcement learning (RL) algorithms [14, 15, 16]. I also have experience collaborating with domain scientists in designing and running online RL algorithms for mobile health interventions [9, 10]; these interventions deliver personalized support through wearables and smart devices to help users in managing a variety of chronic health problems in an "in-the-moment" fashion. My research is driven by a desire to use bandit and RL algorithms to address sequential decision-making problems in socially beneficial, human-centered application areas.

1. Statistical Inference for Data Collected with Bandit and RL Algorithms

Due to their ability optimize treatment delivery, bandit and RL algorithms are increasingly used in real-world sequential decision making problems like digital health interventions, online education, and public policy. When designing experiments using RL algorithms two foremost considerations are (i) *within-experiment learning*, i.e., the algorithm's ability to learn during the experiment and personalize treatment delivery (minimize regret), and (ii) *between-experiment learning*, i.e., being able to use the resulting experimental data to perform statistical inference to answer scientific questions, inform stake-holders, and improve the design of the intervention for future iterations. Much of the RL literature has focused on developing algorithms that optimally minimize regret; only recently has there been more work that also considers the "between-experiment learning" objective. Specifically this work develops methods for statistical inference on data collected with RL algorithms. These statistical inference methods are needed because scientific teams often will not even consider using RL algorithms in their applications if they are unable to construct valid confidence intervals for treatment effects after the experiment is over. An example treatment effect is the difference in average reward under two different treatments.

Since bandit and RL algorithms use previously collected data to inform future treatment decisions, they induce dependence in the collected data, i.e., the state, treatment, reward tuples are dependent over time. This induced dependence generally causes standard statistical inference approaches for i.i.d. data to be invalid on this data-type. For example, under many common bandit algorithms, the dependence between observations can be so severe that standard estimators like the sample mean can fail to be asymptotically normal (first proved in my paper [14]). My PhD research has focused on developing novel statistical inference methods for data collected with bandit and RL algorithms [14, 15, 16].

Inference after Running Bandit Algorithms: In [14], I consider bandit algorithms that select treatments in batches; this reflects how in many digital applications online RL algorithms select treatments for multiple users simultaneously. In this work I prove that the sample mean is asymptotic non-normal on data collected with many common bandit algorithms. Specifically, I show that this non-normality occurs when there is no unique optimal treatment, i.e., the margin or treatment effect between the best treatments is zero (or in a shrinking neighborhood around zero). When there is no unique optimal treatment, standard bandit algorithms like Thompson Sampling and epsilon-greedy in the limit do not learn to favor one treatment over another; rather, the action selection propensities formed by these algorithms continue to fluctuate and do not settle down in the limit. In this setting, the action selection propensities depend on the noise of previously observed rewards and this dependence does not go away in the limit; this dependence leads to the asymptotic non-normality.

In [14] I also introduce the batched ordinary least squares estimator that allows one to form valid confidence intervals for treatment effects when the data is collected by bandit algorithms—even when there is no unique optimal treatment. The key idea is to compute t-statistics for each batch and combine across batches, rather than to compute a single t-statistic using all the data; this mitigates the ill-effects of dependence between batches on statistical inference.

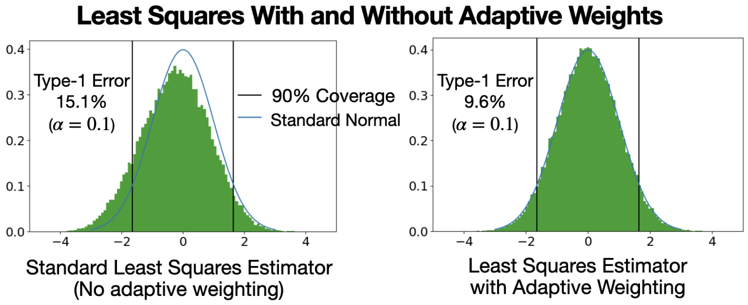


Figure 1: Histograms of z-statistics for least squares estimators for the mean reward under a treatment. The data was collected with a two-arm Thompson Sampling algorithm in a setting with no unique optimal arm.

My next work on inference on bandit data introduced the adaptively weighted M-estimator [15], which facilitates the construction of valid confidence regions for the parameters of M-estimators on data collected with contextual bandit algorithms. The key idea is to include "adaptive" weights which are a function of the action selection propensities; similar to the batched ordinary least squares approach, this has the effect of weighing observations in a data-dependent fashion to preserve asymptotic normality of the M-estimator under adaptive sampling. See Figure 1 for a simple example illustrating the effect of adaptive weighting on a least squares estimator's large sample distribution. *In terms of impact, both these inference approaches introduced above have already been used in practice. The batched ordinary least squares approach was used by political scientists for hypothesis testing after running an experiment with a bandit algorithm that selected between different political survey methods [5]. The adaptively weighted M-estimator approach was used by the World Bank for an education related adaptive experiment in Kenya [3].*

Inference for Longitudinal Data: Contextual bandit environments assume that states are i.i.d. and the expected reward can only depend on the most recent state and action. However, often treatment decisions are made on the same users over time (e.g. in mobile health and online education). In these settings, the contextual bandit environment assumption can be unrealistic; this is because treatment decisions may affect users' future responsiveness to treatments and user outcomes may be non-stationary. In my most recent work[1] [16], I consider longitudinal data environments which allow for non-stationarity and long-term dependencies between each user's outcomes over time. Specifically, I consider inference on longitudinal data after using online RL algorithms that combine or "pool" data across multiple users to inform treatment decisions. These pooling RL algorithms are of great interest because they can potentially learn to select the best treatments faster. In fact, there is so much interest that several clinical trials have already used such pooling RL algorithms [4, 6, 8, 12].

Existing statistical inference methods for longitudinal data that assume independent user data trajectories, can be invalid when applied on data collected with adaptive sampling algorithms that pool data online. This is because by pooling, adaptive sampling algorithms induce dependence between the collected user data trajectories. For example, if an algorithm uses the outcomes of one user to inform future treatment decisions for another user, the data trajectories collected from these two users will not be independent.

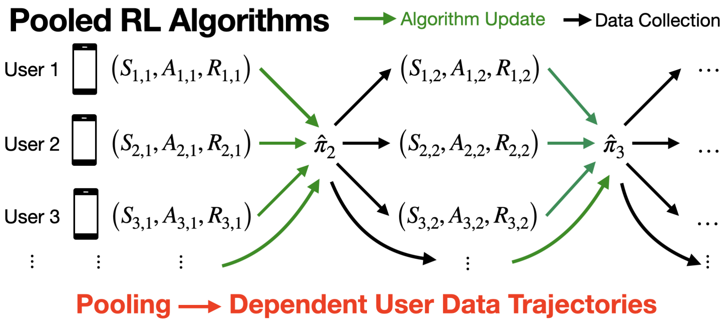

Pooled RL Algorithms
Algorithm Update
Data Collection
User 1
User 2
User 3
$(S_{1,1}, A_{1,1}, R_{1,1})$
$(S_{2,1}, A_{2,1}, R_{2,1})$
$(S_{3,1}, A_{3,1}, R_{3,1})$
$\hat{\pi}_2$
$(S_{1,2}, A_{1,2}, R_{1,2})$
$(S_{2,2}, A_{2,2}, R_{2,2})$
$(S_{3,2}, A_{3,2}, R_{3,2})$
$\hat{\pi}_3$
Pooling → Dependent User Data Trajectories

Figure 2: Illustrating how adaptive sampling algorithms that pool across users induce dependence in the collected user data trajectories.

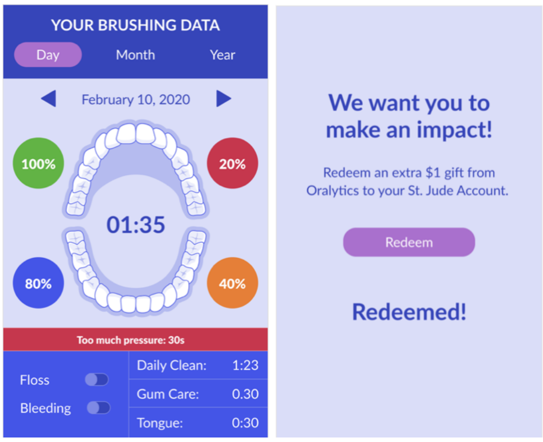
To make the problem tractable, I consider a class of "smooth" pooling RL algorithms, which include Boltzman sampling and stochastic mirror descent algorithms. Note though that these algorithms do not minimize regret optimally with respect to the standard oracle in contextual bandit and Markov decision process environments; however, restricting how regret minimizing an RL algorithm can be generally increases the power of statistical tests one can perform on the resulting collected data [7, 11] (I discuss open questions related to this point later in the Future Research section). Specifically in [16] I develop theory for inference via general Z-estimators on data collected by these "smooth" pooling algorithms. I use ideas from martingale and empirical process theory to show that standard Z-estimators on data collected by algorithms from this class are still asymptotically normal under a particular class of pooling adaptive sampling algorithms, but can have inflated variance. In turn, I introduce the *adaptive sandwich* variance estimator, a modified version of the standard sandwich variance estimator, that corrects for the potential inflation in variance due to the adaptive sampling. *In terms of impact, this work enables researchers to use pooling RL algorithms in mobile health trials. Specifically, my collaborators and I will use the adaptive sandwich variance estimator for a mobile health trial with a pooled RL algorithm that will go into the field in spring 2023 (see discussion of Oralytics further below).*

[1] I am still revising this paper in preparation for submission.

2. Developing and Deploying RL Algorithms

Since 2020, I have been involved with Oralytics, a mobile health micro-randomized trial for oral health [9, 10]. Oral diseases are largely preventable through simple measures like regular brushing and flossing. Despite this, 5-10% of healthcare budgets in industrialized countries are spent on treating dental cavities [1] and nearly one-fifth of U.S. adults 65 or older have lost all their teeth [2]. Oralytics is a study that will investigate the role that data collected through bluetooth enabled toothbrushes can be used to design personalized mobile health interventions to help individuals improve their oral health self-care. The intervention involves sending users engagement messages at opportune moments.

For this project, I collaborate closely with with behavioral scientists, dentists, and software engineers. My work on Oralytics has specifically focused on (i) developing the personalizing RL algorithm that will be used in the trial and (ii) developing the statistical inference approach that will be used for the primary analysis after the experiment concludes (mentioned earlier). Developing the RL algorithm to be used in a mobile health trial involves many challenges, e.g., ensuring that state variables reach the RL algorithm before decision times and ensuring the RL algorithm is able to run stably online autonomously without constant human monitoring. In the process of developing the RL algorithm for Oralytics, in order to help others developing such algorithms for digital interventions, my collaborators and I developed a holistic framework for making RL algorithm design decisions [9], which builds on the predictability, computability, and stability framework [13] for supervised learning problems.



Daily Clean: 1:23
Gum Care: 0.30
Tongue: 0:30

We want you to make an impact!

Redeem an extra $1 gift from Oralytics to your St. Jude Account.

Redeemed!

Figure 3: Oralytics app mockup. Right panel is an example engagement message.

Online RL algorithms use the data collected during the experiment to repeatedly fit approximate models for the users' outcomes; RL algorithms use these models to inform treatment decisions. In applications like mobile health, the models used in online RL algorithms are chosen to appropriately trade-off bias and variance so the algorithm can quickly learn to select effective treatments. For example, in mobile health, if an algorithm sends a user too many messages, this may burden the user and make future treatment messages less effective. However, since the amount of data collected during these experiments is limited and user outcomes are very noisy, it is difficult for algorithms to accurately fit complex models of user outcomes. Thus, often simple algorithms (like bandit algorithms) that do not model the delayed effects of treatment are used in practice [4, 9, 12]. However, since user burden is a significant concern, using such simple algorithms is not ideal. In [10], my collaborators and I develop an approach to modify the reward used by standard contextual bandit algorithms to allow the algorithm to consider user burden. This approach does not increase the complexity of the model used by the bandit algorithm. Specifically, using domain knowledge we penalize the reward with a cost term that is designed to capture the effects of user burden. The cost term can be viewed as a crude proxy for the cost in future value from the Bellman equation in a Markov decision process environment.

3. Future Research

In the future, I plan to extend my current projects and continue to explore new directions particularly in the area of *adaptive experimental design*—which encompasses RL algorithm development, statistical inference after adaptive sampling, and problems at the intersection of of algorithm development and statistical inference. For example, there are many open questions regarding how to design RL algorithms with inference considerations in mind—including formalizing the trade-off between regret minimization and power maximization in different environments, and designing algorithms and statistical tests that are able to optimally achieve that trade-off. Moreover, as mentioned earlier in the section on "Inference for Longitudinal Data", there is a bias variance trade-off in terms of the complexity of the model used by the RL algorithm. It would be interesting to formalize this trade-off and see if that would lead to recommendations as to how to choose model classes in practice. Finally, I hope to continue to work on theoretical problems that are motivated by the challenges faced by scientists running adaptive experiments. I am particularly interested in collaborating with researchers in designing RL algorithms and adaptive experiments for application areas like mobile health, online education, and public policy.

References

[1] Sugars and dental caries. Technical report, World Health Organization, 2017.

[2] B. A. Dye, G. Thornton-Evans, X. Li, and T. Iafolla. *Dental caries and tooth loss in adults in the United States, 2011-2012*. US Department of Health and Human Services, Centers for Disease Control and Prevention, 2015.

[3] B. Esposito and A. Sautmann. Adaptive experiments for policy choice: Phone calls for home reading in kenya. World Bank Group. 2022.

[4] C. A. Figueroa, A. Aguilera, B. Chakraborty, A. Modiri, J. Aggarwal, N. Deliu, U. Sarkar, J. Jay Williams, and C. R. Lyles. Adaptive learning algorithms to optimize mobile applications for behavioral health: guidelines for design decisions. *Journal of the American Medical Informatics Association*, 2021.

[5] M. Offer-Westort, A. Coppock, and D. P. Green. Adaptive experimental design: Prospects and applications in political science. *American Journal of Political Science*, 2021.

[6] J. D. Piette, S. Newman, S. L. Krein, N. Marinec, J. Chen, D. A. Williams, S. N. Edmond, M. Driscoll, K. M. LaChappelle, M. Maly, et al. Artificial intelligence (ai) to improve chronic pain care: Evidence of ai learning. *Intelligence-Based Medicine*, 2022.

[7] D. Simchi-Levi and C. Wang. Multi-armed bandit experimental design: Online decision-making and adaptive inference. *Available at SSRN 4224969*, 2022.

[8] S. Tomkins, P. Liao, S. Yeung, P. Klasnja, and S. Murphy. Intelligent pooling in thompson sampling for rapid personalization in mobile health. 2019.

[9] A. L. Trella, K. W. Zhang, I. Nahum-Shani, V. Shetty, F. Doshi-Velez, and S. A. Murphy. Designing reinforcement learning algorithms for digital interventions: Pre-implementation guidelines. *Algorithms*, 2022.

[10] A. L. Trella, K. W. Zhang, I. Nahum-Shani, V. Shetty, F. Doshi-Velez, and S. A. Murphy. Reward design for an online reinforcement learning algorithm supporting oral self-care. *Thirty-Fifth Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-23)*, 2023.

[11] J. Yao, E. Brunskill, W. Pan, S. Murphy, and F. Doshi-Velez. Power constrained bandits. In *Proceedings of the 6th Machine Learning for Healthcare Conference*, Proceedings of Machine Learning Research, pages 209–259, 2021.

[12] E. Yom-Tov, G. Feraru, M. Kozdoba, S. Mannor, M. Tennenholtz, and I. Hochberg. Encouraging physical activity in patients with diabetes: intervention using an rl system. *Journal of medical Internet research*, 2017.

[13] B. Yu and K. Kumbier. Veridical data science. *Proceedings of the National Academy of Sciences*, 2020.

[14] K. W. Zhang, L. Janson, and S. A. Murphy. Inference for batched bandits. *NeurIPS 2020*, 2020.

[15] K. W. Zhang, L. Janson, and S. A. Murphy. Statistical inference with m-estimators on adaptively collected datas. *NeurIPS 2021*, 2021.

[16] K. W. Zhang, L. Janson, and S. A. Murphy. Statistical inference after adaptive sampling in non-markovian environments. *Preparing for submission*, 2022.