Rethinking experiment design as algorithm design

Jordan W. Suchow **Thomas L. Griffiths**

University of California, Berkeley
{suchow, tom_griffiths}@berkeley.edu

Abstract

As experimentation in the behavioral and social sciences moves from brick-and-mortar laboratories to the web, new opportunities arise in the design of experiments. By taking advantage of the new medium, experimenters can write complex computationally mediated adaptive procedures for gathering data: algorithms. Here, we explore the consequences of adopting an algorithmic approach to experiment design. We review several experiment designs drawn from the fields of medicine, cognitive psychology, cultural evolution, psychophysics, computational design, game theory, and economics, describing their interpretation as algorithms. We then discuss software platforms for efficient execution of these algorithms with people. Finally, we consider how machine learning can optimize crowdsourced experiments and form the foundation of next-generation experiment design.

1 Introduction

The repertoire of experiment designs used in the behavioral and social sciences has hardly changed over the last century [12]. For example, to test a hypothesis an experimenter may randomly assign participants to groups (e.g., treatment or control), manipulate some factor for one group but not the other, and observe whether the participants behave differently as a result. Or, to understand a functional relationship between two variables, an experimenter may ask each participant to judge a set of stimuli that vary along one variable and observe the impact on the other. In some ways, these designs result from constraints on how experiments are typically conducted. Human participants often visit the laboratory for a predetermined amount of time (e.g., an hour) and so it makes sense to have each participant perform a time-consuming task or respond to a number of stimuli. And because experiments are run manually by the experimenter, reconfiguring more than the group assignment or changing the set of stimuli seen by the participant can be challenging.

Recently, behavioral and social scientists have begun to move from brick-and-mortar laboratories to the web, where participants are recruited through crowdsourcing services such as Amazon Mechanical Turk and use browsers to interact with web applications hosted on a server [6, 8, 29, 18, 19]. However, the experiment designs that are used in these virtual laboratories are still largely those developed in the context of physical laboratories – crowdsourcing supports larger and more diverse samples, rapid prototyping, and high-throughput, but has not changed the fundamental structure of experiments.

This is a missed opportunity. Crowdsourced experiments differ from traditional experiments in two important ways. First, participants can be recruited to complete short tasks consisting of as little as a single judgment. Second, the tasks that people are asked to perform can depend on the responses of previous participants. Because experiments are orchestrated by a computer rather than by the experimenter, there is more room for customizing the task that each participant performs. Combining these two properties, a crowdsourced experiment can be understood as a complex, computationally mediated adaptive iterative procedure for gathering data: an algorithm.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

By taking this perspective, it becomes possible to ask what the best algorithms are for learning about human cognition and behavior. Experiment design becomes algorithm design. In this paper, we explore the consequences of adopting this algorithmic approach to experiment design. To begin, we review several experiment designs from the behavioral and social sciences and show how they can be interpretted as algorithms. We then discuss software platforms that can efficiently execute algorithms with people, providing a testbed for experiment design. Finally, we close by describing areas where machine-learning techniques show promise for improving experimentation in the behavioral and social science, optimizing known experiment designs and forming the basis of new ones.

2 Experiments as algorithms

Several experiment designs from the behavioral and social sciences can be understood as implementing algorithms commonly used in machine learning, but with people rather than machines. In this section, we review these designs, identifying the properties of people-based algorithms that must be accommodated by crowdsourcing software and providing background for thinking more broadly about the contributions that machine learning can make in this area.

2.1 Adaptive clinical trials and the multi-armed bandit

The Belmont Report formalizes principles and standards of medical ethics so as to prevent egregious breaches of a doctor's professional duty to the patient while recognizing that even careful research may expose participants to a risk of harm, a risk that must be minimized [13]. Clinical trials, which measure the efficacy and safety of new medical treatments, typically employ randomized, double-blind, placebo-controlled experiments. In these studies, some of the participants are assigned the new treatment, while the others are assigned to a placebo treatment. An ethical dilemma arises whenever partial results of a clinical trial suggest a significant improvement or worsening in the group receiving the new treatment as compared to the control. It has been argued that by continuing the study as originally planned, through their inaction the experimenters harm one set of participants or the other. Adaptive trials, which have grown in popularity over the past several decades, can solve the dilemma by considering the results of an ongoing trial when deciding whether new participants will be assigned to receive the new treatment or the placebo [43].

The multi-armed bandit problem provides a formalization of this dilemma, describing in abstract form the problem faced by a decision-maker who aims to maximize reward in a setting where there are multiple options with unknown value [26]. On each trial, the decision-maker selects an arm and perhaps receives a reward. The decision-maker must balance the benefit of exploiting options known to provide good rewards with the cost of failing to explore the other options, which may provide even better rewards.

A variety of algorithms have been proposed that aim to maximize expected reward. For example, *Thompson sampling* is an algorithm that addresses the exploration–exploitation dilemma by probability matching to the posterior probability that an arm provides the highest reward, and under certain conditions converges to the optimal policy [7]. Other algorithms, such as UCB1 and ϵ-greedy, take different approaches [7].

2.2 Markov chain Monte Carlo with People

A central goal of cognitive psychology is to elucidate the mental representations underlying people's perceptions, inferences, and decisions about the world around them. The psychologist's toolbox includes several methods for estimating these mental representations by asking participants to make judgments about a fixed set of stimuli chosen at the outset of the experiment. For example, one such method, multidimensional scaling (MDS), begins with a matrix of pairwise dissimilarity judgments between the items in a set (e.g. relative distances between major cities in the U.S.) and reduces it to a low-dimensional space preserving the similarity structure (e.g., a 2D map) [39]. Though successful, these methods are inefficient because they do not use information acquired during an experiment to inform its design.

A more efficient technique to reveal human mental representations is based on Markov chain Monte Carlo (MCMC), a class of algorithms from statistical physics that sample from a probability distribution by constructing a Markov chain that converges on the distribution of interest [15]. A Markov chain is a stochastic process defined by a matrix of probabilities of transitions from a given state to any other. Variants of MCMC differ in how each new state of the chain is proposed and then perhaps accepted or rejected. By noting an equivalence between the Barker acceptance function [1] and the Luce choice axiom, a classic decision rule in psychology [25], Sanborn et al. designed an experimental procedure that reveals human mental representations by inserting human decisions into an MCMC algorithm [35]. Specifically, on each iteration of the algorithm, the participant is asked which of two stimuli belong to the category under question. Because under both the Luce choice axiom and the Barker acceptance function stimuli are chosen with probability proportional to their probability under distribution in question, the resultant chain converges on a probability distribution matching the human's mental representations. To validate the technique, Sanborn et al. asked participants to choose which of two stick-figure quadrupeds represents a particular animal species (e.g., a giraffe). Over many iterations of the algorithm, the chain converges on a stationary distribution that visits stick-figure forms in proportion to their probability under the human participant's mental representation of that species. The technique has been applied to uncover diverse mental representations, including those of phonemes [32], facial affects [27], and colliding objects [9].

2.3 Transmission chains and computing the prior

The transmission chain is an experimental technique that, much like the children's game Telephone, passes information from one person to the next in succession [2]. As the information changes hands, it is transformed by the perceptual, inductive, and reconstructive biases of the individuals. This eventually leads to erasure of the information contained in the input, leaving behind a signature of the transformation process itself. For this reason, transmission chains have been used to study language evolution [37, 23] and the effect of culture on memory [2].

Transmission chains can be formally modeled as a Markov chain by assuming that perception, learning, and memory follow the principles of Bayesian inference [22]. Under this analysis, in which agents use Bayes' rule to infer the process that generated observed data, each agent holds a set of hypotheses $\mathcal{H}$ about the process that produced the observed data d. A prior probability distribution $p(h)$ encodes perceptual, inductive, and reconstructive biases in the form of the agent's assigned probability of each $h \in \mathcal{H}$ before observing the data. In context of a transmission chain, each agent uses Bayes' rule to compute the posterior distribution

$$p(h|d) = \frac{p(d|h)p(h)}{\sum_{h \in \mathcal{H}} p(d|h')p(h')} \tag{1}$$

A hypothesis is then sampled from the posterior distribution and used to generate the data passed to the next person in the chain by sampling from the hypothesis's likelihood function. When a transmission chain is set up in this way, iterated learning is equivalent to a form of Gibbs sampling, a widely-used Markov chain Monte Carlo algorithm [17]. The convergence results for the Gibbs sampler thus apply, with the prior $p(h)$ as the stationary distribution of the Markov chain on hypotheses.

2.4 Contrast sensitivity and adaptive, sequential procedures

The study of human sensation and perception is often concerned with measuring the threshold intensity at which a stimulus can just barely be detected, recognized, or identified. For example, the contrast sensitivity function, which tracks human sensitivity to a faint pattern as a function of the pattern's spatial frequency, is a useful diagnostic tool for visual dysfunction [33]. The classic experiment design for measuring a threshold is the *method of constant stimuli*, which begins by measuring human performance at each of a fixed set of stimulus intensities and proceeds by fitting some kind of parametric form to the resultant data in order to estimate threshold [31]. This procedures typically requires several hundred human judgments.

More efficient techniques for measuring thresholds adaptively adjust the intensity of the stimulus between trials, on each iteration recomputing the stimulus placement based on the partial results of the ongoing experiment. Designing an efficient measurement procedure then becomes a question of designing an active learning algorithm that places the stimulus at whichever intensity will be most informative with respect to determining the threshold [38]. The algorithm QUEST, for example, recommends placing each trial at the mode of the posterior probability density function [45]. More recent methods, such as the FAST algorithm, efficiently measures thresholds that vary as a function of another variable [44].

2.5 Automated mechanism design through optimization

Game theory considers the behavior that results when rational decision-makers cooperate and compete in well-defined strategic games. In contrast, mechanism design, sometimes called "reverse game theory", considers the design of arrangements, rules, and procedures under which rational decision-makers will behave in accordance with the designer's stated objective [11]. For example, in the context of auctions, the theory can identify incentive-compatible mechanisms that maximize the seller's expected revenue [28].

Mechanisms are usually hand crafted by a designer who uses experience and intuition as a guide. In contrast, *automated mechanism design* uses computational techniques to craft a mechanism for the specific problem at hand [36]. To do so requires a formal specification of the design space of possible mechanisms and a means of evaluating the objective function for a given mechanism. It then becomes possible to use search and optimization algorithms to find mechanisms that score well. For example, in an early application of this approach, an automated procedure reinvented the Myerson auction [28] and created expected-revenue-maximizing combinatorial auctions [10].

2.6 Interactive evolutionary computation

Preferences can be elicited experimentally by asking a participant to choose which of two options is preferable. For example, in psychology and behavioral economics, studies of temporal discounting ask a person to select between a small sum given now and a larger sum given later [16]. Two-alternative forced-choice designs over all possible pairs of competing forms can find the preferred form when the set of options is small, but the procedure scales poorly when there are many forms to compare. Consider for example *vanity numbers*, telephone numbers with a custom and often easily remembered sequence of digits (e.g., +1 202 456-1111). An algorithm that finds preferred vanity numbers by taking as input pairwise preference judgments across all valid telephone numbers is intractable.

Interactive evolutionary computation can solve problems like finding good vanity numbers by inserting human participants into various parts of an evolutionary process [41]. In interactive genetic algorithms, for example, a human participant may act as the selection mechanism, using aesthetic judgment to select which forms will replicate to form the next generation [41]. In many cases, these algorithms quickly converge on preferred forms.

3 Programming languages for algorithms with people

Executing an algorithm that incorporates human computation requires a means by which to interface with people. Over the past several years, a variety of systems have been created and deployed that provide such interfaces. In fact, having drawn here the link between experiment design and algorithm design, there is a sense in which all web-based experiments are automated algorithms with people. However, most of these web-based experiments use designs that were developed for physical laboratories. Here, we focus on three systems that exploit the change in medium:

TurKit is a Java-based toolkit that executes algorithms that include human computation as a function call [24]. TurKit recruits participants using Amazon Mechanical Turk and uses a crash-and-rerun programming model. Interestingly, it is possible to define higher-order functions that have as sub-elements individual human judgments, such as group voting among a set of alternatives. The toolkit was initially validated on a range of simple tasks such as iterative editing of prose, OCR of highly degraded text, sequential decision making in a group setting, and collaborative sorting of salient imagery [24].

Wallace [40] is a software-based tool for orchestrating experiments in the behavioral and social sciences using crowdsourcing, originally developed to run large-scale experiments on cultural transmission, but which we have more recently extended to a diverse set of experimental paradigms. Wallace provides efficient high-throughput automation for running the full pipeline of experimentation: it recruits participants through crowdsourcing services, obtains their informed consent, arranges them into a network, coordinates their communication, records the data they produce, pays them, and validates and manages the resultant data. Critically, because Wallace both exposes its configuration settings in a standard format and is fully automated, it is possible to write algorithms that use entire Wallace-based experiments as subroutines.

EvoSpace-i [14] is a framework for running interactive evolutionary algorithms to create, for example, art. The framework's architecture is a web application that integrates with the Facebook social graph to recruit people who then evaluate or otherwise interact with the evolving artistic forms.

4 Using machine learning to design better experiments

Machine learning can contribute to rethinking experiment design as algorithm design in two ways: by offering algorithms that can be translated into new experiment designs, and by providing tools for optimizing existing experiment designs.

4.1 Developing new algorithms to run with people

Many of the experimental procedures discussed above are linked to machine learning algorithms – the multi-armed bandit, Markov chain Monte Carlo, Gibbs sampling, etc. This correspondence is no coincidence: machine learning problems often infer the structure of a complex object, be it a reward function or a probability distribution, and this is exactly the problem faced by a scientist who seeks to identify the nature of some aspect of human cognition.

The known correspondences between experiment designs and machine learning algorithms raise the question of whether other ideas from machine learning can form the basis of new experiment designs. There are a few added constraints associated with running algorithms with people. For example, because human behavior is noisy, any learning algorithm must accommodate noise. And because of the timescales associated with recruitment and performance of human participants, these in vivo algorithm may require a different set of computational tradeoffs than algorithms run in silico. Within these constraints, however, it seems likely that there are many algorithms that can be translated into new experiment designs.

Problems that seem particularly pertinent include methods such as manifold learning that can be used to recover complex non-linear embeddings of stimulus domains [42, 34], and methods used for optimization that might make it possible to identify optimal stimuli for eliciting a particular response (for example, the most memorable images [21, 5]). Methods based on stochastic gradient descent [4], for example, might provide a way to find optimal stimuli in complex spaces, provided it is possible to develop tasks that reliably measure the gradient of the behavior of interest.

4.2 Optimizing experimental designs

Machine learning methods can also be used to optimize existing experimental designs. Recent work in machine learning has started to explore "meta-learning" methods – procedures that use machine learning ideas to speed up or otherwise improve machine learning algorithms [3, 20]. For example, Bayesian optimization [30] uses Bayesian inference over functions to better select the next steps in an algorithm that aims to maximize those functions. In the context of experiment design, Bayesian optimization can potentially be used to tune the parameters of experiments in real time as participants produce their responses.

Consider, for example, a basic design decision that affects nearly all crowdsourced experimentation: how much time should the participant be given to complete the task? If that duration is too short, the participant will be unable to complete the task in the alloted time or a speed–accuracy tradeoff may preclude collection of data of sufficient quality. But if that duration is too long, the participant may delay unnecessarily, perhaps waiting until the end of the interval to begin the task, further delaying the overall experiment. This parameter of the experiment's design is often hand-tuned or chosen arbitrarily; machine learning methods provide a more principled approach.

5 Conclusion

An experiment's design is an algorithm for gathering data. From medicine to game theory, popular experiment designs such as adaptive clinical trials and cultural transmission chains are equivalent to familiar algorithms from computer science and machine learning. Platforms such as Wallace, TurKit, and EvoSpace-i can efficiently execute these algorithms with people on the web. In combination, machine learning and crowdsourcing provide the tools needed for a new generation of experiment designs.

Acknowledgments

This work was funded in part by the National Science Foundation (SBE-1408652 to J.W.S and T.L.G; BCS-1456709 to T.L.G).

References

[1] AA Barker. Monte carlo calculations of the radial distribution functions for a proton? electron plasma. *Australian Journal of Physics*, 18(2):119–134, 1965.

[2] Frederic C Bartlett. Remembering: An experimental and social study. *Cambridge: Cambridge University*, 1932.

[3] James S Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *Advances in Neural Information Processing Systems*, pages 2546–2554, 2011.

[4] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[5] Timothy F Brady, Talia Konkle, George A Alvarez, and Aude Oliva. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105(38):14325–14329, 2008.

[6] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5, 2011.

[7] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems*, pages 2249–2257, 2011.

[8] Daniel L Chen, Martin Schonger, and Chris Wickens. otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97, 2016.

[9] Andrew L Cohen and Michael G Ross. Exploring mass perception with markov chain monte carlo. *Journal of Experimental Psychology: Human Perception and Performance*, 35(6):1833, 2009.

[10] Vincent Conitzer and Tuomas Sandholm. Applications of automated mechanism design. In *UAI-03 workshop on Bayesian Modeling Applications*, 2003.

[11] Arthur G Erdman and George N Sandor. *Mechanism design: analysis and synthesis (Vol. 1)*. Prentice-Hall, Inc., 1997.

[12] Ronald Aylmer Fisher. The design of experiments. 1960.

[13] National Commission for the Protection of Human Subjects of Biome Beha Resea and Kenneth John Pres Ryan. *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research-the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research*. US Government Printing Office, 1978.

[14] Mario García-Valdez, Juan J Merelo, Leonardo Trujillo, Francisco Fernández-de Vega, José C Romero, and Alejandra Mancilla. Evospace-i: a framework for interactive evolutionary algorithms. In *Proceedings of the 15th Annual Conference Companion on Genetic and Evolutionary Computation*, pages 1301–1308. ACM, 2013.

[15] Walter R Gilks. *Markov chain monte carlo*. Wiley Online Library, 2005.

[16] Leonard Green and Joel Myerson. A discounting framework for choice with delayed and probabilistic rewards. *Psychological Bulletin*, 130(5):769, 2004.

[17] Thomas L Griffiths and Michael L Kalish. Language evolution by iterated learning with bayesian agents. *Cognitive Science*, 31(3):441–480, 2007.

[18] Todd M Gureckis, Jay Martin, John McDonnell, Alexander S Rich, Doug Markant, Anna Coenen, David Halpern, Jessica B Hamrick, and Patricia Chan. psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, pages 1–14, 2015.

[19] John J Horton, David G Rand, and Richard J Zeckhauser. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, 14(3):399–425, 2011.

[20] Frank Hutter and Youssef Hamadi. Parameter adjustment based on performance prediction: Towards an instance-aware problem solver. In *In: Technical Report: MSR-TR-2005125, Microsoft Research*, 2005.

[21] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 145–152. IEEE, 2011.

[22] Michael L Kalish, Thomas L Griffiths, and Stephan Lewandowsky. Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, 14(2):288–294, 2007.

[23] Simon Kirby, Hannah Cornish, and Kenny Smith. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31):10681–10686, 2008.

[24] Greg Little, Lydia B Chilton, Max Goldman, and Robert C Miller. TurKit: Human computation algorithms on MTurk. 2009.

[25] R Duncan Luce. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2005.

[26] Aditya Mahajan and Demosthenis Teneketzis. Multi-armed bandit problems. In *Foundations and Applications of Sensor Management*, pages 121–151. Springer, 2008.

[27] Jay B Martin, Thomas L Griffiths, and Adam N Sanborn. Testing the efficiency of markov chain monte carlo with people using facial affect categories. *Cognitive Science*, 36(1):150–162, 2012.

[28] Roger B Myerson. Optimal auction design. *Mathematics of Operations Research*, 6(1):58–73, 1981.

[29] David C Parkes, Andrew Mao, Yiling Chen, Krzysztof Z Gajos, Ariel Procaccia, and Haoqi Zhang. Turkserver: Enabling synchronous and longitudinal online experiments. In *Proceedings of the Fourth Workshop on Human Computation (HCOMP'12)*. AAAI Press, 2012.

[30] Martin Pelikan, David E Goldberg, and Erick Cantú-Paz. Boa: The bayesian optimization algorithm. In *Proceedings of the 1st Annual Conference on Genetic and Evolutionary Computation-Volume 1*, pages 525–532. Morgan Kaufmann Publishers Inc., 1999.

[31] Denis G Pelli and Peter Bex. Measuring contrast sensitivity. *Vision Research*, 90:10–14, 2013.

[32] James P Pooley. Exploring phonetic category structure with Markov chain Monte Carlo. 2008.

[33] JG Robson. Spatial and temporal contrast-sensitivity functions of the visual system. *Journal of the Optical Society of America*, 56(8):1141–1142, 1966.

[34] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[35] Adam N Sanborn, Thomas L Griffiths, and Richard M Shiffrin. Uncovering mental representations with markov chain monte carlo. *Cognitive Psychology*, 60(2):63–106, 2010.

[36] Tuomas Sandholm. Automated mechanism design: A new application area for search algorithms. In *International Conference on Principles and Practice of Constraint Programming*, pages 19–36. Springer, 2003.

[37] Thomas C Scott-Phillips and Simon Kirby. Language evolution in the laboratory. *Trends in Cognitive Sciences*, 14(9):411–417, 2010.

[38] Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11, 2010.

[39] Roger N Shepard. Analysis of proximities as a technique for the study of information processing in man. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 5(1):33–48, 1963.

[40] Jordan W. Suchow, Thomas J. H. Morgan, Jessica Hamrick, Michael D. Pacer, Stephan C. Meylan, and Thomas L. Griffiths. Wallace: A platform for simulating cultural evolution in structured populations online. In *Crowdsourcing and Online Behavioral Experiments Workshop at the ACM Conference on Economics and Computation*, 2015.

[41] Hideyuki Takagi. Interactive evolutionary computation: Fusion of the capabilities of ec optimization and human evaluation. *Proceedings of the IEEE*, 89(9):1275–1296, 2001.

[42] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[43] Rieke van der Graaf, Kit CB Roes, and Johannes JM van Delden. Adaptive trials in clinical research: scientific and ethical issues to consider. *JAMA*, 307(22):2379–2380, 2012.

[44] Edward Vul, Jacob Bergsma, and Donald IA MacLeod. Functional adaptive sequential testing. *Seeing and Perceiving*, 23(5):483–515, 2010.

[45] Andrew B Watson and Denis G Pelli. Quest: A bayesian adaptive psychometric method. *Perception & Psychophysics*, 33(2):113–120, 1983.