Adaptive Experimentation at Scale

Ethan Che
Decision, Risk, and Operations Division
Columbia Business School
New York, NY 10027
eche25@gsb.columbia.edu

Hongseok Namkoong
Decision, Risk, and Operations Division
Columbia Business School
New York, NY 10027
namkoong@gsb.columbia.edu

Abstract

In typical experimentation paradigms, reallocating measurement effort incurs high operational costs due to delayed feedback, and infrastructural and organizational difficulties. Challenges in reallocation lead practitioners to employ a few reallocation epochs in which outcomes are measured in large batches. Standard adaptive experimentation methods, however, do not scale to these regimes as they are tailored to perform well as the number of reallocation epochs grows. We develop a new adaptive experimentation framework that can flexibly handle any batch size and learns near-optimal designs when reallocation opportunities are few. By deriving an asymptotic sequential experiment based on normal approximations, we formulate a Bayesian dynamic program that can leverage prior information based on previous experiments. We propose policy gradient-based lookahead policies and find that despite relying on approximations, our methods greatly improve statistical power over uniform allocation and standard adaptive policies.

1 Introduction

Experimentation is the basis of scientific decision-making for engineering solutions, business products, and policy-making. As engineering solutions and policy interventions become more sophisticated, modern experiments increasingly involve many treatment options ("arms") tested over large populations [2, 27]. For example, to improve product design, online platforms test many configurations across millions of users, e.g., hyperparameters of an ML-driven recommendation system. In such scenarios, treatments typically impact a small part of the overall outcome of interest, which are typically defined using key business metrics such as revenue or user satisfaction; although any single product improvement may yield small relative increase in revenue, absolute gains are nevertheless substantial. The relative performance differential between arms can thus be difficult to discern even with a large sampling budget, particularly when the population is stratified into groups [31].

Adaptive allocation of measurement effort can improve statistical power and allow reliable identification of the optimal decision/treatment. Accordingly, adaptive methods—dubbed pure-exploration multi-armed bandit (MAB) algorithms—have received tremendous attention since the foundational works of Thompson, Chernoff, Robbins, and Lai [30, 6, 23, 16]. However, standard frameworks cannot model typical experimentation paradigms in online platforms and scientific studies where adaptive reallocation incurs high operational costs. Although a universal assumption in the MAB literature [4, 18, 29], unit-level continual reallocation of sampling effort is often expensive or infeasible due to organizational cost and delayed feedback, especially in large-scale experiments. Even for online platforms with advanced experimentation infrastructure designed to handle millions to billions of units, engineering difficulties and lack of organizational incentives deter continual reallocation at the unit level [28, 1, 3, 22].

Due to challenges associated with reallocating measurement effort, typical real-world experiments employ a few reallocation epochs in which outcomes are measured for many units in parallel

2022 NeurIPS Workshop on Gaussian Processes, Spatiotemporal Modeling, and Decision-making Systems.

("batches") [7, 15]. Most MAB algorithms do not scale well to batch settings as they are specifically designed to enjoy strong theoretical guarantees as the number of reallocation epochs grows [4, 18, 29]. Motivated by these challenges, we develop and analyze adaptive experimentation methods tailored to a handful of reallocation opportunities. Our methods are near-optimal for the fixed number of reallocation epochs and are adapted to the instance-specific measurement noise and statistical power. Specifically, we formulate a limiting adaptive experiment based on normal approximations and develop a dynamic program (DP) solving for the optimal sequential allocation. Our DP framework is Bayesian and can leverage prior information constructed using the rich reservoir of previous experiments. Computationally, we propose approximate DP methods using black-box ML models (e.g., neural networks), including more myopic yet effective heuristics such as lookahead policies.

2 Asymptotic Sequential Experiment

Our goal is to select the best treatment arm out of K alternatives, using a small (known) number of reallocation epochs (T). Each experimentation epoch $t = 0, ..., T-1$ involves a batch of $B_t = b_t n$ samples, where $b_t > 0$ is a fixed and known constant (that may vary across t) and n is a scaling parameter. If the gap in the average rewards of each arm is $\gg 1/\sqrt{n}$, adaptive experimentation is unnecessary as the best arm can be found after only a single epoch. Conversely, if the gaps are $\ll 1/\sqrt{n}$, then we cannot reliably learn even after many experimentation epochs; one cannot improve upon the uniform allocation (a.k.a. randomized design or A/B testing). We thus focus on the admissible regime where the gaps between arm rewards are $\Theta(1/\sqrt{n})$. Upon allocating a unit to a treatment arm a, the experimenter observes i.i.d. draws of the arm rewards $R_a = \frac{h_a}{\sqrt{n}} + \epsilon_a$, where h_a is an unknown "local parameter" that determines the difference between average arm rewards. Without loss of generality, we set the baseline reward to zero. We assume the noise ϵ_a has mean zero, and that $\text{Var}(\epsilon_a) = s_a^2$ is known and constant. In particular, s_a^2 does not scale with n, so it is crucial to sample arms many times to discern differences between their means $\frac{h_a}{\sqrt{n}}$. Although reward variances are typically unknown, they can be estimated from a small initial batch in practice; empirically, the policies we consider are robust to estimation error in s_a^2, and a rough estimate suffices.

Since modern organizations run hundreds of experiments, it is natural for the experimenter to have a prior distribution $h \sim \nu$ over the relative gaps between treatment effects. At the end of the experiment (epoch $t = T$), our goal is to minimize the *Bayes simple regret* $\mathbb{E}_{h\sim\nu}\mathbb{E}[h_{a^\star} - h_{\widehat{a}}]$ the scaled optimality gap between the selection $\widehat{a}$ and the optimal arm $a^\star$ averaged over the prior. To optimize the Bayes simple regret, the experimenter uses the information collected until the beginning of epoch t to select $\pi_t \in \Delta_K$, the fraction of samples allocated to each of the K treatment arms. In general, the reward distributions are unknown, and it is challenging to calculate posterior distributions that inform Bayesian sampling procedures (e.g., [24]). Instead, we note the importance sampling estimator $\bar{R}^n_{t,a}$ for the average reward approximately follows a normal distribution

$$\sqrt{n}\bar{R}^n_{t,a} := \frac{1}{b_t\sqrt{n}} \sum_{j=1}^{b_t n} \frac{\xi^t_{a,j}}{\pi_{t,a}} R^t_{a,j} \overset{d}{\rightsquigarrow} N(h_a, \frac{s_a^2}{b_t \pi_{t,a}}), \tag{1}$$

where $\xi^t_{a,j}$ is an indicator for whether arm a was pulled for unit j at time t. The allocation π_t controls the effective sample size and thus the level of uncertainty in the sample mean and the ability to distinguish signal from noise (a.k.a. statistical power).

At each epoch t, the experimenter chooses π_t and observes an independent Gaussian measurement distributed as $N(h_a, \frac{s_a^2}{b_t \pi_t})$ for each arm a. We use the asymptotic Gaussian sequential experiment as an approximation to the original batched adaptive epochs and derive near-optimal adaptive experimentation methods for the asymptotic problem. While our framework allows naturally incorporating prior information, we do not assume restrictive distributional assumptions on rewards that are required in typical Bayesian sequential sampling approaches [25]. Instead, the likelihood function of the aggregate rewards $\sqrt{n}\bar{R}^n_{t,a} \mid h$ is derived from the normal approximation (1). Our limiting normal approximation coincides with typical (frequentist) inferential paradigms for confidence intervals and power calculations. While our approach to improving statistical power is new, the corresponding Gaussian sequential experiment (2) was previously studied in the context of robust control [21] and attention allocation [19], and Frazier and Powell [10] studied a single epoch variant.

Based on the observation (1), our main theoretical result derives an asymptotic sequential experiment as $n \to \infty$. At each epoch t, the experimenter chooses π_t and observes an independent Gaussian measurement distributed as $N(h_a, \frac{s_a^2}{b_t \pi_t})$ for each arm a.

Theorem 1. *Let* $\text{BSR}_T(\pi, \nu, \bar{R})$ *be the Bayes simple regret under policy* π*, prior* ν *over* h*, and observation process* $\bar{R} = (\bar{R}_0, \ldots, \bar{R}_{T-1})$*, which is the set of observations used by the policy to determine the sampling allocations. Consider the asymptotic sequential experiment characterized by observations* $V_0, \ldots, V_{T-1}$ *with conditional distributions* $V_t | V_0, \ldots, V_{t-1} \sim N(h, diag(\frac{s_a^2}{b_t \pi_t}))$*. Under regularity conditions for the rewards and the policy,* $\text{BSR}_T(\pi, \nu, \sqrt{n}\bar{R}^n) \to \text{BSR}_T(\pi, \nu, V)$.

We use the asymptotic Gaussian sequential experiment as an approximation to the original batched adaptive epochs and derive near-optimal adaptive experimentation methods for the asymptotic problem. While our framework allows naturally incorporating prior information, we do not assume restrictive distributional assumptions on rewards that are required in typical Bayesian sequential sampling approaches [25]. Instead, the likelihood function of the aggregate rewards $\sqrt{n}\bar{R}^n_{t,a} \mid h$ is derived from the normal approximation (1). Our limiting normal approximation coincides with typical (frequentist) inferential paradigms for confidence intervals and power calculations.

A number of works in Bayesian optimization [11, 32, 12, 13, 17, 14] and experimental design study batched adaptive designs that alleviate the myopia of standard acquisition functions. In contrast to this literature studying continuous design spaces, we focus on allocation of measurement effort over a finite number of arms under limited statistical power. Our setting is characterized by limited extrapolation between arms and a fixed, finite exploration horizon. As our approach maximizes an expected utility function, it is intimately connected to Bayesian experimental design methods [5, 26, 9]. Instead of optimizing expected information gain (EIG), we minimize expected simple regret at the terminal period, a more tractable objective. While our approach to improving statistical power is new, the corresponding Gaussian sequential experiment (2) was previously studied in the context of robust control [21] and attention allocation [19], and Frazier and Powell [10] studied a single epoch variant.

3 Bayesian Adaptive Experimentation

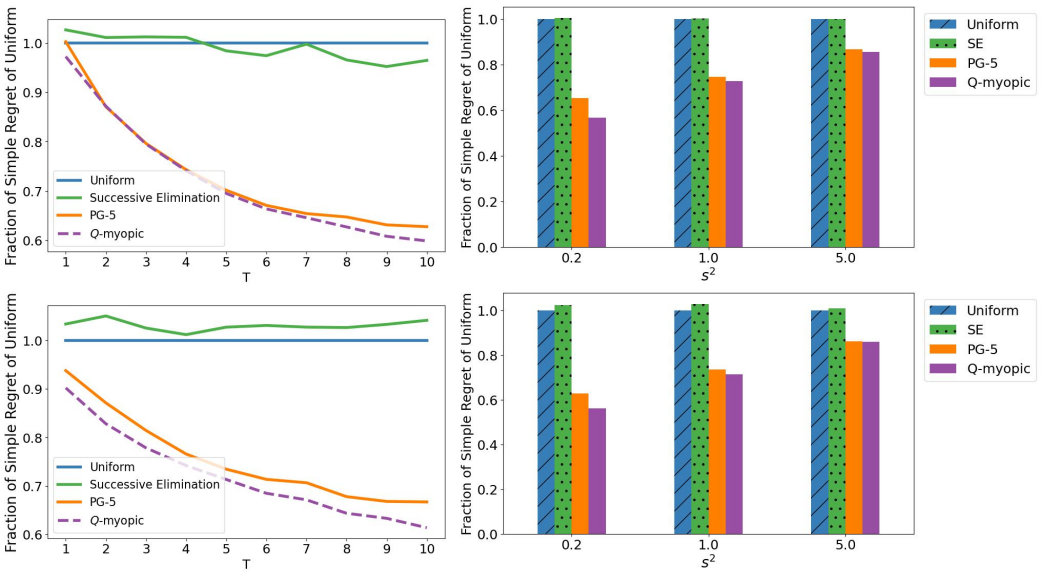
We rewrite the asymptotic sequential experiment as a Markov decision process to find the optimal adaptive design. Our formulation is based on reparameterizing the distribution of sequential observations $V_0, \ldots, V_{T-1}$ using posterior mean and variances defined over standard normal variables. Formally, consider an independent Gaussian prior over the local parameters, $h_a \sim N(\mu_{0,a}, \sigma^2_{0,a})$, under which the trajectory of posterior beliefs $(\mu_{t,a}, \sigma^2_{t,a})$ follows a Markov Decision Process. Using standard normal variables $Z_{t,a} \overset{\text{iid}}{\sim} N(0,1)$, the "states" μ_t, σ_t^2 and "actions" $\pi_t \in \Delta_k$ characterize the asymptotic sequential experiment through the following transitions

$$\mu_{t+1,a} = \mu_{t,a} + \sigma_{t,a} \sqrt{\frac{b_t \pi_{t,a} \sigma^2_{t,a}}{s_a^2 + b_t \pi_{t,a} \sigma^2_{t,a}}} Z_{t,a} \quad \text{and} \quad \sigma^2_{t+1,a} = \left(\frac{1}{\sigma^2_{t,a}} + \frac{b_t \pi_{t,a}}{s_a^2} \right)^{-1}. \tag{2}$$

Our goal is to minimize the Bayes simple regret, which is equivalent to maximizing $\mathbb{E}\left[\max_{a=1,\ldots,K} \mu_{T,a}\right]$, the highest posterior mean at the end of the experiment. At epoch t, conditional on the information $Z_0, \ldots, Z_{t-1} \overset{\text{iid}}{\sim} N(0, I)$ available—summarized via the current state (μ_t, σ_t)—the Q-function for a policy π is given by

$$Q^\pi_t(\mu_t, \sigma_t) = \mathbb{E}^\pi_t[\max_a \mu_{T,a}] = \mathbb{E}^\pi_t \left[\max_a \left\{ \mu_{t,a} + \sum_{s=t}^{T-1} \sigma_{s,a} \sqrt{\frac{b_s \pi_{s,a}(\mu_s, \sigma_s) \sigma^2_{s,a}}{s_a^2 + b_s \pi_{s,a}(\mu_s, \sigma_s) \sigma^2_{s,a}}} Z_{s,a} \right\} \right]. \tag{3}$$

Using dynamic programming to directly solve the policy optimization problem $\max_\pi Q^\pi_0(\mu_0, \sigma_0)$ is computationally intractable even for a moderate number of treatment arms and reallocation epochs. Instead, we propose an approximate DP method based on policy gradients (PG) computed through black-box ML models. We consider an auto-differentiable parameterized policy $\pi_\theta = \{\pi^\theta_t\}_{t=0}^{T-1}$ (e.g., neural networks). We aim to directly optimize the Q-function (3) using stochastic approximation methods: we use stochastic gradient ascent over sample paths $Z_0, \ldots, Z_{T-1}$ to update policy parameters $\theta \leftarrow \theta + \alpha \nabla_\theta Q^{\pi_\theta}_0(\mu_0, \sigma_0)$. For long horizons, training the policy becomes more



(a) Relative gains over uniform allocation for $s_a^2 \equiv 1$ (above: $K = 10$; below: $K = 100$).

(b) Relative gains over uniform allocation for $s_a^2 \in \{\frac{1}{5}, 1, 5\}$ with $T = 5$ (above: $K = 10$; below: $K = 100$).

Figure 1: Pre-limit Bayes simple regret for Gumbel rewards and Gamma$(100, 100)$ prior

difficult due to high variance in stochastic gradient estimates. We hence consider m-lookahead policies (denoted PG-m) trained to optimize the shorter time horizon objective $Q^{\pi_\theta}_{T-m}(\mu_0, \sigma_0)$.

We also consider a simple heuristic, Q-*myopic*, which solves a computationally cheaper approximation of the dynamic program by assuming non-adaptive future allocations. Instead of maximizing the Q function $\mathbb{E}^{\pi}_t[\max_a \mu_{T,a}]$ over policies, Q-myopic solves the open loop problem that considers future sampling allocations that only depend on the available information (μ_t, σ_t) at time t. At each epoch t, π_t is derived by assuming that an identical allocation will be used for the remaining periods horizon regardless of new information obtained in the future $\pi_t^{Q\text{-myopic}}(\mu_t, \sigma_t) = \text{argmax}_{\bar{\pi} \in \Delta_K} \mathbb{E}^{\bar{\pi}}_t[\max_a \mu_{T,a}]$. We again solve for the Q-myopic policy through stochastic approximation methods. Empirically, we find that the open loop approximation provides a highly effective heuristic. One appealing feature is that solving the open-loop problem is easier for longer residual horizons $T - t$. As $T - t \to \infty$, the objective function of the open-loop problem becomes strongly concave and we can characterize the asymptotic sampling policy explicitly from the KKT conditions .

We denote the limiting policy as π^{DTS}, Density Thompson Sampling (DTS), as it determines the allocation based on the partial derivatives of the Gaussian Thompson sampling probabilities. Our result provides a novel connection between optimization-based approaches for sampling (e.g. expected improvement [20]) and probability-matching methods (e.g. Thompson sampling [25]).

Proposition 2. *Let* $\Delta_K^\epsilon = \Delta_K \cap \{p : p \geq \epsilon\}$ *for some* $\epsilon > 0$. *There exists* $t_0 > 0$ *such that* $\forall T - t > t_0$, $(\sum_{s=t}^{T-1} b_s)\mathbb{E}^{\bar{\pi}}_t[\max_a \mu_{T,a}]$ *is strongly concave in* $\bar{\pi} \in \Delta_K^\epsilon$. *As* $T \to \infty$,

$$\pi_{t,a}^{Q\text{-myopic}}(\mu_t, \sigma_t) \to \pi_a^{\text{DTS}}, \text{ where } \pi_a^{\text{DTS}} \propto s_a \left(\frac{\partial}{\partial \mu_a} \pi_a^{\text{TS}}(\mu_t, \sigma_t) \right)^{1/2}$$

4 Empirical results

Our main empirical observations inform the design of effective adaptive experimentation methods when reallocation of sampling efforts is few and expensive. For general reward distributions, the asymptotic Gaussian sequential experiment problem (2) provides a valuable framework for experimental design, even when the scaling parameter n (batch size) is not large. Our approximate dynamic programming policies outperform uniform allocation and even specialized algorithms (e.g., variants of Thompson sampling [24]) that require complete knowledge of the reward distribution. We observe the policy gradient-based policies can generalize to longer horizons than initially trained on. The Q-myopic policy enjoys strong performance despite optimizing a lower bound of the Q function. Both policy gradient and Q-myopic achieve more significant performance gains in harder/underpowered experiments with high effective measurement noise s_a^2/b_t.

As a concrete illustration of our main findings, consider a setting with $K = 10,100$ arms with up to $T = 10$ batches. We consider 100 samples in each batch, where rewards follow a Gumbel distribution with a fixed scale parameter $\beta > 0$ across all arms, implying the measurement variance $s_a^2 = \frac{\pi^2}{6}\beta$. The location parameters μ_a are drawn from an independent $Gamma(100, 100)$ prior, which is known to the experimenter. The Gaussian sequential approximation for each batch gives $\bar{R}_{t,a}^n \sim N\left(\mu_a, s_a^2/(10\pi_t)\right)$. To maintain conjugacy, we also approximate the prior with a normal distribution with the same mean ($\mu_{0,a} = 1$) and standard deviation ($\sigma_{0,a} = 0.1$). We use an adapted version of the successive elimination (SE) algorithm [8]—a popular heuristic in practice—as a key benchmark. In Figure 1a), policy gradient and Q-myopic achieve substantial performance gains over uniform allocation, despite relying on normal approximations of the true reward distribution and the prior distribution. Gains are significant even with few reallocation epochs, and grow with more adaptivity (larger T). In Figure 1b), these gains persist when the noise level is high, even though standard adaptive procedures (SE) struggle to eliminate non-performant arms.

References

[1] A. Agarwal, S. Bird, M. Cozowicz, L. Hoang, J. Langford, S. Lee, J. Li, D. Melamed, G. Oshri, O. Ribas, et al. Making contextual decisions with low technical debt. *arXiv preprint arXiv:1606.03966*, 2016.

[2] D. Agarwal, B. Long, J. Traupman, D. Xin, and L. Zhang. Laser: A scalable response prediction platform for online advertising. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 173–182. ACM, 2014.

[3] E. Bakshy, L. Dworkin, B. Karrer, K. Kashin, B. Letham, A. Murthy, and S. Singh. Ae: A domain-agnostic platform for adaptive experimentation. In *Neural Information Processing Systems Workshop on Systems for Machine Learning*, pages 1–8, 2018.

[4] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

[5] K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

[6] H. Chernoff. Sequential design of experiments. *Annals of Mathematical Statistics*, 30(3):755–770, 1959.

[7] T. Desautels, A. Krause, and J. W. Burdick. Parallelizing exploration-exploitation tradeoffs in gaussian process bandit optimization. *Journal of Machine Learning Research*, 15:3873–3923, 2014.

[8] E. Even-Dar, S. Mannor, and Y. Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7:1079–1105, 2006.

[9] A. Foster, D. R. Ivanova, I. Malik, and T. Rainforth. Deep adaptive design: Amortizing sequential bayesian experimental design. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[10] P. I. Frazier and W. B. Powell. Paradoxes in learning and the marginal value of information. *Decision Analysis*, 7(4):378–403, 2010.

[11] D. Ginsbourger, R. L. Riche, and L. Carraro. A multi-points criterion for deterministic parallel global optimization based on gaussian processes. Technical Report HAL-00260579, 2008.

[12] J. Gonzalez, M. Osborne, and N. Lawrence. Glasses: Relieving the myopia of bayesian optimisation. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, 2016.

[13] S. Jiang, H. Chai, J. González, and R. Garnett. Binoculars for efficient, nonmyopic sequential experimental design. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[14] S. Jiang, D. Jiang, M. Balandat, B. Karrer, J. Gardner, and R. Garnett. Efficient nonmyopic bayesian optimization via one-shot multi-step trees. In *Advances in Neural Information Processing Systems 20*, 2020.

[15] K. Kandasamy, A. Krishnamurthy, J. Schneider, and B. Poczos. Parallelised bayesian optimisation via thompson sampling. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, volume 84, pages 133–142, 2018.

[16] T. L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

[17] R. R. Lam, K. E. Willcox, and D. H. Wolpert. Bayesian optimization with a finite budget: an approximate dynamic programming approach. In *Advances in Neural Information Processing Systems 16*, 2016.

[18] T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge, 2019.

[19] A. Liang, X. Mu, and V. Syrgkanis. Dynamically aggregating diverse information. *Econometrica*, 90(1): 47–80, 2022.

[20] J. Močkus. On bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference Novosibirsk, July 1–7, 1974*, pages 400–404, Berlin, Heidelberg, 1975. Springer Berlin Heidelberg.

[21] M. I. Müller, P. E. Valenzuela, A. Proutiere, and C. R. Rojas. A stochastic multi-armed bandit approach to nonparametric h∞-norm estimation. In *56th IEEE Conference on Decisions and Control*, pages 4632–4637. IEEE, 2017.

[22] H. Namkoong, S. Daulton, and E. Bakshy. Distilled thompson sampling: Practical and efficient thompson sampling via imitation learning. *arXiv:2011.14266 [cs.LG]*, 2020.
[23] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin American Mathematical Society*, 55:527–535, 1952.
[24] D. Russo. Simple bayesian algorithms for best-arm identification. *Operations Research*, 68(6):1625–1647, 2020.
[25] D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, and Z. Wen. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.
[26] E. G. Ryan, C. C. Drovandi, J. M. McGree, and A. N. Pettitt. A review of modern computational algorithms for bayesian optimal design. *International Statistics Review*, 84(1):128–154, 2016.
[27] E. M. Schwartz, E. T. Bradlow, and P. S. Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522, 2017.
[28] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems 28*, pages 2503–2511, 2015.
[29] A. Slivkins. Introduction to multi-armed bandits. *arXiv:1904.07272 [cs.LG]*, 2019.
[30] W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.
[31] N. Wernerfelt, A. Tuchman, B. Shapiro, and R. Moakler. Estimating the value of offsite data to advertisers on meta. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, 1(114), 2022.
[32] J. Wu and P. I. Frazier. Practical two-step look-ahead bayesian optimization. In *Advances in Neural Information Processing Systems 19*, 2019.