Adaptive Experimental Design and Counterfactual Inference

TANNER FIEZ, SERGIO GAMEZ, ARICK CHEN, HOUSSAM NASSIF, and LALIT JAIN*, Amazon

Adaptive experimental design methods are increasingly being used in industry as a tool to boost testing throughput or reduce experimentation cost relative to traditional A/B/N testing methods. This paper shares lessons learned regarding the challenges and pitfalls of naively using adaptive experimentation systems in industrial settings where non-stationarity is prevalent, while also providing perspectives on the proper objectives and system specifications in these settings. We developed an adaptive experimental design framework for counterfactual inference based on these experiences, and tested it in a commercial environment.

ACM Reference Format:
Tanner Fiez, Sergio Gamez, Arick Chen, Houssam Nassif, and Lalit Jain. 2022. Adaptive Experimental Design and Counterfactual Inference. In . ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

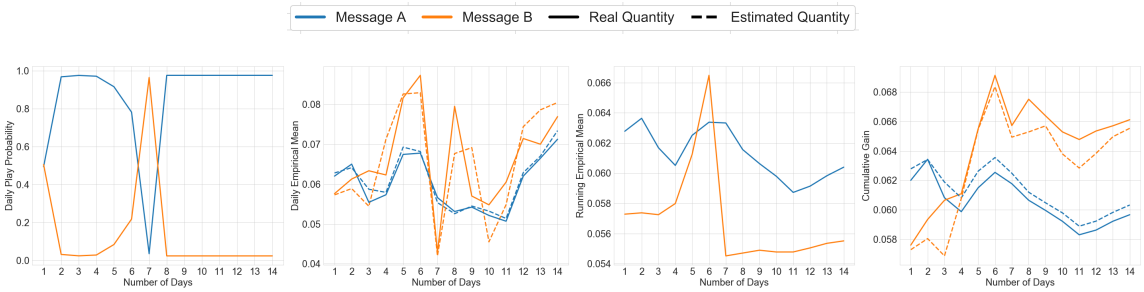
A/B/N testing is a classic and ubiquitous form of experimentation that has a proven track record of driving key performance indicators within industry [11]. Yet, experimenters are steadily shifting toward *Adaptive Experimental Design* (AED) methods with the goal of increasing testing throughput or reducing the cost of experimentation. AED promises to use a fraction of the impressions that traditional A/B/N tests require to yield high confidence inferences or to directly drive business impact. In this paper, we share lessons learned regarding the challenges and pitfalls of naively using adaptive experimentation systems in industrial settings where non-stationarity is the norm rather than the exception. Moreover, we provide perspectives on the proper objectives and system specifications in these settings. This culminates in a high level presentation of an AED framework for counterfactual inference. To provide a robust and flexible tool for experimenters with performance certificates at minimal cost, our methodology combines cumulative gain estimators, always-valid confidence intervals, and an elimination algorithm.

2 A CASE STUDY

Imagine a setting where on a retailer web page, a marketer has been running a message A for the last year and now wants to test whether message B beats A. At the start of the experiment the messages are initialized with a default prior distribution, and then at each round a Thompson sampling bandit dynamically allocates traffic to each treatment, playing each message according to the posterior probability of its mean being the highest [15]. After day 8, the algorithm directs most traffic to message A (see Figure 1). On day 14, the experimenter needs to decide whether A has actually beaten B. They conduct a paired t-test which, somewhat surprisingly, does not produce a significant p-value. As the bandit shifted all traffic to message A, not enough traffic was directed to message B, diminishing the power of the test. The experimenter is forced to conclude that they can not reject the null hypothesis that there is no difference between

*Also with University of Washington.

Manuscript submitted to ACM



(a) Daily play probability (b) Daily mean (c) Running empirical mean (d) Cumulative gain

Fig. 1. Case study of time-variation and adaptive allocations causing Simpson's paradox.

the messages. A few days later, the experimenter, who is still perplexed, looks at the daily means and is then shocked to see that on most days, B tends to have a higher empirical mean than A, which disagrees with the bandit's beliefs.

To understand this behavior, note that in Figure 1c the cumulative success rate of A is exceeding that of B, leading the algorithm to put all its traffic on A. This phenomenon where the total success rate shows a different direction than daily comparisons is referred to as *Simpson's Paradox*, and occurs in settings where the traffic is dynamically allocated to arms whose means change over time [11]. At an intuitive level, the experimenter has perhaps made a Type I error by trusting the algorithm and choosing arm A. Indeed, during the time period from day 8 to day 14, the algorithm decided to put more traffic on arm A, exacerbating Simpson's paradox. Convinced by it's own bad decision, the algorithm then chooses a bad traffic allocation which further exacerbates the problem and leads to a vicious cycle.

3 LESSONS LEARNED

The case study from Section 2, though simple, demonstrates many of the challenges and pitfalls of naively using adaptive experimentation systems in industrial settings where time variation is the norm rather than the exception. We now dive deeper into some of these concerns and also provide thoughts on objectives and specifications in these settings.

Regret Minimization Isn't Enough. The fundamental goal of experimentation is to test hypothesis and deliver results that allow for *future iterations* [3]. As a result, it is important that experimentation procedures give the experimenter the ability to arrive at valid and measurable inferences. In settings where the experimenter wants to learn the best treatment, optimal regret minimization procedures (in stochastic settings) tend to have much lower power (as they are far from a balanced allocation) and can take a significantly longer time to return the identity of the best arm with high probability [2, 4]. In addition regret minimization procedures lead to biased estimates of empirical means [16].

Be Wary of the Batch. Most experimental systems use batched model updates (daily or weekly). In the example from Section 2, the traffic was not constant daily (not shown), so an update on one day can have an undue impact on the rest of the experiment time. In experiments over short horizons, this implies that observations on the first few days can have a disproportionate impact on the traffic allocation and also the subsequent inferences that are made.

Identify the Counterfactual Best. Naively assuming stationarity and using empirical mean estimates can lead to faulty inferences. In general, regret minimization is a well defined objective, however it may fail to deliver the inference that the marketer is intuitively looking for, that is, the identity of the best-arm. Furthermore, in settings where arm means are shifting over time, it is challenging to define this notion as the mean performance of an arm and the identity

of the best arm may change daily. To bridge this gap, our proposed objective is to *identify with high probability the treatment that would have obtained the highest possible reward, if all traffic had been diverted to it.* This counterfactual metric is known as the *cumulative gain*. Figure 1d demonstrates the cumulative gain over time for the case study. With the exception of [1], we believe that this objective has hardly been considered in the best-arm identification literature.

Always Valid Inference. In traditional A/B testing, the experiment horizon is fixed ahead of time (generally based on a minimum detectable effect size the experimenter is interested in detecting) and then a test for significance is used at the end of the experiment. Monitoring, or stopping the experiment, based on p-values computed during the experiment (a process known as *p-hacking*) is heavily frowned upon as it leads to Type 1 error inflation [9]. Recent work in the experimental space that builds upon ideas of Robbins [14], has lead to generalizations of the p-value known as *always-valid p-values* which are slightly inflated and can safely be sequentially monitored [6, 7, 9]. This capability is critically important in practice to allow for early stopping and to ensure valid inferences are drawn.

The Best of Three Worlds. Though optimal regret minimization procedures fail to provide valid inferences and tend to identify the best arm more slowly, we still would like to minimize the opportunity cost of experimentation. Thus experimentation systems should try to provide the best of three worlds: identification of the counterfactual best, mitigation of opportunity cost, robustness to arbitrary time variation. In completely adversarial settings we can't hope to have all three [1], but real life settings mostly live somewhere between fully stochastic and fully adversarial.

4 ROBUST AND ADAPTIVE COUNTERFACTUAL INFERENCE

We now present an AED methodology for counterfactual inference that is simple, mitigates many of the concerns raised regarding non-stationarity, and adheres to the preferred objectives and system specifications.

Experimentation Setting. We consider an experimentation setting over $k \geq 2$ arms with potentially non-stationary daily success rates and batch feedback. On any day $t \geq 1$, each customer that arrives is shown an arm $i \in [k]$ with probability $p_{i,t}$, where the sampling distribution $p_t = (p_{1,t}, \cdots, p_{k,t}) \in \Delta^k = \{p_t \in \mathbb{R}^k : p_{i,t} \geq 0 \, \forall i \in [k], \sum_{i=1}^k p_{i,t} = 1\}$ is chosen by an algorithm dependent on the observations prior to day t. Denote by $n_{i,t}$ the number of impressions for arm $i \in [k]$ on day t and the total day traffic by $n_t = \sum_{i \in [k]} n_{i,t}$. Let $r_{i,t}$ and $\widehat{\mu}_{i,t} := r_{i,t}/n_{i,t}$ denote the total reward and empirical mean on day t for any arm $i \in [k]$, respectively. We further assume that the mean of an arm $i \in [k]$ on any day $t \geq 1$ is fixed over the day and denote it by $\mu_{i,t} \in [0, 1]$. Thus, conditional on the allocation $n_{i,t}$, $r_{i,t} \sim \mathrm{B}(n_{i,t}, \mu_{i,t})$.

Estimation: Cumulative Gain. As discussed previously, assessing the performance of arms using empirical mean estimates can lead to faulty inferences in the presence of dynamic traffic allocations and non-stationarity due to biases and the ill-defined nature of the underlying quantity being estimated. To overcome this challenge, we consider the counterfactual metric known as the cumulative gain that measures the performance of an arm had it received all of the impressions. Formally, for any arm $i \in [k]$, the *cumulative gain* after $t \geq 1$ days is $G_{i,t} := \sum_{s=1}^t n_s \mu_{i,s}$. We can build an estimator for this metric using inverse probability weighting. Indeed, an unbiased cumulative gain estimator is $\widehat{G}_{i,t} = \sum_{s=1}^t r_{i,s}/p_{i,s}$. Unlike an empirical mean, the cumulative gain estimator will never suffer from *Simpson's paradox*.

Algorithm: Elimination on Cumulative Gain via Sequential Monitoring. Given the cumulative gain metric, our objective is to identify the arm with the maximum cumulative gain with high probability while also minimizing regret. Motivated by the experimentation setting, metric and objective, we have arrived at the elimination style algorithm presented in Algorithm 1 as a procedure for AED and counterfactual inference. As input, Algorithm 1 takes a set of arms $[k]$, a confidence parameter δ (normally set to 0.1), and a "settling period" τ. On each day, an active set of arms $\mathcal{A}$ is maintained and each is shown with equal probability $1/|\mathcal{A}|$. At the end of each day, after the settling period, it removes any arms that it can verify are statistically worse than an existing arm based on the cumulative gain metric.

Algorithm 1 Successive Elimination on Cumulative Gain

Input Arm set $[k]$, Confidence $\delta \in (0,1)$, Settling Time $\tau \geq 1$
Initialize $\mathcal{A} \leftarrow [k], t \leftarrow 1$
while $|\mathcal{A}| > 1$ **do**
  On day t, set $p_{i,t} = 1/|\mathcal{A}|$ for all $i \in \mathcal{A}$ and for each customer $s \leq n_t$, show arm i with probability $p_{i,t}$
  if $t \geq \tau$ **then**
    $\mathcal{A} \leftarrow \mathcal{A} \setminus \{j \in \mathcal{A} : \exists i \in \mathcal{A}, \widehat{G}_{i,t} - \widehat{G}_{j,t} - \phi(i,j,t,\delta/k) > 0\}$
  $t \leftarrow t+1$
Return $\mathcal{A}$

Specifically, Algorithm 1 eliminates using an always-valid confidence interval [6, 8, 9]. That is, for each pair of arms $i, j \in [k]$ an always-valid confidence interval guarantees that

$$\mathbb{P}(\exists t \geq 1, i, j \in [k] : |(\widehat{G}_{i,t} - \widehat{G}_{j,t}) - (G_{i,t} - G_{j,t})| \geq \phi(i,j,t,\delta)) \leq \delta.$$

If each n_t is sufficiently large and each arm receives enough traffic, we can invoke the CLT and as an approximation employ the mixture sequential probability ratio test (MSPRT) [9, 14] to define the always-valid confidence interval

$$\phi(i,j,t,\delta) := \sqrt{(V_t(i,j)+\rho)\log\left((V_t(i,j)+\rho)/(\rho\delta^2)\right)} \quad \text{where} \quad V_t(i,j) = \textstyle\sum_{s=1}^{t} n_s\left(\widehat{\mu}_{i,s}(1-\widehat{\mu}_{i,s})/p_{i,s} + \widehat{\mu}_{j,s}(1-\widehat{\mu}_{j,s})/p_{j,s}\right)$$

and $\rho > 0$ is a fixed constant. Motivated by the existence of the always valid confidence interval, Algorithm 1 eliminates an arm $j \in [k]$ on some day $t \geq 1$ when there exists an arm $i \in [k]$ such that $\widehat{G}_{i,t} - \widehat{G}_{j,t} - \phi(i,j,t,\delta) > 0$.

AED System Guarantees. In the stochastic stationary setting or the constant gap setting, Algorithm 1 reduces to a version of the successive elimination algorithm [5]. In this case we have a guarantee that the best arm will be returned with probability greater than $1-\delta$ in a number of samples not exceeding $O(\log(k/\delta)\sum_{i=1}^{k}\Delta_i^{-2})$ and with an instance-dependent regret of no more than $O(\log(k/\delta)\sum_{i=1}^{k}\Delta_i^{-1})$, both of which are near-optimal [7, 10, 12]. In the general non-stationary setting it is more difficult to make a strong statement about Algorithm 1's performance, beyond that if an arm is eliminated then there exists an arm with a higher cumulative gain in the active set at that day. Note that the downside of elimination in a time-varying setting is that an arm that is eliminated because it is sub-optimal today, could potentially be the best performing arm in the future. However, from a practical perspective, eliminating arms once they become sub-optimal is an easily interpretable solution that allows practitioners to hone in on a winner. In addition, as the number of arms shrink, the remaining arms acquire more samples, increasing power over time.

5 DISCUSSION & FUTURE DIRECTIONS

In this work, we discuss the challenges of applying AED methods in practice, provide thoughts on objectives and specifications of such systems, and present the approach we have arrived at using AED for counterfactual inference. Our methodology is perhaps most closely linked to the best-of-both-worlds setting and the P1 algorithm [1], where the experimenter in each round plays each arm with some positive probability and declares a winner when there is one arm whose lower confidence bound is greater than the upper confidence bound of each other arm. We take a far more aggressive approach that effectively guarantees that in the stochastic setting we will recover the best possible sample complexity, but give up on strong guarantees in the adversarial setting. Based on our experiences, this appears to be reasonable. This work brings up a number of interesting future work directions including exploring combinatorial settings for multivariate testing, and experimentation dynamics leveraging priors on the probability of launching successful treatments to more effectively balance identification and regret objectives [13].

REFERENCES

[1] Yasin Abbasi-Yadkori, Peter Bartlett, Victor Gabillon, Alan Malek, and Michal Valko. Best of both worlds: Stochastic & adversarial best-arm identification. In *Conference on Learning Theory*, pages 918–949. PMLR, 2018.

[2] Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *COLT*, pages 41–53, 2010.

[3] Ari Biswas, Thai T Pham, Michael Vogelsong, Benjamin Snyder, and Houssam Nassif. Seeker: Real-time interactive search. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 2867–2875, 2019.

[4] Rémy Degenne, Thomas Nedelec, Clement Calauzenes, and Vianney Perchet. Bridging the gap between regret minimization and best arm identification, with application to a/b tests. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1988–1996, 2019.

[5] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(Jun):1079–1105, 2006.

[6] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

[7] Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439, 2014.

[8] Kevin G Jamieson and Lalit Jain. A bandit approach to sequential experimental design with false discovery control. *Advances in Neural Information Processing Systems*, 31:3660–3670, 2018.

[9] Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. Peeking at a/b tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1517–1525, 2017.

[10] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016.

[11] Ron Kohavi, Diane Tang, and Ya Xu. *Trustworthy online controlled experiments: A practical guide to a/b testing*. Cambridge University Press, 2020.

[12] Zhaoqi Li, Lillian Ratliff, Houssam Nassif, Kevin Jamieson, and Lalit Jain. Instance-optimal pac algorithms for contextual bandits. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[13] Sareh Nabi, Houssam Nassif, Joseph Hong, Hamed Mamani, and Guido Imbens. Bayesian meta-prior learning using Empirical Bayes. *Management Science*, 68(3):1737–1755, 2022.

[14] Herbert Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970.

[15] Neela Sawant, Chitti Babu Namballa, Narayanan Sadagopan, and Houssam Nassif. Contextual multi-armed bandits for causal marketing. In *Workshops of International Conference on Machine Learning (ICML)*, 2018.

[16] Jaehyeok Shin, Aaditya Ramdas, and Alessandro Rinaldo. On the bias, risk and consistency of sample means in multi-armed bandits. *arXiv preprint arXiv:1902.00746*, 2019.