Best of Three Worlds: Adaptive Experimentation for Digital Marketing in Practice

Tanner Fiez
Amazon
Seattle, WA, USA

Houssam Nassif
Meta
Seattle, WA, USA

Lalit Jain
Amazon
Seattle, WA, USA

1 INTRODUCTION

A/B/N testing is a classic form of experimentation that has a proven track record within industry [8]. 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 precise and correct inference or to directly drive business impact. This paper shares lessons learned regarding the challenges and pitfalls of naively using AED in industrial settings. Moreover, we provide perspectives on the proper objectives and system specifications in these settings. This culminates in a presentation of an AED framework which simultaneously guarantees fast counterfactual inference while minimizing experimentation cost by combining cumulative gain estimators, always-valid confidence intervals, and an elimination algorithm.

2 REAL-WORLD PITFALLS AND LESSONS

We now present an experimentation case study and then generalize to discuss lessons learned using AED systems and viewpoints on the proper system specifications in industry.

2.1 Case Study: Adaptive Designs & Inference

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. Fearful of incurring a large amount of loss from A/B testing opportunity cost, the marketer chooses to use an AED method, namely Thompson Sampling (TS). At the start of the experiment, the messages are initialized with a default prior distribution, and then at each round the bandit dynamically allocates traffic to each treatment, playing each message according to the posterior probability of its mean being the highest. 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 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 that lead to the traffic allocation it produced.

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

2.2 Lessons Learned & System Specifications

The previous simple case study demonstrates many of the challenges and pitfalls of naively using AED systems in industrial settings with non-stationarity. We now dive deeper into such concerns and provide thoughts on industrial objectives and specifications.

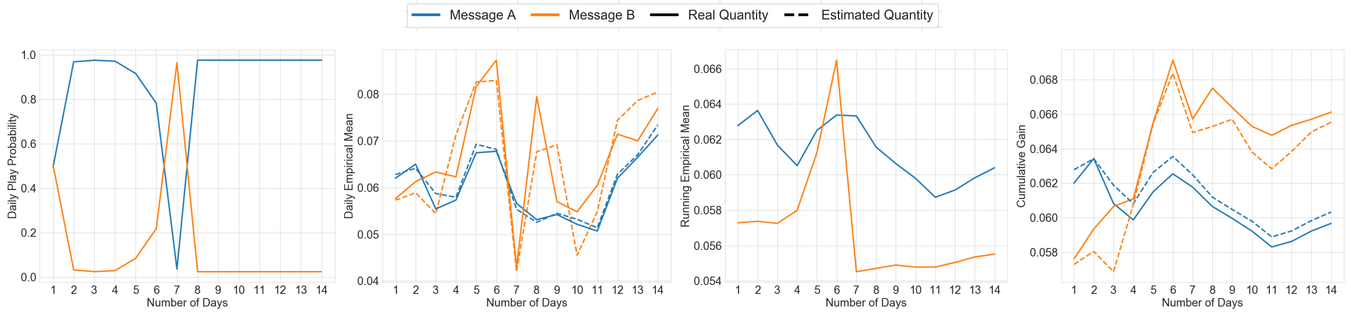
Regret Minimization Isn't Enough. The fundamental goal of experimentation is to test hypothesis and deliver results that allow for *future iterations* [2]. As a result, it is important that experimentation procedures give the experimenter the ability to arrive at valid and measurable inferences. In settings where the experimenter wants to learn the best treatment, optimal regret minimization procedures take a significantly longer time to return the identity of the best arm with high probability [3] and lead to biased mean estimates.

Stochastic Bandit Algorithms Often Fail. While it is common in industrial systems to deploy regret-minimizing algorithms based on underlying stationarity assumptions, these algorithms fail with regularity in any given experiment even for the sole purpose of accruing an optimization metric. Often these failures go unnoticed due to the absence of a suitable comparison. We highlight in our experiments that stochastic bandit algorithms can fail to maximize the accumulation of an optimization metric as a result of dynamic traffic allocations in combination with an estimation based on the observed adaptively collected data in time-varying environments.

Identify the Counterfactual Best. In settings where arm means are shifting over time, it is challenging to define the notion of a "best-arm" 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/N testing, the experiment horizon is fixed ahead of time, with a significance test at the end of the experiment. Monitoring p-values during the experiment is heavily frowned upon as it leads to Type 1 error inflation [7]. Work in the experimental space has lead to generalizations of the p-value known as *always-valid p-values* that can safely be sequentially monitored [6, 7]. This capability is critical in practice.

The Best of Three Worlds (BOTW). 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, and robustness to arbitrary time variation. In completely adversarial settings



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

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

we can't hope to have all three [1], but real life settings mostly live somewhere between fully stochastic and fully adversarial.

3 OUR APPROACH

We now sketch our methodology motivated by the above discussion. At a high level, our approach consists of a) accurate estimation in a time varying setting, b) anytime inference, c) an elimination procedure. We demonstrate the effectiveness of our method both for identification and regret minimization in several natural settings both in theory and practice.

3.1 Experimentation Setting

We consider a setting with k arms running for T days beginning on day $t = 1$. On day $t \in [T]$, arm $i \in [k]$ receives $n_{i,t}$ impressions and $n_t = \sum_{i \in [k]} n_{i,t}$ is the total amount of traffic on that day. We assume that the underlying reward distribution of an arm $i \in [k]$ on day $t \in [T]$ is fixed over the period of a day (but not over the experiment) and given by a Bernoulli distribution with mean $\mu_{i,t} \in (0, 1)$. Finally, we let $r_{i,t}$ and $\widehat{\mu}_{i,t} := r_{i,t}/n_{i,t}$ denote the total reward (success) and empirical mean on day $t \in [T]$ for any arm $i \in [k]$, respectively.

3.2 Estimation with Time Variation

We now discuss estimation and inference with time-variation.

3.2.1 *Empirical Means.*

The (running) empirical mean of arm $i \in [k]$ after T days of an experiment is given by $\widehat{\mu}_i := (\sum_{t=1}^{T} r_{i,t})/(\sum_{t=1}^{T} n_{i,t})$. Given the standard assumptions of fixed horizon A/B/N testing, the empirical mean is an unbiased estimator of the underlying performance. However, when the underlying performance of an arm exhibits daily time-variation, the mean performance is not a well-defined metric and the empirical mean can be a problematic estimator. Specifically, the estimate $\widehat{\mu}_i$ is subject to *Simpson's Paradox* [8]. In the context of experimentation, Simpson's paradox refers to a circumstance in which the daily empirical mean of an arm $i \in [k]$ is higher than that of an arm $j \in [k]$ on each given day ($\widehat{\mu}_{i,t} > \widehat{\mu}_{j,t}\ \forall\ t \in [T]$), *but* the empirical mean of arm j is higher than that of arm i over the course of an experiment ($\widehat{\mu}_j > \widehat{\mu}_i$). As we saw in Case Study 1 (Figure 1c), in experimentation where the traffic allocation is changing over time, this paradox often arises.

3.2.2 *Cumulative Gain.*

As the above discussion implies, the empirical mean estimator has many negative characteristics that make it inappropriate for time-varying settings with adaptive traffic allocation. Part of the challenge is that in time-varying settings the notion of "the best performing arm" may be poorly defined since the best-arm may change from day-to-day. To overcome this, we instead try to answer the following counterfactual: "*how much reward would this arm have accrued if it had received all of the traffic*". For any arm $i \in [k]$, the cumulative gain (CG) after T days is defined as

$$G_{i,T} := \textstyle\sum_{t=1}^{T} n_t \mu_{i,t}. \qquad (1)$$

The corresponding cumulative gain rate variant is $\overline{G}_{i,T} :== G_{i,T}/\bar{n}_T$ where $\bar{n}_T := \sum_{t=1}^{T} n_t$ is the total experiment traffic.

Cumulative Gain Estimator. Assume that on each day $t \in [T]$ of the experiment, a probability vector $p_t = (p_{1,t}, \cdots, p_{k,t}) \in \Delta_k$ is selected and each visitor $s_t \in [n_t]$ on day $t \in [T]$ is shown an arm $I_{s_t} \in [k]$ that is selected with probability $\mathbb{P}(I_{s_t} = i) = p_{i,t}$ and a reward r_{s_t} is observed. A natural and unbiased cumulative gain estimator is given by inverse propensity weighing [5]:

$$\widehat{G}_{i,T} = \textstyle\sum_{t=1}^{T} (r_{i,t}/p_{i,t}). \qquad (2)$$

The cumulative gain estimator will never suffer from Simpson's paradox by definition, unlike the empirical mean estimator. As shown in the case study (Fig 1c-1d), using the cumulative gain would have prevented misleading inferences.

3.2.3 *Always-Valid Inference.*

The high risk of error rate inflation from fixed horizon p-values motivates adopting *always-valid confidence intervals* [6, 7] on the cumulative gain gaps between arms as a tool to yield always-valid inferences. In this context, an always-valid confidence interval $C(i, j, t, \delta)$ for a pair of arms $i, j \in [k]$ with error tolerance $\delta \in (0, 1)$ guarantees

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

To obtain the always-valid confidence interval, we apply the MSPRT[1] using the plugin estimators $\widehat{\mu}_{i,t}$ and $\widehat{\mu}_{j,t}$ for the unknown arm means $\mu_{i,t}$ and $\mu_{j,t}$ on each day in an estimate of the variance to get

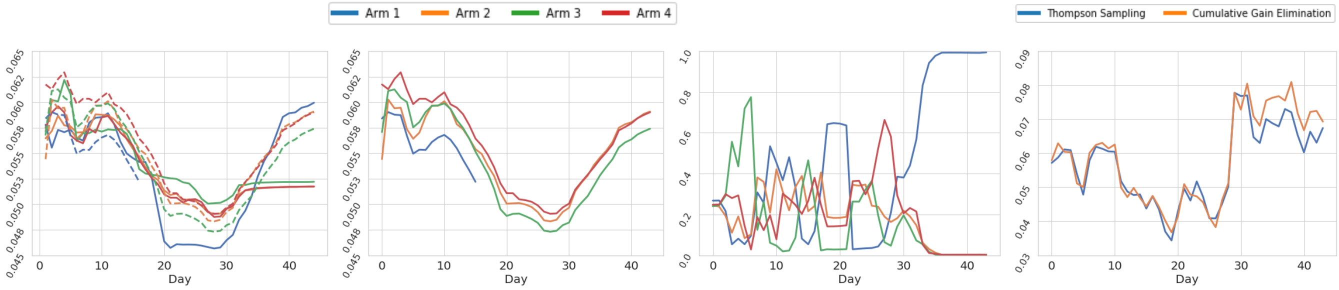
$$\mathbb{P}(\exists\ t \geq 1, i, j \in [k] : |\widehat{G}_{i,j,t} - G_{i,j,t}| \geq C(i, j, t, \delta)) \leq \delta,$$

$$\text{with} \quad C(i, j, t, \delta) := \sqrt{(\hat{V}_{i,j,t} + \rho) \log((\hat{V}_{i,j,t} + \rho)/(\rho\delta^2))} \qquad (3)$$

where $\rho > 0$ is a fixed constant and

$$\hat{V}_{i,j,t} = \textstyle\sum_{\tau=1}^{t} n_\tau \big(\widehat{\mu}_{i,\tau}(1 - \widehat{\mu}_{i,\tau})/p_{i,\tau} + \widehat{\mu}_{j,\tau}(1 - \widehat{\mu}_{j,\tau})/p_{j,\tau}\big).$$

[1] For reference, see Eq. 14 in [6].



(a) TS & CGSE empirical means

(b) CGSE cumulative gain rates

(c) TS daily probability allocation

(d) Algorithm daily empirical mean

Figure 2: Experiment: Thompson sampling catastrophically fails on production data and shifts all traffic to the worst arm.

Algorithm 1 Cumulative Gain Successive Elimination (CGSE)

```
Input Arm set $[k]$, error tolerance $\delta \in (0, 1)$
Initialize Active arm set $\mathcal{A} \longleftarrow [k]$, day $t \longleftarrow 1$
while $|\mathcal{A}| > 1$ do
    Set $p_{i,t} = 1/|\mathcal{A}|$ for all $i \in \mathcal{A}$ and $p_{i,t} = 0$ for all $i \in [k] \setminus \mathcal{A}$
    For each arrival $s_t \in [n_t]$ show arm $I_{s_t} \sim p_t$
    Collect observations $\{r_{s_t}, I_{s_t}, p_t\}_{s=1}^{n_t}$
    $\mathcal{A} \longleftarrow \mathcal{A} \setminus \{j \in \mathcal{A} \text{ s.t. } \exists\, i \in \mathcal{A} : \widehat{G}_{i,j,t} - C(i, j, t, \delta/k) > 0\}$
    $t \longleftarrow t + 1$
Return $\mathcal{A}$
```

3.3 Adaptive Counterfactual Inference

We present an AED method that gives sample efficient identification of the counterfactual optimal treatment, while simultaneously minimizing regret in experiments where stationarity is not guaranteed.

3.3.1 Algorithm Description. Algorithm 1 (CGSE) is an elimination based method on the cumulative gain. On each day, an active set of arms $\mathcal{A}$ is maintained and each is shown to users with equal probability $1/|\mathcal{A}|$. At the conclusion of each day, any arms that can be concluded to not have the maximum cumulative gain up through the current day among the active set based on the always-valid confidence interval are removed and never sampled again. This procedure controls the cumulative gain estimator variance and sample complexity by keeping the sampling probabilities uniform across the set of active arms, while the regret is controlled by ceasing to give any traffic to provably sub-optimal arms.

3.3.2 Guarantees for Correct Inference. Define the cumulative gain rate gap relative to an arm $i^* \in [k]$ for any arm $j \in [k]$ at any day $t \geq 1$ as $\Delta_{j,t} := \overline{G}_{i^*,t} - \overline{G}_{j,t}$. The assumption below implies that the counterfactual optimal arm i^* is time-independent. This is a mild restriction that allows for daily-time variation among all arms and does not require that i^* always has the highest daily mean.

Assumption 1. *There exists an arm i^* such that for each arm $j \in [k] \setminus \{i^*\}$ the cumulative gain rate gap $\Delta_{j,t} > 0$ for all $t \geq 1$.*

This following result gives a strong guarantee in a general time-varying setting for the correctness of CGSE.

Proposition 1. *CGSE with $\delta \in (0, 1)$ returns arm i^* with probability at least $1 - \delta$ under Assumption 1.*

3.3.3 Guarantees in Stochastic Environments. In the stochastic stationary setting, CGSE reduces to a closely related version of the classical successive elimination algorithm [4]. Thus, it obtains the known guarantees for the algorithm in this situation and in settings with constant performance gaps, which include **near-optimality with high probability for both sample complexity and regret**.

3.3.4 Experiments. Fig. 2 shows the results of the experiment from an online production environment. Traffic was split equally between TS and CGSE. After 2 weeks, CGSE eliminates Arm 1 (Fig. 2b), while TS was allocating nearly 100% of the traffic to this arm at the end of the experiment (Fig. 2c). This may appear to be an example where the arm switched from being the worst performer and became the best performer, but we can validate that this is not the case. Consider Fig. 2a, in this plot the solid lines represent the empirical means of each arm as estimated from the TS algorithm, versus the dashed lines with empirical means estimated from CGSE. Since TS is giving little traffic to Arms 2-4 from day 30 onward, we see that there is a huge bias downwards in their empirical means, compared to SE which is uniformly allocating traffic. This suggests that TS's confidence in giving all of its traffic to Arm 1 is misplaced and instead CGSE was wise to eliminate Arm 1 early. Specifically, we see that the performance of all arms moves up as TS begins to switch its allocation to Arm 1 and this reinforcing feedback loop causes continued flawed allocations by TS. This is a real-world example of Simpson's paradox. As Fig. 2d shows, this has an impact on the total successes observed. Indeed the daily empirical means of the total successes observed by each algorithm shows that CGSE has a higher overall success rate and thus minimizes regret more effectively toward the end of the experiment.

REFERENCES

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

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

[3] Rémy Degenne, Thomas Nedelec, Clement Calauzenes, and Vianney Perchet. 2019. 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*. 1988–1996.

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

[5] D. G. Horvitz and D. J. Thompson. 1952. A Generalization of Sampling Without Replacement from a Finite Universe. *J. Amer. Statist. Assoc.* 47, 260 (1952), 663–685.

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

[7] Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. 2017. 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*. 1517–1525.

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