Redefining Real-Time Experimentation: Bayesian Sequential A/B Testing's Role in the Tech Industry

Richie Lee
MSc Thesis: Econometrics & Management Science
Erasmus School of Economics, Rotterdam
505917KL@student.eur.nl

ABSTRACT

With the ever-growing scale of experimentation in Tech, Sequential A/B testing stands out as one of its most vital components, serving the need for fast, real-time causal insights with strong statistical guarantees. This study highlights the comparison of Bayesian A/B testing with prevalent methodologies, its adaptation to tech-specific challenges, and introduces novel guidelines for its implementation in online experimentation environments. Moreover, we demonstrate its outstanding speed in accumulating statistical power, alongside strategies to effectively manage the corresponding trade-offs that most notably include type-I error inflation risks and prior sensitivity.

CCS CONCEPTS

• **Mathematics of computing** → **Bayesian computation**; **Hypothesis testing and confidence interval computation**.

KEYWORDS

Sequential A/B Testing, Real-World Application, Tech Industry, Experimentation, Statistical Power, Early Stopping Methodologies

ACM Reference Format:
Richie Lee. 2024. Redefining Real-Time Experimentation: Bayesian Sequential A/B Testing's Role in the Tech Industry. In *Proceedings of Msc Econometrics & Management Science (MSc Thesis).* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

MSc Thesis, February, 2024,

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

As one of the leading applications of applied statistics in Tech, A/B testing has evolved to become widely recognized as the golden standard for identifying and measuring causal effects, playing a pivotal role in data-driven decision-making [14]. In the online big data domain, traditional fixed horizon A/B tests often rely on large sample sizes to ensure robustness of their findings. However, prolonged experiments come with significant drawbacks, most notably the opportunity costs related to slowing innovation cycles and extending harmful experiments. Practitioners are often falling into malpractices such as *Peeking*, i.e. to continuously monitor results and make ad-hoc decisions prematurely, which thereby compromises statistical validity. To meet evolving needs, industry professionals have rapidly adopted sequential A/B testing, giving rise to nowadays well-established methods such as *Group sequential testing* (Spotify [28], Booking.com [30]) and *Always valid inference* (Uber [5], Netflix [19], Optimizely [12]). The key objectives that these methods emphasise include maximising statistical power, optimising the speed at which this power is achieved, and preserving reliable control over the false discovery rates (type-I errors).

This paper sheds light on an alternative sequential testing solution that has had limited exposure in the world of online experimentation: *Bayesian A/B testing* – a fundamentally different solution for the same problem. The study is guided by two central research questions: (1) In which specific sequential A/B testing use-cases is Bayesian A/B testing favored over current industry-standard solutions?, and (2) When and how should practitioners exercise caution regarding the trade-offs inherent to Bayesian A/B testing, and under what conditions can these concerns safely be dismissed?

The primary objective is to identify and demonstrate practical use-cases for Bayesian A/B testing, with a focus on seizing its sample-efficient power opportunities, while adeptly handle the associated trade-offs. Additionally, desirable properties include probabilistic interpretability [40], futility stopping [31], and validity under optional stopping / unlimited interim testing [6]. A key emphasis is placed on robustifying the methodology, not only

by benchmarking the method against state-of-the-art alternatives, but also by testing the method on real industry data, sourced from globally leading food delivery platform Just Eat Takeaway.com. All in all, with the support of these associated insights, we strive to facilitate easier generalisation and application of its findings through a comprehensive trade-off overview and actionable recommendations as a foundation for potential new Bayesian A/B testing use-cases in the Tech industry and beyond.

Contribution: This study stands out as one of the first evaluations of Bayesian A/B testing that combines real-world industry data with an explicit focus on the method's risks and trade-offs under both early stopping rules and continuous monitoring. Ultimately, this aims to generalise past literature findings to broader applications.

With practicality in mind, as a second novelty, we gather all the insights and knowledge in an template that is designed to lower the entry barrier to Bayesian A/B testing, providing essential tools and information not just for effective implementation, but also for taking an important step back first, to critically assess whether the Bayesian approach makes sense given the user's unique sequential A/B testing needs.

Outline: The paper is structured as follows. Section 2 gives background around sequential A/B testing and an overview of popular solutions in this field. Section 3 introduces the methodology. Our experiments and benchmarks are described in Section 4 and accompanying key findings in Section 5 and summarized in Section 6 – accompanied a flowchart that captures all the accumulated insights to support Bayesian A/B testing implementation and sequential testing method selection.

2 RELATED WORK

Currently in the Tech sector, despite earlier optimistic perspectives on the methodology, such as those by [6] (Microsoft), its industry exposure remains scarce with public paper contributions being limited to recent related work from [36] and [3]. This situation is and has been notably different in other disciplines like clinical trials or psychology, where the demand for "optimal" experimentation is just as strong – striking an interesting, perhaps promising paradox. The work in [1] was one of the first to showcase the relevance of Bayesian A/B testing in clinical trial settings, and ever since, its has retained its presence in the field in applications such as phase II clinical trials for oncology and drug development [18, 21, 32]. These research stages involve human participants and usually have sample sizes between 100 and 300. This is in stark contrast to online experimentation scenarios, which often involve large datasets and easier access to asymptotic results. This underscores a fundamental difference in the resources and needs of the two fields.

The key advantage of sequential testing lies in its ability to balance the trade-off between efficiency and reliability. However, it is vital to acknowledge that unavoidable noise and variance in data can significantly increase the risk of false discovery with each additional examination if not addressed. This inflation of false discovery rate (type-I errors) is commonly referred to as *Peeking* and discussed in [11]. Currently, the Tech industry predominantly recognises Group sequential testing [9, 22, 23] and Always valid inference [12] as its most popular and well-established solutions to interim testing.

2.0.1 Group Sequential Testing. Conceptually, Group Sequential Testing (GST) can be considered a more sophisticated Bonferroni correction that brings similar family-wise type-I error controls while sacrificing less power. This improvement is possible by accounting for the serial correlations that comes with filtration in sequential testing. O'Brien & Fleming [22] introduce an implementation framework for GST that sets a maximum limit ($\mathcal{J}$) on the total number of interim evaluations. Then, using time-varying critical values ($\alpha_{OF,j}$), it allows for efficient and flexible allocation of the error rate across multiple looks at the data. For composite hypothesis testing, $\alpha_{OF,j}$ is defined as follows [17]:

$$\alpha_{OF,j} = 2 \cdot (1 - \Phi(\frac{\Phi^{-1}(1 - \frac{\alpha}{2})}{\sqrt{j/\mathcal{J}}})) \tag{1}$$

To enable early stopping capabilities, we can pair these adjusted critical values with classical p-values to as follows:

$$\text{For each interim test} \begin{cases} p\text{-value} < \alpha_{OF,j} & \text{STOP \& reject } H_0 \\ j = \mathcal{J} & \text{STOP \& accept } H_0 \\ \text{otherwise} & \text{continue sampling} \end{cases} \tag{2}$$

Overall, this embodies a nuanced trade-off between the risk of type-I error inflation due to early and possibly spurious null rejections and the loss of statistical power from overly conservative test thresholds.

2.0.2 Always Valid Inference. The Mixture Sequential Probability Ratio Test (mSPRT) represents a hallmark in the evolution of Always Valid Inference (AVI), a concept that has increasingly drawn interest from the technology sector for its robustness and flexibility in statistical testing. This method, particularly highlighted in the work by [12], stands at the confluence of Bayesian and frequentist statistical paradigms, offering a unique approach to hypothesis testing that rigorously controls the type-I error rate across multiple testing points.

At the core of mSPRT is the use of likelihood ratios, which leverage the martingale property to ensure that the expected value of the test statistic remains constant over time, given past observations. This mSPRT implementation, based on [33], utilizes an integral over all possible parameter values (θ) under both null and alternative hypotheses, weighted by a prior density, to compute the test statistic $\hat{\Lambda}_n$ as follows:

$$\hat{\Lambda}_n = \int_0^\infty \prod_{i=1}^{n} \frac{f(x_i|H_1 : \theta > 0)}{f(x_i|H_0 : \theta = 0)} \pi(\theta) d\theta \tag{3}$$

with likelihoods $f(\cdot)$ and half-normal prior density $\pi(\theta)$. This critical value can subsequently be used to derive "always valid p-values" as follows:

$$p\text{-value}_{AVI,n} = \min\left(1, \frac{1}{\hat{\Lambda}_n}\right) \tag{4}$$

Due to the complex nature of the integral required to compute $\hat{\Lambda}_n$, mSPRT often necessitates numerical methods for its evaluation causing a common reliance on computational techniques to achieve its inferential objectives. However, in return, it achieves optional

stopping capabilities, enabling full flexibility in when to terminate the experiment without statistical validity violations. This flexibility is particularly appealing in sequential analyses where ability to continuously monitor results is as critical as the decisions themselves.

3 METHODOLOGY

We aim to determine a treatment's effect and its direction. In this paper, we focus on a one-sided hypothesis:

$$\mathbf{H_0} : \delta \leq 0 \quad \mathbf{H_1} : \delta > 0 \tag{5}$$

Here, δ represents the difference between the treatment group mean (μ_T) and the control group mean (μ_C), reflecting the treatment's causal impact in a well-conducted randomized controlled trial.

Bayesian inference stands out for its ability to include prior knowledge—such as expert opinions, historical data, or initial assumptions—into the analysis, which, when combined with new empirical evidence, leads to a statistically sound update of beliefs reflected in the posterior probability. This methodology relies on Bayes' theorem to merge prior beliefs and new data, thereby offering a refined measure of certainty or uncertainty about an outcome, encapsulating both pre-existing knowledge and freshly gathered information. In the context of Bayesian hypothesis testing, Bayes' theorem can be rewritten to the following form:

$$\underbrace{\frac{P(H_1 \mid \text{data})}{P(H_0 \mid \text{data})}}_{\text{post odds}} = \underbrace{\frac{P(H_1)}{P(H_0)}}_{\text{prior odds}} \cdot \underbrace{\frac{P(\text{data} \mid H_1)}{P(\text{data} \mid H_0)}}_{\text{Bayes factor}} \tag{6}$$

Yielding clear distinctions between the posterior odds (post odds), prior odds and a data driven component, known as the *Bayes factor* respectively. When dissecting Equation 6, we first consider the prior odds, which express our prior belief about either hypothesis being true (not to be mistaken for the priors that are present inside the Bayes factor). This belief is typically based on e.g. domain knowledge, expert opinions, assumptions, complementary research, etc. When accurate, prior odds can contribute to statistical power with external information which becomes especially attractive when facing sample size limitations due to research circumstances. Throughout this study, we set all prior odds to a default uninformative value of one, which simplifies analyses without loss of generality.

Secondly, the data-driven component in Bayesian hypothesis testing, the Bayes factor that we define as follows:

$$BF_{H1|H0} = \frac{P(data|H_1)}{P(data|H_0)} = \frac{\int_{\theta \in H_1} p(data|\theta)p(\theta|H_1)d\theta}{\int_{\theta \in H_0} p(data|\theta)p(\theta|H_0)d\theta} \tag{7}$$

Bayes factors are ratios of *marginal likelihood* densities, which are integrals that incorporate likelihood densities under all plausible values of the variable of interest θ under its respective hypothesis, where the observed data over this parameter space is weighted by a prior density $P(\theta|H)$.

Once the Bayes factor is obtained, we can use the corresponding posterior odds from which, through its fractional construction, it becomes straightforward to calculate *relative* posterior probabilities $P(H_0|data)$ and $P(H_1|data)$, as shown in [6]. This paper closely follows [36]'s implementation of the Bayesian A/B testing methodology specified in [6], which leverage a normal (conjugate) prior. For the one-sided hypothesis testing setup, we can conveniently obtain the exact Bayes factor analytically using:

$$BF_{H_1|H_0} = \frac{1 - \Phi(-\mu_1'/\sigma_1')}{1 - \Phi(-\mu_1/\sigma_1)} \cdot \frac{\Phi(-\mu_0/\sigma_0)}{\Phi(-\mu_0'/\sigma_0')} \cdot \sqrt{\frac{\sigma_0^2 + \sigma^2}{\sigma_1^2 + \sigma^2}} \cdot \exp\left(-\frac{1}{2}\left(\frac{(\mu_1 - y)^2}{\sigma^2 + \sigma_1^2}\right) + \frac{1}{2}\left(\frac{(\mu_0 - y)^2}{\sigma^2 + \sigma_0^2}\right)\right) \tag{8}$$

With $\Phi(\cdot)$ denoting the normal CDF; observed mean and variance y and σ^2; prior distributions $N(\mu_i, \sigma_i)$; and μ_i' being calculated as $\frac{y\sigma_i^2 + \mu_i\sigma^2}{\sigma_i^2 + \sigma^2}$ with σ_i' as $\sqrt{\frac{\sigma^2\sigma_i^2}{\sigma^2 + \sigma_i^2}}$ for $i \in \{0, 1\}$. In practice, for numerical reasons, converting expression 8 to a logarithmic is more practical. This, and the full derivation of the marginal likelihoods are presented in Appendix B.

Once calculation of the Bayes Factor is concluded, we can use its corresponding posterior probability to conclude H_1 to be the more probable amongst the two hypotheses when $BF_{H_1|H_0} > 1$, and vice versa. In this context, we can also interpret the numerical value to reflect strength of the evidence, where more extreme values, approaching zero or infinity, show stronger certainty towards H_0 and H_1 respectively.

3.1 Early Stopping & Optional Stopping

In sequential A/B testing, we differentiate between two types of early termination protocols, termed Early Stopping (ES) and Optional Stopping (OS).

3.1.1 Early Stopping. To extend the continuous monitoring tests to enable early stopping, we introduce a stopping rules with a hyperparameter, typically denoted by $\mathcal{K}$, that can be interpreted as the minimum degree of certainty that we require from the Bayes factor (or post odds) before committing to accepting or rejecting a hypothesis early. We follow [6], with a symmetrical design with equal stringency for H_0 and H_1. This would yield the following stopping rule:

$$\text{For each interim test:} \begin{cases} BF_{H_1|H_0} > \mathcal{K} & \text{STOP \& Reject } H_0 \\ BF_{H_1|H_0} < \frac{1}{\mathcal{K}} & \text{STOP \& Accept } H_0 \\ \text{otherwise} & \text{continue sampling} \end{cases} \tag{9}$$

Through simple operations, we can improve intuition by linking $\mathcal{K}$ to an interpretable probability using $P(data|H_1) = \frac{\mathcal{K}}{\mathcal{K}+1}$. To enable direct comparison with the traditional frequentist critical value, $\alpha = 0.05$, we choose to set $\mathcal{K} = 19$ corresponding to 95% certainty in our analysis.

In contrast to frequentist methodologies that can only terminate experiments through H_0 rejection, Bayesian A/B testing facilitate *Futility stopping* which contributes to efficiency when H_0 is true [31]. Futility stopping rules do not inflate the type-I error rate; actually, they decrease the type-I error rate. However, this feature may decrease the power through type-II errors instead [40]. Additionally, Bayesian A/B testing supports continuous monitoring and

optional stopping [27], which [35] justifies using the Stopping Rule Principle. This principle implies that statistical inference ought to be independent of the choice of when to terminate data collection. In other words, unlike p-values with binding hypothesis rejection regions, in Bayesian hypothesis testing, we are offered flexibility in our conclusions where the integrity of Bayes factors is preserved, irrespective of the decisions and actions that follow it. This also handily removes the need to estimate the sample size or number of planned peeks a priori, as we would for GST.

3.1.2 Optional Stopping. In optional stopping, the experiment is not automatically concluded upon reaching a critical decision threshold of $\frac{1}{\mathcal{K}}$ or $\mathcal{K}$. Instead, termination can occur based on an optional trigger. Overall, assuming prior odds of 1 without loss of generality, we can denote the desired type-I control at evaluation in the Bayesian A/B testing context as:

$$\epsilon_{\tau,OS} = P(H_1|H_0, post\ odds) \leq \frac{1}{\mathcal{K}+1} := \alpha \tag{10}$$

Where $\epsilon_{\tau,OS}$ denoting the type-I error under optional stopping that we would incur at evaluation τ, and α the Frequentist interpretation of the corresponding desired type-I error control. The overall error rate is determined by assuming the worst-case scenario $\arg\max_{\tau}(\epsilon_{\tau,OS})$. Alternatively, under early stopping rules we assess its risks through a family-wise type-I error using

$$\begin{aligned}\epsilon_{\tau,ES} &= P(H_1|H_0, Post\ odds, \{BF_1, \cdots, BF_{\tau-1}\} \in \left(\frac{1}{\mathcal{K}}, \mathcal{K}\right)) \\ &= P(H_1|H_0, Post\ odds, t=\tau) \leq \frac{1}{\mathcal{K}+1} := \alpha \end{aligned} \tag{11}$$

Note that unlike in Frequentist hypothesis testing where $\epsilon_{t,ES} \leq \arg\max_{\tau}(\epsilon_{\tau,OS})$, in Bayesian A/B testing, analytically this does not necessarily have to be the case, due to its ability to futility stop. Therefore, when deciding between OS or ES Bayesian testing, it is important to be mindful of the practical context that determines whether family-wise or regular type-I error rate is more relevant.

3.2 Type-I error control

In contrast to frequentist sequential testing methods with widely accepted views on the risks and challenges type-I error inflation brings, the Bayesian A/B testing community have yet to reach a common agreement on this issue, which has led to an ongoing controversy in the field. This subsection will focus on insights from past simulation studies and the widely acknowledged viewpoint on peeking validity from [6].

Through Monte Carlo simulations, results vary from the positive, as in [8] and [25], who assert that "peeking is no issue at all" (Peeking-immunity), to studies that reject this statements by pointing out significant risks for varying reasons. Examples of these risks include significant estimation biases [16], dangers of confirmation bias [26], and a higher likelihood of incorrect research implementation [39].

The work of [6] identifies a key misunderstanding on Bayesian Sequential Testing as a possible cause for the differing views. Specifically, it clarifies two limitations of Bayesian tests: (1) they do not guarantee type-I error bounds, and (2) they cannot ensure the unbiasedness seen in frequentist maximum likelihood estimation under Bayesian optional stopping. However, in [6], both analytical derivations and simulation results demonstrate that Bayesian tests can effectively accommodate continuous monitoring, when adhering to the following stopping rules:

Theorem 3.1 (Appropriate stopping rules, as per [6]). *optional stopping is valid if (1) the early stopping criteria solely depend on historical, non-counterfactual data and (2) does not cherry-pick or exclude any observations during estimations.*

3.2.1 Prior influence. When considering the second key challenge in Bayesian A/B testing, prior specifications, it's essential to differentiate between two distinct elements: (1) the prior odds, and (2) the priors that influence marginal likelihoods, which subsequently contribute to the Bayes factor. Each of these operates independently, while carrying its own form of influence and corresponding interpretation.

Firstly, the prior odds, defined as $\frac{P(H1)}{P(H0)}$. These priors represent our initial belief about a hypothesis being true or not, which ultimately boils down to a manual bias of some capacity towards either H_0 or H_1. This inference method is typically regarded as more subjective and are most influential in smaller samples, when the data-driven evidence, contained in the Bayes factors, is still more conservative. Asymptotically, as the Bayes factors converge to zero or diverge to infinity, relative influence of the prior odds reduces to none.

Secondly, to strike balance in both type-I and II error control, the two priors $p(\theta|H_0)$ and $p(\theta|H_1)$ located in the Bayes factor's marginal likelihoods should be specified with the aim to fit the treatment effects distributions, under their respective hypothesis, as closely as possible.

Recall that the marginal likelihoods where they reside are derived from observed data, averaged over a parameter space that encompasses all plausible values, where this averaging process is weighted according to prior beliefs – i.e. $P(y|H_i) = \int_{\theta \in H_i} p(y|\theta)p(\theta|H_i)d\theta$. With this in mind, to understand it's eventual effect on the overall inference, consider the Bayes Factor defined as $BF_{H_1|H_0} = \frac{P(data|H_1)}{P(data|H_0)}$. Through its fractional form, it allows both the numerator and denominator to affect the test's sensitivity towards either H_0 or H_1. It does so by each prior lowering its respective marginal likelihood value through prior weighting ($\int_{\theta \in H_i} p(y|\theta)p(\theta|H_i) \leq \int_{\theta \in H_i} p(y|\theta)$).

Under severe prior misspecification, as one marginal likelihood approaches 0, it dramatically amplifies the Bayes factor, swiftly propelling the value either towards 0 (favoring H_0) or towards infinity (rejecting H_0). This creates an asymmetry where the numerator $P(data|H_1)$ and the denominator $P(data|H_0)$ assume different roles in controlling power (type-II error) and type-I error respectively, as detailed in Table 1.

Table 1: Marginal likelihood prior misspecification risks for Bayesian A/B test performance

Ground truth	H_0: $\delta \leq 0$	H_1: $\delta > 0$
$P(\theta\|H_0)$ misspecified	Type-I error ↑	
$P(\theta\|H_1)$ misspecified		Power ↓

4 EXPERIMENTS

The main objective of our simulation study is to assess the efficacy of Bayesian A/B testing in conditions that mirror real-world applications. The success of experiments in our case studies is measured by two main factors: the quality of the evidence and the sample efficiency of the experiment. To assess quality, we rely on classification accuracy metrics such as type-I error and empirical statistical power. These metrics are derived from Monte Carlo simulations, where each iteration intends to mimic a sequential A/B test with early stopping. We initiate this process by setting parameters such as the sample size (N), effect size (δ), and the total number of simulations (n_{test}). Throughout n_{test} iterations, we simulate A/B tests by generating data, dividing it into control and treatment groups, and applying an effect—positive for assessing power and negative for evaluating type-I error. We conduct these tests incorporating early stopping criteria. The efficacy of our approach is determined by calculating the proportion of tests that reject the null hypothesis, mathematically represented as $\frac{n_{\text{rejections}}}{n_{\text{test}}}$, to quantify the statistical power or type-I error rate.

In our simulations, we create scenarios with control and treatment groups of equal size, each containing N observations, drawn from a normal distribution. The treatment effect is modeled as a uniform shift in the mean, quantified by treatment effect size δ. As a result, the control group follows a $\mathcal{N}(0,1)$ distribution, while the treatment group is modeled as $\mathcal{N}(\delta,1)$. Type-I error rate is evaluated under a negative effect $-\delta$ of equal size. To understand the magnitude of the effect in context, it's informative to measure it relative to the inherent variability (standard deviation σ) present in the underlying data. This becomes especially helpful when aiming to standardise effect sizes across both simulations and industry data experiments.

To replicate an early stopping scenario, the simulation incorporates sequential data analysis where evaluations are conducted periodically after a predetermined number of new samples are observed. As an extension, some A/B tests also introduce a *minimum sample threshold* to mitigate the risk of misleading results due to potentially excessive small sample variability at the early stages of the experiment.

The choice of critical value threshold for Bayesian methods will introduce hyperparameter $\mathcal{K}$, which strives to, but occasionally does not provide the same level of reliability in false discovery control as a frequentist method would. To ensure a fair comparison between the Bayesian and frequentist approaches, we set $\mathcal{K} = 19$, aligning it with a 95% threshold, analogous to a frequentist α level of 0.05, which is used for both of the benchmark models.

For prior specification, our default choice is $\mathcal{N}(0,2)$, a conservative estimate in terms of effect size reflecting the expected difference in the absence of treatment effects ($\delta_{prior} = 0$), akin to an A/A test scenario. As for the variance, we apply the same rationale yielding us $\sigma^2_{prior} = 2\sigma^2$, where given the abundance data typically available in the Tech sector, we consider it realistically attainable for most cases[1]. This deliberate selection of a less-than-optimal prior is designed to highlight the fundamental characteristics of Bayesian A/B testing and to demonstrate its baseline potential.

To enhance the relevance of our study to real-world contexts, we include scenarios using actual data from Just Eat Takeaway.com. Rather than drawing from a theoretical distribution, we sample (with replacement) from this real dataset.

4.1 Data

The data used in this study is sourced from Just Eat Takeaway.com (JET), a publicly-listed global technology company and market leader in food delivery services. JET's platform facilitates connection between nearly a billion food orders and over 690.000 partnered restaurants yearly, serving a customer base of 90 million across 20 countries, as per 2022 [13]. The core industry dataset is derived from a censored order-level variable, carefully chosen to represent a continuous metric within a large-scale data environment. The data generating process encapsulates multiple complex characteristics, such as time-varying variance and non-stationarity, that with the variable's real-world context in mind, are largely driven by seasonality. The data also reveals a notable increase in positively-biased outliers, zeros caused by measurement errors, and an overall non-standard distribution. More details about the schema of the data used can be found in Appendix D.

4.2 Benchmarks

In our simulations, we evaluate Bayesian A/B testing, incorporating both early stopping and optional stopping strategies, against the two leading sequential testing approaches in technology: GST and AVI.

Table 2: Sequential A/B testing method characteristics highlighting pros (✓) & cons (×)

Characteristic	Bayes	AVI	GST
Potentially conservative	✓	×	×
Type-I error control	×	✓	✓
Arbitrary stopping rules (unlimited peeking)	✓	✓	×
No pre-planning sample size & number of tests	✓	✓	×
Simplicity	×	×	✓
Futility stopping (H_0 early termination)	✓	×	×
Probabilistic interpretability (compatibility)	✓	×	×
Tech industry widespread adoption	×	✓	✓
Prior(s)	✓/×		

For GST the implemented variation, [22], requires a priori specification of a maximum sample size $\hat{N}$ and number of interim evaluations $\mathcal{J}$, which can restrict flexibility in dynamic research environments. The selected sample sizes are derived from a simple rule-of-thumb expression $\hat{N} = \frac{16\sigma^2}{\delta^2}$ [15] with sample variance σ^2 and, normally predicted but now known, treatment effect δ. We default to limit interim testing to a maximum of $\mathcal{J} = 100$ equal-spaced peeks.

The mSPRT framework provides flexibility for continuous data monitoring in A/B testing with its adaptable stopping rules, allowing for unlimited sampling without pre-planning, but requires larger sample sizes and presents a complex statistical design. Bayesian

[1] Under a set of assumptions, reproducing a similar variance estimate is not difficult as we only require a sample variance σ^2 to compute: $\sigma^2_{prior} = \text{Var}[T-C] \overset{def}{=} \text{Var}[T] + \text{Var}[C] - 2\cdot\text{Cov}[T,C] \overset{randomisation}{=} \text{Var}[T] - \text{Var}[C] \overset{No\ effect}{=} 2\cdot\text{Var}[C] \overset{n\to\infty}{=} 2\sigma^2$

A/B testing, on the other hand, integrates prior knowledge to potentially enhance analysis depth and improve power and false discovery rates, especially in limited sample size scenarios, offering interpretability and the option for futility stopping. However, it lacks guaranteed control over type-I error, which may worsen with poorly specified priors, demanding significant effort in prior specification.

5 RESULTS

In this section, we present our evaluation of Bayesian A/B testing, alongside the established benchmarks. The leading objective will be to equip the reader with a comprehensive trade-off overview that will aid readers to discern the relevance of Bayesian A/B testing for their specific use-cases, paired with actionable recommendations around its implementation.

5.1 Power curves for varying effect sizes

The first plots in Figure 6 present power curves that assess the speed of effect detection and the long-term sustained statistical power for treatment effects of 0.01 ($\frac{1}{100}\sigma$), conceptually corresponding small effect sizes. All results were found to generalise positively for larger effect sizes, and thus the presented results can be considered a minimum performance we can expect. These plots illustrate baseline scenarios for sequential A/B testing within a controlled environment.

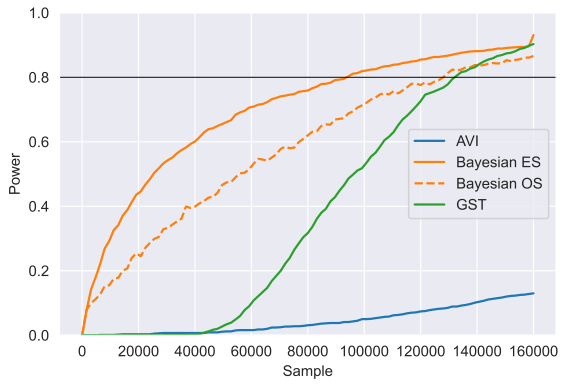

Figure 1: Power curves for Bayesian, AVI and GST

Figure 6 demonstrates capability of the Bayesian A/B tests under both Early stopping rules and optional stopping for a considerable power gain for detecting effects as rapidly as possible. However, this advantage comes with notable drawbacks at smaller effect sizes: an elevated type-I error rate that could compromise the method's reliability, and for early stopping, a potential sacrifice of long-term power as unlike the frequentist alternative, its power will not converge to 100% due to potential type-II errors caused by futility stopping. The main trade-off is the increased risk of type-errors, which were found to be predominantly dependent on sample size, relative to effect size. Overall, it was found that expected performance could reliably be projected with this ratio. When considering the common standard of 5% type-I error and 80% power, we can derive the following table, using the following risk assessment:

$$\text{Risk} = \begin{cases} \checkmark & \text{if type-I error} < 0.05 \text{ and power} > 0.8 \\ \checkmark^* & \text{if type-I error} < 0.2 \text{ and power} > 0.8 \\ \times & \text{otherwise} \end{cases} \tag{12}$$

Table 3: Risks of Bayesian A/B testing based on projected power and type-I error, for varying sample and effect sizes. The exact results can be found in Appendix C

effect size	$\frac{1}{1000}\sigma$	$\frac{1}{200}\sigma$	$\frac{1}{100}\sigma$	$\frac{1}{20}\sigma$	$\frac{1}{10}\sigma$	$\geq \frac{1}{4}\sigma$
n = 1000	×	×	×	×	✓*	✓
n = 5000	×	×	×	✓	✓	✓
n = 10000	×	×	×	✓	✓	✓
n = 50000	×	×	✓*	✓	✓	✓
n = 100000	×	✓*	✓*	✓	✓	✓

This table highlights which ranges tend to fail to meet the requirements, require (reasonably achievable) good configurations to meet requirements or meet the requirements irrespective of the configurations, with configurations including number of interim tests, prior specifications and potential minimum sample thresholds. Overall, the performance remains consistent when extended to industry data analyses, as displayed in Appendix C.

5.2 Peeking and Mitigation

In current literature, we observe the topic of type-I error inflation risks to be one of the main sources of controversy, with a continuing discourse to the present [4, 35]. We evaluate this characteristic by running the same experiments as in Figure 6 with varying number of interim evaluations, under both early stopping rules and optional stopping rules.

Table 4: Type-I error. Asterisk indicates values that are typically not acceptable by industry standards

Peeks	10	100	1000	10000
ES	0.5	7*	16*	27.6*
OS	0.3	2	3.9	7*

These findings emphasize the importance of distinguishing between an early stopping and optional stopping strategy, as they both present different trade-offs in terms of power and corresponding risks. Overall, there's risks associated with both. Type-I error inflation is significantly more harmful, than it is for optional stopping.

Upon closer inspection, it becomes apparent that all incorrect conclusions, type-I and type-II error alike, are incurred in the earlier stages of the experiment as displayed in Figure 2. This is because here, the sensitivity to small sample variability is the highest.

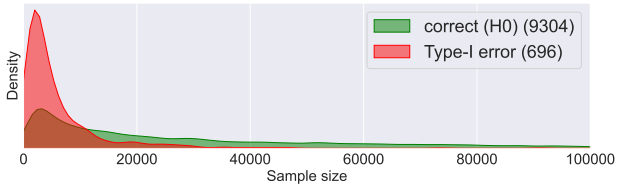

Figure 2: experiment length distributions under ES

On a longer horizon, the risk of false findings should decline following the assertion in [6] that Bayesian tests are consistent, i.e. as we observe more data, posterior $P(H_1|Data)$ converges to 1 if H_1 is true and to 0 otherwise. Consequently, this implies that decisions made from more data are always preferred. Using this property, a simple, yet effective solution to mitigate the error inflation can be introducing a *Minimum Sample Threshold* (MST) that prohibits the A/B test from terminating before this threshold is passed, regardless of any outcomes. To illustrate this, we rerun the experiments in Figure 6 with a set of arbitrary MSTs, yielding the following results:

Table 5: Performance under varying MSTs

Minimum sample threshold	0	10000	25000	50000
Type-I error	7*	2.6	0.5	0.1
Power	91.1	98.1	99.4	100

Figure 5 shows the main benefit in reducing both type-I and II error. The average improvement that MST's bring to early stopping false discovery rate ϵ_{ES} can be expressed as $\mathbb{E}[\epsilon_{ES}|n > MST] - \mathbb{E}[\epsilon_{ES}]$, which given Bayesian testing's consistency property is greater or equal to zero. On the flip side, pursuing the implementation and the length of the MST should be weighed against the overall opportunity cost of not terminating experiments in sample sizes below the MST.

5.3 Prior Sensitivity

The discussion on prior sensitivity in Bayesian hypothesis testing predominantly revolves around the influence of the priors present in the marginal likelihoods. A key insight is that as the sample size increases, the impact of observed data, as captured through the likelihood function, becomes more pronounced—evident through its decreasing variance. Conversely, the influence of priors remains constant by design, leading to their relative impact on the analysis converging to 0 as $n \to \infty$ as displayed below.

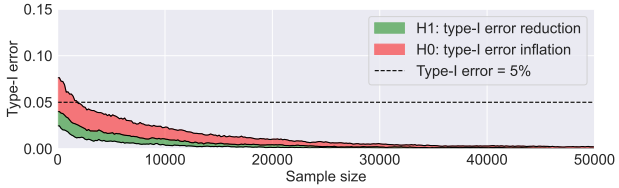

Figure 3: OS type-I error under prior misspecification.

Figure 3 shows that for optional stopping, priors misspecifications can both improve and impair error rates. The magnitude of the effect is influenced by the prior's absolute bias and variance, and the direction of the effect is driven by which prior is misspecified, relative to the underlying ground truth following Table 1. However, whether pursuing prior benefits or minimising its risks, note that it is only meaningful in the early experiment sample ranges where its overall influence is yet to be diluted. As a more conservative approach, MSTs can be useful to reduce the priors' involvement overall if desired.

In the adoption of early stopping rules, we shift our focus to family-wise type-I error rate whose the dynamics of prior specification change considerably as highlighted below:

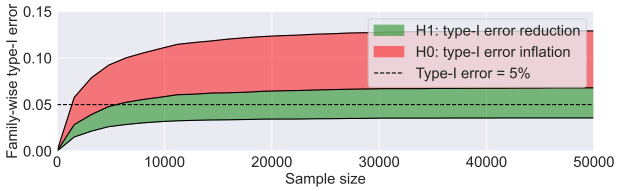

Figure 4: ES type-I error under prior misspecification.

Comparing Figures 3 and 4 reveals a key difference between OS and ES Bayesian A/B testing: transient error effects in OS versus lasting impacts of family-wise outcomes in ES. This highlights that, unlike for OS, prior sensitivity under early stopping rules remains important for all experimentation durations.

6 CONCLUSION

Seeking to fulfil our interpretation of the primary objectives of Sequential testing—accelerating decision-making, maximising power and maintaining type-I error control—we explored Bayesian A/B testing, a methodology still niche in Tech but with promising potential under the right circumstances. The study was guided by two key objectives: demonstrating Bayesian A/B testing's potential, and equipping researchers with the necessary insights to assess when and how the method and its associated trade-offs can fit the readers' specific use-cases.

When combining our findings, with the known properties of the three sequential A/B testing methods (Table 2), when limiting our scope to prioritising core A/B test usefulness characteristics, we can capture the rationale behind our method recommendations as follows:

Amongst all case studies, when put against industry-favourites AVI and GST, Bayesian A/B testing emerged as a convincing winner in terms of rapid power acquisition, demonstrating a particularly promising role for early stopping use-cases that emphasise speed. This effectiveness, however, comes with an important consideration — the nuanced trade-off of increased type-I error risks under early stopping rules in early experiment phases (contradicting optimistic views in literature claiming *peeking-immunity*). For optional stopping this downside is considerably less pronounced.

When it comes to evaluating the risks associated with the methodology, our findings suggest that the relationship between effect size

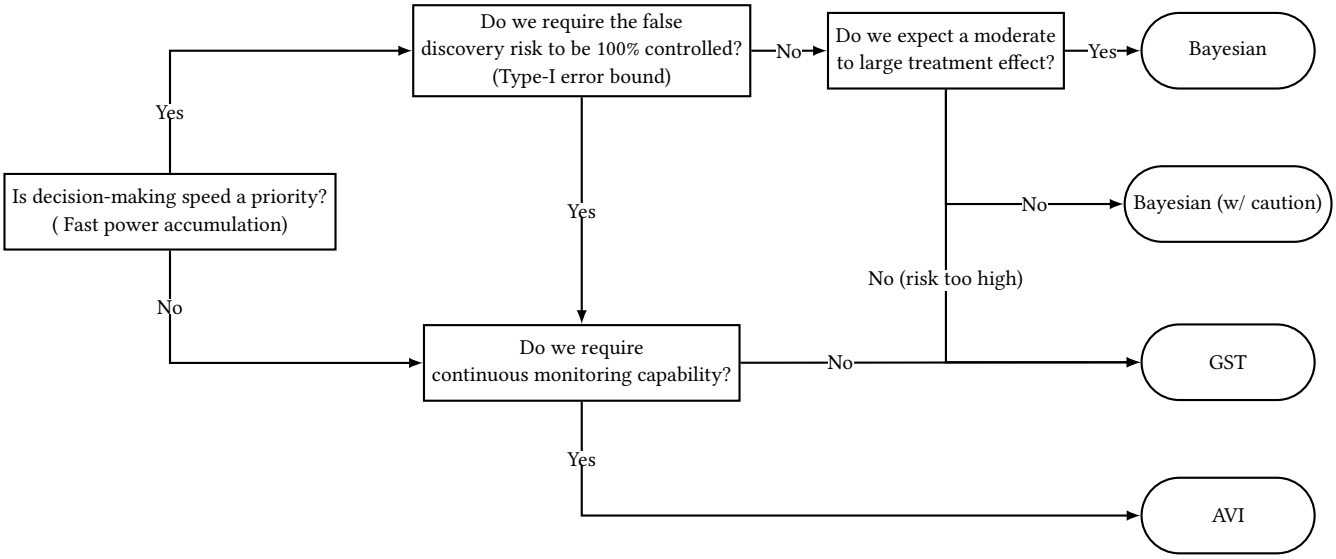

Figure 5: Decision criteria driving sequential A/B testing recommendation

and sample size is the primary driver of performance. Situations with favourable ratios, where the effect size is large relative to the sample size, consistently reduce or in some cases completely dismiss drawbacks like prior misspecification risks and type-I error inflation (peeking). The positive prior influence potential was found to be relatively small due to an overall prior influence that dilutes over time. Therefor, we shift the focus to not breaking the A/B test, rather than seeking benefit from it, as the marginal benefit of having perfect priors over a decent conservative one was often found to be negligible when moving past excessively small sample sizes. Implementing minimum sample thresholds that limit premature termination, has shown great impact in reducing both type-I and type-II errors with little downsides in return, making it a valuable extension to the majority of Bayesian A/B tests. Complementing all the findings, our real-world tests with data from JET show optimistic signs for robustness and wider applicability to more complex data sets, with little to no inconsistencies when moving from simulated to industry datasets.

Lastly, to reinforce the practicality and relevance of Bayesian A/B testing across different experimental contexts, we present a simple template in Figure 5. This template synthesizes all our key insights, providing an accessible methodology to assist researchers in assessing the suitability of Bayesian A/B testing for their unique needs.

In the light of practical applications, when considering the power and risk trade-off we observed, Bayesian A/B testing can become exceptionally valuable in circumstances where the implications of failing to detect treatment effects or delaying decision-making outweigh the potential risks that false positives bring. Such situations could include stringent deadlines, limited sample availability, or (unintentionally) severely harmful experiments. While it is consistently first in terms of speed across all experiments, its effectiveness is especially notable for moderate to large effect sizes, where the associated risk concerns around prior misspecification and type-I error inflation are markedly reduced. Furthermore, the method's adaptability for optional stopping and continuous monitoring fits well with the industry's common need for agile, real-time decision-making. Lastly, Bayesian inference's probabilistic nature aligns well with decision science and game theory, facilitating strong explainability and therefore, smoother collaboration between researchers and non-technical audiences.

In conclusion, Bayesian A/B testing emerges as a valuable approach in situations demanding sample-efficient experimentation. Beyond boosting experimental metrics like power, Bayesian A/B testing's distinctive features can also enrich overall business processes, enabling real-time decision-making, continuous evaluations, and collaboration between researchers and applied professionals, through its intuitive probabilistic explainability. Our comprehensive analysis and practical guide strive to empower industry practitioners to assess when and how to engage with Bayesian A/B testing effectively, which we support with industry datasets originating from online platform JET. Nevertheless, while it offers outstanding power advantages, we advocate critically assessing the circumstances and remaining conscious of its potentially severe type-I error trade-offs.

6.1 Future research

In our study, we rigorously tested the hypothesis that δ is either less than or equal to zero or greater than zero, yielding robust results for Bayesian A/B testing within this specific context. It's critical to understand that these findings are highly reliable within the tested hypothesis framework, but their applicability may vary under different hypotheses due to the inherent variability of the Bayes factor. Our comprehensive analysis, leveraging both simulations and real-world data from JET, intentionally simplified the treatment effect to focus on core dynamics, although this approach might limit direct real-world application. We strongly advise that our results be applied with caution in differing scenarios, emphasizing that

our conclusions are firmly grounded but must be contextualized to fit other complex real-world situations accurately.

6.1.1 Complementing Bayesian A/B testing with variance reduction (CUPED). Another promising area of future research is extending Bayesian A/B testing with variance reduction techniques such as, most notably, CUPED [7]. Intuitively, we expect these to greatly complement Bayesian A/B testing, as reducing variance indirectly increases relative effect size, which based on our findings, can greatly improve the power/type-I error trade-offs that the method typically brings. Here, it would also be interesting investigating how much Bayesian A/B testing gains, relative to other already established CUPED-consuming early stopping approaches [33].

6.1.2 Decision-theoretic designs. By blending Bayesian testing's probabilistic metrics with context-driven utility functions, [40] shows how the method can enable a strong and intuitive connection to practical requirements, for example using utility or loss functions. Particularly when exploiting the flexibility that customising hypotheses, prior specifications and stopping rules bring, it empowers users to directly and precisely accommodate for decision consequences like costs, risks and returns when designing the early stopping logic. Examples of use-cases include canary testing [20] for online experimentation or [2, 32, 34] in other disciplines. Due to its contributions to Bayesian inference's impact potential, we consider investigating decision-theoretic designs promising for the method's industry impact potential.

6.1.3 Bridging Bayesian A/B testing, test martingales and e-value methodologies. While relatively, if not entirely new in the tech space, this is a promising statistical field that has seen great growth in recognition in recent years [29, 37, 38]. Following [10, 24], we can modify the Bayes factors to include a simple H_0 marginal likelihood, such that it becomes an instance of an *e-value*. In this specific set-up, it will preserve this property for all arbitrary stopping times to yield, resulting in an *e-process*, the sequential testing equivalent that promises to complete type-I error inflation control. To our knowledge, this intersection of Bayesian statistics and recent literature advancements remains fairly unexplored in industry at the time of writing.

REFERENCES

[1] Donald A Berry. 1985. Interim analyses in clinical trials: classical vs. Bayesian approaches. *Statistics in medicine* 4, 4 (1985), 521–526.

[2] Donald A Berry and Chih-Hsiang Ho. 1988. One-sided sequential stopping boundaries for clinical trials: A decision-theoretic approach. *Biometrics* (1988), 219–227.

[3] Srivas Chennu, Andrew Maher, Christian Pangerl, Subash Prabanantham, Jae Hyeon Bae, Jamie Martin, and Bud Goswami. 2023. Rapid and Scalable Bayesian AB Testing. In *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 1–10.

[4] Rianne De Heide and Peter D Grünwald. 2021. Why optional stopping can be a problem for Bayesians. *Psychonomic Bulletin & Review* 28 (2021), 795–812.

[5] Deb et al. 2018. Under the Hood of Uber's Experimentation Platform. https://www.uber.com/en-SE/blog/xp/.

[6] Alex Deng, Jiannan Lu, and Shouyuan Chen. 2016. Continuous monitoring of A/B tests without pain: Optional stopping in Bayesian testing. In *2016 IEEE International conference on data science and advanced analytics (DSAA)*. IEEE, 243–252.

[7] Alex Deng, Ya Xu, Ron Kohavi, and Toby Walker. 2013. Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the sixth ACM international conference on Web search and data mining*. 123–132.

[8] Ward Edwards, Harold Lindman, and Leonard J Savage. 1963. Bayesian statistical inference for psychological research. *Psychological review* 70, 3 (1963), 193.

[9] KK Gordon Lan and David L DeMets. 1983. Discrete sequential boundaries for clinical trials. *Biometrika* 70, 3 (1983), 659–663.

[10] Peter Grünwald, Rianne de Heide, and Wouter M Koolen. 2020. Safe testing. In *2020 Information Theory and Applications Workshop (ITA)*. IEEE, 1–54.

[11] 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.

[12] Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. 2022. Always valid inference: Continuous monitoring of a/b tests. *Operations Research* 70, 3 (2022), 1806–1821.

[13] Just Eat Takeaway.com N.V. 2022. Full Year 2022 Results. https://www.justeattakeaway.com/newsroom/en-WW/223461-full-year-2022-results.

[14] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1168–1176.

[15] Ron Kohavi, Alex Deng, and Lukas Vermeer. 2022. A/b testing intuition busters: Common misunderstandings in online controlled experiments. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3168–3177.

[16] John Kruschke. 2013. Optional Stopping in Data Collection: Power of a Single Test. https://doingbayesiandataanalysis.blogspot.com/2013/11/optional-stopping-in-data-collection-p.html.

[17] Daniel Lakens, Friedrich Pahlke, and Gernot Wassmer. 2021. Group sequential designs: A tutorial. (2021).

[18] J Jack Lee and Diane D Liu. 2008. A predictive probability design for phase II cancer clinical trials. *Clinical trials* 5, 2 (2008), 93–106.

[19] Michael Lindon, Chris Sanden, and Vaché Shirikian. 2022. Rapid regression detection in software deployments through sequential testing. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3336–3346.

[20] Michael Lindon, Chris Sanden, Vache Shirikian, Yanjun Liu, Minal Mishra, and Martin Tingley. 2024. Sequential A/B Testing Keeps the World Streaming. (2024). https://netflixtechblog.com/sequential-a-b-testing-keeps-the-world-streaming-netflix-part-1-continuous-data-cba6c7ed49df

[21] Richard M Nixon, Anthony O'Hagan, Jeremy Oakley, Jason Madan, John W Stevens, Nick Bansback, and Alan Brennan. 2009. The Rheumatoid Arthritis Drug Development Model: a case study in Bayesian clinical trial simulation. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry* 8, 4 (2009), 371–389.

[22] Peter C O'Brien and Thomas R Fleming. 1979. A multiple testing procedure for clinical trials. *Biometrics* (1979), 549–556.

[23] Stuart J Pocock. 1977. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 64, 2 (1977), 191–199.

[24] Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. 2023. Game-theoretic statistics and safe anytime-valid inference. *Statist. Sci.* 38, 4 (2023), 576–601.

[25] Jeffrey N Rouder. 2014. Optional stopping: No problem for Bayesians. *Psychonomic bulletin & review* 21 (2014), 301–308.

[26] Adam N Sanborn and Thomas T Hills. 2014. The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic bulletin & review* 21 (2014), 283–300.

[27] Felix D Schönbrodt, Eric-Jan Wagenmakers, Michael Zehetleitner, and Marco Perugini. 2017. Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological methods* 22, 2 (2017), 322.

[28] Schultzburg and Ankergren. 2023. Choosing Sequential Testing Framework: Comparisons and Discussions. https://engineering.atspotify.com/2023/03/choosing-sequential-testing-framework-comparisons-and-discussions/.

[29] Glenn Shafer. 2021. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society Series A: Statistics in Society* 184, 2 (2021), 407–431.

[30] Nils Skotara. 2023. Sequential Testing at Booking.com. https://booking.ai/sequential-testing-at-booking-com-650954a569c7.

[31] Steven Snapinn, Mon-Gy Chen, Qi Jiang, and Tony Koutsoukos. 2006. Assessment of futility in clinical trials. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry* 5, 4 (2006), 273–281.

[32] Nigel Stallard, John Whitehead, and Simon Cleall. 2005. Decision-making in a phase II clinical trial: a new approach combining Bayesian and frequentist concepts. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry* 4, 2 (2005), 119–128.

[33] Erik Stenberg. 2019. Sequential A/B Testing Using Pre-Experiment Data.

[34] Steffen Ventz and Lorenzo Trippa. 2015. Bayesian designs and the control of frequentist characteristics: a practical solution. *Biometrics* 71, 1 (2015), 218–226.

[35] Eric-Jan Wagenmakers, Quentin F Gronau, and Joachim Vandekerckhove. 2019. Five bayesian intuitions for the stopping rule principle. (2019).

[36] Runzhe Wan, Yu Liu, James McQueen, Doug Hains, and Rui Song. 2023. Experimentation Platforms Meet Reinforcement Learning: Bayesian Sequential

Decision-Making for Continuous Monitoring. *arXiv preprint arXiv:2304.00420* (2023).
[37] Ruodu Wang and Aaditya Ramdas. 2022. False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84, 3 (2022), 822–852.
[38] Ian Waudby-Smith and Aaditya Ramdas. 2023. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B Methodological* (2023).
[39] Erica C Yu, Amber M Sprenger, Rick P Thomas, and Michael R Dougherty. 2014. When decision heuristics and science collide. *Psychonomic bulletin & review* 21 (2014), 268–282.
[40] Tianjian Zhou and Yuan Ji. 2023. On Bayesian sequential clinical trial designs. *The New England Journal of Statistics in Data Science* (2023), 1–16.

A DERIVATION BAYES FACTORS FOR NORMAL CONJUGATE PRIORS

This paper closely follows [36]'s implementation and derivation of [6]'s Bayesian A/B testing methodology, which leverage a normal (conjugate) prior. For the one-sided hypothesis testing setup, we can conveniently obtain the exact Bayes factor analytically using

$$BF_{H_1|H_0} = \frac{1-\Phi(-\mu_1'/\sigma_1')}{1-\Phi(-\mu_1/\sigma_1)} \cdot \frac{\Phi(-\mu_0/\sigma_0)}{\Phi(-\mu_0'/\sigma_0')} \cdot \sqrt{\frac{\sigma_0^2+\sigma^2}{\sigma_1^2+\sigma^2}} \cdot \exp\left(\frac{-1}{2}\left(\frac{(\mu_1-y)^2}{\sigma^2+\sigma_1^2}\right) + \frac{1}{2}\left(\frac{(\mu_0-y)^2}{\sigma^2+\sigma_0^2}\right)\right) \tag{13}$$

with $\Phi(\cdot)$ denoting the normal CDF; observed mean and variance y and σ^2; prior distributions $N(\mu_i, \sigma_i)$; and μ_i' being calculated as $\frac{y\sigma_i^2+\mu_i\sigma^2}{\sigma_i^2+\sigma^2}$ with σ_i' as $\sqrt{\frac{\sigma^2\sigma_i^2}{\sigma^2+\sigma_i^2}}$ for $i \in \{0, 1\}$.

The marginal likelihoods that underpin these can be derived by taking the product of the likelihood $p(y|\mu)$ and prior density $p(\mu|H)$, summed over all values of μ that are defined on the range specified by the associated hypothesis. The likelihood is assumed to be normal $\mathcal{N}(y, \sigma^2)$ and the prior a truncated normal $\mathcal{TN}(\mu_i, \sigma_i^2)$ for $\mu > 0$ under H_1 and $\mu \leq 0$ for H_0:

$$P(y|H_0) = \int_{\mu\leq 0} p(y|\mu)p(\mu|H_0)d\mu \tag{14}$$

$$= \frac{1}{\Phi\left(\frac{-\mu_0}{\sigma_0}\right) - \Phi\left(\frac{-\infty-\mu_0}{\sigma_0}\right)} \frac{1}{2\pi\sigma_0\sigma} \cdot \int_{-\infty}^{0} \exp\left(-\frac{1}{2}\left(\frac{(y-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{\sigma_0^2}\right)\right) d\mu \tag{15}$$

$$= \frac{1}{\Phi\left(\frac{-\mu_0}{\sigma_0}\right)} \frac{1}{2\pi\sigma_0\sigma} \exp\left(-\frac{1}{2}\left(\frac{(\mu_0-y)^2}{\sigma^2+\sigma_0^2}\right)\right) \cdot \int_{-\infty}^{0} \exp\left(-\frac{1}{2}\frac{\sigma^2+\sigma_0^2}{\sigma_0^2\sigma^2}\left(\mu - \frac{y\sigma_0^2+\mu_0\sigma^2}{\sigma_0^2+\sigma^2}\right)^2\right) d\mu \tag{16}$$

$$= \frac{1}{\sqrt{2\pi(\sigma_0^2+\sigma^2)}} \exp\left(-\frac{1}{2}\frac{(\mu_0-y)^2}{\sigma^2+\sigma_0^2}\right) \frac{\Phi(-\mu_0'/\sigma_0')}{\Phi(-\mu_0/\sigma_0)} \tag{17}$$

with μ_i' being calculated as $\frac{y\sigma_i^2+\mu_i\sigma^2}{\sigma_i^2+\sigma^2}$ and σ_i' as $\sqrt{\frac{\sigma^2\sigma_i^2}{\sigma^2+\sigma_i^2}}$. Similarly, we can derive the marginal likelihood under H_1, which will then yield:

$$P(y|H_1) = \int_{\mu>0} p(y|\mu)p(\mu|H_1)d\mu \tag{18}$$

$$= \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma^2)}} \exp\left(-\frac{1}{2}\frac{(\mu_1-y)^2}{\sigma^2+\sigma_1^2}\right) \frac{1-\Phi(-\mu_1'/\sigma_1')}{1-\Phi(-\mu_1/\sigma_1)} \tag{19}$$

Then, through the ratio of 18 and 17 respectively, we obtain the Bayes factor expression for composite hypothesis testing with normal conjugate priors as in 8.

For numerical reasons, we opt to calculate logarithms of the Bayes factors instead. The version that was eventually implemented is defined as follows:

$$\log(BF_{H_0|H_1}) = \log(P(y|H_1)) - \log(P(y|H_0)) \tag{20}$$

With log marginal likelihoods:

$$\log(P(y|H_0)) = -\frac{1}{2}\log(2\pi(\sigma_0^2+\sigma^2)) - \Phi\left(-\frac{\mu_0}{\sigma_0}\right) + \Phi\left(-\frac{\mu_0'}{\sigma_0'}\right) + -\frac{1}{2}\frac{(\mu_0-y)^2}{\sigma^2+\sigma_0^2} \tag{21}$$

$$\log(P(y|H_1)) = -\frac{1}{2}\log(2\pi(\sigma_1^2+\sigma^2)) + \log\left(1-\Phi\left(-\frac{\mu_1'}{\sigma_1'}\right)\right) + -\log\left(1-\Phi\left(\frac{\mu_1}{\sigma_1}\right)\right) - \frac{1}{2}\frac{(\mu_1-y)^2}{\sigma^2+\sigma_1^2} \tag{22}$$

B EXPERIMENT CONFIGURATIONS

$Control \sim \mathcal{N}(0, 1)$ and $Treatment \sim \mathcal{N}(\delta, 1)$ with

$$\delta = \begin{cases} -0.01 & \text{if } H_0 = \text{True} \\ 0.01 & \text{if } H_1 = \text{True} \end{cases} \tag{23}$$

Appropriate maximum sample sizes were derived using the rule-of-thumb estimation method in [14].

$$\hat{N} = \frac{16\sigma^2}{\delta^2} = 160000 \tag{24}$$

	Prior H_0	Prior H_1	Sample size ($\hat{N}$)	Number of peeks	$\mathcal{K}$	α	MC iterations
Figure 1	N(0, 2)	N(0, 2)	160000	100	19		10000
Figure 1 (GST)			160000	100		0.05	10000
Figure 1 (AVI)		N(0, 2)	160000	100		0.05	10000
Figure 2	N(0, 2)	N(0, 2)	160000	100	19		10000
Figure 3	N(0, 2)	N(0, 2)	160000	100	19		10000
Table 3	N(0, 2)	N(0, 2)		100	19		1000
Table 4	N(0, 2)	N(0, 2)	160000_{MST}	$100 - \lfloor \frac{\text{MST}}{1600} \rfloor$	19		1000
Table 5	N($\delta - 1$, 2)	N($\delta + 1$, 2)	160000	10_000 (OS)	19		10000

C MONTE CARLO SIMULATION RESULTS

Monte Carlo simulations (1000) for varying combinations of effect size and sample size limits. Each configuration represents type-I error and power as follows ($power, \epsilon_{ES},$). Effect sizes are reflected

in terms of relative size to the standard deviation of the underlying raw data σ.

effect size	$\frac{1}{1000}\sigma$	$\frac{1}{200}\sigma$	$\frac{1}{100}\sigma$	$\frac{1}{20}\sigma$	$\frac{1}{10}\sigma$	$\geq \frac{1}{4}\sigma$
n = 1000	(0.05, 0.50)	(0.07, 0.42)	(0.07, 0.38)	(0.32, 0.11)	(0.76, 0.1)	(1, 0)
n = 5000	(0.15, 0.52)	(0.22, 0.45)	(0.21, 0.34)	(0.93, 0)	(1, 0)	(1, 0)
n = 10000	(0.20, 0.44)	(0.35, 0,33)	(0.31, 0.25)	(1, 0)	(1, 0)	(1, 0)
n = 50000	(0.30, 0.44)	(0.49, 0.25)	(0.65, 0.09)	(1, 0)	(1, 0)	(1, 0)
n = 100000	(0.55, 0.44)	(0.79, 0.19)	(0.91, 0.09)	(1, 0)	(1, 0)	(1, 0)

Table 6: *(Power, Type-I error)* performance under varying effect and sample sizes (Simulated DGP)

effect size	$\frac{1}{1000}\sigma$	$\frac{1}{200}\sigma$	$\frac{1}{100}\sigma$	$\frac{1}{20}\sigma$	$\frac{1}{10}\sigma$	$\geq \frac{1}{4}\sigma$
n = 1000	(0.04, 0.47)	(0.06, 0.42)	(0.03, 0.37)	(0.28, 0.08)	(0.99, 0.1)	(1, 0)
n = 5000	(0.06, 0.65)	(0.21, 0.56)	(0.48, 0.47)	(0.92, 0)	(1, 0)	(1, 0)
n = 10000	(0.08, 0.58)	(0.26, 0,42)	(0.59, 0.27)	(1, 0)	(1, 0)	(1, 0)
n = 50000	(0.31, 0.51)	(0.41, 0.32)	(0.88, 0.15)	(1, 0)	(1, 0)	(1, 0)
n = 100000	(0.68, 0.5)	(0.87, 0.26)	(0.99, 0.08)	(1, 0)	(1, 0)	(1, 0)

Table 7: *(Power, Type-I error)* performance under varying effect and sample sizes (Industry data)

Power accumulation over time can also be assessed through power curve comparisons, utilizing a difference of $\delta = 0.01\sigma$, where σ represents the standard deviation of the underlying simulated or industry data.

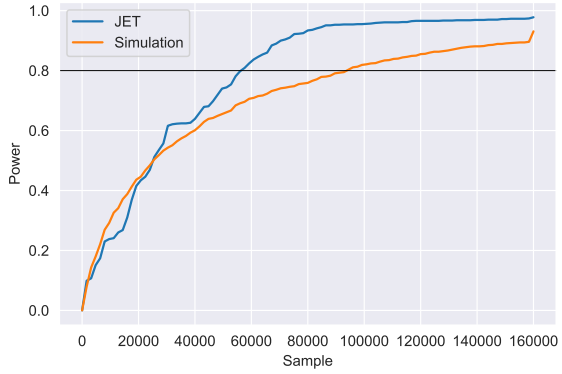

Figure 6: Power curves for Bayesian for simulated & industry data

The effect size was calculated using variance derived from the entire dataset; however, because for the industry dataset this variance is time-varying and relatively lower at the start, the model finds inference easier early on. This enhances the model's performance in initial stages through this specific study design.

D DATA SCHEMA

More details related to the schema and characteristics of JET data.

variable	description
order_id	Id of the order.
restaurant_id	Id of the restaurant.
city_id	Id of the city.
datetime	Time and date when order was placed.
gmv	Gross marginal value of the order placed.

Table 8: *gmv* is our variable of interests, has a high season component, and changes dependent on *city_id* and *restaurant_id*.

Received 9 February 2024