Accelerating Experimentation: Bayesian Sequential A/B Testing's Role in the Tech Industry

Master Thesis: Econometrics & Management Science - Quantitative Marketing & Business Analytics, Erasmus University Rotterdam (Erasmus School of Economics)

Author: Richie Lee (505917KL), Date: 26/04/2024
Supervisor: Nick Koning, Second Assessor: Jesse Hemerik

ABSTRACT

This study highlights the comparison of Bayesian A/B testing with other prevalent sequential A/B testing methodologies, its adaptation to tech-specific challenges, and introduces guidelines for its implementation in online experimentation environments. Moreover, we demonstrate its promising 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.[1]

CCS CONCEPTS

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

KEYWORDS

Sequential A/B testing, Bayesian statistics, Bayes factor, early stopping, continuous monitoring, peeking, prior sensitivity

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. For this reason, practitioners are often incentivised to fall into malpractices such as *Peeking*, i.e. to continuously monitor results and make ad-hoc decisions, 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 [29], Booking.com [31]) 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 main objective is to identify and demonstrate practical applications for Bayesian A/B testing, emphasising its ability to efficiently samples to maximize power, while effectively managing the associated trade-offs. Additionally, desirable properties include probabilistic interpretability [41], futility stopping [32], and validity under optional stopping / continuous monitoring [6]. Additionally, the method is benchmarked

[1]The content of this thesis is the sole responsibility of the author and does not reflect the view of the supervisor, second assessor, Erasmus School of Economics or Erasmus University.

against state-of-the-art alternatives and testing it with real industry data from the global food delivery platform, *Just Eat Takeaway.com*. This evaluation assesses the impact of unmet assumptions and sets a baseline for the minimum expected outcomes from basic Bayesian A/B testing implementations [6, 37] before tailoring to specific practitioner needs.

Ultimately, by leveraging these insights, we aim to simplify the generalisation and application of our findings. We provide a trade-off overview and actionable recommendations to lay the groundwork for new Bayesian A/B testing use-cases in the tech industry and beyond.

Contribution: This study expands the focus of existing literature on applied Bayesian A/B testing, traditionally centered on statistical validity and accuracy, to now dedicate attention to the speed of power accumulation and associated risk trade-offs. The experiments demonstrated strong performance in decision-making speed, with manageable risks under both early and optional stopping rules, particularly when large effect sizes or minimum sample size requirements were in place. We analysed both simulated and industry data, the latter still being uncommon in Bayesian A/B testing literature as of this writing.

As a second contribution, we present a framework aimed at simplifying the adoption of Bayesian A/B testing for industry practitioners. This framework provides tools that help researchers determine if the Bayesian approach is preferable under their specific conditions compared to alternative sequential testing methods. Additionally, it highlights the risks to consider during its deployment, ensuring practitioners are well-informed of potential challenges.

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 experimental design and benchmarks are described in Section 4 and corresponding key findings in Section 5 and summarized in Section 6 – accompanied by a flowchart that captures all the accumulated insights to support Bayesian A/B testing implementation and sequential testing method selection.

2 RELATED WORK

This section introduces sequential A/B testing and provides an overview of the most common industry solutions. We review the widely adopted variants, *Alpha spending* and *Always valid inference*. Additionally, we discuss *E-values*, which, although not yet deployed in industry, are included to offer a more comprehensive overview of relevant sequential testing methods.

Despite earlier positive outlooks on the methodology, such as those detailed by [6], the presence of Bayesian A/B testing in the tech sector remains limited. Published industry contributions to the literature are currently limited to two recent related works: [37] and [3]. This situation is and has been notably different in other disciplines like clinical trials [41] or psychology [21], where the demand for "optimal" experimentation is just as strong – striking an interesting asymmetry to investigate. 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, has retained its presence in the field in applications such as phase II clinical trials for oncology and drug development [18, 22, 33]. These research stages involve human participants and usually have sample sizes between 100 and 300. This is in 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, 23, 24] and Always valid inference [12, 19] 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 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 come with the data generating process for sequential testing (filtration). O'Brien & Fleming [23] 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 as follows:

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

Overall, it starts more stringent and relaxes over time, based on the 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 adopted Mixture Sequential Probability Ratio Test (mSPRT) is a popular variant of Always Valid Inference (AVI), which has attracted significant interest in the technology sector due to its robustness and flexibility in statistical testing. Highlighted in studies by [12, 19], mSPRT combines Bayesian and frequentist methods to create a hybrid approach for hypothesis testing. This method controls the type-I error rate across an unlimited number of 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 [34], applies 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:

$$\hat{\Lambda}_n = \int_{\theta \in \Theta} \prod_{i=1}^{n} \frac{f_\theta(x_i)}{f_0(x_i)} \pi(\theta) d\theta, \quad (3)$$

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:

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

Due to the complexity of the integrals required to compute $\hat{\Lambda}_n$, mSPRT frequently requires numerical methods for evaluation. This leads to a reliance on intensive computational techniques to meet its inferential goals. 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.

2.0.3 E-values. 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 [30, 38, 39]. This hypothesis testing approach revolves around substituting classical p-values with E-values, which are designed to quantify evidence in the data against the null hypothesis. More specifically, when a test statistic E satisfies:

$$E \geq 0 \text{ and } \mathbb{E}_p[E] \leq 1 \text{ for } \forall\, \mathrm{p} \in H_i, \quad (5)$$

[10] shows we obtain type-I error control, which is preserved in sequential testing setting. In repeated evaluations, or more formally, discrete stochastic processes, this practice of safe optional continuation is referred to as an E-process.

[10] highlights various ways of constructing E-values satisfying the definition in Equation 5. One such way is through a mixture of likelihood densities, exemplified by the mSPRT test statistic that was introduced in Equation 3. An alternative interpretation highlighted in [25] is recognising the similarities to Bayesian hypothesis testing, where likelihood ratio is viewed as a Bayes factor with a null hypothesis that required to be simple. For a deeper dive into more sophisticated applications of the methodology, readers are referred to the overview paper [10].

3 METHODOLOGY

This section fleshes out the implemented Bayesian A/B testing methodology in terms of both theory and the more specific configurations for this study. We highlight the distinction between early and optional stopping, alongside a review of the currently ongoing discussion on the type-I error control capabilities of these approaches.

This study aims to determine the direction of treatment effects, prioritising sign prediction rather than estimating magnitude. This approach leads to the formulation of the following one-sided hypotheses:

$$\mathbf{H_0} : \delta \leq 0 \quad \mathbf{H_1} : \delta > 0. \quad (6)$$

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. For this study, both groups' data consist of i.i.d. samples from a normal distribution for the simulations, and non-parametric industry data sets for real-life evaluations. Variance is assumed to not be affected by the simulated treatment.

Bayesian inference is distinguished by its ability to incorporate prior knowledge—such as expert opinions, complementary research, historical data, and initial assumptions—that may not be captured in the data itself. Using Bayes' theorem, this integration of prior knowledge with new empirical evidence results in a statistically valid update of beliefs, which is then reflected in the posterior probability. In the context of Bayesian hypothesis testing, Bayes' theorem can be rewritten to the following form:

$$\underbrace{\frac{P(H_1 \mid \text{data})}{P(H_0 \mid \text{data})}}_{\text{post odds}} = \underbrace{\frac{P(H_1)}{P(H_0)}}_{\text{prior odds}} \cdot \underbrace{\frac{P(\text{data} \mid H_1)}{P(\text{data} \mid H_0)}}_{\text{Bayes factor}}, \quad (7)$$

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 7, 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). When adequately 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, defined as follows:

$$BF_{H1|H0} = \frac{P(data|H_1)}{P(data|H_0)} = \frac{\int_{\theta\in H_1} p(data|\theta)p(\theta|H_1)d\theta}{\int_{\theta\in H_0} p(data|\theta)p(\theta|H_0)d\theta}. \quad (8)$$

Bayes factors are calculated as ratios of *marginal likelihood* densities. These densities are integrals that weigh the likelihood densities across all plausible values of the parameter θ under each hypothesis. The weighting is done using the prior density $P(\theta|H)$ over this parameter space.

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 [37]'s implementation of the Bayesian A/B testing methodology specified in [6], which assumes a normal prior distribution for δ [2], enabling a conjugate prior with an analytical solution to be used when testing a one-sided hypothesis setup. The corresponding Bayes factor can be derived as follows:

$$BF_{H_1|H_0} = \frac{1-\Phi(-\mu_1'/\sigma_1')}{1-\Phi(-\mu_1/\sigma_1)} \cdot \frac{\Phi(-\mu_0/\sigma_0)}{\Phi(-\mu_0'/\sigma_0')} \cdot \sqrt{\frac{\sigma_0^2+\sigma^2}{\sigma_1^2+\sigma^2}} \cdot \exp\left(-\frac{1}{2}\left(\frac{(\mu_1-y)^2}{\sigma^2+\sigma_1^2}\right)+\frac{1}{2}\left(\frac{(\mu_0-y)^2}{\sigma^2+\sigma_0^2}\right)\right), \quad (9)$$

with $\Phi(\cdot)$ denoting the normal CDF; observed mean y and assumed to be known variance 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 9 to a logarithmic Bayes factor is typically more practical. This, and the full derivation of the marginal likelihoods are presented in Appendix C.

Once calculation of the Bayes Factor is completed, 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). This section explains how these methods operate, and how their respective insights are designed to be interpreted.

3.1.1 Early Stopping. To extend the continuous monitoring tests to enable ES, we introduce a stopping rule 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, 28], with a symmetrical design with equal stringency for H_0 and H_1, visualised in Appendix A. This results in the following stopping rule:

$$\text{For each interim test: } \begin{cases} BF_{H_1|H_0} > \mathcal{K} & \text{STOP \& Reject } H_0 \\ BF_{H_1|H_0} < \frac{1}{\mathcal{K}} & \text{STOP \& Accept } H_0 \\ \text{otherwise} & \text{continue sampling} \end{cases} \quad (10)$$

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 facilitates *Futility stopping* which contributes to efficiency when H_0 is true [32]. 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 [41].

3.1.2 Optional Stopping. Bayesian A/B testing can also support continuous monitoring and OS [28], which was analytically proven in [6] and justified in [36] 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

[2]Note that for the normality assumption, as well as other assumptions, should be acknowledged to be challenging to satisfy in practice. The findings of this basic implementation following [6, 37] is therefore instead intended to present a generalisable illustration of the methods minimum potential, rather than diving into specific extensively optimised use-cases.

integrity of Bayes factors is preserved, irrespective of the decisions and actions that follow it. This implies that under OS, the experiment is not automatically concluded upon reaching a critical decision threshold of $\frac{1}{\mathcal{K}}$ or $\mathcal{K}$. Instead, termination can be decided manually.

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:

$$\epsilon_{\tau,OS} = P(H_1|H_0, \textit{post odds}) \leq \frac{1}{\mathcal{K}+1} := \alpha, \tag{11}$$

where $\epsilon_{\tau,OS}$ denoting the type-I error under OS that we would incur at evaluation τ, with desired control α. The reported overall error rate is determined by assuming the worst-case scenario $\arg\max_{\tau}(\epsilon_{\tau,OS})$.

Alternatively, under early stopping (ES) rules, we assess its risks through type-I error using

$$\begin{aligned}\epsilon_{\tau,ES} &= P\left(H_1|H_0, \textit{Post odds}, \{BF_1, \cdots, BF_{\tau-1}\} \in \left(\frac{1}{\mathcal{K}}, \mathcal{K}\right)\right) \\ &= P(H_1|H_0, \textit{Post odds}, t=\tau) \leq \frac{1}{\mathcal{K}+1} := \alpha. \end{aligned} \tag{12}$$

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 which approach to experiment evaluation is most applicable.

3.2 Type-I error control

In this section, we examine the literature on peeking in Bayesian A/B testing. The goal is to highlight the difference between early stopping (ES) and optional stopping (OS). Additionally, we present a theoretical evaluation of the impact of prior misspecification on Bayes factors and consequently type-I error rates.

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 [26], who claim 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 [27], harmful prior specifications [4], and a higher likelihood of incorrect research implementation [40].

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) while they do offer mitigation of type-I errors to an usually controlled extend, they do not *guarantee* a type-I error *bound*, and (2) they cannot ensure the treatment effect *measurement 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 and OS (not to be mistaken for ES), when adhering to the following stopping rules: 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. This section elaborates on how the priors present in the Bayesian inference in Equations 7 and 8 affect the methodology. More specifically, while the magnitude of harm or benefit is hard to predict a priori as explored in section 5.3, the direction of the effects can be mathematically inferred and is therefore discussed here.

When considering the second key challenge in Bayesian A/B testing, prior specifications, it's important 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 with different corresponding properties.

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 is 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 are defined as observed data, averaged over a parameter space that covers all plausible values. This averaging process allocates its weights 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 its 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 amplifies the Bayes factor's influence, swiftly pushing its 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 displayed 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		Type-II error ↑

The normal prior distribution has two notable consequences. Firstly, because it spans non-zero values across the entire parameter space, the marginal likelihood approaches but never reaches zero. Secondly, as sample sizes increase, the data should approach a normal distribution following central limit theorem. However, instead of solely aiming to match a specific distribution, we believe that the main objective of setting priors should be to simply fit the observed data as well as possible.

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. This section outlines the experimental design, the data and provides a brief overview of three main methods and their respective defining characteristics, highlighting the differences and trade-offs of the models that are compared.

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. First we set 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 that is positive for assessing power and negative for evaluating type-I error. 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 of size δ. As a result, the control group follows a $\mathcal{N}(0, 2)$ distribution, and treatment group $\mathcal{N}(\delta, 2)$. Type-I error rate is evaluated under a negative effect $-\delta$ of equal size. The relative magnitude of the effect δ is adjusted to the context, where we scale 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 small sample variability at the early stages of the experiment.

The selection of a critical value threshold in Bayesian methods introduces the hyperparameter $\mathcal{K}$, which aims to provide reliable control over false discoveries. However, it is not guaranteed to offer the same level of certainty in controlling false discoveries as frequentist type-I error bound 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 $\mathcal{N}(0, 2)$, reflecting the expected sample mean 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 $\sigma^2_{prior} = 2\sigma^2$. In practice, one would estimate this variance in absence of treatment, which we consider reasonably attainable for most online experimentation cases[3]. 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 industry data from Just Eat Takeaway.com. For each evaluation, a representative industry dataset is created by first sampling N observations (with replacement) from the real dataset. For the selected

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

large N, this sample is assumed to largely carry over the characteristics of the underlying true dataset. The sampled data is then split, into a control and treatment group with a simulated treatment effect. To improve the comparison of simulation and industry data evaluations, we ensure that the effect size relative to the underlying data's standard deviation is the same for both experiments. The time and order of the observations is preserved, and we can thus then recreate sequentially observed data, as you would in practice.

4.1 Data

To improve our checks for complex data situations where usual assumptions may fail, we included a dataset from Just Eat Takeaway.com (JET). JET is a global leader in food delivery services, offering a practical setting to test our methods. 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 dataset we used contains several complex features, including time-varying variance and non-stationarity, which are primarily influenced by seasonal changes. Additionally, the data shows a significant presence of positively-biased outliers and zeros that result from measurement errors, contributing to a non-standard distribution overall. More details about the schema of the data used can be found in Appendix D.

While this evaluation does not ensure that the findings will apply to noisier data, it offers an indication of the performance for more complex datasets.

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	×	✓	✓
No pre-planning sample size & number of tests	✓	✓	×
Simplicity	×	×	✓
Built-in futility stopping	✓	×	×
Probabilistic interpretability (compatibility)	✓	×	×
Tech industry widespread adoption	×	✓	✓
Prior(s)	✓/×		

For GST the implemented variation, [23], requires a priori specification of a maximum sample size $\hat{N}$ and number of interim evaluations $\mathcal{J}$, which imposes a restriction flexibility in dynamic research environments. The selected sample sizes are derived from a simple rule-of-thumb expression $\hat{N} = \frac{16\sigma^2}{\delta^2}$ following [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 and peeking without pre-planning, at the expense of larger sample size requirements and a more complex statistical design. Bayesian 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 assess Bayesian A/B testing against established benchmarks, illustrating its benefits through power curve analyses conducted with simulated data. We address its main drawbacks—type-I error control and prior sensitivity—while noting that analyses on both simulated and industry data yielded largely similar results; detailed results for industry data are included in the appendix. We then synthesise all findings into a concise guide, helping readers navigate the trade-offs and select the most suitable sequential A/B testing method.

The scope of this study will be around composite hypothesis testing for difference δ in means:

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

5.1 Power curves for varying effect sizes

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

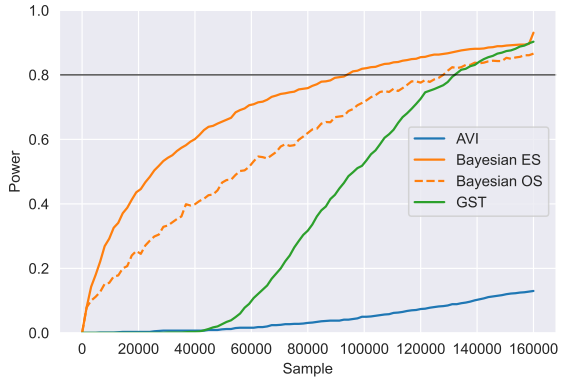

Figure 1: Power curves for Bayesian, AVI and GST

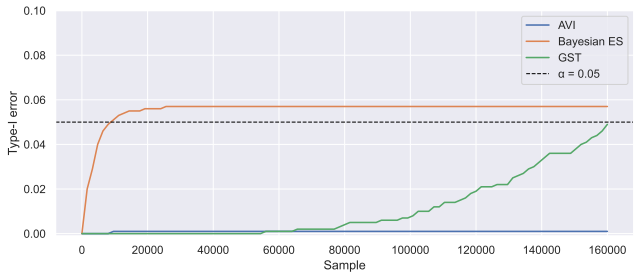

Figure 2: Type-I error curves for Bayesian, AVI and GST

Figure 1 demonstrates the effectiveness of Bayesian A/B testing under both early stopping (ES) and optional stopping (OS) rules, showcasing its capacity for significant power gains in detecting effects quickly. However, as depicted in Figure 2, a main drawback is the lack of a type-I error rate upper bound, which becomes apparent when dealing with small effect and sample sizes. This increase in error rates can hurts the method's reliability, particularly with ES, which might result in diminished long-term power and inflated type-I error rates. This is an issue that the frequentist methods do not suffer from.

One strong predictor for the magnitude of risks associated with this trade-off is effect size relative to sample size. When considering the common standard of 5% type-I error and 80% power, we can visualise the risk assessment as follows:

$$\text{Risk} = \begin{cases} \checkmark & \text{if type-I error} < 0.05 \text{ and power} > 0.8 \\ \checkmark^* & \text{if type-I error} < 0.2 \text{ and power} > 0.8 \\ \times & \text{otherwise} \end{cases} \quad (13)$$

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 E

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	×	×	×	×	$\checkmark^*$	✓
n = 5000	×	×	×	✓	✓	✓
n = 10000	×	×	×	✓	✓	✓
n = 50000	×	×	$\checkmark^*$	✓	✓	✓
n = 100000	×	$\checkmark^*$	$\checkmark^*$	✓	✓	✓

This table highlights which ranges tend to fail to meet the requirements (×), require (reasonably achievable) good configurations to meet requirements ($\checkmark^*$) 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 E.

5.2 Peeking and Risk Mitigation

In current literature, we observe the topic of type-I error inflation risks to be one of the main sources of controversy still ongoing today [4, 36]. To evaluate its relevance, we run a simple evaluation by repeating the experiments in Figure 1 with varying number of interim evaluations, under both ES rules and OS rules.

Table 4: Type-I error. Asterisk indicates values intended type-I error control (5%)

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

These findings emphasise the importance of distinguishing between an ES and OS strategy, as they both present different trade-offs in terms of power and corresponding risks. Overall, while there's risks associated with both, ES stands out as becomes more susceptible to Type-I error inflation is significantly than OS.

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 3. This is because here, the sensitivity to small sample variability and consequent spurious findings is the highest.

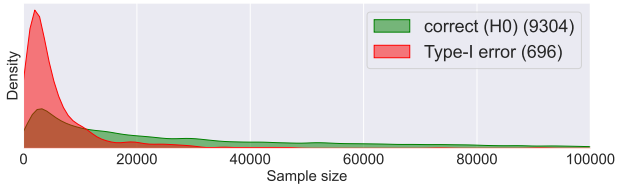

Figure 3: experiment length distributions under ES

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. A practical approach to mitigating the error inflation is the implementation of a *Minimum Sample Threshold* (MST) that prohibits the A/B test from terminating before this threshold is passed, irrespective of the outcomes prior. To demonstrate this, we re-conducted the experiments shown in Figure 1 with various arbitrary MSTs, obtaining the following outcomes:

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 focus on prior sensitivity in Bayesian hypothesis testing revolves around how priors affect marginal likelihoods. As sample sizes increase in Bayesian analysis, the growing influence of data reduces the relative impact of priors until their influence is entirely negated. The upcoming evaluations aim to clearly demonstrate the progression and scale of this effect that starts strong and dilutes over time under both optional stopping (OS) and early stopping (ES) scenarios.

Figures 4 and 5 illustrate the type-I error rates incurred when concluding an experiment at various sample sizes, under both optional stopping (OS) and early stopping (ES) rules respectively. Each figure explores three scenarios: a baseline uninformative prior (middle), misspecification of H_0 (top), and misspecification of H_1 (bottom). The gaps show how incorrectly specifying priors can affect results, here demonstrated by a bias of one full standard deviation.

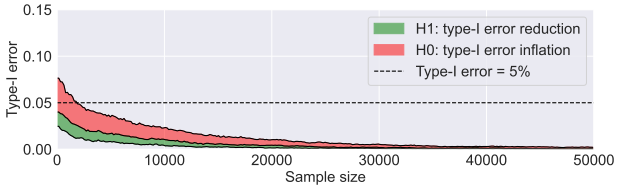

Figure 4: OS type-I error under prior misspecification.

Figure 4 shows that 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 small sample ranges where its overall influence is yet to be diluted. As a more conservative approach, MSTs can be useful to reduce the marginal likelihood priors' involvement overall if desired.

In the adoption of ES rules, the type-I error rate dynamics in the face of prior specification change considerably as highlighted below:

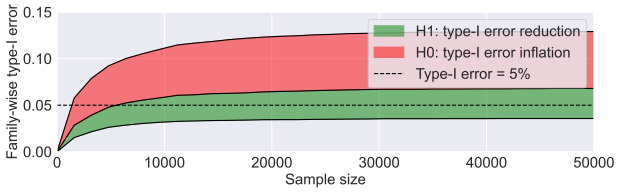

Figure 5: ES type-I error under prior misspecification.

Comparing Figures 4 and 5 demonstrates a key difference between optional stopping (OS) and early stopping (ES) when considering prior influence. Under OS, the influence of priors diminishes as sample sizes increase. In contrast, under ES, early risks from priors have lasting effects due to strict stopping rules, continuously impacting the experiment and necessitating cautious prior management from the start.

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: exploring 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), we can enrich our method recommendations, centered around common sequential testing needs, as in Figure 6.

Amongst all case studies, when put against industry-favourites AVI and GST, Bayesian A/B testing emerged the most efficient experimentation method in terms of sample size, demonstrating a particularly promising role for early stopping use-cases that emphasise speed. The finding of this effectiveness, however, comes with an important consideration — the 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 harmful. Although Bayesian A/B testing maintains acceptable average type-I error rates, it is more challenging to control compared to the formal bounds provided by alternatives AVI and GST. Note that due type-I error rates that were not perfectly standardised, there may shifts in relative performances when the experiment would be reproduced differently.

When it comes to evaluating the risks associated with the methodology, our findings suggest that the relationship between effect size and sample size is the primary driver and predictor 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). Without accounting for use-case specific context, the positive prior influence potential was found to be relatively small due to an overall prior influence that dilutes over time. Therefor, particularly for ES, we shift the focus for priors to not breaking the A/B test, rather than seeking benefit from it. This is because the marginal advantage of having perfect priors over reasonably conservative ones tends to be negligible once the sample sizes exceed very small amounts. 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 no notable issues that would arise when moving from simulated to the selected industry datasets from Just Eat Takeaway.com.

Lastly, to support the practicality and relevance of Bayesian A/B testing across different experimental contexts, we present a simple template in Figure 6. This template combines all our key insights, providing an accessible methodology that assists researchers in assessing compatibility of Bayesian testing with their unique needs.

In the light of practical applications, considering the observed trade-offs between power and risk, Bayesian A/B testing proves particularly valuable in situations where the consequences of not detecting treatment effects or delaying decisions outweigh the potential risks associated with false positives. 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 usefulness 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. Additionally, the adaptability of this method for continuous monitoring suits the industry's growing demand for agile, real-time decision-making. Furthermore, the probabilistic nature of Bayesian inference complements decision science and game theory well, enhancing explainability and thus facilitating better collaboration between researchers and non-technical stakeholders.

In conclusion, Bayesian A/B testing emerges as a valuable approach in situations that demand 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. Our analyses 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 limit our scope to composite hypotheses H_0: $\delta \leq 0$ and H_1: $\delta > 0$, with promising results for Bayesian A/B testing. Using both simulations and real-world data from JET, we intentionally simplified the treatment effect to focus on core characteristics. Perfectly standardising benchmark methods without introducing distracting complexity proved challenging due to fundamental differences in configuration possibilities and overall functionality between Bayesian and frequentist approaches. As such, while our conclusions

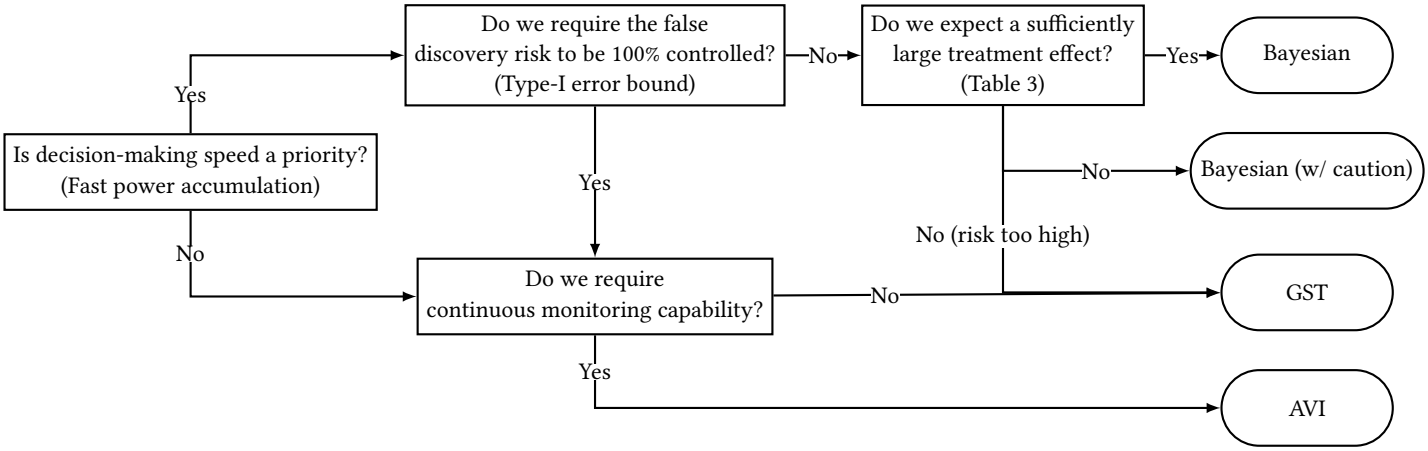

Figure 6: Decision criteria driving sequential A/B testing recommendation

are grounded, We still recommend applying our results cautiously and emphasize the need for additional contextualization deploying the methodology to complex real-world applications.

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 [34].

6.1.2 Decision-theoretic designs. By blending Bayesian testing's probabilistic output with decision science, [41] highlights 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. Examples of use-cases include canary testing [20] for online experimentation or [2, 33, 35] in other disciplines. Due to its contributions to Bayesian A/B testing's inference quality, we consider decision-theoretic designs an extension with promising practical impact potential.

6.1.3 Bridging Bayesian A/B testing, test martingales and e-value methodologies. Although E-values are becoming popular in academic settings ([10, 25]), their adoption in the online experimentation environment remains limited. As attention shifts from applied data science and reviewing sequential A/B testing to innovating and understanding novel methods in applied settings, a more comprehensive study is crucial. For instance, investigating various E-value construction methods and drawing clear parallels to Bayes factors and Always Valid Inference could facilitate a potential integration into commercial A/B testing tool kits in the future.

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] Dora Matzke, Sander Nieuwenhuis, Hedderik van Rijn, Heleen A Slagter, Maurits W van der Molen, and Eric-Jan Wagenmakers. 2015. The effect of horizontal eye movements on free recall: a preregistered adversarial collaboration. *Journal of Experimental Psychology: General* 144, 1 (2015), e1.

[22] 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.

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

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

[25] 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.

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

[27] 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.

[28] 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.

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

[30] 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.

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

[32] 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.

[33] 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.

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

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

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

[37] 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).

[38] 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.

[39] 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).

[40] 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.

[41] 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 VISUALISATION BAYESIAN A/B TEST UNDER EARLY STOPPING

Given the symmetric decision rule:

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

Under early stopping rules, the experiment is terminated once one of the critical region boundaries is passed. The upper rejection region represents rejecting H_0 (accepting H_1), and the bottom one is used for futility stopping. Each sequential test is considered a single Monte Carlo iteration and can be visualised as follows. Here we opt for $\mathcal{K} = 19$ and $\delta = 0.1$.

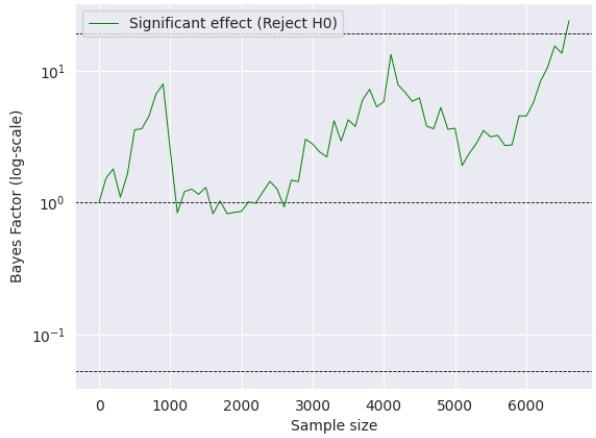

Figure 7: Visualisation of sequential Bayesian A/B tests for positive effect

B DERIVATION BAYES FACTORS FOR NORMAL CONJUGATE PRIORS

This thesis closely follows [37]'s implementation and derivation of [6]'s Bayesian A/B testing methodology, which leverages a normal (conjugate) prior. For the one-sided hypothesis testing setup, we can conveniently obtain the exact Bayes factor analytically using

$$BF_{H_1|H_0} = \frac{1-\Phi(-\mu_1'/\sigma_1')}{1-\Phi(-\mu_1/\sigma_1)} \cdot \frac{\Phi(-\mu_0/\sigma_0)}{\Phi(-\mu_0'/\sigma_0')} \cdot \sqrt{\frac{\sigma_0^2+\sigma^2}{\sigma_1^2+\sigma^2}} \cdot \exp\left(\frac{-1}{2}\left(\frac{(\mu_1-y)^2}{\sigma^2+\sigma_1^2}\right) + \frac{1}{2}\left(\frac{(\mu_0-y)^2}{\sigma^2+\sigma_0^2}\right)\right) \tag{14}$$

with $\Phi(\cdot)$ denoting the normal CDF; observed mean y and variance σ^2 assumed to be known; 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:

$$P(y|H_0) = \int_{\mu \leq 0} p(y|\mu)p(\mu|H_0)d\mu \tag{15}$$

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

$$= \frac{1}{\Phi\left(\frac{-\mu_0}{\sigma_0}\right)} \frac{1}{2\pi\sigma_0\sigma} \exp\left(-\frac{1}{2}\left(\frac{(\mu_0-y)^2}{\sigma^2+\sigma_0^2}\right)\right) \cdot \int_{-\infty}^{0} \exp\left(-\frac{1}{2}\frac{\sigma^2+\sigma_0^2}{\sigma_0^2\sigma^2}\left(\mu - \frac{y\sigma_0^2+\mu_0\sigma^2}{\sigma_0^2+\sigma^2}\right)^2\right) d\mu \tag{17}$$

$$= \frac{1}{\sqrt{2\pi(\sigma_0^2+\sigma^2)}} \exp\left(-\frac{1}{2}\frac{(\mu_0-y)^2}{\sigma^2+\sigma_0^2}\right) \frac{\Phi(-\mu_0'/\sigma_0')}{\Phi(-\mu_0/\sigma_0)} \tag{18}$$

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:

$$P(y|H_1) = \int_{\mu > 0} p(y|\mu)p(\mu|H_1)d\mu \tag{19}$$

$$= \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma^2)}} \exp\left(-\frac{1}{2}\frac{(\mu_1-y)^2}{\sigma^2+\sigma_1^2}\right) \frac{1-\Phi(-\mu_1'/\sigma_1')}{1-\Phi(-\mu_1/\sigma_1)} \tag{20}$$

Then, through the ratio of 19 and 18 respectively, we obtain the Bayes factor expression for composite hypothesis testing with normal conjugate priors as in 9.

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

$$\log(BF_{H_0|H_1}) = \log(P(y|H_1)) - \log(P(y|H_0)) \tag{21}$$

With log marginal likelihoods:

$$\log(P(y|H_0)) = -\frac{1}{2}\log(2\pi(\sigma_0^2+\sigma^2)) - \Phi\left(-\frac{\mu_0}{\sigma_0}\right) + \Phi\left(-\frac{\mu_0'}{\sigma_0'}\right) + - \frac{1}{2}\frac{(\mu_0-y)^2}{\sigma^2+\sigma_0^2} \tag{22}$$

$$\log(P(y|H_1)) = -\frac{1}{2}\log(2\pi(\sigma_1^2+\sigma^2)) + \log\left(1-\Phi\left(-\frac{\mu_1'}{\sigma_1'}\right)\right) + - \log\left(1-\Phi\left(\frac{\mu_1}{\sigma_1}\right)\right) - \frac{1}{2}\frac{(\mu_1-y)^2}{\sigma^2+\sigma_1^2} \tag{23}$$

C EXPERIMENT CONFIGURATIONS

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

$$\delta = \begin{cases} -0.01 & \text{if } H_0 = \text{True} \\ 0.01 & \text{if } H_1 = \text{True} \end{cases} \tag{24}$$

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

$$\hat{N} = \frac{16\sigma^2}{\delta^2} = 160000 \tag{25}$$

The following table highlights configurations for the main experiments including: prior specifications, maximum sample sizes, total number of evaluations, critical value parameters ($\mathcal{K}$, α) and number of Monte Carlo iterations (MC).

	Prior H_0	Prior H_1	Sample size	N_{peeks}	$\mathcal{K}$	α	MC
Fig 1	N(0, 2)	N(0, 2)	160000	100	19		10000
Fig 1 (GST)			160000	100		0.05	10000
Fig 1 (AVI)		N(0, 2)	160000	100		0.05	10000
Figure 2	N(0, 2)	N(0, 2)	160000	100	19		10000
Fig 3	N(0, 2)	N(0, 2)	160000	100	19		10000
Tab 3	N(0, 2)	N(0, 2)		100	19		1000
Tab 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

D INDUSTRY 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 6: *gmv* is our variable of interests, has a high season component, and changes dependent on *city_id* and *restaurant_id*.

E 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 7: *(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 8: *(Power, Type-I error)* performance under varying effect and sample sizes (Industry data)

Power accumulation over time can be evaluated by comparing power curves, using a difference of $\delta = 0.01\sigma$, where σ denotes the standard deviation of the underlying simulated or industry data.

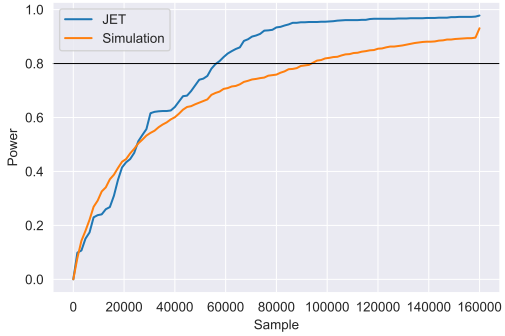

Figure 8: Power curves for Bayesian for simulated & industry data

The effect size was determined using variance from the entire dataset. However, for the industry dataset, where the variance is time-varying and initially lower, the A/B test shows heightened sensitivity early on. This enhances model performance in the initial stages due to the study design.