Effect Size Estimation for Duration Recommendation in Online Experiments: Leveraging Hierarchical Models and Objective Utility Approaches

Yu Liu*, **Runzhe Wan***, **James McQueen, Doug Hains, Jinxiang Gu, Rui Song**

Amazon.com Inc

{liuyu0jlu, runzhe.wan}@gmail.com, {jmcq, dhains, gjinxian, ruisong}@amazon.com

Abstract

The selection of the assumed effect size (AES) critically determines the duration of an experiment, and hence its accuracy and efficiency. Traditionally, experimenters determine AES based on domain knowledge. However, this method becomes impractical for online experimentation services managing numerous experiments, and a more automated approach is hence of great demand. We initiate the study of data-driven AES selection for online experimentation services by introducing two solutions. The first employs a three-layer Gaussian Mixture Model considering the heteroskedasticity across experiments, and it seeks to estimate the true expected effect size among positive experiments. The second method, grounded in utility theory, aims to determine the optimal effect size by striking a balance between the experiment's cost and the precision of decision-making. Through comparisons with baseline methods using both simulated and real data, we showcase the superior performance of the proposed approaches.

Introduction

Sample size determination (SSD) plays a pivotal role in online experiments, answering the critical question of "how long should an experiment run?" (Richardson et al. 2022). In the spheres of online A/B testing and related classic application such as clinical trials, it is paramount to ascertain the minimal sample size during the experiment planning phase. If the sample size is too small, it may not accurately represent the whole population being studied, thereby introducing both bias in the results that limit drawing conclusions or generalizing the inferences. It can also compromise the statistical power of the experiment, leading to inaccuracies in detecting meaningful effects (Ramsey and Schafer 2012; Lenth 2001). Conversely, if the sample size is selected as too large, it could unnecessarily extend the duration of the experiment and inflate the associated costs, such as the human and hardware resources and opportunity costs (Wan et al. 2023). Therefore, selecting an appropriate sample size is imperative to maintain a balance between decision accuracy and resource utilization.

Among the numerous methodologies for SSD in the literature (Kelley and Rausch 2006; Adcock 1997; Lindley 1997; Weiss 1997), one crucial step is to specify the assumed effect size (the absolute or percent lift from the treatment over the control). For example, it is a critical component in statistical power analysis used for SSD, in either the frequentist or the Bayesian setting (Du and Wang 2016). First of all, we need to distinguish the *true effect size* of an experiment and the *assumed effect size (AES)*. The first is an objective unknown quantity, while the latter is a subjective manually specified number. In traditional power analysis problems (e.g., in clinical trials), AES mainly reflects (i) the experimenter's expectation (e.g., what level of improvement would be regarded as acceptable) and (ii) the trade-off between the opportunity cost of running a longer experiment and the accuracy. AES does not necessarily relate to the true effect size (or its estimate). Moreover, using observed power based on the true effect size of an experiment to determine its sample size is inappropriate, due to its 1-1 mapping to p-value (Hoenig and Heisey 2001). For example, in case where an experiment lacks a noticeable effect size, the resulting observed power tends to be small. Using this value in SSD can lead to overestimating the required experiment duration.

Traditionally, the AES is established by domain experts or experimenters, rather than by the experimentation services itself. For instance, Lenth (2001) offers general guidance for selecting appropriate AES. However, within the sphere of online experiments - a setting where tens of thousands of experiments are conducted annually - a significant number of experimenters may lack the requisite domain knowledge or statistical acumen to define an appropriate AES. Additionally, statistical consultants are often resource-constrained, limiting their ability to offer personalized guidance to each team. In such scenarios, harnessing the vast amount of data gleaned from similar past experiments can be invaluable in determining the AES. Employing data-centric methodologies, such as meta-analysis, can automatically direct experimenters towards an appropriate AES for power analysis. This empowers them to make decisions grounded in robust, empirical evidence.

Contribution. To the best of our knowledge, this paper is the *first* work on data-driven effect size recommendation for large-scale online experiments in the literature. Our contributions are three-fold:

- We design a three-layer Gaussian mixture model for the distribution of observed effect sizes across experiments,

*These authors contributed equally.

which considers experiment heteroskedasticity and identifies positive ones. We develop an EM-type algorithm for parameter estimation;
- We also propose a utility-maximization approach to determine the optimal AES. This optimization of utility function aims to achieve a balance between experiment cost and accuracy;
- We run the first large-scale empirical analysis for effect size recommendation in the literature, using real data from Amazon. The analysis clearly demonstrates the effectiveness of the proposed methods.

Related Work

Our work is closely related to the SSD problem, for which numerous methodologies have been studied in the literature. For example, one approach involves selecting a sample size to achieve a narrow confidence interval for the standardized mean difference (Kelley and Rausch 2006). Alternatively, sample sizes can be determined based on utility theory (Adcock 1997; Lindley 1997) or to ensure the Bayes factor exceeds a predetermined value (Weiss 1997). However, in all these works, the effect size is assumed as a hyper-parameter that has been pre-specified by experimenters, which is not practical for online experimentation services.

To our knowledge, the only work related to AES estimation is Du and Wang (2016). The authors use a random-effect meta-analysis model to quantify the uncertainty of power analysis due to the uncertainty over the effect size. However, the main focus is still on SSD, and no systematic study on AES recommendation is done.

Preliminaries and Notations

Online A/B experiments are widely used to compare customer responses between the existing feature A and the new feature B, guiding online companies' feature launch decisions. To reduce the noise and established the causal relationship in online experiments, it is standard to track customers' outcomes (e.g. ad clicks) after a customer is triggered based on some event (e.g., displaying ads).

Let the duration of the experiment be defined in terms of a certain number of weeks, which helps mitigate the influence of daily and weekend fluctuations. To maintain the simplicity, we assume the experiments have the same duration t and omit the time subscript t in notations. Denote n_g as the total number of customers who are triggered up to week t in group $g \in \{T, C\}$ (T for the treatment and C for control). Denote sets of those customers who are triggered up to week t as I_g. Denote Y_i as the observed outcome for customer i up to week t, and $\bar{Y}_g = \sum_{i \in I_g} \frac{1}{n_g} Y_i$ is the sample mean of responses for customers in group g. Assuming that Y_i are independent and identically distributed (i.i.d.) with mean u_g and finite variance σ_g^2, as $n_\cdot \to \infty$, we can apply the central limit theorem (CLT) (Casella and Berger 2021):

$$\bar{Y}_g | \mu_g \sim \mathcal{N}(\mu_g, \frac{\sigma_g^2}{n_g}), g \in \{T, C\}.$$

Define the average treatment effect (ATE) $\mu = \mu_T - \mu_C$, often estimated by the sample mean difference. Taking into account the independence among different customers, the sample mean difference follows as

$$\bar{Y}_T - \bar{Y}_C | \mu_T, \mu_C \sim \mathcal{N}(\mu, \sigma^2) \tag{1}$$

where $\mu = \mu_T - \mu_C$ and $\sigma^2 = Var(\bar{Y}_T - \bar{Y}_C) = \frac{\sigma_T^2}{n_T} + \frac{\sigma_C^2}{n_C}$.

Assume all σ_g^2 are known, the Null hypothesis under Frequentist setting and its alternative for one-sided test are

$$H_0 : \mu \leq 0 \quad H_1 : \mu > 0.$$

When $n_\cdot \to \infty$, *frequentist power* is computed as:

$$\text{power}(\delta, \sigma_\cdot, n_\cdot, \alpha) \approx \Phi(z_\alpha + \frac{\delta}{\sigma}), \tag{2}$$

where α is Type I error. δ is AES. $z_\cdot$ represents the standard normal quantile. $\Phi(.)$ stands for the cumulative distribution function of the standard normal distribution. To achieve a desired power of b for the alternative hypothesis, considering for a given effect size $\delta > 0$ and an allocation rate of $n_T/n_C = p$, the minimum sample size n_T, n_C are selected to satisfying $power(\delta, \sigma_\cdot, n_\cdot, \alpha) \geq b$. The required minimum duration can be determined through sample size prediction (Richardson et al. 2022).

Standardized effect sizes, such as Cohen's d (Cohen 2013), are often used in meta-analysis across different experiments, serving to standardize variations in ATE magnitudes. If the estimators of standardized effect size follow an asymptotic normal distribution, all models presented in this paper can seamlessly accommodate the use of standardized effect sizes. Consequently, a dedicated discussion on this topic is omitted in this paper.

Pooled Effect Size

Problem Setup. Consider m past similar experiments, where δ_i is the true effect size for experiment i, and the observed effect size d_i serves as the estimator for the true effect size. d_i is commonly derived from the observed customer outcomes, e.g. Equation (1). We assume that observed effect sizes $\{d_i\}_{i=1}^m$ are independent.

The following model is used to estimate the AES (Du and Wang 2016):

$$\begin{aligned} \delta_i &\sim \mathcal{N}(\mu_0, \tau^2) \\ d_i &= \delta_i + e_i \\ e_i &\sim \mathcal{N}\left(0, \sigma_i^2\right), \end{aligned} \tag{3}$$

where we further assume $\{\delta_i\}_{i=1}^m$ and $\{e_i\}_{i=1}^m$ are mutually independent. σ_i^2 varies across different experiments due to *heteroscedasticity*, evident in Equation (1), where the variance of mean difference depends on σ_g^2 and n_g from each experiment. σ_i^2 can be estimated by $\hat{\sigma}_i^2 = \frac{\hat{\sigma}_{T,i}^2}{n_{T,i}} + \frac{\hat{\sigma}_{C,i}^2}{n_{C,i}}$. The estimated AES is the Maximum Likelihood Estimation (MLE) estimator of μ_0, which is a weighted linear combination of d_i, thus it was also referred to as the pooled effect size.

In an online company setting, not every experiment yields a measurable effect size. This stems from the nature of online experiments with their short cycles, which was designed

to encourage the exploration of innovative ideas. However, the drawback is that a significant number of ideas may not achieve statistical significance. We categorize experiments into three groups: 1) true positive, 2) true negative, and 3) flat experiments which have no significant effect. Using d_i from all experiments to train the model (3) will lead to an underestimation of AES, consequently elongating the required duration. The guideline is to train the model (3) using experiments where we know there is an underlying true positive effect. While it might seem intuitive to use hypothesis testing to categorize experiments into significant positive, negative or flat ones, the existence of Type I and Type II errors poses a challenge in precisely determining which experiments truly have such effects.

Three-Layer Heteroscedastic GMM

In the previous section, we highlighted that the main challenge of the state-of-the-art method lies in the accurately identifying experiments with true positive effect. It is natural to consider the categorization of experiments as latent variables and use the Gaussian Mixture Model (GMM).

Motivation. We first illustrate our motivation of using GMM in Figure 1. Figure 1a presents the distribution of the *observed* effect sizes for certain outcome of interest over 3,300 experiment run within Amazon. Although at the first glance, this graph seems to support a normal distribution, we argue that it indeed illustrates the challenge of this problem. In Figure 1b, we simulate the *true* effect sizes following a two-Layer GMM, and in Figure 1c we increases random errors to each of them to generate the *observed* effect size, which looks like a single mode Gaussian distribution. Therefore, under the assumption that there exists latent clusters of positive, flat and negative experiments, it is actually infeasible to identify those positive ones with simple rules; instead, a more principled statistical model as GMM should be used. The two-layer GMM didn't account for the required heterogeneity in our setting. Thus, we propose a three-layer heteroscedastic GMM.

Model. Assume $K = 3$ is the number of clusters corresponds to clusters containing the negative, flat and positive experiments, and m is the total number of past experiments. Without loss of generality, we assume the means of these clusters are decreasing from K to 1. Therefore, $k = 2$ represents the cluster of flat experiments.

Latent variable $\boldsymbol{z} = \{z_1, z_2, \ldots, z_m\}$, $z_i \in \{1, \ldots, K\}$ represent the index class sampled from the categorical distribution parameterized by $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$. μ_k, τ_k^2 and π_k correspond to the mean, variance, and weight of the k-th Gaussian component. Denote $\boldsymbol{\theta} = \{\mu_1, \ldots, \mu_K, \tau_1, \ldots, \tau_K\}$. We have

$$\begin{aligned} z_i &\sim Categorical(k, \boldsymbol{\pi}) \\ \delta_i | z_i = j &\sim \mathcal{N}(\mu_j, \tau_j^2) \\ d_i | \delta_i &\sim \mathcal{N}(\delta_i, \sigma_i^2), \end{aligned} \tag{4}$$

where, the third layer arises from the heteroskedasticity among different experiments. Note that the model above is equivalent to:

$$\begin{aligned} z_i &\sim Categorical(k, \boldsymbol{\pi}) \\ d_i | z_i = j &\sim \mathcal{N}(\mu_j, \tau_j^2 + \sigma_i^2), \end{aligned} \tag{5}$$

where σ_i^2 are known and can be estimated using observed outcomes of the experiment i, e.g. Equation (1). We can further pre-specify the means of k = 2 cluster as zero, i.e. $\mu_2 = 0$.

EM algorithm. We propose the following EM algorithm to estimate the unknown parameters in the model (5).We first derive a few formulas that are essential for developing the EM algorithm. The marginal distribution is

$$f(d_i | \boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k f_k(d_i | \boldsymbol{\theta}),$$

where $f_k(d_i|\boldsymbol{\theta})$ is the density of Gaussian distribution $\mathcal{N}(\mu_k, \tau_k^2 + \sigma_i^2)$. The conditional probability function for z_i given the data d_i is:

$$k_i(z_i | d_i, \boldsymbol{\theta}, \boldsymbol{\pi}) = \frac{\prod_{k=1}^{K} \Big(\pi_k f_k(d_i|\boldsymbol{\theta})\Big)^{\mathbb{I}[z_i=k]}}{\sum_{k=1}^{K} \pi_k f_k(d_i|\boldsymbol{\theta})}.$$

$\mathbb{I}[.]$ represents the indicator function. Due to the independence among experiments, the joint distribution is

$$\begin{aligned} p(\boldsymbol{d}, \boldsymbol{z} | \boldsymbol{\theta}, \boldsymbol{\pi}) &= \prod_{i=1}^{m} f(d_i | \boldsymbol{\theta}, \boldsymbol{\pi}) k_i(z_i | d_i, \boldsymbol{\theta}, \boldsymbol{\pi}) \\ &= \prod_{i=1}^{m} \prod_{k=1}^{K} \Big(\pi_k f_k(d_i|\boldsymbol{\theta})\Big)^{\mathbb{I}[z_i=k]}, \end{aligned}$$

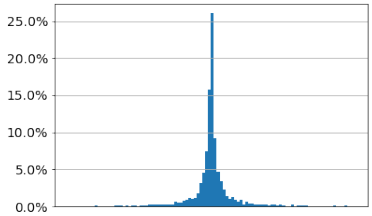
where $\boldsymbol{d} = (d_1, \ldots, d_m)$.

E-step. calculates the conditionally expected log-likelihood, where the expectation is conditioned on the parameters from the previous iteration step $p\ (\boldsymbol{\theta}^{(p)}, \boldsymbol{\pi}^{(p)})$ and the expectation is taken over the latent assignments.

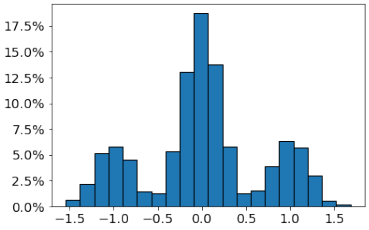
$$\begin{aligned} & Q(\boldsymbol{\theta}, \boldsymbol{\pi} | \boldsymbol{\theta}^{(p)}, \boldsymbol{\pi}^{(p)}) \\ = & \sum_{i=1}^{m} \mathbb{E}_{z|d} \Big[log\Big(p_i(d_i, z_i | \boldsymbol{\theta}, \boldsymbol{\pi}) \Big) | \boldsymbol{\theta}^{(p)}, \boldsymbol{\pi}^{(p)} \Big] \\ = & \sum_{i=1}^{m} \sum_{k=1}^{K} \mathbb{E}_{z|d} \Big[\mathbb{I}[z_i = k] | \boldsymbol{\theta}^{(p)}, \boldsymbol{\pi}^{(p)} \Big] \\ & \cdot \Big[log(\pi_k) + log(f_k(d_i|\boldsymbol{\theta})) \Big] \end{aligned}$$

Denote the posterior probability of each Gaussian mixture component k given each observation d_i as:

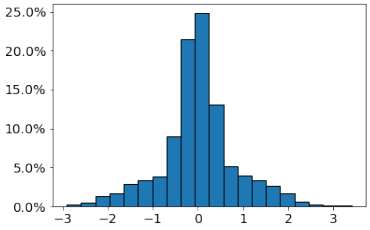
$$\begin{aligned} \omega_{i,k}^{(p)} &= \mathbb{E}_{z|d} \Big[\mathbb{I}[z_i = k] | \boldsymbol{\theta}^{(p)}, \boldsymbol{\pi}^{(p)} \Big] \\ &= \frac{\pi_k^{(p)} f_k(d_i | \boldsymbol{\theta}^{(\boldsymbol{p})})}{\sum_{j=1}^{K} \pi_j^{(p)} f_j(d_i | \boldsymbol{\theta}^{(\boldsymbol{p})})} \end{aligned} \tag{6}$$



(a) Histogram of observed effect sizes for certain outcome of interest among 3,300 real experiments.



(b) The simulated *true* effect sizes, with three clusters corresponding to positive, flat, and negative.



(c) The simulated *observed* effect sizes, with three clusters corresponding to positive, flat, and negative.

Figure 1: Illustration of the motivation of using GMM. The x-axis is the effect size and y-axis is the frequency. Data in (a) are from real experiments. The x-axis in this plot is not annotated owing to business confidentiality. (b) was simulated through a two-layer Gaussian Mixture Model (GMM) with mean values of $(-1, 0, 1)$, variances of $(0.2^2, 0.2^2, 0.2^2)$, and component weights of $(0.2, 0.6, 0.2)$. (c) was simulated from the same model as (b) with variances of $(0.7^2, 0.3^2, 0.7^2)$.

M-step. involves finding the optimal values for $\boldsymbol{\theta}, \boldsymbol{\pi}$ by maximizing the log-likelihood derived in the E-step. The parameters in the $(p+1)$-th iteration are updated as follows:

$$\pi_j = \frac{1}{m}\sum_{i=1}^{m}\omega_{i,j}^{(p)}, \tag{7}$$

where $\hat{\mu}_j$ and $\hat{\tau}_j^2$ are solved simultaneously by:

$$\mu_j = \Big[\sum_{i=1}^{m}\frac{\omega_{i,j}^{(p)} d_i}{(\sigma_i^2+\tau_j^2)}\Big] \Big/ \Big[\sum_{i=1}^{m}\frac{\omega_{i,j}^{(p)}}{(\sigma_i^2+\tau_j^2)}\Big] \tag{8}$$

$$\sum_{i=1}^{m}\frac{\omega_{i,j}^{(p)}}{\sigma_i^2+\tau_j^2} = \sum_{i=1}^{m}\omega_{i,j}^{(p)}\frac{(d_i-\mu_j)^2}{(\sigma_i^2+\tau_j^2)^2} \tag{9}$$

The EM algorithm is summarized in Algorithm 1. By setting $K = 3$ and $\mu_2 = 0$, the mean for positive components μ_1 serves as the estimate of AES. To avoid getting trapped in stationary point(Wu 1983), we repeat the EM algorithm multiple times with randomly initialized starting points.

Singularity issues in Gaussian mixture model. When the covariance matrix is singular, the variances becomes zero, it leads to a spiky Gaussian component that "collapses" into a single point. There are several papers discussing how to address this issue. The first approach involves setting a lower bound on variance (Hathaway 1985). The second approach introduces a penalty term in the log-likelihood function to prevent the variance of a specific component from becoming too small (Chen, Tan, and Zhang 2008; Ridolfi and Idier 2001). In addition, Chen, Tan, and Zhang (2008); Chen (2017) introduce the conditions and provide proofs for the consistency of the MLE under finite Gaussian mixture model. Using the second approach, the marginal log-likelihood takes the following form:

$$log(f(\boldsymbol{d}|\boldsymbol{\theta},\boldsymbol{\pi})) + log(p_m(\boldsymbol{\tau^2})) = \sum_{i=1}^{m}\sum_{k=1}^{K}[log(\pi_k)$$

$$+log(f_k(d_i|\boldsymbol{\theta}))] - \frac{1}{m}\sum_{k=1}^{K}(\frac{1}{\tau_k^2} + log(\tau_k^2)), \tag{10}$$

Algorithm 1: EM algorithm solving Three-Layer Heteroscedastic GMM

Input: $\boldsymbol{d} = (d_1, \ldots, d_m)$, $\boldsymbol{\sigma^2} = (\sigma_1^2, \ldots, \sigma_m^2)$, K

Output: $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_K, \tau_1, \ldots, \mu_K)$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$

1: Let converge = False.
2: Initialize model parameters in previous iteration as $\boldsymbol{\theta}^{(p)}, \boldsymbol{\pi}^{(p)}$.
3: **while** Not converge **do**
4: Compute values of $\omega_{i,k}^{(p)}$ using Equation (6), $\forall i, \forall k$.
5: Update $\boldsymbol{\theta}, \boldsymbol{\pi}$ based on Equations (7), (8) and (9)
6: Compute the marginal log-likelihood of the data $log(f(\boldsymbol{d}, |\boldsymbol{\theta}, \boldsymbol{\pi}))$ given the new parameters.
7: **if** Change in log-likelihood/m is less than the specified tolerance **then**
8: converge = True
9: **else**
10: $\boldsymbol{\theta}^{(p)} = \boldsymbol{\theta}, \boldsymbol{\pi}^{(p)} = \boldsymbol{\pi}$
11: **end if**
12: **end while**
13: **return** solution.
14: Repeat Step 1-12 multiple times using different initial points to obtain the global optimum.

where the penalty function $p_m(\boldsymbol{\tau^2})$ represents the product of K inverted Gamma distributions. Equation (9) can be revised to incorporate this penalty.

Extensions. This framework can be extended to any K with $K > 3$, using information criteria to determine the optimal K. Subsequently, the estimation of AES for $K > 3$ is a weighted average of the means of these positive components, where weights are also estimated in GMM.

Utility Theory: Bayesian Optimal Effect Size

In the previous section, we proposed a modeling approach that focuses on estimating the mean effect size. Such an ap-

proach is easy to explain to experimentation service users. However, in many real-world scenarios, users prioritize identifying the most suitable parameters for achieving maximum gains rather than just maximizing the accuracy of estimation. Motivated by the real needs, we propose a utility-maximization framework as our second approach.

The SSD problem is essentially a trade-off between information gain and decision accuracy. We define an utility (reward) function that considers all the related components, with the goal of maximizing the overall gain from experimentation. The utility include the following components:

1. Weekly experimentation cost c: we assume there is a fixed and pre-specified *weekly cost for running the experiment*, which may include the *personnel and hardware* to support this specific experiment to collecting more samples, the *opportunity cost for the experimentation service*, and the *opportunity cost for the experimenters*.
2. The *customers impact on the treatment group during the experiment*. For example, if the treatment has clear negative impacts on customers, running it longer will result in greater losses for company .
3. The *customers impact from making the launch recommendation* on weeks t. This term concerns the decision accuracy, i.e., we would like to launch those new features that indeed have positive customer impacts.

We demonstrate this approach with the problem of finding an one-size-fits-all default effect size. We assume there are m historical experiments. Our goal is to learn a default effect size that works best for all experiments on average, and we estimate the expectation using the empirical average over past experiments. The objective of finding the appropriate AES becomes maximizing the following utility function:

$$\begin{aligned}
&\arg\max_{\mu>0} \sum_{i=1}^{m} \mathbb{E}_{\delta_i\sim\mathcal{N}(d_i',\sigma_i^2)}\Bigg[\underbrace{-c_i\cdot(T_i(\mu)-1)}_{\text{Opportunity Cost}} \\
&+\underbrace{\left(\delta_i\cdot N_{i,1:T_i(\mu),T}\right)}_{\substack{\text{Impact during the experiment}\\ \text{(relative to control)}}}+\underbrace{u_2(\delta_i,H,T_i(\mu))\mathbb{I}(\pi(\boldsymbol{Y}_{T_i(\mu),i})=1)}_{\text{Launch impact (relative to control)}}\Bigg] \\
&=\arg\max_{\mu>0}\sum_{i=1}^{m}\Bigg[\underbrace{-c_i\cdot(T_i(\mu)-1)}_{\text{Opportunity Cost}} \\
&+\underbrace{\left(d_i'\cdot N_{i,1:T_i(\mu),T}\right)}_{\substack{\text{Impact during the experiment}\\ \text{(relative to control)}}}+\underbrace{u_2(d_i',H,T_i(\mu))\mathbb{I}(\pi(\boldsymbol{Y}_{T_i(\mu),i})=1)}_{\text{Launch impact (relative to control)}}\Bigg]
\end{aligned} \tag{11}$$

where the notations are as follows:

- δ_i is the per-customer per-week treatment effect for experiment i.
- d_i' is the posterior mean of the effect size at the end of the experiment and σ_i^2 is the posterior variance
- The subscript $i = 1, ..., m$ indexes the past similar experiments.
- $T_i(\mu)$ is the recommended duration week for experiment i, given AES μ. In Frequentist hypothesis testing, $T(\mu)$ is the minimum number of weeks required for its $power(\mu, \sigma_\cdot, n_\cdot, \alpha)$ in Equation (2) to exceed a predetermined threshold, such as 80%.
- $N_{i,1:t,g}$ denotes the total number of customers in group $g \in \{T, C\}$ observed up to week t for experiment i.
- c_i is the weekly opportunity cost for experiment i. c_i is proportional to the total sample sizes of each experiments.
- $u_2(\delta, H, t)$ represents the impact of launching the treatment feature on the whole population at week t, given a pre-specified time horizon H. Typically, we study one year impact, i.e. $H = 52$ weeks. One example of estimating one year launch impact is: $u_2(\delta, H, t) = \delta * (H - t) * (\sum_{g\in\{T,C\}} N_{i,1:t,g})$
- π is the decision rule. $\pi(.) = 1$ indicates the experiment is launched, 0 otherwise.
- $\boldsymbol{Y}_{t,i}$ contains all observed outcomes from customers triggered up to week t for experiment i.

The equality in Equation (11) is due to the linearity of expectation when u_2 is a linear function in the ATE (which is true in our case).

Optimization. Equation (11) is a one-dimensional optimization problem, for which we can efficiently find a good solution. We use grid search in our prototype.

Opportunity cost. One challenge to this approach is that how to select the opportunity cost c. This value is provided by business team or estimated through analysis of past launch experiments. Choosing a larger c will result in a greater AES, thus a shorter duration, as the longer experiments incurs higher costs. The impact of c has been studied in Figure 4 in Wan et al. (2023).

Experiments

We have proposed two approaches to select AES given past similar experiments. In this section, we compare the performance of both approaches against baseline methods.

Accuracy Comparison with Simulation

In this section, we compare the accuracy of different AES estimators. Since the ground truth is unknown in real data, we use simulation for this study. As the utility-based approach requires more information of experiments and its primary objective differs from the other approaches, we postpone its analysis to the next section.

Dataset. We simulate *observed* effect size from three-layer heteroscedastic GMM model in (4), with $K = 3$, $(\mu_1, \mu_2, \mu_3) = (2, 0, -2)$, $(\tau_1, \tau_2, \tau_3) = (0.5, 0.5, 0.5)$, $(\pi_1, \pi_2, \pi_3) = (0.2, 0.6, 0.2)$, $\sigma_i^2 \sim$ *Inverse-Gamma*$(3, 0.7)$, $m = 200$. The histogram of simulated d_i is shown in Figure 2. Due to the presence of heteroscedasticity σ_i^2 alongside the variance in each Gaussian component τ_k^2, it is challenging to distinguish individual components through visual inspection alone.

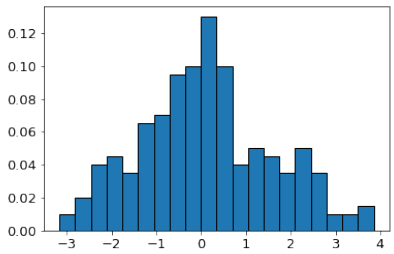


Figure 2: Histogram of simulated observed effect size d_i.

Metrics. To quantify the accuracy of AES estimators, we repeat the simulations $iter = 50$ times and use the Mean Square Error (MSE) and Mean Absolute Error (MAE):

$$MSE = \sum_{i=1}^{iter} \frac{(Estimation_i - Actual_i)^2}{iter},$$

$$MAE = \sum_{i=1}^{iter} \left| \frac{Estimation_i - Actual_i}{iter} \right|.$$

Methods and Results. Table 1 compares the MSE and MAE for the pooled effect size (Pooled-MLE), standard two-layer GMM (Two-layer GMM) and the proposed three-layer heteroscedastic GMM (Three-layer GMM). For Pooled-MLE, the MLE of parameter μ_0 is computed using the model in Equation (3) on all positive observed effect sizes d_i. Two-layer GMM uses Gaussian Mixture in *sklearn* package (Pedregosa et al. 2011) with $K = 3$. Three-layer GMM, the proposed method, uses the proposed EM-Algorithm 1 with the penalty term (10), $K = 3$ and $\mu_2 = 0$. Each simulation uses 10 different starting points to avoid local optimum (One of these starting points is initialized using the k-means clustering). Both Two-layer GMM and Three-layer GMM uses the estimated mean for the positive components as AES and $tolerance = 10^{-3}$. Table 1 shows that Three-layer GMM performs slightly better than Two-layer GMM. The p-value from a two-sample t-test is 0.036, indicating the Three-layer GMM has significantly better accuracy. As expected, Pooled-MLE is underestimated. GMM outperforms Pooled-MLE notably in cases where we lack information about whether the experiment's effect is a true positive or not. The Boxplot in Figure 3 illustrates a similar conclusion.

	MSE	MAE
Pooled-MLE	0.709	0.842
Two-layer GMM	0.017	0.180
Three-layer GMM	0.003	0.137

Table 1: Accuracy Comparison: MSE and MAE comparison among pooled effect size, standard two-layer GMM and the proposed three-layer heteroscedastic GMM.

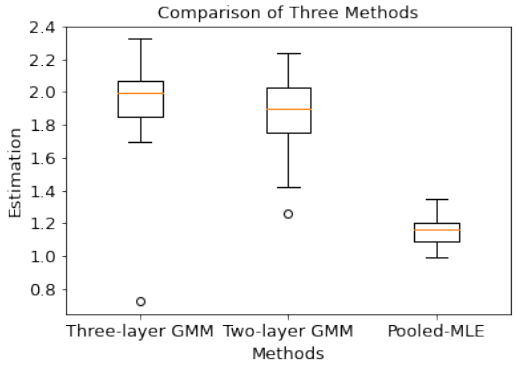


Figure 3: Boxplots comparing the AES estimations among pooled effect size, standard two-layer GMM and the proposed three-layer heterocasdestic GMM. Ground truth is 2.

Meta-Analysis with Real Experiments

In the previous section, we evaluated the estimation accuracy of different methods. However, the accuracy is just one facet to consider. Recall that the choice of the AES is always a trade-off between experimentation cost and accuracy. Therefore, it is hard to define a single optimal solution. In this section, we compare the two proposed approaches with baseline methods on real-world experimental data in terms of their empirical performance over a few metrics.

Dataset and setup. We collect a dataset of 3,300 historical experiments conducted in the past two years within Amazon, each having a duration of 4 weeks. d_i represents the observed standardized effect size derived from a specific standardization formulation used in the company at week 4. Given the absence of the ground-truth effect size, we adopt a heuristic yet easy-to-explain approach that uses the observed effect size at the end of the 4 weeks as the empirical ground-truth effect size.

We set the maximum duration as 4 weeks for the analysis. For each estimated AES obtained from different methods, we plug in them, along with the observed sample size and sample variance (or their predicted versions as described in Richardson et al. (2022)), into power Equation (2) to calculate the statistical power at week $i = 1, 2, 3, 4$. We define the recommended duration as the minimum number of weeks ($\leq$ 4) required to attain a power of 80%. In Frequentist setting, we use the one-sided two-group Welch's t-test (Welch 1947) as decision policy $\pi(.)$. This implies that the decision policy $\pi(.) = 1$ for launch if the p-value is less than the significance level of 0.05 and the ATE is positive.

Methods. We compare the proposed estimated AESs from Three-Layer GMM and utility theory-based optimal effect size (Utility-maximization) with those from two-layer GMM and Pooled-MLE. The settings for Two-layer GMM, Three-layer GMM and Pooled_MLE are the same as the previous section. Utility-maximization uses grid-search to find the optimal effect size within the set $\{0.02\%, 0.04\%, \ldots, 2\%\}$.

Metrics. We consider the following metrics:

1. The percentage of empirical false positives (proportion

	Estimated AES (%)	(Empirical) False Positive	(Empirical) False Negative	Avg Weeks	Avg Opportunity Cost (D)	Avg Launch Impact (D)	Avg Impact During Exper (D)	**Avg Reward** (D)
Pooled-MLE	0.05%	0.0%	0.0%	3.99	1.0	2.61	0.01	1.61
Two-layer GMM	0.22%	0.29%	0.29%	3.63	0.41	2.42	-0.0	2.01
Three-layer GMM	0.15%	0.13%	0.11%	3.83	0.65	2.55	0.01	1.91
Utility-maximization	1.58%	1.17%	0.99%	2.3	0.02	2.22	0.0	**2.2**

Table 2: Meta-analysis results. Recall that all utility-related metrics share the same unit D, the meaning of which is omitted due to confidentiality.

	Estimated AES	(Empirical) False Positive	(Empirical) False Negative	Avg Weeks	Avg Opportunity Cost (10^4)	Avg Launch Impact (10^4)	Avg Impact During Exper (10^4)	**Avg Reward** (10^4)
Pooled-MLE	0.782%	0.13%	0.1%	4.0	2.8	5.88	0.0	3.08
Two-Layer GMM	1.041%	0.13%	0.27%	3.89	2.53	5.83	0.01	3.31
Three-layer GMM	**1.005**%	0.13%	0.17%	3.92	2.6	5.87	0.01	3.28
Utility-maximization	1.100%	0.13%	0.37%	3.82	2.41	5.8	0.01	**3.41**

Table 3: Simulated data analysis results (Ground truth is 1).

of incorrectly detecting significantly positive effect when the true effect is flat/negative) and the percentage of empirical false negatives (proportion of incorrectly detecting flat/negative effect when the true effect is indeed positive). These two metrics reflect the (empirical) Type-I and Type-II errors.

2. The average utility in Equation (11) and its three components, including the opportunity cost, the impact during the experiment, and the launch impact.

Results. We present results from 3300 experiments in Table 2. The pooled MLE has the smallest decision errors. However, it comes at the expense of requiring longer experiment duration, leading to higher costs. The proposed Utility-maximization method outperforms other methods and generates the highest average cumulative reward, with a desired balance between the cost of experimentation and making correct launch decisions.

More simulations. Due to confidentiality constraints, the above dataset cannot be shared. To ensure reproducibility, we conducted a simulation study with 3000 simulated experiment trajectories at weeks $1, \dots 4$: (1) Simulate weekly sample size following beta-geometric distribution (Richardson et al. 2022). The beta distribution parameters, α and β, are drawn from uniform distributions: $\alpha \sim Uniform(0.1, 1)$, $\beta \sim Uniform(4, 60)$. Each arm assumes a total of 10K customers. (2) Sample the observed effect size $d_{i,t}$ for experiment i at week t from the three-layer GMM model (4) with three Gaussian components $(\mu_1, \mu_2, \mu_3) = (-1, 0, 1)$, $(\tau_1, \tau_2, \tau_3) = (0.3, 0.5, 0.3)$, $(\pi_1, \pi_2, \pi_3) = (0.2, 0.6, 0.2)$. Given $\sigma_g^2 = 500$, σ_i^2 are computed using sample size and Equation (1). (3) Total weekly opportunity cost is $c = 4 \cdot 10^6$, decomposed to each experiment according to their final weeks' sample sizes. The observed effect size on last week $\{d_{i,4}\}_{i=1}^m$ are used to fit MLE and GMM methods. Utility-maximization uses grid-search to find the optimal effect size within the set $\{0.1, 0.2, \dots, 5\}$. Results in Table 3 demonstrate that the Three-layer GMM yields the most precise estimation (ground truth is 1), whereas the utility-based approach achieves the highest utility.

Conclusion

AES places an central role in duration recommendation, yet little attention has been drawn to its specification, particularly for large online experimentation services. In this paper, we propose two approaches to estimate AES from a large number of historical online experiments. The first approach introduces a novel three-layer GMM to account for experiment heteroscedasticity. The second approach finds the optimal AES that maximizes the expected utility. Through simulations and a large-scale meta-analysis using real experiments from Amazon, we conclude that the first proposed ensures a high estimation accuracy and the second proposed approach leads to a significant gain in the expected utility. With the provided flexibility, one can choose either of these approaches to achieve their specific goal.

For the GMM-based method, exploring other hierarchical models and leveraging experiment-specific features to recommend personalized effect size are important next steps. Bayesian non-parametric models (Orbanz and Teh 2010) or structures that have been explored in bandits (Wan, Ge, and Song 2021) are good starting points. For the utility-based method, we can also easily extend it to provide personalized effect size (and hence personalized duration) recommendation, by replacing the end-of-horizon posteriors of the m experiments with the posterior for the target experiment at the duration recommendation time point. Besides, exploring the performance with other ATE estimators such as covariate adjusted estimators (Masoero, Hains, and McQueen 2023) is also an interesting next step.

References

Adcock, C. 1997. Sample size determination: a review. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(2): 261–283.

Casella, G.; and Berger, R. L. 2021. *Statistical inference*. Cengage Learning.

Chen, J. 2017. Consistency of the MLE under Mixture Models. *Statistical Science*, 32(1): 47–63.

Chen, J.; Tan, X.; and Zhang, R. 2008. Inference for normal mixtures in mean and variance. *Statistica Sinica*, 443–465.

Cohen, J. 2013. *Statistical power analysis for the behavioral sciences*. Academic press.

Du, H.; and Wang, L. 2016. A Bayesian power analysis procedure considering uncertainty in effect size estimates from a meta-analysis. *Multivariate behavioral research*, 51(5): 589–605.

Hathaway, R. J. 1985. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *The Annals of Statistics*, 13(2): 795–800.

Hoenig, J. M.; and Heisey, D. M. 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55(1): 19–24.

Kelley, K.; and Rausch, J. R. 2006. Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological methods*, 11(4): 363.

Lenth, R. V. 2001. Some practical guidelines for effective sample size determination. *The American Statistician*, 55(3): 187–193.

Lindley, D. V. 1997. The choice of sample size. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(2): 129–138.

Masoero, L.; Hains, D.; and McQueen, J. 2023. Leveraging covariate adjustments at scale in online A/B testing. In *The KDD'23 Workshop on Causal Discovery, Prediction and Decision*, 25–48. PMLR.

Orbanz, P.; and Teh, Y. W. 2010. Bayesian Nonparametric Models. *Encyclopedia of machine learning*, 1: 81–89.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Ramsey, F.; and Schafer, D. 2012. *The statistical sleuth: a course in methods of data analysis*. Cengage Learning.

Richardson, T. S.; Liu, Y.; McQueen, J.; and Hains, D. 2022. A bayesian model for online activity sample sizes. In *International Conference on Artificial Intelligence and Statistics*, 1775–1785. PMLR.

Ridolfi, A.; and Idier, J. 2001. Penalized maximum likelihood estimation for univariate normal mixture distributions. In *AIP Conference Proceedings*, volume 568, 229–237. American Institute of Physics.

Wan, R.; Ge, L.; and Song, R. 2021. Metadata-based multi-task bandits with bayesian hierarchical models. *Advances in Neural Information Processing Systems*, 34: 29655–29668.

Wan, R.; Liu, Y.; McQueen, J.; Hains, D.; and Song, R. 2023. Experimentation platforms meet reinforcement learning: Bayesian sequential decision-making for continuous monitoring. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*.

Weiss, R. 1997. Bayesian sample size calculations for hypothesis testing. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 46(2): 185–191.

Welch, B. L. 1947. The generalization of 'STUDENT'S'problem when several different population varlances are involved. *Biometrika*, 34(1-2): 28–35.

Wu, C. J. 1983. On the convergence properties of the EM algorithm. *The Annals of statistics*, 95–103.