Optimizing duration of online experiments via Bayesian early termination

Lorenzo Masoero, Melany Gualavisi
Amazon
{masoerl, melanygd}@amazon.com

Determining appropriate experimental duration remains a challenging problem in online experimentation. While experimenters ideally would know in advance how long to run experiments in order to inform confident business decisions, many factors affecting conclusiveness of their results are difficult to predict prior to the experiment. Consequently, experimentation services develop "in-flight" tools that suggest early termination based on observed data. We present a novel optimization framework to guide duration and early termination of online experiments, in particular with the goal of detecting with confidence whether an online experiment can be interrupted earlier than planned on the basis of partial evidence gathered from early results. We formulate a general optimization problem yielding an algorithm recommending early termination of ongoing online experiments when certain conditions of partial "early" results, obtained from collecting evidence from a shorter duration than originally planned, are met. In our algorithm, the specific parameters determining whether the experiment at hand can be cut short are functions of both experiment metadata, desired launch criteria specified by the experimenter, and past historical results. Using production data from real online experiments, our metadata-aware approach demonstrates that we can achieve high rates of early termination — up to more than 10% of total experiments run — while maintaining low error rates. Here, errors represent cases where partial results from early termination suggest outcomes that differ from what would be observed if the experiment were run to completion, and our approach successfully minimizes such misaligned recommendations. Our approach reveals intuitive patterns: high-stringency experiments use conservative thresholds while lenient criteria allow aggressive early decisions. This work demonstrates that principled metadata-aware optimization can dramatically improve early termination systems while maintaining statistical rigor.

1. Introduction Online A/B testing has become fundamental to data-driven decision-making, with thousands of experiments running continuously across digital services. While standard practice typically involves running experiments for fixed durations (e.g., 4 weeks), there is increasing demand to identify cases where experiments could conclude earlier, either due to clear success or futility. Early termination can reduce experimental costs, accelerate innovation cycles, and minimize user exposure to potentially harmful treatments. Early Launch Decision (ELD) tools [Deng et al., 2016, Wan et al., 2023] address this need by continuously monitoring experimental results and providing recommendations to terminate experiments before their planned duration. These tools suggest early termination when data shows strong evidence regarding launch criteria metrics, either indicating the experiment will meet success criteria or show strong evidence of harm. We here focus on the setting in which experimenters adopt a Bayesian decision making framework. We provide a general algorithm for early termination of online experiments. Our contributions are threefold: (1) We develop a theoretical framework for duration-dependent early termination thresholds that accounts for experimental uncertainty evolution; (2) We present two optimization approaches—week-specific thresholds and differential evolution accounting for metric characteristics; (3) We demonstrate significant improvements using production data while maintaining error rates below target levels.

2. Methodology and statistical framework We use ω to denote an individual experiment and m to denote a metric under investigation. For experiment ω and metric m, let δ_{ω_m} represent the unknown treatment effect of the intervention. We employ a conjugate Bayesian approach to model the average treatment effects, where both (i) the prior distribution of the treatment effect and (ii)

the likelihood model describing the observed effect follow a normal distribution. In this framework, the posterior distribution describing the average treatment effect follow normal distributions via conjugacy. Let $\mu_{\omega_m}^{(d)}$ and $\sigma^{(d)}\omega_m$ denote the posterior mean and standard deviation at duration d.

The Probability of Positive Return (PPR) is: $\text{PPR}_{\omega_m}^{(d)} = P(\delta_{\omega_m} > 0 \mid \mathcal{D}^{(d)}) = \Phi\left(\frac{\mu_{\omega_m}^{(d)}}{\sigma_{\omega_m}^{(d)}}\right)$. Given a "hurdle" value h_{ω_m} – e.g. encoding a notion of cost that the intervention has to overcome to be successful, the posterior odds are: $\text{PO}_{\omega_m}^{(d)} = \frac{1-P(\delta_{\omega_m} > h_{\omega_m} | \mathcal{D}^{(d)})}{P(\delta_{\omega_m} > h_{\omega_m} | \mathcal{D}^{(d)})}$.

Early Launch Decision Algorithm The ELD algorithm provides recommendations based on launch criteria $\text{LC}_\omega = \{\text{LC}_{\omega_m}, m \in \mathcal{M}_\omega\}$. Each criterion LC_{ω_m} is defined by threshold λ_{ω_m}, direction $d_{\omega_m} \in \{\text{GE}, \text{LE}\}$, and hurdle h_{ω_m}. Given thresholds Θ, ELD proceeds in two steps: We first compute direction-agnostic recommendation:

$$\text{ELD}_{\text{no-dir},\omega_m}^{(d)} = \begin{cases} -1 & \text{if } \text{PO}_{\omega_m}^{(d)} < (\theta_{\omega,m}^{(d)})^{-1} \\ 0 & \text{if } (\theta_{\omega,m}^{(d)})^{-1} \leq \text{PO}_{\omega_m}^{(d)} \leq \theta_{\omega,m}^{(d)} \\ 1 & \text{if } \text{PO}_{\omega_m}^{(d)} > \theta_{\omega,m}^{(d)} \end{cases} \tag{1}$$

After adjusting for launch criterion direction (flipping signs if experimenters seek a negative effect on the metric of interest), we aggregate across metrics to produce final recommendation. We evaluate ELD by comparing early recommendations against final results. Define:

- $\text{TotalEarlyRec}^{(d)}(\Omega, \Theta)$: total flagged experiments at duration d
- $\text{ErrorRateOnFlagged}^{(d)}(\Omega, \Theta)$: error rate among flagged experiments

Our scoring function balances decision rate and accuracy:

$$\text{Score}^{(d)}(\Omega, \Theta) = \frac{\text{ErrorRateOnFlagged}^{(d)}(\Omega, \Theta)}{\text{TotalRateEarlyRec}^{(d)}(\Omega, \Theta)} \tag{2}$$

For durations $D = \{d_1, \ldots, d_K\}$, we aggregate scores and solve:

$$\Theta^* \in \arg\max_{\Theta} \left\{ \sum_{d \in D} \text{Score}^{(d)}(\Omega, \Theta) \right\} \tag{3}$$

subject to $\text{ErrorRateOnFlagged}^{(d)}(\Omega, \Theta) \leq \epsilon$ for all d, and monotonicity constraint $\theta^{(d_1)} \geq \theta^{(d_2)}$ when $d_1 \leq d_2$. The key insight of our approach is that optimal thresholds should depend on experiment metadata, particularly launch criteria specifications. We formulate a general optimization framework where thresholds $\theta^{(d,\mathbf{x})}$ are functions of duration d and metadata vector $\mathbf{x}$. For launch criteria metadata, we consider the statistical threshold λ_{ω_m} that defines the stringency of success criteria. Experiments with lenient criteria ($\lambda = 0.33$) should use different thresholds than those with stringent criteria ($\lambda = 0.8$). This principle extends to other metadata: experiment domain, sample size, effect size expectations, and business criticality. We cluster experiments based on metadata similarity and optimizing group-specific thresholds $\theta^{(d,g)}$ where g represents metadata clusters. For launch criteria, we define groups by stringency: low ($\lambda \in [0, 0.4]$), medium ($\lambda \in (0.4, 0.6]$), and high ($\lambda \in (0.6, 1]$). We use differential evolution to solve the constrained optimization problem across the metadata-duration space, with 5-fold cross-validation ensuring robustness.

Experiments and Results Our analysis uses a sample of experimental data from a large-scale online experimentation services. After filtering for complete results at durations 7, 14, 21, and 28 days and removing experiments with incomplete data or anomalous results, our dataset contains 2,941 unique experiments with 6,105 rows. The baseline system employs fixed thresholds across all weeks: $\theta_m = 5.1$ for one primary metric and $\theta_m = 14$ for others. This approach identifies 93 experiments (3.11%) for early decisions with an error rate of 10.75% using our alternative error definition (not counting inconclusive terminations as errors).

We then define a metadata-aware approach to demonstrate the power of tailoring thresholds to experiment characteristics. By clustering experiments based on launch criteria stringency and optimizing group-specific thresholds, we achieve substantial improvements. The learned thresholds

reveal intuitive patterns: high-stringency experiments (requiring strong evidence) use more conservative thresholds, while low-stringency experiments allow more aggressive early decisions. This metadata-aware approach increases early decisions to 11.24% (926 experiments), representing a 261% improvement over fixed thresholds while reducing error rates to 8.04%.

Table 1: Comparison of threshold optimization approaches

	Fixed Threshold	**Week-Specific Threshold**	**Differential Evolution**
Early decision rate	3.11%	7.83%	11.24%
Early launches	54	131	181
Early terminations	39	103	155
Error rate	10.75%	9.40%	8.04%

Our results demonstrate three critical insights: (1) Duration-dependent thresholds significantly outperform fixed approaches, with more sophisticated optimization yielding better performance; (2) The differential evolution approach excels at identifying inconclusive experiments—48.4% of its early terminations prove inconclusive at week 4, representing valuable resource savings; (3) Error definition fundamentally shapes evaluation—while traditional definitions show increasing error rates with coverage, our alternative definition reveals improving accuracy.

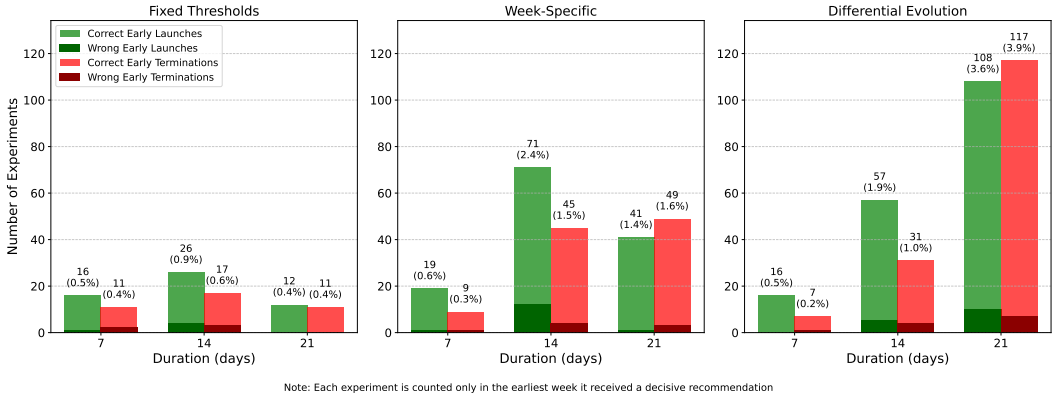

Figure 1: Weekly early decisions flagged by different approaches. Fixed thresholds (left), week-specific thresholds (middle), and differential evolution (right).

Conclusion We presented a novel optimization framework for early termination decisions in online experimentation, with the key insight that optimal thresholds should be functions of both experimental duration and metadata characteristics, particularly launch criteria specifications. Our metadata-aware approach achieves substantial improvements—261% increase in early termination coverage (3% to 11%) while improving precision (11% to 8% error rate). The framework's core principles of duration-dependent uncertainty and constrained optimization apply broadly to sequential decision-making in experimentation, contributing to more effective and resource-conscious experimentation practices while maintaining statistical rigor. Future work will be devoted towards further tailoring metadata for even better results.

References

A. Deng, J. Lu, and S. Chen. Continuous monitoring of A/B tests without pain: Optional stopping in Bayesian testing. *arXiv preprint arXiv:1602.05549*, 2016.

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