Leveraging Structural Information about Treatments to Improve Adaptive Experimentation

Jennah Gosciak[1,2] Blake Mason[2] Hammaad Adam[2] Mahnaz Islam[2]

September 2, 2026

Adaptive experiments are increasingly used in industry. They allow experimenters to test many more treatments than in a typical randomized experiment. A defining feature is the use of an exploration phase in which the goal is to identify a subset of the most effective treatments to test in a subsequent validation phase. There are a variety of simple algorithms that companies can use to explore a large set of treatments. However, these algorithms still require relatively large sample sizes and budgets to obtain decent performance. They also make relatively few assumptions.

For example, they typically make few assumptions about the *structure* of the treatment effects themselves – i.e., the underlying model that explains how treatment attributes relate to the treatment effects. In practice, this can be useful and robust across a range of real data settings. But companies often do have access to a rich set of descriptive information *about the treatments*, which they might be able to exploit to efficiently explore the treatment space. In other words, in settings where we might be fairly confident that structure matters, can it help us explore more efficiently?

In this paper, we investigate whether accounting for structure meaningfully improves exploration in adaptive experiments, given the practical constraints companies face and the types of experiments run. We define structure broadly: structure can refer to a linear additive model on a vector of treatment attributes or a Gaussian Process that models correlation based on treatment arm similarity. We propose two modifications to the Sequential Halving (Sh) algorithm [2], a simple and widely used fixed-budget algorithm for best-arm identification: (1) we predict arm rankings using available feature information, and (2) we allocate the sample budget to minimize the worst-case estimation error [1, 6]. We show via simulations in realistic experiment settings that these modifications can improve the quality of selected treatments relative to Sh. However, the improvements come at a cost; they can yield *worse* results when assumptions about structure are miscalibrated or incorrect. We frame these results as a bias-variance trade-off, made especially salient through the lens of Gaussian process (GP) estimation. Practitioners can weigh this trade-off when deciding whether to leverage structure in their own experiments. We also highlight the benefits of GPs for experimentation more broadly. In settings where stronger assumptions about structure are reasonable, GPs offer a more informed prior, yielding more precise estimates.

1 Adapting SH to Account for Structure

The conventional approach to Sh proceeds in rounds based on a fixed sample budget. In each round, sample is allocated equally among all "active" arms. At the end of each round, arms are ranked by their sample means and the bottom half are permanently eliminated. After all rounds have concluded, the surviving treatment(s) are considered the winners.

SH makes no assumptions about the structure of treatments. However, correctly assuming structure on the treatments can reduce the problem complexity.[1] Consider a simple UI experiment. Treatment effects might reasonably follow a linear additive model – e.g., $\theta_1 \text{COLOR} + \theta_2 \text{BOLDING} + \theta_3 \text{POSITION}$. To generalize, in a variety of experiment settings, we can model the

[1]Department of Information Science, Cornell University

[2]Amazon

[1]As Fiez et al. [1], Soare et al. [6] demonstrate, the lower bound on the sample complexity for linear bandit algorithms in the worst case scales with dimension d as opposed to the number of treatment arms K. The benefits of linear bandit algorithms are thus most pronounced when $d \ll K$.

mean treatment effects by some unknown function of observable attributes $f^*(x)$. Given the form of $f^*(x)$, we can then exploit this structure to improve SH. Specifically, we propose two modifications: (1) arm ranking and elimination decisions and (2) sample allocation in each round. These simple algorithm modifications, derived from information-theoretic lower bounds, are straightforward to incorporate into existing SH architectures. We describe our modifications to SH in Table 1.

1. **Arm ranking and elimination** Instead of ranking arms by their sample means, we rank them by model predictions based on treatment attributes. This method can share information across arms that are similar, yielding more precise and accurate rankings. We focus on predictions from Gaussian Process (GP) models. GPs are flexible, account for nonlinearities in the data, and require fewer assumptions.
2. **Sample allocation** SH allocates the sample budget uniformly across all arms. This can be inefficient, particularly if there are redundancies in the data. If we assume there is structure, it may be more efficient to sample only the most informative arms as opposed to all arms. To formalize this, we draw on the theory of $\mathcal{XY}$-optimal experimental design [1, 6], which seeks to minimize the worst-case prediction error. In each round of SH, we compute the optimal $\mathcal{XY}$ allocation, and use this solution to allocate the sample budget.

2 GP Estimation Provides a Flexible Approach

In addition to modifying SH to account for structure, we also model structure using Gaussian Processes (GPs). GPs are a flexible modeling tool because they are nonparametric, impose minimal modeling assumptions, and seamlessly generalize from existing approaches to Bayesian inference (e.g., Kessler [3] propose Bayesian inference to counteract the winner's curse). In fact, GP predictions are functionally Bayesian posterior estimates, with a prior on the correlation structure across arms in addition to the marginal arm variance.

For K treatment arms, feature vectors $\mathbf{x}_1, \ldots, \mathbf{x}_K$, and a 0 mean prior, the GP posterior is [5]:

$$\boldsymbol{\mu}_{\text{post}} = \mathcal{K}(\mathcal{K} + \Gamma)^{-1}\hat{\boldsymbol{\delta}}, \tag{1}$$

where $\mathcal{K}$ is the kernel matrix with $\mathcal{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ and $\Gamma = \text{diag}(\hat{\tau}_1^2, \ldots, \hat{\tau}_K^2)$. If we set $\mathcal{K} = \sigma^2_{\text{prior}} I$ (independent prior on each arm), the posterior reduces to the single-dimensional case, estimating Bayesian posterior means per-arm with no assumed correlation.

For the choice of $\mathcal{K}$, we use a radial basis function (RBF) kernel, $k(x_i, x_j) = \exp\left(-\|x_i - x_j\|^2/2\ell^2\right)$ (results are similar if using a Matern kernel). This sets up a useful trade-off for practitioners to consider. The GP fit is governed by the choice of lengthscale ℓ. Lengthscale ℓ controls the smoothness of the fit and the amount of correlation between arms. As $\ell \to 0$, the kernel converges to the identity: arms are treated as independent and the GP functionally returns sample mean estimates. As ℓ grows, arm similarity matters more and estimates borrow information from neighboring arms, which can yield more precise GP predictions. Eventually, when $\ell \gg 0$, the GP converges to a flat, constant function, shrinking the estimates to the prior mean. Assuming a larger lengthscale reduces variance and improves estimation quality, but the assumption can also lead to bias if the approximation is wrong. **The GP framing thus provides a useful mechanism for clarifying the bias-variance trade-off that practitioners face in deciding whether to use structure.**

3 Discussion and Takeaways

Drawing on several real-data experiment settings, we simulate the performance of GP-RAGE relative to SH. Simulations assume that prior information, such as the signal variance and prior mean, is known. This isolates the effect of the lengthscale ℓ, which we vary separately, on algorithm performance. In practice, the signal variance and prior mean are additional hyperparameters that experimenters would need to set and test. The choice of these hyperparameters can substantially

affect performance as well. For example, incorrectly setting the prior signal variance (also known as the kernel amplitude) can either shrink estimates toward the prior mean (without any regard for the data) or render the prior meaningless. Based on simulations, a subset of which are shown in Figure 1, we synthesize the following takeaways about leveraging structure:

Structure Can Modestly Improve Decision Quality We find that accounting for structure does modestly improve the quality of selected treatments by several percentage points. Specifically, we measure the average percentile ranking of the winning treatment under the marginal prior distribution of treatment effects. This metric does not penalize an algorithm for returning a "good" arm over the best arm – arms of similar quality can have comparable rankings on average.

Improvements in performance are most noticeable when the data is relatively smooth, simulated from a GP with moderate and large lengthscale ratios $\ell/d_{\text{med}} = 0.7, 1.5$. We define ℓ relative to the median pairwise distance d_{med} across arms.[2] Performance is highest when the lengthscale ratio matches the smoothness of the data. When the lengthscale is too low, GP-RAGE behaves similarly to SH – the GP essentially treats arms as independent – and therefore we observe little difference in performance.

There Are Substantial Risks to Oversmoothing The simulations demonstrate that there is a clear risk to *over-smoothing* – choosing a lengthscale ratio that is too large for the data. When $\ell_{\text{est}} > \ell_{\text{data}}$, there are steep performance decreases. This underscores that practitioners need to be careful when deciding whether to use structure. Such decisions should err conservatively; larger lengthscales should only be chosen when there is clear evidence that the treatment estimates are smooth. At lower lengthscales, GP-RAGE may not offer sufficient benefits over existing methods like SH to justify the added complexity. As we discuss below, estimating structure from historical data is already challenging. Experimenters may find that risks are severe, especially given uncertainty about whether sufficient structure exists in the data.

Leveraging Structure May Vary by Experiment Type A promising direction is that structure may be more or less plausible in different kinds of experiment settings. Specifically, we find that improvements vary based on characteristics about the type of experiment. This is an important practical consideration for real-world adoption. Figure 1 contrasts two examples: a simple UI change vs. hyperparameter tuning. In industry, many experiments test simple UI changes, often in a factorial setting. These experiments are characterized by binary or categorical treatment attributes that result in relatively high-dimensional features. In factorial settings with binary attributes, the benefit of accounting for structure is low. There are still a substantial number of features relative to the number of treatments, which does not reduce the complexity of the problem substantially and limits borrowing across arms. As a result, the benefits are especially reduced in *fractional* factorial experiments, where only a subset of the feature combinations is tested (e.g., testing 20 treatments as opposed to all 64 combinations).

In contrast to UI experiments, we consider hyperparameter tuning. In hyperparameter tuning experiments, the goal is typically to identify what kind of deterministic configurations perform the best from a set of continuous or integer-valued treatment attributes. These experiments are easier to set up and typically lower-dimensional. This makes it easier for GP-RAGE to exploit structure and performance increases are much more pronounced.

Making Decisions in Practice Our simulation results hinge on the validity of these assumptions in real-world settings. Critically, estimating structure from historical experiments is not straightforward. One can fit different models to historical treatment estimates, but how should we evaluate whether these models correctly capture structure in the data? We propose two methods for testing

[2]Normalizing by d_{med} makes ℓ comparable across experiments with different numbers of arms and attributes.

the validity of structure, and pose offline evaluation from historical data as an open question for future work.

First, we consider whether model predictions reduce error using Leave-One-Out Cross-Validation (LOOCV). We perform LOOCV by predicting the treatment effect for a held-out arm using data for all other arms. LOOCV is useful for assessing whether structural information has genuine out-of-sample predictive value. It asks: *if we evaluate an unseen treatment arm with this model, can we predict its effect from features alone?* While LOOCV may be a helpful first diagnostic, it may impose overly stringent standards: out-of-sample generalization is not necessarily needed for GP-RAGE to exploit the structure. Weak correlations across arms, even if these correlations do not generalize to new, unseen treatment arms may still be useful. LOOCV does not reflect the kinds of realistic experiment settings that practitioners mainly face: all treatment arms are observed, but estimates are uncertain due to sampling noise. Additionally, it can be more unstable in settings where the number of treatment arms is low.

Next, we consider a version of sample-splitting that assesses model fit across different simulated customer samples, as opposed to eliminating an entire treatment arm. Conveniently, Kessler et al. [4] propose a version of sample splitting that only relies on sufficient statistics like sample means and variances. Sample splitting asks: *if we re-ran this experiment with different customers, would using a structural model consistently produce better estimates of the treatment effects relative to the status quo (pooled sample means)?* Sample-splitting thus models, more realistically, uncertainty in replicating the experiment over time, with different customer samples. We prefer the results from sample-splitting, though sample-splitting and LOOCV should ideally align when structure exists in the data.

Both methods are useful for assessing the validity of assumptions about structure. Testing these assumptions is critical: without assuming some structure, investing in algorithms like GP-RAGE will make no difference; assuming too much structure, and experimenters risk worse performance. Open questions for future work include: are sample-splitting and LOOCV sufficient and useful methods for evaluating structure on historical data, are there specific challenges posed by the use of *adaptively* collected historical data for such evaluations, and how do the number of treatment arms vs. the number of total experiments affect the evaluation quality.

References

[1] Tanner Fiez, Lalit Jain, Kevin G Jamieson, and Lillian Ratliff. 2019. Sequential experimental design for transductive linear bandits. *Advances in neural information processing systems* 32 (2019).

[2] Zohar Karnin, Tomer Koren, and Oren Somekh. 2013. Almost optimal exploration in multi-armed bandits. In *International conference on machine learning*. PMLR, 1238–1246.

[3] Ryan Kessler. 2022. Overcoming the Winner's Curse: Leveraging Bayesian Inference to Improve Estimates of the Impact of Features Launched via A/B Tests. (2022).

[4] Ryan Kessler, James McQueen, and Miikka Rokkanen. 2025. Evaluating A/B Testing Methodologies via Sample Splitting: Theory and Practice. *arXiv preprint arXiv:2512.03366* (2025).

[5] C. E. Rasmussen and C. K. I. Williams. 2006. Gaussian processes for machine learning: book webpage.

[6] Marta Soare, Alessandro Lazaric, and Rémi Munos. 2014. Best-arm identification in linear bandits. *Advances in neural information processing systems* 27 (2014).

Algorithm	Arm Ranking Method	Per-Round Allocation	Motivation
SH	Pooled sample means	Uniform	• Baseline method used currently in pproduction
SH-OLS	OLS fitted values	Uniform	• Improve ranking of arms for elimination by estimating arm rewards using OLS
RAGE	OLS fitted values	$\mathcal{XY}$-allocation	• Improve ranking of arms for elimination by estimating arm rewards using OLS • Allocate sample to reduce the worst-case estimation error
GP	GP posterior means	Uniform	• Relax linearity assumption and estimate arm rewards via GP posterior means
GP-RAGE	GP posterior means	$\mathcal{XY}$-allocation	• Relax linearity assumption and estimate arm rewards via GP posterior means • Allocate sample to reduce the worst-case estimation error

Table 1. Summary comparison of proposed algorithms and the modifications to the original SH approach. Following Fiez et al. [1], we use the term RAGE to denote any algorithm that uses an $\mathcal{XY}$-optimal allocation. RAGE uses OLS fitted values; GP-RAGE uses GP posterior means. Both allocate sample according to the optimal $\mathcal{XY}$-allocation.

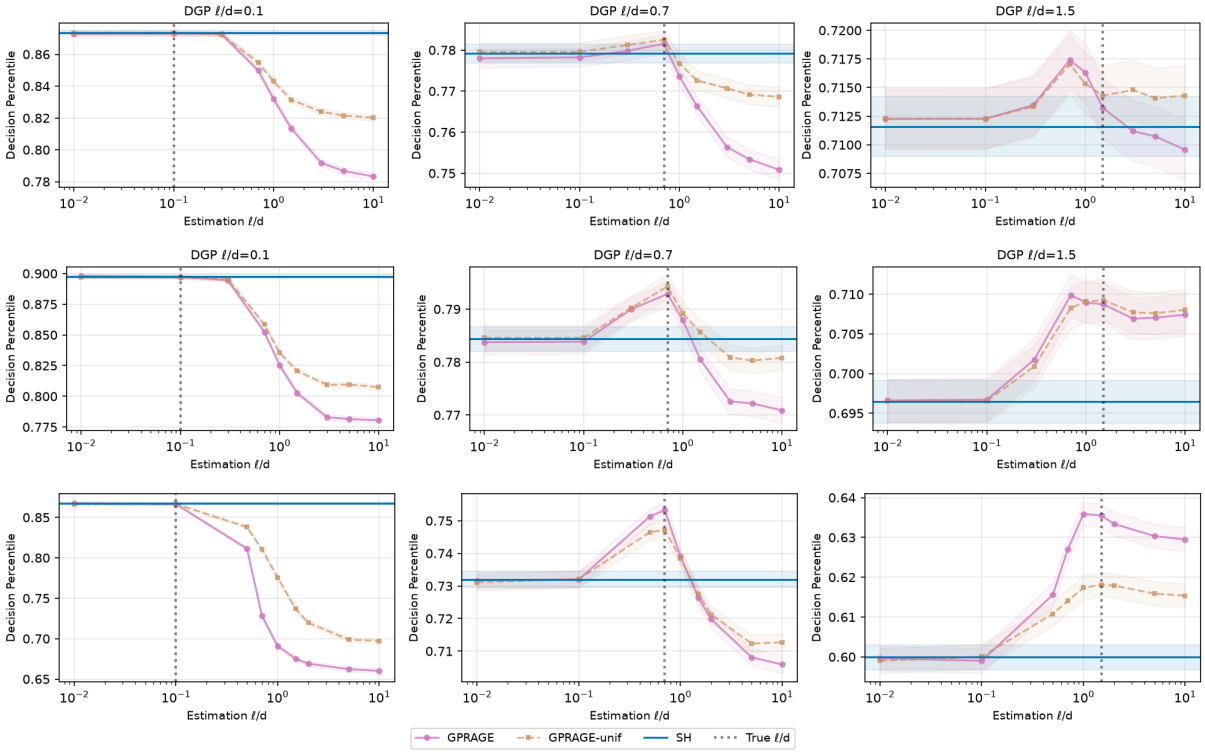


Fig. 1. (Top) The average best percentile metric for a simulated UI experiment (6 binary treatment attributes, 20 total treatments). (Middle) The average best percentile metric for a simulated UI experiment (6 binary treatment attributes, 60 total treatments). (Bottom) The average best percentile metric for a simulated example of hyperparameter tuning (2 numeric treatment attributes, 16 total treatments). We simulate the behavior of different algorithms varying the GP lengthscale ℓ relative to the median pairwise distance d_{med}: $\ell/d_{\text{med}} = 0.1, 0.7, 1.5$. Overall, GP-RAGE can outperform SH, but primarily when the lengthscale is well-calibrated. Larger lengthscales can substantially reduce performance when the true lengthscale ratio $\ell/d_{\text{med}} = 0.1, 0.7$. We generate treatment effects from an RBF kernel and GP-RAGE uses an RBF kernel; results are similar if assuming a linear or Matérn kernel.