Estimating the Returns from an Experimentation Program

Simon Ejdemyr*† Martin Tingley* Yian Shang* Travis Brooks*

October 17, 2024

Abstract

We describe the development, validation and implementation of a Bayesian hierarchical model used by Netflix for estimating the returns to experimentation. The model provides a trusted source for estimates of the cumulative returns from A/B test launches in various product innovation areas, and integration into Netflix's flagship experimentation UI has facilitated a better understanding of the rate of innovation in these different areas. This understanding can help company leaders prioritize the most promising innovation programs, or pivot innovation strategies in areas that show diminishing returns.

1 Introduction

Teams at Netflix run thousands of A/B tests every year to improve the product. This testing is organized around different *experimentation programs*, each of which comprises a set of similar tests. For example, we have programs supporting UI improvements, better signup flows, and more optimized personalization algorithms. A simple framing for the goal of each program is that it aims to maximize values on a target metric that reflects broader company priorities.

We describe how Netflix quantifies the extent to which programs achieve this goal. This means estimating the *cumulative impact* of a program—that is, the sum of treatment effects from launched treatments in the program—over a given time period. This information is useful for two reasons. First, if a program shows diminishing returns, experimenters can change levers such as p-value thresholds or sample sizes to improve the program's success going forward. Second, estimates of past program performance can help leaders redirect resources across programs to those that show the greatest potential to improve business outcomes.

However, estimating the cumulative impact of a program is not as simple as naively adding up the estimated average treatment effects (ATEs) of launched treatments. Such estimates suffer from the winner's curse, which is an upward bias due to selection on statistical significance or estimate magnitude (Andrews et al., 2019; Gelman and Carlin, 2014; Smith and Winkler, 2006). For intuition, suppose a product team tests 1,000 interventions that have no actual effect. With a false positive rate of α, just by chance $\alpha/2$ of these interventions will be positive and statistically significant, inducing an upward bias in the estimation of cumulative effects.

To overcome this challenge, we implement a Bayesian hierarchical model that shrinks the estimated effects of winning treatments. The model has two hierarchical levels—one for experiments within the same testing area, and one for treatments within the same experiment—and allows for either Gaussian or fat-tailed priors on the true treatment effect distribution. We show that model estimates are consistent with gold standard estimates from holdback tests (large retests of launched treatments). We also discuss the model's integration into Netflix's flagship experimentation UI, which has facilitated a better understanding of how different programs improve north star metrics, aided annual reviews, and streamlined goal tracking within the company.

This work is an important component of broader initiatives at Netflix that leverage meta-analysis to enhance the impact of experimentation at the company. In particular, meta-analysis of historical experiments can identify levers for improving experimentation *strategy*. For example, traditional null hypothesis testing typically is a non-optimal experimentation strategy if the goal is to optimize north star metrics (Azevedo et al., 2020). Yet, without estimates of the cumulative impact on these metrics, it is difficult—at least at a data-driven company like Netflix—to argue for a change in strategy.

2 Setup

Consider an experimentation program comprised of J experiments, each labeled by j. Each experiment contains K_j treatments, labeled k, plus a control. We define the treatment set $\mathcal{T}$, encompassing all treatments tested in the program, as:

$$\mathcal{T} = \{(j,k) : j \in \{1,\ldots,J\}, k \in \{1,\ldots,K_j\}\}.$$

*Netflix, Inc.
†Corresponding author, sejdemyr@netflix.com.

A subset of these treatments is rolled out to all users, forming the launch set $\mathcal{L} \subseteq \mathcal{T}$. Selection into $\mathcal{L}$ is determined by test-specific launch criteria, such as the ATE estimate being statistically significant and positive (see next section).

For each (j,k), let $\delta_{jk} = \alpha_{jk} - \alpha_{j0}$ represent the true ATE, or the difference between the experiment's k-th treatment group (with true mean α_{jk}) and its control group (α_{j0}). These treatment effects are estimated on a metric material to the company's long-term success, such as revenue or engagement. Under the assumption of additivity,[1] the cumulative treatment effect is the sum of the true effects of treatments in $\mathcal{L}$:

$$\Delta = \sum_{(j,k)\in\mathcal{L}} \delta_{jk}.$$

This quantity represents the returns from an experimentation program accruing to the company's users, and is our estimand of interest.

3 The winner's curse

At Netflix, a treatment is typically launched if its ATE estimate is statistically significant, positive, and the largest among the experiment's treatment groups. We therefore assume that the following indicator function determines whether treatment k in test j will be launched (is a "winner"):

$$\text{Winner}(j,k) = \begin{cases} 1 & \text{if } \frac{\hat{\delta}_{jk}}{\text{stderr}(\hat{\delta}_{jk})} > Z_{1-\frac{\alpha}{2}} \text{ and } k = \arg\max_{k'\in\{1,\ldots,K_j\}} \hat{\delta}_{jk'}, \\ 0 & \text{otherwise} \end{cases}$$

where $\hat{\delta}_{jk}$ is the ATE estimate and $Z_{1-\frac{\alpha}{2}}$ is the critical value for a desired false positive rate α. If $\hat{\delta}_{jk}$ is based on a non-penalized conventional ATE estimator like (regression-adjusted) difference-in-means, then selection on winners induces an upward bias known as the winner's curse:

$$\mathbb{E}[\hat{\delta}_{jk} \mid \text{Winner}(j,k) = 1] \geq \delta_{jk}.$$

For a proof, see Smith and Winkler (2006). This means that, when the launch set $\mathcal{L}$ selects on winners, a naive cumulative effect estimator based on non-penalized $\hat{\delta}_{jk}$ will also have an upward bias—that is, $\hat{\Delta} = \sum_{(j,k)\in\mathcal{L}} \hat{\delta}_{jk} \geq \Delta$. If uncorrected, this upward bias not only confounds estimates of the cumulative impact of launches, but it also provides poor incentives for experimenters, since one can increase $\hat{\Delta}$ by running A/A tests alone.

4 Bayesian hierarchical modeling

The objective of our Bayesian model is to rectify the biases introduced by the winner's curse and achieve better estimates for Δ, at a fraction of the costs of holdback tests. The model is fit to all treatments in the experimentation program to obtain a posterior estimate $\hat{\delta}^*_{jk}$ for $(j,k) \in \mathcal{T}$. To estimate the returns to launches, we then compute

$$\hat{\Delta}^* = \sum_{(j,k)\in\mathcal{L}} \hat{\delta}^*_{jk}$$

where inclusion of treatment k from test j into the launch set $\mathcal{L}$ could be based on any launch criteria, even if those criteria normally would induce the winner's curse.

In particular, for each experiment j, we store the unadjusted ATE estimates in the vector $\boldsymbol{Y}_j = [\hat{\delta}_{j1}, \ldots, \hat{\delta}_{jK_j}]$. We also construct the structural covariance matrix $\boldsymbol{\Sigma}_j$ from the unadjusted ATE standard errors as well as the sample sizes for the groups in the experiment. The off-diagonals of $\boldsymbol{\Sigma}_j$ are positive because the treatment groups within the experiment share a common control group. Our hierarchical model takes the following form:

$$\begin{aligned} \boldsymbol{Y}_j \mid \boldsymbol{\delta}_j &\sim N(\boldsymbol{\delta}_j, \boldsymbol{\Sigma}_j) \\ \delta_{jk} \in \boldsymbol{\delta}_j \mid \mu_j, \sigma_\delta &\sim N(\mu_j, \sigma_\delta^2) \\ \mu_j \mid \theta, \sigma_\mu &\sim N(\theta, \sigma_\mu^2) \\ \sigma_\delta &\sim \ s \cdot t_v^+ \\ \sigma_\mu &\sim \ s \cdot t_v^+ \\ \theta &\sim N(0, s^2) \end{aligned}$$

[1] Additivity implies that interactions between the treatments in $\mathcal{L}$ are accounted for. This condition is satisfied if each new launched treatment is compared against all prior launches—that is, if for experiment j all preceding ATEs have been absorbed into the experiment's baseline mean α_{j0}:

$$\alpha_{j0} = \alpha_0 + \sum_{j'<j,(j',k)\in\mathcal{L}} \delta_{j'k},$$

where α_0 is the true mean prior to realizing any ATE in $\mathcal{L}$.

where $\boldsymbol{Y}_j$ and $\boldsymbol{\Sigma}_j$ are data per above; $\boldsymbol{\delta}_j = [\delta_{j1}, \ldots, \delta_{jK_j}]$ is the vector of true ATEs; μ_j is the experiment-wide true treatment effect around which each of the experiment's per-treatment group ATE is distributed with a variance of σ_δ^2; θ is the grand true treatment effect across all experiments around which each experiment-wide effect is distributed with a variance of σ_μ^2; s is a scale hint (e.g., the standard deviation of the unadjusted ATEs across all treatments in $\mathcal{T}$); and t_v^+ is a half-t distribution with v degrees of freedom.

This model is fit using MCMC in Stan (Stan Development Team, 2018), and typically converges in a few minutes for data sizes in the thousands. Posterior samples are used to obtain both $\hat{\Delta}^*$ and $\hat{\text{Var}}(\hat{\Delta}^*)$.

5 Validation and productization at Netflix

Model performance is validated against holdback tests, large retests of launched treatments that do not suffer from the winner's curse. The results in Figure 1 show that the naive approach (adding up unadjusted ATEs) overstates the cumulative effect from holdbacks by 37%. In contrast, the Bayesian estimate consistently tracks the holdbacks.

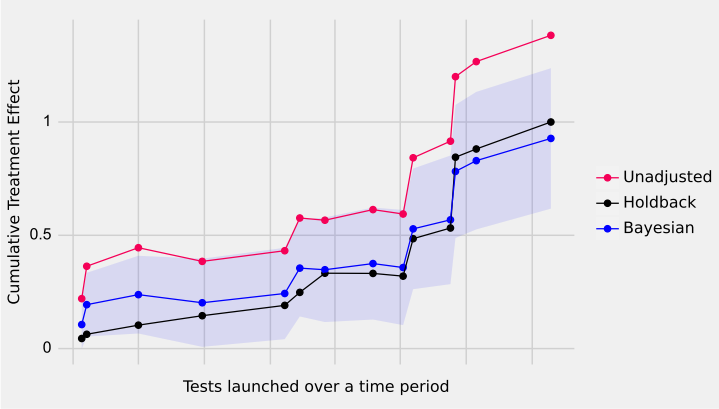
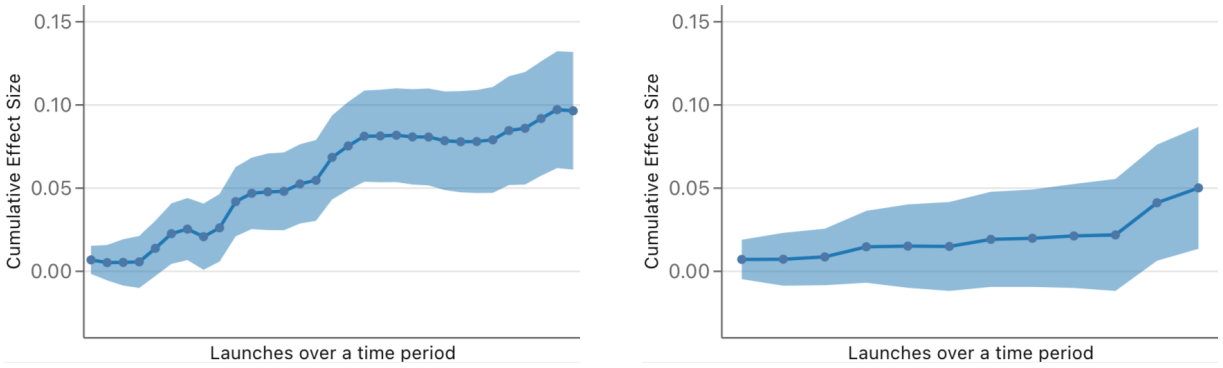


Figure 1: Validation against holdback tests (cumulative treatment effect in black), with unadjusted frequentist estimates shown in red and Bayesian estimates (with 95% credible intervals) in blue.

Given these validation results, we productized this estimator on the platform. Figure 2 shows examples from Netflix's A/B testing UI. In this example, the first innovation program drove a larger return on the metric of interest ($\hat{\Delta}^* = 0.1$) than the second program ($\hat{\Delta}^* = 0.05$) over the time period considered for these assessments.



(a) Experimentation Program 1 ($\hat{\Delta}^* = 0.1$)

(b) Experimentation Program 2 ($\hat{\Delta}^* = 0.05$)

Figure 2: Cumulative effect estimates on an important metric from Netflix's A/B UI (some information withheld).

This work shows the promise of meta-analytic approaches for improving experimentation quality and velocity. In a body of follow-up work, we are leveraging the results presented here for *optimizing* the returns from experimentation. Without estimates of cumulative returns (our optimization target), such work would be difficult and often leaves us with traditional statistical approaches for decision making, like null hypothesis testing with $p < 0.05$, which are suboptimal for improving long-term north star metrics within technology companies.

References

Isaiah Andrews, Toru Kitagawa, and Adam McCloskey. 2019. *Inference on winners.* Technical Report. National Bureau of Economic Research.

Eduardo M Azevedo, Alex Deng, José Luis Montiel Olea, Justin Rao, and E Glen Weyl. 2020. A/b testing with fat tails. *Journal of Political Economy* 128, 12 (2020), 4614–000.

Andrew Gelman and John Carlin. 2014. Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science* 9, 6 (2014), 641–651.

James E Smith and Robert L Winkler. 2006. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science* 52, 3 (2006), 311–322.

Stan Development Team. 2018. *The Stan Core Library.* http://mc-stan.org