Surrogate Metrics as Filters

Kenneth Hung and Michael Gill

Meta
{kenhung, michaelgill}@meta.com

October 5, 2023

1 Introduction

Experimentation has long been an integral part of technology companies [5]. But a common challenge is when changes in the goal metric is difficult to detect. Some potential causes include

- *Low signal-to-noise ratio.* As a product matures, the easy changes have been mostly considered. Additional incremental changes tend to lead to smaller improvements in the metric, while the noise stayed in similar levels.
- *Long-term outcomes.* Some metrics like user retention are inherently long-term outcomes. To measure these, an long-term experiment is often needed which occupies more experimental bandwidth.
- *Complex experimental designs.* In the absence of stable unit treatment value assumption (SUTVA), the effects require a more complex experimental design for proper measurement, e.g. budget-split design for marketplace experiments [8], cluster randomized experiments [4, 7]. On one hand, these complex designs are less likely to suffer from biases [1]; on the other hand, they tend to have larger standard errors.

A common idea is to utilize proxy or surrogate metrics [5]. For example, in advertising, number of clicks can act as a surrogate for the number of conversions; short-term engagement metrics may act as a surrogate for long-term retention; imperfect unit-randomized experimental outcome can be a surrogate for the outcome in the better experimental design. However, the validity of a surrogate metric comes as the main concern, and relies on various assumptions. For example, in building a proxy for long-term retention, we implicitly assume that this relationship is unaffected by the studied treatment; a variable that lies on the causal pathway between treatment and goal metric can be a surrogate, but we need to have a correct mental causal model.

In more unfortunate scenarios, we may even encounter a “surrogate paradox” — a surrogate metric may be improved by the treatment when the goal metric deteriorated. For example, using number of clicks as a surrogate for number of conversions may lead to adopting ranking models that favor clickbaits, that would actually reduce conversions. [9] provided criteria for when the paradox may happen, but the checks are mostly on an experiment-by-experiment basis. A notable exception, the meta-analytic approach [6] suggests to look at a collection of similar experiments to assess whether the surrogate paradox is happening. However, if we encounter an experiment that does not seem similar to this collection, we may feel uncomfortable claiming that the surrogate is valid.

2 Surrogate Metrics as Filters

Large-scale experimentation platforms typically emphasize on scalability, and do not utilize much knowledge about the causal mechanism in how the surrogate metrics is related to the goal metric. Often times, even the experimenters are unsure or it is difficult to obtain a consensus on the mechanism, e.g. more clicks may lead to more conversions, but the conversions turn out to be of low quality, it may lead to fewer clicks in the future.

To bypass this challenge, we propose to use surrogate metrics as only filters, while retaining validity by requiring an experiment using the goal metric. In other words, we aim to make experiments more detectable by economizing the experimental bandwidth.

Suppose we desire a false positive rate of α, e.g. 5%. Each experiment will have two stages:

- *Stage 1.* Smaller experiments based on surrogate metric, where γ is the fraction of units in this stage. If the p-value is smaller than some threshold β (not necessarily α), we move on the Stage 2.
- *Stage 2.* Larger experiments based on goal metric. We perform a statistical test at level α for this stage. There are fewer experiments reaching this stage, so experiments can be bigger than status quo and have higher power.

With a two-stage experiment, the definitions of power and minimum detectable effect (MDE) need to be altered accordingly: power is the probability that an experiment moves past Stage 1 and rejects the null hypothesis in Stage 2; MDE is the smallest effect size such that power (per the updated definition) meets some predetermined requirement, e.g. 80%. Note that both are data-dependent, specifically on the joint distribution of true effects in the goal metric and surrogate metric.

The parameters γ and β can then be optimized either maximize the average power or minimize the MDE. Note that these two objectives may require different choices of γ and β.

This idea is similar to futility stopping [2]. In futility stopping, the goal metric is also observed in Stage 1, and is used to compute the conditional power for Stage 2. In contrast, in our approach, a surrogate metric is observed in lieu of the goal metric in Stage 1, and the filtration uses a principled threshold that trades potential premature stopping of some experiments for higher power in the remaining experiments.

3 Opportunities

We focus on the scenario where Stage 1 and Stage 2 occupy the same experimentation bandwidth, i.e. Stage 2 is left with $(1-\gamma)$ of the original number of units. Similar analysis can be performed if Stage 2 has a separate experimentation bandwidth[1]. Under a simple bivariate Gaussian model [3], we can normalize such that the status quo estimator variance for both metrics is 1, and assume the true effects in the surrogate (θ_S) and goal metric (θ_G) distribute as

$$\begin{pmatrix}\theta_G\\ \theta_S\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_G^2 & \rho\sigma_G\sigma_S\\ \rho\sigma_G\sigma_S & \sigma_S^2\end{pmatrix}\right),$$

Hence for Stage 1, with the bandwidth γ, we observe

$$\hat{\theta}_S \mid \theta_S \sim N(\theta_S, 1/\gamma).$$

An experiment proceeds to Stage 2 if $\hat{\theta}_S > c_1(\beta,\gamma) := z_{1-\beta}\sqrt{1/\gamma}$, so the number of experiments entering Stage 2 is reduced by a factor of $\mathbb{P}(\hat{\theta}_S > c_1)$, sharing the bandwidth $1-\gamma$. In Stage 2, we will hence observe

$$\hat{\theta}_G \mid \theta_G \sim N(\theta_G, \mathbb{P}(\hat{\theta}_S > c_1)/(1-\gamma)),$$

where we reject if $\hat{\theta}_G > c_2(\beta,\gamma) := z_{1-\alpha}\sqrt{\mathbb{P}(\hat{\theta}_S > z_{1-\beta}\sqrt{1/\gamma})/(1-\gamma)}$.

This allow us to evaluate the two updated objectives: average power and MDE, given by

$$\text{average power}(\beta,\gamma) := \mathbb{P}(\hat{\theta}_S > c_1(\beta,\gamma), \hat{\theta}_G > c_2(\beta,\gamma)).$$

The power conditional on a specific θ_G can analogously be defined as

$$\text{conditional power}(\theta_G,\beta,\gamma) := \mathbb{P}(\hat{\theta}_S > c_1(\beta,\gamma), \hat{\theta}_G > c_2(\beta,\gamma) \mid \theta_G),$$

[1] For example, unit-randomized experiments vs cluster-randomized experiments

which allows us to extend the definition of MDE as

$$\mathrm{MDE}(\beta, \gamma) := \min\{\theta_G : \text{conditional power}(\theta_G, \beta, \gamma) \geq 0.8\}.$$

Now the potential power gain (Figure 1) and MDE reduction (Figure 2) depend solely on three inputs: signal-to-noise ratios of the goal metric (σ_G) and the surrogate metric(σ_S), and the correlation of the true effects in these two metrics[2] (ρ).

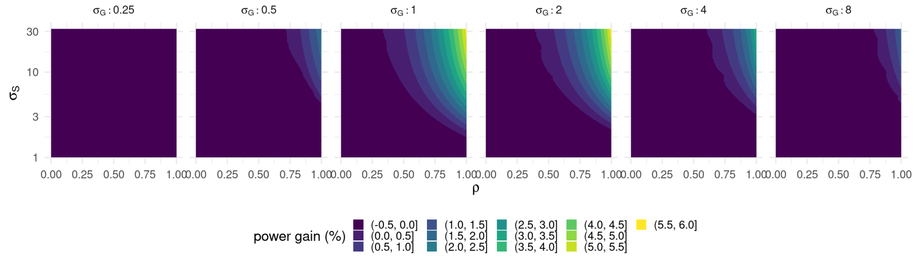

Figure 1: Power gain, as a function of the signal-to-noise ratio of the goal metric (σ_G) and the surrogate metric (σ_S), and the correlation of the true effects in the two metrics (ρ). When σ_G is small, there is little hope of increasing the power; when σ_G is large, there is little room for increasing the power. Finding a surrogate metric with larger signal-to-noise ratio (σ_S) or stronger correlation (ρ) always yields higher power.

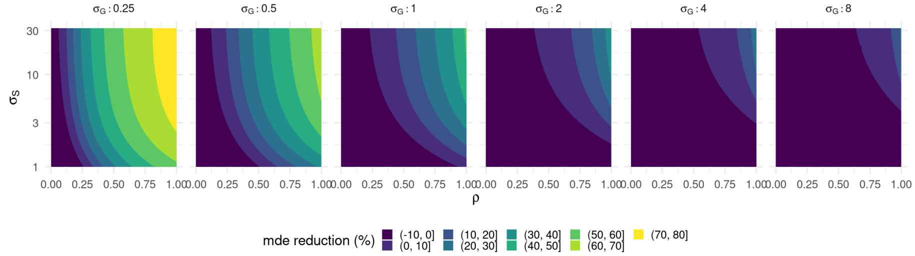

Figure 2: MDE reduction, as a function of the signal-to-noise ratio of the goal metric (σ_G) and the surrogate metric (σ_S), and the correlation of the true effects in the two metrics (ρ). Finding a surrogate metric with larger signal-to-noise ratio (σ_S) or stronger correlation (ρ) always yields smaller MDE.

[2] Not to be confused with the correlation of the sampling noise

References

[1] Thomas Blake and Dominic Coey. “Why Marketplace Experimentation is Harder than It Seems: The Role of Test-Control Interference”. In: *Proceedings of the Fifteenth ACM Conference on Economics and Computation*. EC ’14. Palo Alto, California, USA: Association for Computing Machinery, 2014, pp. 567–582. ISBN: 9781450325653. DOI: 10.1145/2600057.2602837. URL: https://doi.org/10.1145/2600057.2602837.

[2] Yen Chang et al. “Futility stopping in clinical trials, optimality and practical considerations”. In: *Journal of Biopharmaceutical Statistics* 30.6 (Nov. 2020), pp. 1050–1059. URL: https://doi.org/10.1080/10543406.2020.1818253.

[3] Tom Cunningham and Josh Kim. “Interpreting Experiments with Multiple Outcomes”. July 2020.

[4] Dean Eckles, Brian Karrer, and Johan Ugander. “Design and Analysis of Experiments in Networks: Reducing Bias from Interference”. In: *Journal of Causal Inference* 5.1 (2017), p. 20150021. DOI: doi:10.1515/jci-2015-0021. URL: https://doi.org/10.1515/jci-2015-0021.

[5] Somit Gupta et al. “Top Challenges from the first Practical Online Controlled Experiments Summit”. In: *SIGKDD Explorations*. Ed. by Ankur Teredesai Hanghang Tong Xin Luna Dong and Reza Zafarani. Vol. 21. 1. 2009.

[6] Marshall M. Joffe and Tom Greene. “Related Causal Frameworks for Surrogate Outcomes”. In: *Biometrics* 65.2 (June 2009), pp. 530–538. URL: https://doi.org/10.1111/j.1541-0420.2008.01106.x.

[7] Brian Karrer et al. “Network experimentation at scale”. In: *Proceedings of the 27th acm sigkdd conference on knowledge discovery & data mining*. 2021, pp. 3106–3116.

[8] Min Liu, Jialiang Mao, and Kang Kang. “Trustworthy and Powerful Online Marketplace Experimentation with Budget-Split Design”. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. KDD ’21. Virtual Event, Singapore: Association for Computing Machinery, 2021, pp. 3319–3329. ISBN: 9781450383325. DOI: 10.1145/3447548.3467193. URL: https://doi.org/10.1145/3447548.3467193.

[9] Tyler J. VanderWeele. “Surrogate Measures and Consistent Surrogates”. In: *Biometrics* 69.3 (2013), pp. 561–565. DOI: https://doi.org/10.1111/biom.12071. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.12071. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12071.